Detecting Influencers in Social Media Discussions

Sara Rosenthal

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2015

ABSTRACT

Detecting Influencers in Social Media Discussions

Sara Rosenthal

In the past decade we have been privileged to witness the creation and revolution of social media on the World Wide Web. The abundance of content available on the web allows us to analyze the way people interact and the roles they play in a conversation on a large scale. One such role is influencer in the conversation. Detecting influence can be useful for successful advertisement strategies, detecting terrorist leaders and political campaigning.

We explore influence in discussion forums, weblogs, and micro-blogs using several components that have been found to be indicators of influence. Our components are author traits, agreement, claims, argumentation, persuasion, credibility, and certain dialog patterns. In the first portion of this thesis we describe each of our system components. Each of these components is motivated by social science through Robert Cialdini's "Weapons of Influence" [Cialdini, 2007]. The weapons of influence are Reciprocation, Commitment and Consistency, Social Proof, Liking, Authority, and Scarcity. We then show the method and experiments for classifying each component.

In the second part of this thesis we classify influencers across five online genres and analyze which features are most indicative of influencers in each genre. The online genres we explore are Wikipedia Talk Pages, LiveJournal weblogs, Political Forum discussions, Create Debate debate discussions, and Twitter microblog conversations. First, we describe a rich suite of features that were generated using each of the system components. Then, we describe our experiments and results including using domain adaptation to exploit the data from multiple online genres. Finally, we also provide a detailed analysis of a single weapon of influence, social proof, and its impact in detecting influence in Wikipedia Talk Pages. This provides a single example of the usefulness of providing comprehensive components in the detection of influence.

The contributions of this thesis include a system for predicting who the influencers are in online discussion forums. We provide an evaluation of a rich set of features inspired by social science. In our

system, each feature set used to detect influence is complex and computed by a system component. This allows us to provide a detailed analysis as to why the person was chosen as an influencer. We also provide a comparison of differences across several online discussion datasets and exploit the differences across the different genres to provide further improvements in influence detection.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisor, Kathleen McKeown. This would not have been possible without her. Kathy has been more than just a mentor to me in my thesis and research. I have felt her support throughout my time at Columbia in all matters. I take comfort in knowing that she is someone I can always turn to for excellent advice.

This thesis would also not be possible without Owen Rambow, who was instrumental in much of the earliest research. I appreciate how he has continued to be available and very much involved throughout my time at Columbia. I would also like to thank the rest of my thesis committee, Luis Gravano, Julia Hirschberg, and David Park. I appreciate the time spent in reading my thesis, the valuable feedback, and the excitement in my topic I have felt from them.

I have had the great fortune of working with many colleagues and students, many of who have been instrumental in this thesis. Many thanks to Or Biran, Jacob Andreas, Chris Kedzie, Swabha Swayamdipta, Greg Barber, Suvarna Bothe, Ethan Robert Grant, Yousuf Hauke, Mukund Jha, Elizabeth Kierstead, Alisa Krivokapvic, William Lipovsky, Coleman Moore, Kapil Thadani, Lakshmi Vikraman, and Sahil Yakhmi.

I would not have enjoyed my time at Columbia without the great friends and colleagues I have made in the NLP group and in the rest of the department, particularly my WICS friends. I have enjoyed our conversations about research and everything else. Thanks to all that I have not yet mentioned including (but not limited to) Apoorv Agarwal, Mohamed AlTantawy, Daniel Bauer, Hila Becker, Dana Dachman-Soled, Fadi Biadsy, Erica Cooper, Bob Coyne, Heba Elfardy, David Elson, Noura Farra, Weiwei Guo, Kevin Lerman, Weiyun Ma, Chris Hidey, Rivka Levitan, Jessica Ouyang, Kristen Parton, Yves Petinot, Vinod Prabhakaran, Arthi Ramachandran, Karl Stratos, and John Zhang. I would also like to thank the rest of CCLS and the CS department faculty and staff,

most of whom have helped me in some way during my time here.

Of course this thesis would not have been possible without the loving support of my friends and family. My mother has done so much, watching my children every week and more, so I could finish this thesis. Of course my mother has had my stepfather Pinchus by her side, pitching in at all times. I couldn't have done it without them. All of my siblings, Donny, Ephraim, Benny, Reuven, Rivka, and Esti. Just about every one has helped babysit at one point. Thank you to all my extended family. I would like to thank my grandparents. I would especially like to thank my grandfather who funded part of my masters degree. I would like to thank my in-laws, brother in-laws and sister in-laws, nieces and nephews, cousins (especially Laurel), aunt and uncles. My closest friends have not seen me as often as we would all like. Faigy, Gitty, Rochel Tila, Shany, Shiffy, Shoshana, thanks for still being there for me. I hope we will have more time to spend together now! Although it has taken everyone here to accomplish this, I couldn't have done this without my best friend, my husband, Dovid, and my children, Avraham and Malky. They are my light and my happiness. This would all mean nothing without them.

Finally, to my father, who although he is not here, I know he is looking down at me and is so very proud of this, and all my other accomplishments. I recall how proud he was when I got into the Columbia masters program. He past away during my first week at Columbia, but I know he has been here in spirit all the way through.

To my mother and father

Part I

Introduction

Chapter 1

Introduction

> *All the weapons of influence discussed in this book work better under some conditions than under others. If we are to defend ourselves adequately against any such weapon, it is vital that we know its optimal operating conditions in order to recognize when we are most vulnerable to its influence.*
>
> Robert B. Cialdini, *Influence: The Psychology of Persuasion*

In the past decade we have been privileged to witness the creation and revolution of social media on the World Wide Web. Media content, whether it be images and video, networking, or online discussions have become a daily and integral part of our lives. Lengthy conversations between groups of people are available publicly as never before through weblogs, discussion forums, micro-blogs, and social networking sites.

The abundance of content available on the web allows us to analyze the way people interact and the roles they play in a conversation on a large scale. One such role is the person who influences others in the conversation. Detecting influencers can be useful for successful advertisement strategies, detecting terrorist leaders, political campaigning, and promoting grassroots movements. Furthermore, it can be also used for less obvious tasks such as grading students based on helpful participation in a virtual discussion and, similarly, promoting influential employers in a corporation.

This thesis describes a method for automatically detecting influencers in multiple online genres using several system components that are motivated by social science. In the rest of the introduction,

we will define influencers, explain our approach for detecting influencers, discuss related work in influencer detection, and provide an overview of our contributions. Finally, we will end with an outline of the rest of this thesis.

1.1 Defining Influencers

Before one can begin to explore how to detect influencers, one must first understand what makes someone an influencer. There are two different, but related, forms of influence: situational Influence, and global influence. Situational influence refers to influence by an individual as evident by their participation in the conversation and global influence refers to how influence spreads among a community. The focus of this research is on situational influence. It is important to explore situational influence because a person's influence can vary across discussions.

In any conversation where participants express their beliefs some people are more influential than others. An influencer can alter the opinions of their audience, resolve disagreements where no one else can, be recognized by others as one who makes important contributions, and often continue to influence a group even when not present. Other conversational participants often adopt their ideas and even the words they use to express their ideas. These forms of *personal influence* [Katz and Lazarsfeld, 1955] are part of what makes someone an opinion leader.

Robert B. Cialdini [2007] defines several principles, or "weapons of influence", that a person can employ to influence others: Reciprocation, Commitment and Consistency, Social Proof, Liking, Authority, and Scarcity:

Reciprocation

People tend to return favors. For example, if a person borrows an item from a friend they will feel obligated to return the courtesy in the future. Similarly, if a colleague, Mary, is in favor of Bob's idea, Bob will feel obligated to favor Mary's future ideas.

Commitment and Consistency

People tend to want to commit and be consistent with things they have said or written in the past. For example, if a person agrees with a colleague's initial proposal of an idea they are more likely to continue to support it.

Social Proof

People tend to follow what others do; this is most common when a person is uncertain of a course of action. For example, people are more likely to purchase clothing that is similar to what their friends wear.

Liking

A person is more likely to be influenced by someone they like. For example, people are more likely to buy a product from someone they know and like.

Authority

People of authority tend to be obeyed more often. For example, people are more likely to be influenced by their superior in the workplace.

Scarcity

Things that are perceived to be scarce, or in limited supply generate more demand. For example, if a product is rare it becomes more desirable and people will be influenced to pay more for it.

These principles inspire each of the components we use to detect situational influence: Opinion, Claims, Argumentation, Persuasion, Agreement, Author Traits, Dialog Patterns, and Credibility. We discuss the relationship between each of the principles and our components in the appropriate chapters. As Cialdini describes (in opening quote to this chapter), certain weapons may work better in different scenarios, or in our case, in different online genres. Finally, it is important to note that we are interested in finding people that display the *characteristics* of someone who is influential regardless of *causality*. This means that we do not know if the influencer actually changes the mind of the people they have influenced, but rather they display the characteristics of someone who could.

1.2 Approach

The research in this thesis combines the areas of computational linguistics and computational social science to automatically detect influencers within a conversation. Notably, there is a focus on using the context in the discussion to understand why a person is influential. Our approach first involves detecting the components of influence and then detecting influencers across multiple online genres.

1.2.1 Components of Influence

As described in the definition, we motivate influencer detection through Robert Cialdini's weapons of influence [Cialdini, 2007]: Reciprocation, Commitment and Consistency, Social Proof, Liking, Authority, and Scarcity. These principles inspire each of the components we use to detect situational influence. We explore multiple online genres in several components (e.g. claims and agreement) as well as comparing and contrasting the differences between online genres. Each component is motivated by one or more weapons of influence:

Opinion Detection

Opinion Detection includes subjectivity (subjective or objective) and polarity (positive, negative, or neutral) detection. Opinion detection is a subcomponent used in agreement and claim detection. Opinion is motivated by reciprocation and liking. When a person gives compliments it is likely to cause reciprocation. This is also motivation for using opinion in Agreement detection. People are more likely to be influenced by someone they like, even if the compliments received from that person are false.

Author Traits

Author Traits are characteristics derived about the person from demographics (age, gender, and religion) and political affiliation. Including author traits in influencer detection is motivated by social proof, liking, and scarcity. Social proof entails that a person will be influenced by others in their surroundings. It is most effective when a person perceives the people in their surroundings to be similar to them, such as through demographics. The association to a person through age, location, or gender will cause liking. Scarcity is also motivation because when an opportunity becomes less available or restricted it causes a psychological reactance to the loss of freedom. This becomes evident across groups of people that are affected by the restriction. For example, teenagers have been found to be most affected by scarcity.

Agreement and Disagreement

This component detects agreement, disagreement, or the lack thereof between two posts. Agreement is motivated by reciprocation, commitment and consistency, and scarcity. If someone says something positive, a person will be more likely to agree with them and if someone

says something negative, a person will be likely to disagree with them. Furthermore, once someone agrees with someone else they are likely to agree again for the sake of commitment and consistency. In addition, scarcity, in the form of a restriction, is likely to cause a negative reaction which will cause others to react by defying and disagreeing with the restriction and the person who placed it.

Claims

Claims are an opinionated assertion indicating the truth of something. Since by definition, the claim must be opinionated, claim detection is motivated by liking and reciprocation, as described earlier under opinion detection. It is also motivated by commitment and consistency as a person needs to stand by their beliefs in order for them to hold value. This is motivation for using committed belief in opinion detection.

Argumentation

Argumentation is a justification towards a claim. Social science experiments have found that argumentation is indicative of influencers. For instance, an experiment examining requests to cut a waiting line found that if the person followed their request with a reason, people were more likely to concede [Langer, 1989; Cialdini, 2007].

Persuasion

Persuasion is a claim followed by argumentation, grounding, or reiteration. We have already described and motivated claim and argumentation. Grounding is an appeal to an external source, knowledge or authority to support the claim. It is motivated by both the liking and authority weapons of influence, which discuss that familiarity and association to the well-known cause influence. Reiteration is a restatement or paraphrase of the original claim. It is motivated by commitment and consistency as a person feels the need to repeat their arguments to be consistent.

Credibility

Credibility is an indication of authority or the lack of. It is motivated by liking and authority. Clearly, those that are in authority tend to be considered more credible. In addition, being credible will cause a person to be liked.

Dialog Patterns

Dialog Patterns are based on the thread structure of the discussion. Each dialog pattern is motivated by different weapons of influence. Most are motivated by social proof. For example, the person who has initiative (started the conversation) will be followed. Similarly if someone is responded to often they are more likely to keep being responded to. This is also motivated by commitment and consistency, and reciprocation. Finally, a person will want to respond quickly of fear of being ignored or missing out. This is scarcity.

1.2.2 Influencers Across Genres

In the second part of our approach, we describe our supervised method for using the system components to detect situational influence across online genres. Influencers are predicted per person within a discussion. In other words, for each participant X in each thread Y, the system answers the following question: *Is X an influencer in Y*? We detect influencers across five online data sources that have been annotated for influencers: Wikipedia Talk Pages, LiveJournal weblogs, Political Forum discussions, Create Debate debate discussions, and Twitter microblog conversations. First, we describe a rich suite of features that were generated using each of the system components. Then, we describe our experiments and results including using domain adaptation [Daumé, 2007] to exploit the data from multiple online genres. Finally, we also provide a detailed analysis of a single weapon of influence, social proof, and its impact in detecting influencers in Wikipedia Talk Pages. This provides a single example of the usefulness of providing comprehensive components in the detection of influencers.

1.3 Related Work in Influence

Influence has been defined and studied for several decades in the social sciences but automatically detecting influence has only recently begun to be explored using social networks and computational linguistics. We discuss related work in all of these areas in the following subsections. There is also a considerable amount of related work for each of our system components (e.g. opinion, agreement, and author trait detection). The related work on each component is left to the appropriate chapters.

1.3.1 Influence in Social Networks

Influence can spread among a community over time. This is known as social or global influence and can occur on a small scale (e.g. a classroom of students) or a large scale (e.g. the Twitter social network). Katz and Lazarsfeld [1955] explain the social aspect of influence as a "two-step flow" model where the opinion leaders act as an intermediary between the mass media (e.g. news articles) and society.

Influence in social networks has origins in social science. Early work originated with Stanley Milgram's [Travers *et al.*, 1969] small-word experiment. It was later popularized by Malcolm Gladwell on his theory of viral contagion and the tipping point [Gladwell, 2002]. More recently, social networks have been analyzed automatically on a larger scale (e.g. [Watts and Dodds., 2007; Bakshy *et al.*, 2011; Barbieri *et al.*, 2013; Huang *et al.*, 2012; Goyal *et al.*, 2011; Myers *et al.*, 2012]) by analyzing how influence spreads through the network. Watts and Dodds [2007] simulate the two-step flow of influence [Katz and Lazarsfeld, 1955] to examine how influence is diffused in a network. They found that influence in a network is driven by a mass of individuals as opposed to the influencers. In more recent work, Bakshy *et al* [2011] investigated influence in Twitter by tracking the diffusion of millions of events over a two month interval. Their experiment found that influencers tend to be people who have been influential in the past and who have a large number of followers. They also found that URLs rated as interesting by Mechanical Turk workers were more likely to spread, but among the interesting ones, they were unable to determine which ones would actually spread.

People tend to perceive those in their surroundings to be similar to them. This is know as *homophily*. Although not directly related, there has also been some work in distinguishing between peer-based influence and homophily [Aral *et al.*, 2009; La Fond and Neville, 2010]. This research explores whether social contagion (how the information spreads) is due to influence or homophily. It finds that homophily accounts for the majority of behavioral contagion. More recent work [Bamman *et al.*, 2012] has also explored the effect of gender identity and homophily. They found that in general, homophily is correlated with the language use of a particular gender. In our work we detect situational influence by exploring who is influential within a single thread, and not how the information spreads. This allows us to focus on the effect of social proof on influence.

Most related to our work, Aral and Walker [2012] analyze the impact of demographics on influ-

ence by identifying influential people on Facebook. They find influential people by examining how a viral message spreads through the network. They found interesting patterns among demographics: Men are more influential than women, influencers are more likely to be married, and people who state their relationship as *it's complicated* are most susceptible to influence. They also found that older people tend to be more influential than younger people and that people are the most influential to their peers. We have similar findings in age and gender in our analysis. In contrast to our work, they did not use the demographics to predict influence nor do they predict influence within a discussion. Similarly, Dow *et al* 2013 investigate how photos on Facebook are shared and which demographics are more likely to share a particular photo. For example, Obama supporters were more likely to share a victory photo of Obama. This is an indication that the topic of the discussion is important.

It has been found that social media networks differ across platforms. Twitter is more fragmented and unevenly distributed among popular users having millions of followers while most users have very few followers [Kwak *et al.*, 2010; Wu *et al.*, 2011]. Facebook, on the other hand, follows a more traditional structure with a power law distributions of the friend network and average separation of six degrees [Ferrara, 2012]. We find that the content within online discussions from different sources differs as well.

While exploring influence using social networks is a different problem from ours, it is a motivation for exploring influence over time and we think our work can be complementary to this research as we will discuss in future work in Chapter 13.

1.3.2 Influence in Conversations

Early work in exploring influence has been examined in spoken conversations [Bales *et al.*, 1951; Scherer, 1979; Brook and Ng, 1986; Ng *et al.*, 1993; Ng *et al.*, 1995; Reid and Ng, 2000; Bales, 1969]. These studies have been performed on small groups of individuals (3-8 people), mainly university students, discussing topics ranging from euthanasia, television programs on homosexuality, legal aspects of commercial surrogacy, and capital punishment. Through this work, it has long been established that there is a correlation between the conversational behavior of a discourse participant and how influential he or she is perceived to be by the other discourse participants. Specifically, factors such as frequency of contribution, proportion of turns, and number of successful interruptions have been identified as being important indicators of influence. Furthermore, context

System	Features	Method	Data	Results
Quercia *et al* [2011]	LIWC	Linear Regression	Twitter	N/A
Nguyen *et al* [2013b]	Topic Shifts, Turn features	Bayesian Model, Ranking	Crossfire TV Show Wikipedia Talk Pages	.83% F $55\% \pm .4\%$ F
Reinks [2007]	Dialog structure and Topic Initiations (obtained manually)	NB, J48, SVM	AMI Meeting Corpus	70.6% A
Young *et al* [2011]	unigrams and bigrams	Utterance Level, NB, SVM, MaxEnt	Hostage Negotiation Transcripts	44.5% F
Prabhakaran *et al* [2013]	Dialog Acts, Dialog Structure, Overt Displays of Power	SVM	Enron E-mail Corpus	22.6% F
Our System	Opinion, Persuasion, Agreement, Author Traits, Credibility, Dialog Structure	SVM, Domain Adaptation	Wikipedia Talk Pages LiveJournal Create Debate Political Forum Twitter	56.7% F 81.6% F 34.2% F 74.1% F 93.8% F

Table 1.1: Related Work in Influencer Detection in Conversations. Our work is shown in the last row as a comparison. Results in F-score (F) or Accuracy (A)

has also found to be important [Reid and Ng, 2000]. Particularly, influence is correlated with prototypical utterances and more likely to be used in successful interruptions. Detecting influence in online conversations differs from spoken conversations as there is no interruption in written dialog; nonetheless, many aspects of this work support our approach for automatically detecting influence in online conversations.

In computational linguistics, several authors have detected influencers in a single conversation using the actual discussion [Quercia *et al.*, 2011; Nguyen *et al.*, 2013b; Young *et al.*, 2011; Prabhakaran and Rambow, 2013]. This work has explored detecting influencers using features such as dialog structure, persuasion, and topic control. Influence has been explored in in Wikipedia Talk Pages, Twitter, Presidential Debates, the ICSI meeting corpus [Janin *et al.*, 2003], and CNN's

Crossfire transcripts.

Quercia *et al* [2011] look at influencers' language use in Twitter contrasted to other users' groups and find some significant differences. However, their analysis and definition relies quite heavily on the particular nature of social activity on Twitter. [Rienks, 2007] discusses detecting influencers in a corpus of conversations. While he focuses entirely on non-linguistic behavior, he does look at (verbal) interruptions and topic initiations which can be seen as corresponding to some of our Dialog Patterns Language Uses. The best cross-validation experiment is using unigram features with a Naive Bayes learner achieving a 44.5% F-measure. More recently, Nguyen *et al* [2013b] identify influencers in the Crossfire TV Show and Wikipedia Talk Pages using topic shiftsIncluding topic shifts in our system would be interesting future work. However, we decided to not explore it at this time as our discussions tend to be on one topic.

There has also been work exploring influence on the utterance level. Young *et al* [2011] detect what they call "persuasion" in dialog, but their definition actually directly relates to influence as their model is based off of Robert B. Cialdini's principles of influence [Cialdini, 2007] who also used the term persuasion when referring to influence. Their work was performed on a micro text corpus of hostage negotiation transcripts. The first difference in their approach is that they annotated the corpus for the principles of influence whereas we asked our annotators to find the influencers in the document using a given guideline explaining behavior likely to be found in an influencer. They do not build system components based off of the principles, but rather use unigrams and bigrams as features. Young *et al* [2011] do not distinguish between the weapons of influence in their results making it impossible to determine their performance on each component alone.

Table 1.1 summarizes the related work in detecting influencers in a conversations and shows how it compares to our work. It is difficult to directly compare the systems due to the different datasets and features, but there are some conclusions that can be drawn from the various research. First, it is clear that the difficulty of influence detection varies across conversation. For example, it is hard to detect influencers in the Enron e-mail corpus and Create Debate, but much easier to detect influencers in the Crossfire TV show and Twitter. Finally, most related work detects influencers in only one genre (or at most two). In contrast, we detect influencers in 5 online genres and apply domain adaptation to exploit our various datasets.

A closely related area of research has been predicting power relations in dialog [Prabhakaran

and Rambow, 2014; Danescu-Niculescu-Mizil *et al.*, 2012; Strzalkowski *et al.*, 2013; Prabhakaran *et al.*, 2014b; Prabhakaran and Rambow, 2013]. This can include several types of power relationships, such as hierarchical and administrative power, as well as influence. Prior work on predicting different power relations tends to be at a broader level using more general features such as dialog structure and topic control [Prabhakaran and Rambow, 2014; Prabhakaran and Rambow, 2013; Danescu-Niculescu-Mizil *et al.*, 2012]. In addition, Prabhakaran *et al* have explored the role of gender in power within the Enron e-mail corpus [Prabhakaran *et al.*, 2014b] and using topic shifts to predict power in the presidential candidate debates [Prabhakaran *et al.*, 2014a]. Other work has explored the effect of accommodation on Power in Wikipedia Admins and Supreme Court Justices [Danescu-Niculescu-Mizil *et al.*, 2012]. Finally, Strzalkowski [2013] predict power using some more complex features such as topic control, disagreement, and dialog behavior in the MPC chat corpus.

One final related work is that of Pursuit of Power [Swayamdipta and Rambow, 2012]. It attempts to find people who pursue power regardless of their success. Prior work in pursuit of power uses Gaussian Mixture Models and Naive Bayes along with similar dialog structure features as described here and lexical bag of word features per author. The lexical features are not useful, probably because of the large amount of text per author on the same topic. This is motivation for excluding bag of word features for influence detection as well.

1.3.3 Summary

As we have described, there has been prior work that has 1) studied influence within social science, 2) analyzed the impact of influence within the social network, and 3) detected influence in conversations. In contrast to prior work, we automatically detect influence using several components, all motivated by social science. This allows us to provide a detailed analysis of *why* the person is influential. We also explore the variation among influencers in multiple online genres.

1.4 Overview of Contributions

The main contributions of this thesis are:

Link to Social Science: The first contribution of our research is to provide a link between

social science and our method for detecting influencers. This is important because in addition to showing that our system is successful, we also show that it validates the earliest research in analyzing influencers. The research in social science was performed manually on a small scale. In contrast, we develop a system that automatically detects influencers.

Detecting Influencers: Of course the major contribution of this thesis is a system that can predict influencers. We predict influencers within a discussion by asking if each person in the discussion is or is not an influencer. The distinctive characteristics of our approach is the use of system components, a comprehensive analysis of influencers, and a cross-genre analysis as described in the next three contributions.

System Components: In addition to detecting influencers, we also create several components that are stand-alone systems. They are Opinion, Claims, Argumentation, Persuasion, Agreement, and Author Traits. These systems are useful for detecting influencers as well as other tasks. For example, opinion detection is useful for understanding what people think about restaurants, their political views, and products they buy. Agreement detection can be useful for understanding how conflicts arise and are resolved. Claim detection is useful for identifying claims that are not trustworthy. Finally, author trait detection can be used in author profiling. As part of this thesis we are also making all of these systems available for public use as described in Appendix B.

Comprehensive Analysis of Influencers: Using the system components to detect influencers has a key advantage. In addition to predicting who the influencer is, we can also explain *why* they are the influencer. For example, perhaps they were persuasive and credible. This could also be used to give people advice on how to be more influential. A person's conversation can be analyzed to determine which components they use. This could then be turned into advice to use a component more to improve their influence. Furthermore, we also provide a detailed analysis of one weapon of influence, social proof, within Wikipedia Talk pages. This detailed analysis provides indication of the usefulness of each weapon of influence.

Cross-Genre Analysis: One key advantage of our research in comparison to prior work is that we compare how influencers differs across multiple online genres. The online genres we explore are comprehensive in the types of social media being explored. They are: LiveJournal (weblogs), Wikipedia Talk Pages (task oriented discussion forum), Political Forum (discussion forum), Create Debate (debate forum), and Twitter (microblog). Analyzing multiple genres is important. As Cialdini

describes in the quote in the beginning of this chapter influencers can vary in different situations. We find that this is the case across online genres as well. In addition, in our cross-genre analysis, we exploit data from multiple online genres to improve performance, particularly in genres with little annotated data.

Annotation Manuals and Corpora: Through our research in detecting influencers and its components we have had the opportunity to develop several annotation tools and datasets. We provide rich annotation manuals as well as several annotation systems using Amazon's Mechanical Turk and stand-alone web annotation systems. We have created several automatically labeled datasets as well as manually annotated datasets. These can be useful for future researchers in many areas. We have developed datasets in multiple online genres for opinions, claims, argumentation, persuasion, agreement, author traits (age, gender, religion, and politics), and influencers.

1.5 Outline

This thesis is structured into five main parts: The introduction, system components, influencer detection, conclusions, and appendices. Part I consists of this introduction (Chapter 1), related work for influencer detection (Chapter 1), and an explanation of the datasets (along with examples) used to detect influencers (Chapter 2).

In Part II of this thesis, we describe in great depth the system components that are used to detect influencers. They are Opinion (Chapter 4), Agreement (Chapter 5), Persuasion (Chapter 6), Author Traits (Chapter 7), and the direct Dialog and Credibility features (Chapter 8). The chapter for each component begins with motivation for including it in influencer detection using social science. Then, if applicable, related work is described. We follow with the method and experiments for the component. Finally, each chapter concludes with a brief summary and future work.

In Part III of this thesis, we describe the system for detecting influencers. First, in Chapter 10, we describe the features generated from each component in Part II. Then, we discuss the method we use for influence detection and domain adaptation in Chapter 10 and our experiments and results in Chapter 11. This includes evaluating the impact of each individual component, all components, and the best components per genre, using all genres, and with domain adaptation. The final Chapter (12) in Part III examines the impact of the weapon of social proof in greater detail within Wikipedia Talk

Pages.

The conclusion of this thesis (Part IV), summarizes our main contributions and discusses future directions for this research area. The future directions are analyzing trends, influencers and causality, social network and context, and author profiling.

Finally, Part VI is the appendices. First, where we describe the annotation manuals for several of our system components as well as influencer detection (Appendix A), then we provide the locations to access our datasets and code resources (Appendix B), and finally a glossary of relevant terms (Appendix C).

Chapter 2

Data

> *There were 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.*
>
> Eric Schmidt, *at Google's 2010 Atmosphere Convention*

In this chapter we describe the main datasets used throughout the thesis. Here, we describe the type of datasets, and leave comprehensive details and statistics related to individual components as well as influence detection to future chapters. Our main datasets are LiveJournal, Wikipedia Talk Pages, Create Debate, Political Forum, and Twitter. A screenshot of a discussion from each dataset is shown in Figure 2.1. Online discussions can vary significantly in style, content, and audience. Our datasets target these variations to obtain a more comprehensive analysis of online social media. We include weblogs (LiveJournal), microblogs (Twitter), classic discussion forums (Wikipedia Talk Pages, Political Forum), and debate style discussion forums (Create Debate). Some of our datasets vary on topic, but we also specifically focus on data related to politics (Political Forum, Create Debate, Twitter). The demographics, such as common age and gender, of the datasets can vary as well. LiveJournal bloggers tend to be women and Wikipedia Talk Page contributors tend to be men. Twitter users tend to be younger.

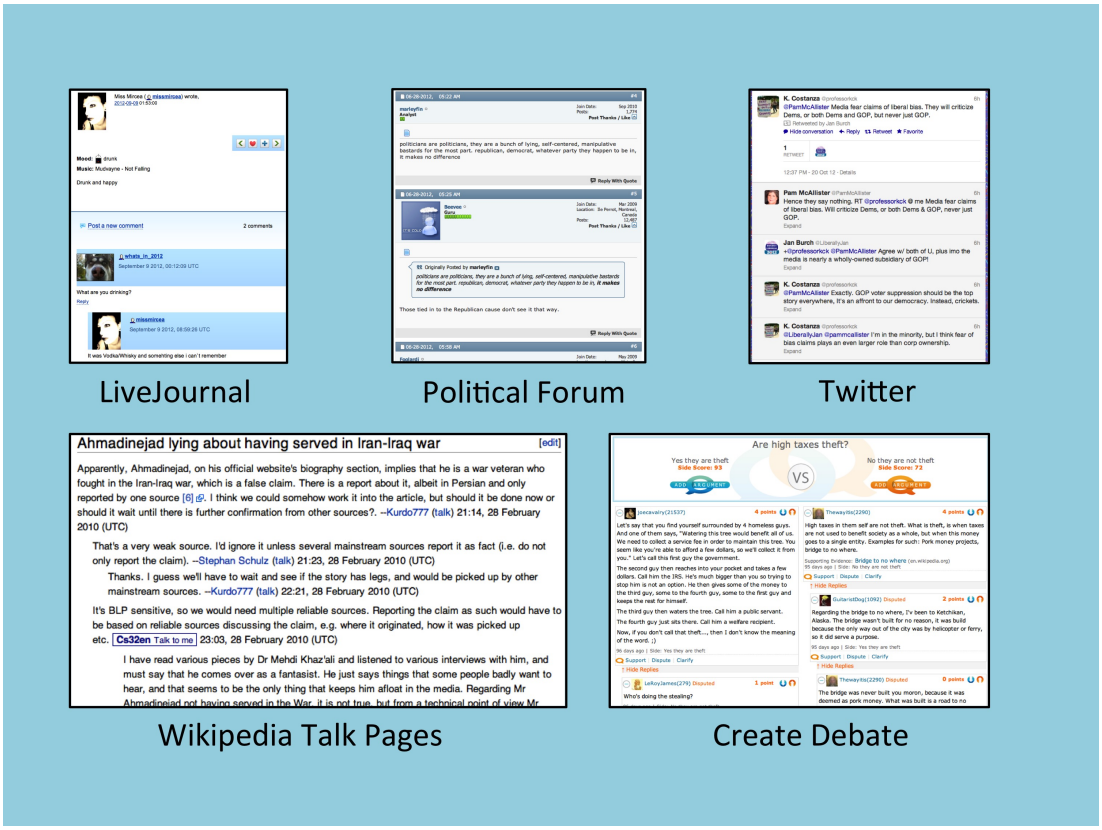


Figure 2.1: A screenshot of a discussion from each dataset

2.1 LiveJournal

We have explored LiveJournal weblogs for Influence and several of its components (sentiment, claim, agreement, and author traits). LiveJournal is a virtual community in which people write about their personal experiences in a weblog and can typically be considered as a public diary. A LiveJournal entry is composed of a post (the top-level content written by the author) and a set of comments (written by other users and the author). Every comment structurally descends either from the post or from another comment. A key advantage to LiveJournal in comparison to other web-logging sites is that the blogs tend to be public, the user profile pages are rich in information such as age and geographic location, and they are easily downloadable via a well structured xml format.

An example of influence in a LiveJournal blog is shown in Table 2.1. The discussion is regarding the “Mexico City Policy”. Poconell, the owner of the blog is the influencer. The other participants agree with Poconell, making him the most influential in the discussion. The owner of the blog is

P1 by poconell <pc_1>He really does make good on his promises! </pc_1><pa_1>Day three in office, and the Global Gag Rule (A.K.A"The Mexico City Policy") is gone!</pa_1>I was holding my breath, hoping it wouldn't be left forgotte. He didn't wait. <pc_2>He can see the danger and risk in this policy, and the damage it has caused to women and families.</pc_2><pc_3>I love that man!</pc_3>
P2 by thalialunacy <a_1>I literally shrieked 'HELL YES!' in my car when I heard. :D:D:D</a_1>
P3 by poconell <a_2>Yeah, me too</a_2>
P4 by lunalovepotter <pc_4><a_3>He is SO AWESOME!</a_3></pc_4><pa_4>Right down to business, no ifs, ands, or buts! :D</pa_4>
P5 by poconell <pc_5>It's amazing to see him so serious too!</pc_5><pa_5>This is one tough, no-nonsense man!</pa_5>
P6 by penny_sieve My icon says it all :)
P7 by poconell <pc_6>And I'm jealous of you with that President!</pc_6><pa_6>We tried to overthrow our Prime Minister, but he went crying to the Governor General. </pa_6>

Table 2.1: **LiveJournal Influence Example**: A LiveJournal discussion thread displaying poconell as the influencer. Replies are indicated by indentation (for example, P2 is a response to P1). The following components are visible in this example: aAttempt to Persuade ($\{pc_i, pa_i\}$), Claims (pc_i), Argumentation (pa_i), Agreement (a_i), Disagreement (d_i), and the five Dialog Patterns Language Uses

often the influencer in LiveJournal because people will rarely disagree with someone in their personal space.

2.2 Wikipedia Talk Pages

Wikipedia is the well known and largest free collaborative encyclopedia on the web. Each Wikipedia page has a discussion forum associated with it where proposed edits for the page are discussed. Therefore, it tends to be argumentative because each person is trying to get his/her idea included in the page, making it an ideal choice for studying influence detection. Wikipedia differs from other online social media in that it tends to be well written. The posts in a Wikipedia talk page thread may or may not structurally descend from other posts: direct replies to a post typically descend from it.

Other posts can be seen as descending from the topic of the thread. One of our main data sources, we use Wikipedia talk pages in claim, sentiment, agreement, argumentation, persuasion, and influence detection.

An example of an influencer in a Wikipedia Talk Page discussion is shown in Table 2.2. The discussion is regarding leukemia classification on the Wikipedia lymphoma page. A participant (Arcadian) starts the thread with a proposal and a request for support from other participants. Emmanuelm later joins the conversation arguing against Arcadian's proposal. There is a short discussion, and Arcadian defers to Emmanuelm's position. Therefore, the influencer is EmmanuelM.

2.3 Create Debate

Create Debate is a unique kind of discussion forum whose format is a debate style where people can post on either a *for* or *against* side of an argument. The posts are separated according to the side chosen. We have used discussions from Create Debate to detect agreement and influence. The discussions used in agreement detection are not topic specific. We restricted the data used in influence detection to discussion related to politics.

An example of an influencer in a Create Debate discussion is shown in Table 2.3. The debate is about investigating torture claims against President Bush. There are two influencers in this example: CupioMinimus and Libertarian1. Libertarian1 appears to be the leader of the *against* side of the debate as three people like his post. CupioMinimus appears to be the leader of the *for* side of the debate. People like his two posts and he also provides supporting evidence.

2.4 Political Forum

Political Forum is a discussion forum website used to discuss political topics. It follows a threading structure similar to Wikipedia that is better enforced; direct replies descend from individual posts and other posts can be seen as descending from the topic of the thread. The website was crawled to gather forums discussing general politics as well as politics related to the Republican primary during the time period of November 2011-present. Political Forum data is only used in influence detection.

An example of an influencer in a Political Forum discussion is shown in Table 2.4 where user Polly Minx makes a statement in support of unions and is initially met with disagreement. Polly

Minx defends her statements, and receives additional support from toddwv. Additionally Polly Minx is quoted by both opposition and support. Therefore, Polly Minx is the influencer.

2.5 Twitter

Twitter is the most popular micro-blogging site on the web where people write short blurbs up to 140 characters. A threading structure can be generated using tweets via retweets (tweeting something written by someone else) and mentions (mentioning another user). We have gathered Twitter datasets that are being used as training in our opinion and author trait detection systems as well as for influence detection. The website was crawled to gather tweets discussing general politics as well as politics related to the Republican primary during the time period of November 2011-present for influence detection. The author trait data was gathered in November 2014 using users that self labeled their political party and religion on Twitter directory sites such as tweepz.com and twellow.com. The sentiment data ranges from several time periods. Those tweets were filtered to ensure at least one opinionated word existed by looking up the words in SentiWordNet [Baccianella *et al.*, 2010]. The sentiment dataset is discussed in further detail in the Sentiment in Twitter Semeval task description paper [Nakov *et al.*, 2013; Rosenthal *et al.*, 2014c].

A Twitter conversation is shown in the Table 2.5, displaying professorkck as the influencer. professorkck is the influencer because he sends a message to PamMcAllistar about the media's fears of bias which receives a response and others agree with it as well. In addition, several people retweet his messages.

P1 by Arcadian $<pc_1>$There seems to be a much better list at the National Cancer Institute than the one we've got.$</pc_1><pa_1>$It ties much better to the actual publication (the same 11 sections, in the same order).$</pa_1>$I'd like to replace that section in this article. Any objections?
P2 by JFW $<pc_2><a_1>$Not a problem.$</a_1></pc_2>$Perhaps we can also insert the relative incidence as published in this month's wiki Blood journal
P3 by Arcadian I've made the update. I've included template links to a source that supports looking up information by ICD-O code.
P4 by Emmanuelm Can Arcadian tell me why he/she included the leukemia classification to this lymphoma page? It is not even listed in the Wikipedia leukemia page! $<pc_3>$I vote for dividing the WHO classification into 4 parts in 4 distinct pages: leukemia, lymphoma, histocytic and mastocytic neoplasms.$</pc_3><pa_3>$Remember, Wikipedia is meant to be readable $</pa_3>$ by all. Let me know what you think before I delete the non-lymphoma parts.
P5 by Arcadian Emmanuelm, aren't you the person who added those other categories on 6 July 2005?
P6 by Emmanuelm $<d_1>$Arcadian, I added only the lymphoma portion of the WHO classification. You added the leukemias on Dec 29th.$</d_1>$Would you mind moving the leukemia portion to the leukemia page?
P7 by Emmanuelm $<pc_4>$Oh, and please note that I would be very comfortable with a "cross-coverage" of lymphocytic leukemias in both pages.$</pc_4>$My comment is really about myeloid, histiocytic and mast cell neoplasms who share no real relationship with lymphomas.
P8 by Arcadian $<pa_5><a_2>$To simplify the discussion, I have restored that section to your version. $</a_2></pa_5>$You may make any further edits, and $<pc_6>$I will have no objection.$</pc_6>$
P9 by JFW The full list should be on the hematological malignancy page, and the lymphoma part can be here. $<pc_7>$It would be defendable to list ALL and CLL here.$</pc_7><pa_7>$They fall under the lymphoproliferative disorders.$</pa_7>$

Table 2.2: **Wikipedia Talk Page Influence Example**: A Wikipedia discussion thread displaying Emmanuelm as the influencer. Replies are indicated by indentation (for example, P2 is a response to P1). The following components are visible in this example: Attempt to Persuade ($\{pc_i, pa_i\}$), Claims (pc_i), Argumentation (pa_i), Agreement (a_i), Disagreement (d_i), and the five Dialog Patterns Language Uses (eg. Arcadian has positive Initiative).

P1 by KingofPopForever Obama Urged to Investigate Bush Torture Claims. Should He?
P2 by Libertarian1 (No 3) $<pc_1>$While im sure liberals would love for that to happen, it simply will do no good. $</pc_1><pa_1>$you'd have to put on trial every military organization that either took part in such a crime being commited. $</pa_1><pc_2>$And we all know the governent doesn't rat itself out. $</pc_2>$
P3 by chatturgha (Yes 1) $<d_1>$While he's at it, he should investigate the possible tens of thousands of innocent Iraqi civilians that were murdered during the second Iraq war, all on Bush's hands. $</d_1><pc_3>$Honestly, I believe in torture... but only in torture of the deserving. $</pc_3>$
P4 by garry77777 (No 1) "he should investigate the possible tens of thousands of innocent Iraqi civilians that were murdered during the second Iraq war, all on Bush's hands." $<d_1>$I must disagree with your numbers, best estimates put the actual number at 1.2 million people. $</d_1>$
P5 by chatturgha (Yes 2) Okay then, he killed MORE people then just tens of thousands. And you're disagreeing with me... why?
P6 by CupioMinimus (Yes 1) $<pc_4>$Of course he should but he won't.$</pc_4><pa_4>$No one gets into power in the west unless the real PTB have got leverage on them. That's why none of our leaders do anything to rock the boat. $</pa_4><pc_5>$Stray but a little and it's character assassination.$</pc_5>$Not always with 'character' either ;]
P7 by CupioMinimus (Yes 3) President Barack Obama is obliged to order a criminal investigation under the Convention against Torture to which the US is a party. A 107 page report by Human Rights Watch presents evidence warranting criminal investigations of Bush and senior officials, for ordering acts of torture. Supporting Evidence: Getting Away With Torture (*LINK*)
P8 by ThePyg (No 1) $<d_2>$While I disagree with many aspects of the war, waterboarding, to me, shouldn't be something that's "investigated" as "torture". $</d_2>$Our military have done what they can to protect the US citizens. Sure, $<pc_6>$I don't think they did it right$</pc_6>$, $<pa_6>$but to punish them for all they've done for OUR protection is... disturbing. $</pa_6>$
P9 by Phreekshow (Yes 1) I do not look at it as a mark against the military who were doing what they were ordered to do by the Commander in Chief. $<pc_7><d_2>$maybe if Americans were able to experience waterboarding they would change their minds on whether it is torture. $</d_2>$ $</pc_7>$

Table 2.3: **Create Debate Influence Example**: A Create Debate discussion thread displaying CupioMinimus and Libertarian1 as the influencer. Replies are indicated by indentation (ex: P2 is a response to P1). In addition, each post also indicates its side in the debate (Yes/No) and the number of likes associated with the post. The following components are visible in this example: Attempt to Persuade ($\{pc_i, pa_i\}$), Claims (pc_i), Argumentation (pa_i), Disagreement (d_i), and the five Dialog Patterns.

P1 by Polly Minx This article mentions (some) contents of our Second Bill of Rights proposals including our larger-than-expected turnout (<a_1>an estimated 35,000 </a_1>), Schultz's remarks and the reaction to the announcement that Mitt Romney had selected Paul Ryan as his VP nominee.
P2 by Leatherface <d_1><pc_1>When your ideas are on the fringe of society, and far beyond the bounds of common sense, don't expect to be welcomed, even by the drooling retards of todays Democrap party.</d_1></pc_1><pc_2><d_5>Your ideas are incompatible with liberty, </pc_2> </d_5>
P3 by Polly Minx <$d_{1,2,4}$><pc_3>35,000 attendees do not a "fringe" constitute. </pc_3> </$d_{2,4}$><pc_4>Neither does the AFL-CIO constitute a "fringe" organization. </pc_4></d_1>
P4 by Leatherface Trumpka is a Mafia Don, and 35,000 is about the size of his band of parasites. <d_2>Not impressed, Polly.</d_2>
P5 by Yosh Shmenge <pc_5><a_1>The thugs at the AFL-CIO "claim" the mob's size at 35,000. </a_1><pc_5><pa_5>Read the OP.</pa_5><pc_6>The real figure is undoubtedly less. </pc_7><pc_8>Big Labor is an enemy of the people and they are Obama's lap dog.</pc_8>
P6 by toddvw <d_3><pc_9>The plight of the middle-class is tied to unions. </pc_9> </d_3><pc_{10}>Most of us don't like the idea of working for $5.00 a day with no benefits <pa_{11}>which seems to be the right-wing's wetdream.</pa_{11}></pc_{10}>
P7 Mac-7 <d_3><pc_{12}>Actually the fate of the middle class is tied to jobs that have moved to China and Mexico <pc_{12}></d_3>thanks in large part to Bill Clinton.
P10 by fiddlerdave <d_4>We have a thread here on PF ballyhooing that maybe 500 people showing up fo Romney/Ryan in Florida was a sign of "Major Excitement" over the new ticket! </d_4>:lol: <pc_{15}>Republicans are SOOOOO pathetically desperate! </pc_{15}>
P11 by toddvw <d_5><pc_{16}>Quite the opposite, unions have been a vessel for freedom for quite some time now.</pc_{15}></d_5><pa_{16}>Otherwise, we end up with wage slaves and child labor like back in the 1800s/early 1900s. </pa_{16}>Do you like vacation pay and sick leave? Would you prefer to work for in horrifying work conditions with your 12 year old son laboring you?
P12 by Mac-7 If libs expect government to provide them with full employment and a living wage why do they support politicians like Obama who wants to flood the country with Mexican refugees? The far lefties make so sense. </pc_{17}>

Table 2.4: **Political Forum Influence Example**: A Political Forum discussion thread displaying Polly Minx as the influencer. Replies are indicated by indentation (ex: P2 is a response to P1). The following components are visible in this example: Attempt to Persuade ($\{pc_i, pa_i\}$), Claims (pc_i), Argumentation (pa_i), Agreement (a_i), Disagreement (d_i), and the five Dialog Patterns Language Uses (eg. Polly Minx has the Initiative).

P1 by professorkck $<a_{1,4}>$@PamMcAllister $<pc_1>$Media fear claims of liberal bias. $</pc_1><pc_2>$ They will criticize Dems, or both Dems and GOP, but never just GOP $</pc_2></a_{1,4}>$
P2 by PamMcAllister $<a_1>$Hence they say nothing. RT @professorkck @ me Media fear claims of liberal bias. Will criticize Dems, or both Dems & GOP, never just GOP. $</a_1>$
P3 by LiberallyJan +@professorkck @PamMcAllister $<a_1>$Agree w/ both of U,$</a_1><a_2>$ $<pc_3>$plus imo the media is nearly a wholly-owned subsidiary of GOP! $</pc_3></a_2>$
P4 by professorkck $<a_3>$RT @PamMcAllister $<a_2><pc_4>$Exactly. GOP voter suppression should be the top story everywhere,$</pc_4></a_2><pa_4>$It's an affront to our democracy. Instead, crickets. $</pa_4></a_3>$
P5 by professorkck @LiberallyJan @pammcallister I'm in the minority, but I think fear of bias claims plays an even larger role than corp ownership.
P6 by hejjet $<a_3>$RT @professorkck: @PamMcAllister Exactly. GOP voter suppression should be the top story everywhere, It's an affront to our democracy. Instead, crickets. $</a_3>$
P7 by ATridentTruth $<a_3>$RT @professorkck: @PamMcAllister Exactly. GOP voter suppression should be the top story everywhere, It's an affront to our democracy. Instead, crickets. $</a_3>$
P8 by PamMcAllister, $<a_3>$RT @professorkck: @PamMcAllister Exactly. GOP voter suppression should be the top story everywhere, It's an affront to our democracy. Instead, crickets. $</a_3>$
P9 by paul_paule1 $<a_3>$RT @professorkck: @PamMcAllister Exactly. GOP voter suppression should be the top story everywhere, It's an affront to our democracy. Instead, crickets. $</a_3>$
P10 by phyl1127 $<a_3>$RT @professorkck: @PamMcAllister Exactly. GOP voter suppression should be the top story everywhere, It's an affront to our democracy. Instead, crickets. $</a_3>$
P11 by professorkck @LiberallyJan @pammcallister $<pc_5>$GOP knows how to play media. $</pc_5>$
P12 by LiberallyJan +@professorkck @PamMcAllister $<pc_6>$Whatever the reason, the "media" is virtually useless for actual journalism. $</pc_6>$WE must be the media now!
P13 by LiberallyJan $<a_4>$RT @professorkck: @PamMcAllister Media fear claims of liberal bias. They will criticize Dems, or both Dems and GOP, but never just GOP. $</a_4>$

Table 2.5: **Twitter Influence Example**: A Twitter discussion displaying professorkck as the influencer. Replies are indicated by indentation (ex: P2 is a response to P1).The following components are visible in this example: Attempt to Persuade ($\{pc_i, pa_i\}$), Claims (pc_i), Argumentation (pa_i), Agreement (a_i), and the five Dialog Patterns.

Part II

Components of Influence

Chapter 3

Introduction

In this portion of the thesis we describe each of our system components. The output of these components are used to detect the influencer in a discussion. Our main components are:

Author Traits are characteristics derived from demographics and political affiliation
Agreement and Disagreement are the agreement, disagreement, or lack of between two posts
Claims are sentences that contain an opinionated assertion indicating the truth of something
Argumentation is justification towards a claim
Persuasion is a Claim followed by Argumentation
Credibility is an indication of authority or the lack of credibility
Dialog Patterns are based on the thread structure of the discussion

We also describe two subcomponents which are used within several other components:

Lexical Style includes features based on general and social media contextual structure
Opinion Detection includes subjectivity and polarity detection

As described in the introduction, we motivate influence detection through Robert Cialdini's weapons of influence [Cialdini, 2007]: Reciprocation, Commitment and Consistency, Social Proof, Liking, Authority, and Scarcity. In the following chapters, motivation for including each component in influence detection is provided by one or more weapons of influence. We also explore multiple online genres in several components (e.g. claims and agreement) as well as comparing and contrasting

the differences between online genres.

In the rest of this part of the thesis, we first describe the subcomponents. We discuss lexical-style in the following section and opinion in Chapter 4. Next, we discuss agreement in Chapter 5. Then, we discuss claim, argumentation, and persuasion in Chapter 6. We follow with author traits in Chapter 7. Finally we conclude with the direct feature components, credibility and dialog patterns in Chatper 8.

3.1 Lexical Style

General		Social Media	
Feature	**Example**	**Feature**	**Example**
Capitalized Words	Hello	Emoticons	:)
Out of Vocabulary	duh	Acronyms	LOL
Punctuation	.	Repeated Questions	???
Repeated Punctuation	#@.	Exclamation Points	!
Punctuation Count	5	Repeated Exclamations	!!!!
Question Marks	?	Word Lengthening	sweeeet
Ellipses	...	All Caps	HAHA
Avg Word Length	5	Links/Images	www.url.com

Table 3.1: List of lexical-stylistic features and examples. These features are computing based on the normalized occurrence within a body of text. The features can be divided into types: **general**: ones that are common across online and traditional genres, and **social media**: one that are far more common in online genres

We include several lexical-stylistic features (see Table 3.1) that can occur in all datasets. We divide these features into two groups, **general**: ones that are common across online and traditional genres, and **social media**: one that are far more common in online genres. Examples of general style features are exclamation points and ellipses. Examples of social media style features are emoticons and word lengthening. Word lengthening is a common phenomenon in social media where letters are repeated to indicate emphasis (e.g. sweeeet). It is particularly common in opinionated words [Brody and Diakopoulos, 2011]. Even though social media features are common in online genres when compared to newswire, they still occur infrequently accounting for less than 2% of the words in each sentence. This means that most sentences have 0-1 social media features. We use these lexical-stylistic features throughout several system components, including demographics, sentiment, agreement, and claim detection.

The social media features are indicative of the informal writing style that occurs in online discussions. We hypothesize that this writing style can be used to distinguish between people of different author traits, different sort of opinions, as well as agreement and disagreement. Similarly,

general lexical style can be indicative as well. For example, questions can indicate a lack of claim and a lack of agreement. The use of such features will be motivated further in each chapter.

Chapter 4

Opinion

> *Success is a lousy teacher. It seduces smart people into thinking they can't lose.*
>
> Bill Gates, *The Road Ahead (1995)*

In online forums and micro-blogs, people write from the heart about their personal experiences, likes and dislikes. The following sentence from Twitter is a typical example: "*Tomorrow I'm coming back from Barcelona...I don't want! :(((*". The ability to detect the sentiment expressed in social media can be useful for understanding what people think about the restaurants they visit, the movies they see, the political viewpoints of the day, and the products they buy. These sentiments can be used to provide targeted advertising, automatically generate reviews, and make various predictions such as political outcomes and the success of a product.

In this chapter, we describe a sentiment detection algorithm for social media that classifies sentence phrases for subjectivity {subjective,objective} and polarity {positive,negative,neutral}. We test the performance of our system in four genres: Twitter, LiveJournal, Wikipedia Talk Pages, and classic newswire articles via the MPQA [Wiebe *et al.*, 2005]. Our system builds on previous work in the development of sentiment detection algorithms for the more formal news genre, notably the work of Agarwal *et al*, [2009] but we adapt it for the language of social media.

Opinion detection is a fundamental system used in several components of detecting influence. Its usefulness is evident in several weapons of influence. For instance, speaking positive compliments and praise are likely to cause **reciprocation**. The role of opinion in reciprocation is strong motivation

for including subjectivity and polarity in agreement detection. In addition, people are more likely to be influenced by someone that they **like**, even if the compliments are false [Drachman *et al.*, 1978]. In fact, on an even more granular level it has been found that the pronoun *"we"* indicates positive association and *"they"* indicates negative association [Cialdini *et al.*, 1976] which is motivation for using part of speech (POS) as features in opinion detection.

In the rest of this chapter, we describe related work, the data and annotation process, our method, and experiments and results. We also describe our participation in public evaluations focused on sentiment analysis in Twitter (Semeval: Sentiment Analysis in Twitter Task [Nakov *et al.*, 2013; Rosenthal *et al.*, 2014c]) and sentiment towards a topic (TAC KBP 2013[1],2014[2]). We conclude with a discussion analyzing several publicly available dictionaries as possible improvements to our current system.

4.1 Related Work

There has been a significant amount of work on sentiment detection. There has been work that has explored subjectivity {subjective,objective} and polarity {positive,negative,neutral} in newswire articles and online genres.

The earliest work explored opinion detection in newswire articles (e.g. [Yu and Hatzivassiloglou, 2003; Wilson *et al.*, 2005; Kim and Hovy, 2004]), and reviews such as product and movie reviews [Pang and Lee, 2004; Turney, 2002; Beineke *et al.*, 2004]. This research built a strong foundation for sentiment detection in social media. Although the newswire and social media genres are very different due to informal language, the underlying methodology remains the same. Such methodologies include sentiment annotation, developing sentiment lexicons [Wiebe *et al.*, 2005], handling negation [Kim and Hovy, 2004] and classification using supervised learning.

More recently, sentiment detection has been explored in online discussions such as weblogs, discussion forums, and microblogs. There is also work on targeted sentiment by exploring sentiment analysis towards a topic or entity. Finally, there is a closely related area of work called emotion detection [Mishne, 2005; Yang *et al.*, 2007] where several emotions are categories (e.g. happy,

[1]http://www.nist.gov/tac/2013/

[2]http://www.nist.gov/tac/2014/

sad) which has some similarities with sentiment detection. The main difference between sentiment detection is that emotion detection involves many more categories some of which are positive (e.g. excited), negative (e.g. disgusted), and neutral (e.g. bored). In the rest of this section we focus on the related work in online genres in more depth due to its closer relationship and relevance to our work.

In general, early approaches to sentiment detection in weblogs and discussion forums tend to classify the entire document (blog post, or discussion forum). The majority of these papers use lexical features [Godbole *et al.*, 2007; Yu and Kübler, 2011; Mei *et al.*, 2007], but only a few of the early papers gear their system towards social context (e.g. [Chesley *et al.*, 2006]). Chesley *et al* [2006] classify the polarity and subjectivity of verbs and adjectives in blog posts and use lexical features, POS, and Wiktionary. Mejova *et al* [2012] explore polarity in blogs, reviews, and Twitter. On the surface this work appears to be similar to ours, however, they use an off the shelf sentiment system and do not adapt it to online genres in any way. They run different experiments to analyze how adaptable the sources are to one another which is similar to our cross-genre experiments. Read [2005] explores the dependency of domain, topic, and time on polarity classification in usenet newsgroups, which are similar to discussion forums. They overcome this by using sentences labeled with emoticons which tend to be independent of domain, topic, and time.

Several recent papers have explored sentiment analysis in Twitter [Go *et al.*, 2009; Barbosa and Feng, 2010; Bermingham and Smeaton, 2010; Agarwal *et al.*, 2011; Pak and Paroubek, 2010]. Go *et al* [2009] classify tweets containing emoticons as positive or negative using n-grams and POS. Pak and Paroubek [2010] use a similar approach, but they also classify subjectivity. Barbosa and Feng [2010] detect subjectivity and polarity using a polarity dictionary that they extended to include web vocabulary and tweet-specific social media features. Bermingham and Smeaton [2010] compare polarity detection in Twitter to blogs and movie reviews using lexical features such as n-grams and POS. Agarwal *et al* [2011] perform polarity sentiment detection on the entire tweet using features that are somewhat similar to ours: the DAL, lexical features (e.g. POS and n-grams), and social media features (e.g. slang and hashtags) and classify their input text using tree kernels. The most recent work is based on the public SemEval evaluations [Nakov *et al.*, 2013; Rosenthal *et al.*, 2014c]. These evaluations have explored sentiment analysis in Twitter on the phrase-level and message-level. Most notably, the best system in 2013, NRC-Canada [Mohammad *et al.*, 2013], used supervised learning with features such as POS tags trained on Twitter [Owoputi *et*

al., 2013], n-grams, negation, position, and length, and built a Twitter hashtag lexicon achieving an F-score of 88.93 in phrase-level polarity and an F-score of 69.02 in message-level polarity. They were also the best team in the phrase-level task in 2014. In 2014, the best system in the message level task was TeamX [Miura *et al.*, 2014]. They followed a similar approach to NRC-Canada, but included more external lexicons and word sense disambiguation features.

There has been previous work that focuses on sentiment towards an entity or topic. This can be useful for influence detection to infer that the people in the conversation share an opinion on the same things. One such system is Godbole *et al* [2007] where they determine whether the sentiment towards an entity within a corpus is positive or negative and how it changes over time. Mei *et al* [2007], model sentiment towards the main topics in a document. Jiang *et al* [2011] perform sentiment towards a topic in Twitter. Nasukawa and Yi [2003] capture sentiment towards the topics in a document by exploring all entities where there is a semantic relationship. They explore the dependency between two entities to determine their semantic relationship. There have been evaluations focusing on sentiment towards a topic in TAC KBP's 2013 and 2014 Sentiment Slot Filling Task [3]. Our participation in these tasks is described at the end of this chapter. Sentiment towards a topic was also included in the most recent Semeval Sentiment Analysis in Twitter task in 2015 [Rosenthal *et al.*, 2015]. In this task the participants were given a topic and asked to provide the polarity of the tweet in relation to the topic. For example, in the tweet "Saturday without Leeds United is like Sunday dinner it doesn't feel normal at all (Ryan)" the overall sentiment of the tweet is negative, but the sentiment towards the topic *Leeds United* is positive. The best results for this task was 50.51% indicating it is considerably more challenging than detecting the overall sentiment where the best team achieved an F-score of 64.84%.

4.2 Data

Our opinion corpus consists of four datasets; MPQA [Wiebe *et al.*, 2005], LiveJournal [Rosenthal *et al.*, 2014c], Twitter [Nakov *et al.*, 2013; Rosenthal *et al.*, 2014c], and Wikipedia Talk Pages. The LiveJournal and Wikipedia datasets consists of 2,200 sentences , the MPQA consists of over 6000 labeled sentences and the Twitter dataset contains over 10,000 tweets. Further statistics for

[3] http://www.nist.gov/tac/2013/,http://www.nist.gov/tac/2014/

Corpus	Positive	Negative	Neutral	Subjective	Objective
Twitter	7897	4364	647	22885	11869
LiveJournal	1323	1095	369	3280	5132
Wikipedia	947	1283	469	2900	4937
MPQA	2076	4578	5992	15002	22948

Table 4.1: Dataset statistics for sentiment phrases.

Source	Example
Twitter	Why would you [still]- wear shorts when it's this cold?! I [love]+ how Britain see's a bit of sun and they're [like 'OOOH]+ LET'S STRIP!'
LiveJournal	[Cool]+ posts , dude ; very [colorful]+ , and [artsy]+ .
MPQA	Large-scale industry [wishes to]o counter the [expected rise]o with so-called benchmark covenants.
Wikipedia	I am [not]- the one who put in any [garbage]- about occupying regime.

Table 4.2: Example of polarity for each source of messages. The target phrases are marked in [. . .], and are followed by their polarity (positive (+), negative (-), or neutral (o).

each dataset are shown in Table 4.1. At least 10% of the data from each corpus was held out for testing. The sentences have been annotated for subjectivity {subjective,objective} and polarity {positive,negative,neutral} on the phrase and sentence level. An example of an annotated sentence in each dataset is shown in Table 4.2. Each sentence was annotated on Mechanical Turk by 3-5 workers at 3-5 cents a hit. The annotations for each sentence were then combined using intersection where the word had to be marked the same way by 2/3 of the workers in order to be kept with the provided subjectivity and polarity. Further details regarding the annotation process including instructions and examples can be found in Appendix A.2.

4.3 Lexicons

Several lexicons are used in our system. We use the Dictionary of Affect and Language (DAL) to assign polarity and subjectivity scores to each word. However, the DAL only contains 8742 terms. In

order to increase coverage we augment it with several other lexicons: WordNet, Wiktionary, and an in-house emoticon lexicon. We also use the SentiWordNet for polarity and subjectivity scores. These lexicons are described in the rest of this section. Their usage is explained in the following Features section.

4.3.1 DAL

The Dictionary of Affect and Language (DAL) [Whissel, 1989] is an English language dictionary of 8742 words built to measure the emotional meaning of texts. In addition to using newswire, it was also built from individual sources such as interviews on abuse, students' retelling of a story, and adolescent's descriptions of emotions. It therefore covers a broad set of words. Each word is given three scores (pleasantness - also called evaluation (w_{ee}), activeness (w_{aa}), and imagery (w_{ii})) on a scale of 1 (low) to 3 (high). We explore whether a phrase is subjective or objective. We compute subjectivity and polarity for each chunk phrase, c, using the DAL by computing the sum of the AE Space Score's ($|\sqrt{{w_{ee}}^2 + {w_{aa}}^2}|$) [Agarwal *et al.*, 2009] of each word, w, within the chunk.
Subjectivity is computed as:

$$\text{sub}(c) = \begin{cases} \text{objective} & \text{if } \sum_{w=1}^{n} |\sqrt{{w_{ee}}^2 + {w_{aa}}^2}| < \alpha \\ & \text{and } \sum_{w=1}^{n} w_{ii} > 1 \\ \text{subjective} & \text{otherwise} \end{cases}$$

Polarity is computed as:

$$\text{pol}(c) = \begin{cases} \text{negative} & \text{if } \sum_{w=1}^{n} \sqrt{{w_{ee}}^2 + {w_{aa}}^2} < \alpha \\ \text{positive} & \text{if } \sum_{w=1}^{n} \sqrt{{w_{ee}}^2 + {w_{aa}}^2} > \alpha \\ \text{neutral} & \text{otherwise} \end{cases}$$

In other words, a chunk phrase is considered subjective (sub(c)) if it is positive, negative, or neutral, and has low imagery (ii) compared to a given threshold. The threshold, $\alpha = 0.7$, was computed empirically in prior work [Agarwal *et al.*, 2009] A chunk phrase is considered positive if it has a high AE Space Score and negative if it has a low AE space score compared against the threshold α.

4.3.2 WordNet

The DAL does cover a broad set of words, but we will still often encounter words that are not included in the dictionary. We address this by augmenting it with additional lexicons. The first being WordNet [Fellbaum, 1998]. WordNet organizes words based on senses. In addition to being a dictionary and thesaurus, it maintains words relations such as synonymy, hyponymy, and antonymy. We follow the original approach [Agarwal *et al.*, 2009] and exploit these relationships to improve coverage of the DAL.

If a word exists in WordNet, the DAL scores of the synonyms of its first sense are used in its place. In addition to the original approach, if there are no synonyms we look at the hypernym. We then compute the average scores ($ee, aa,$ and ii) of all the words and use that as the score for the word.

4.3.3 Wiktionary

Wiktionary is an online dictionary maintained by volunteers. In contrast to WordNet, Wiktionary includes informal terms and acronyms that are common in social media. This makes it an ideal additional dictionary to supplement the more formal words that are found in WordNet and the DAL.

We first examine all "form of" relationships for the word, such as "doesnt" is a "misspelling of" "doesn't", and 'tonite" is an "alternate form of" "tonight". If no "form of" relationships exist, we take all the words in the definitions that have their own Wiktionary page and look up the scores for each word in the DAL. (e.g., the verb definition for *LOL* (laugh out loud) in Wiktionary is *"To laugh out loud"* with *"laugh"* having its own Wiktionary definition; it is therefore looked up in the DAL and the score for "laugh" is used for *"LOL"*.) We then compute the average scores ($ee, aa,$ and ii) of all the words and use that as the score for the word.

4.3.4 Emoticon Dictionary

emoticon	:)	:D	<3	:(	;)	:P	:-)	^_^	:/	=)
definition	happy	laughter	love	sad	wink	silly	happy	happy	crying	surprised

Table 4.3: Popular emoticons and their definitions in the DAL

Corpus	DAL	NNP (Post DAL)	Word Length-ening	Word-Net	Wiktionary	Emoticons	Punctuation & Numbers	Not Cov-ered
Twitter	17.5%	43.7%	0.9%	6.1%	5.4%	0.1%	1.3%	25.0%
LiveJournal	65.7%	3.1%	1%	12.3%	8.2%	0.3%	1.4%	8.1%
MPQA	54%	6.3%	0.1%	19.5%	9%	0.0%	1%	10.2%
Wikipedia	68.1%	2.8%	0.1%	13.7%	7.1%	0.0%	1.4%	6.7%

Table 4.4: Stacked coverage (words are searched for in each lexicon until it is found from left to right starting with the DAL) of the lexicons in the training corpora.

We created a lexicon to map common emoticons to a definition in the DAL. We looked at over 1000 emoticons gathered from several lists on the internet[4] and computed their frequencies within a LiveJournal blog corpus. (In the future we would like to use an external Twitter corpus as well due to the high frequency of emoticons on Twitter). We kept the 192 emoticons that appeared at least once in the LiveJournal corpus and manually mapped each emoticon to a single word definition. The top 5 emoticons and their definitions are shown in Table 4.3. When an emoticon is found in a sentence we look up its definition in the DAL.

4.3.5 SentiWordNet

SentiWordNet [Baccianella *et al.*, 2010] is a sentiment lexicon built on top of WordNet. SentiWordNet assigns three scores to each word in WordNet: positive, negative, and objective. The scores range from [0.0,1.0]. 0.0 indicates a completely negative score and 1.0 indicates a completely positive score. The sum of the three scores is 1 for each word. We use SentiWordNet in addition to the DAL as an alternate score.

4.4 Features

Each sentence is preprocessed for POS tags [Phan, 2006b] and chunks [Phan, 2006a]. We use the DAL along with the other lexicons and a negation state machine to determine the subjectivity or

[4] www.chatropolis.com, www.piology.org, en.wikipedia.org

polarity of a chunk. If a word is a proper noun and is not found in the DAL we automatically mark it as objective. Otherwise, we look up emoticons in the emoticon lexicon and look up all other words that are not found in WordNet followed by Wiktionary in order to find synonyms within the DAL. We also shorten words that are lengthened to see if we can find the shortened version in the lexicons (e.g. sweeeet $\rightarrow$ sweet). We shorten out-of-vocabulary words with more than one repeated letter using regular expressions and attempt to find the same word with no or one repeating letter in the dictionaries. The coverage of the lexicons for each corpus is shown in Table 4.4. It shows some clear differences among the datasets which could impact the effectiveness of the system. For example, Twitter has the poorest coverage. This is not surprising as Twitter has a lot of @ mentions and hashtags which will not be found in any lexicon. The MPQA's coverage is also relatively poor; we expect this is due to the more sophisticated vocabulary and proper nouns commonly found in newswire articles. The MPQA and Wikipedia have no emoticons and very little word lengthening indicating they are more formal in nature. The negation state machine [Agarwal *et al.*, 2009] flips the polarity of words that are preceded by a negative word (e.g. not, never).

The marked sentences are used to create the lexical features described in the original work [Agarwal *et al.*, 2009]. This consists of 45 features based on combining the n-grams (1-3 words), syntactic type (NP, VP, PP, JJP, and other), position to target (chunk to the right, left, and target itself), and subjectivity/polarity ($\{$objective, subjective$\}$ or $\{$positive,negative,neutral$\}$) of the chunk. The min and max pleasantness of the sentence are also included as features.

In addition to the lexical features, we include the most common n-grams and apply feature selection to them. We experimented with a different number of n-grams (e.g. $\{0,100,250,1000,5000\}$) and found that in the online discussions, using a large amount of n-grams did not improve results. We opted to keep the top 100 n-grams. We also use the mean SentiWordNet [Baccianella *et al.*, 2010] score of the phrase as a distinct feature that is used in addition to the prior polarity of the phrase computed from the DAL scores.

Lastly, we also include the lexical stylistic features listed in Section 3.1, Table 3.1. The count values of each feature were normalized by the number of words in the phrase. The percentage of lexical-stylistic features that are $\{$positive,negative,neutral$\}$ is shown in Figure 4.1. For example, emoticons tend to indicate a positive phrase in Twitter and LiveJournal. On the other hand, they are non-existent in the more formal genres of MPQA and Wikipedia Talk Pages. Each stylistic feature

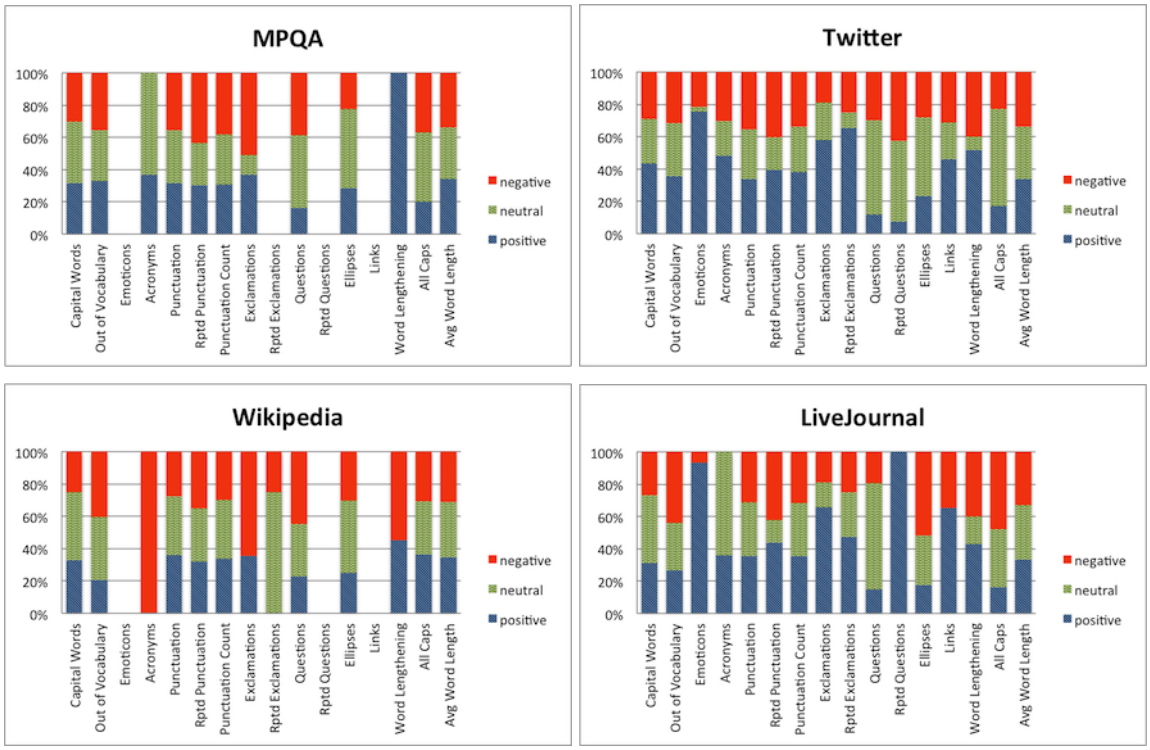


Figure 4.1: Percentage of lexical-stylistic features that are {positive,negative,neutral} in each corpus: MPQA, Twitter, Wikipedia, and LiveJournal. The positive percentage is on the top, the neutral percentage is in the middle, and the negative percentage is on the bottom.

accounts for less than 2% of the sentence but at least one of the stylistic features exists in 65.2% of the tweets, 62.8% of the LiveJournal sentences, 57.7% of the Wikipedia Talk Page sentences, and 52.6% of the MPQA sentences.

4.5 Experiments and Results

We ran cross-validation experiments to tune our systems and show results on a held out test set for each genre using balanced and unbalanced datasets for subjectivity (Table 4.5) and polarity (Table 4.6). We experimented with several classifiers (e.g. SVM and Logistic Regression) and found that Logistic Regression always performed best. We also experimented with a varying number of n-grams (e.g. 100, 250, 500, ...) and found that 100 consistently performed best. All experiments shown were run using Logistic Regression in Weka [Hall *et al.*, 2009] and include the top 100

	Unbalanced				Balanced			
Experiment	LJ	MPQA	Twitter	Wiki	LJ	MPQA	Twitter	Wiki
majority	54.5	69.0	64.8	57.7	50.0	50.0	50.0	50.0
n-grams+DAL	76.7	75.9	84.0	77.8	**74.3**	73.1	81.8	70.9
+Wordnet	77.1	77.2	86.8	79.2	73.5	74.2	81.5	70.6
+Dictionaries+NNP	77.1	**78.7**	87.9	78.8	74.0	**75.9**	**82.3**	**71.1**
+Lexical Style	**77.8**	77.5	**88.2**	**81.5**	73.5	70.4	81.2	70.6

Table 4.5: Subjective vs Objective classification accuracy using a test set. The baseline is the majority class. The best result for each column is highlighted in bold. Dictionaries refers to Wiktionary, Emoticons, and SentiWordNet. NNP refers to proper nouns being marked as objective.

n-grams. The results shown using just WordNet in the third row of Tables 4.5 and 4.6 are equivalent to prior work [Agarwal *et al.*, 2009]. We found that adding additional dictionaries relevant to online discussions, and the lexical-stylistic features in general achieved a 2% absolute boost in performance in subjectivity and polarity as shown in the fourth and fifth rows of Tables 4.5 and 4.6. However, this did not tend to be a statistically significant improvement. We did find that the dictionaries geared towards social media (e.g. Wiktionary) could be used in place of standard dictionaries for similar results. The lexical-stylistic features did not have as significant an impact as we initially expected. We believe this is partly due to the fact that several social media features are not related to either polarity. For example, capital words occur evenly in positive and negative phrases in both Twitter and LiveJournal. Although not shown in the table, we found that the SentiWordNet features provided us with a boost in performance in comparison to our earlier work where it was not included. Further information regarding this in Twitter can be found in the semeval evaluation section below (Section 4.6.1). Due to the improvement from including SentiWordNet, we feel that focusing on improving and adding new sentiment dictionaries will improve our overall results. We discuss this in further detail in the discussion section (Section 4.7).

In addition to our experiments in each genre, we also ran a series of cross-domain experiments using polarity detection where we used one genre for training and another one for testing. As Table 4.7 indicates, all of the cross-domain experiments were promising except for using online genres to

	Unbalanced				Balanced			
Experiment	LJ	MPQA	Twitter	Wiki	LJ	MPQA	Twitter	Wiki
majority	54.5	69.0	64.8	57.7	50.0	50.0	50.0	50.0
n-grams+DAL	81.0	68.7	81.6	70.6	81.5	71.4	80.0	65.5
+Wordnet	81.5	74.8	82.3	70.6	80.1	73.5	80.9	72.1
+Dictionaries+NNP	82.0	**76.1**	**83.9**	**73.6**	**84.4**	71.4	82.2	**73.6**
+Lexical Style	**83.9**	75.5	**83.9**	**73.6**	83.4	**72.1**	**82.5**	70.6

Table 4.6: Positive vs Negative classification accuracy using a test set. The baseline is the majority class. The best result for each column is highlighted in bold. Dictionaries refers to Wiktionary, Emoticons, and SentiWordNet. NNP refers to proper nouns being marked as objective.

predict sentences from newswire articles (MPQA). The LiveJournal sentences were easy to predict regardless of the training corpus. Interestingly, the MPQA data performed best on out-of-domain training sets, especially those with a small amount of training data (Livejournal and Wikipedia). We expect this is mainly due to its larger size.

Finally, in Table 4.8, we show experiments for running subjectivity and polarity detection on all the datasets as one corpus. Although we have focused on positive versus negative polarity, our system also has the capability of performing 3-way classification of positive versus negative versus neutral. Examples of neutral opinion are phrases such as “wishes to” and “expected rise” as shown in Table 4.2. Predicting neutral sentiment can be useful in some settings (e.g. reviews) but tends to cause a decrease in performance. This is the case in our data as shown in Table 4.8. Since neutral opinion is not necessary in detecting influencers in both our agreement and claim systems we opt to exclude it in our model for detecting opinion. Furthermore, we also choose to use balanced data from all the corpora in our classification models for our influence work. This refers to the subjectivity and positive versus negative polarity systems shown in the second row of Table 4.8 with an accuracy of 77.1% and 81.3% respectively.

Experiment	Twitter	LiveJournal	MPQA	Wikipedia
Twitter	82.5%	83.4%	63.3%	72.1%
LiveJournal	77.6%	83.4%	62.6%	73.6%
MPQA	78.5%	83.9%	72.1%	75.6%
Wikipedia	71.6%	81.5%	61.2%	70.6%

Table 4.7: Positive vs Negative accuracy (in %) for cross-domain experiments. Each row refers to an experiment where one corpus (X) was used as the training set and the column corpus (Y) was used as the test set. For example, X=MPQA, and Y=Wikipedia has an F-Score of 75.6%. All features were used in these experiments on balanced training sets and unbalanced test sets.

Experiment	Subjectivity	Positive vs Negative	Positive vs Negative vs Neutral
Majority	63.1%	50.9%	37.9%
Balanced	**77.1%**	**81.3%**	64.4%
Unbalanced	84.4%	81.5%	73.4%

Table 4.8: Subjectivity and Polarity accuracy (in %) using all datasets together. All features were used in these experiments. Highlighted experiments indicate systems used in predicting influence components.

4.6 Public Evaluations

In this section we describe our participation in two public evaluations: Semeval Sentiment Analysis in Twitter and TAC KBP Sentiment Slot Filling. It is important to note that the evaluations were not the focus of our work. We only spent a few weeks preparing for the task and thus, this is reflected in our results, particularly in comparison to the other teams. We provide this information with the aim of introducing room for improvement that is further reflected upon in the discussion section.

4.6.1 SemEval: Sentiment Analysis in Twitter

We participated in the SemEval Sentiment Analysis in Twitter subtask A in 2013 and 2014 [Nakov *et al.*, 2013; Rosenthal *et al.*, 2014c]. Subtask A is the polarity detection of phrases in Twitter, LiveJournal, and SMS messages. The Twitter (excluding sarcasm) and LiveJournal corpora are identical to those discussed earlier in the Data section of this chapter.

Data Set	Majority	2013	2014
Twitter Dev	38.14	77.6	81.5
Twitter Test	42.22	N/A	76.54
Twitter Sarcasm	39.81	N/A	61.76
SMS	31.45	73.3	74.55
LiveJournal	33.42	N/A	78.19

Table 4.9: A comparison between the 2013 and 2014 results for Subtask A using the SemEval Twitter training corpus. All results exceed the majority baseline of the positive class significantly.

The full results in the participation of SemEval 2014: Sentiment Analysis in Twitter, Subtask A, are shown in Table 4.9 [Rosenthal and McKeown, 2013; Rosenthal *et al.*, 2014b]. All results are shown using the average F-measure of the positive and negative class. The results are compared against the majority baseline of the positive class. We do not use neutral/objective as the majority class because it is not included in the average F-score in the Semeval task. Our system outperforms the majority baseline significantly in all classes.

Our submitted system was trained using 3-way classification ({positive,negative,neutral}). It included all the lexicons described in this section and the top 100 n-grams with feature selection. In contrast to 2013, in 2014, it included SentiWordNet and the position of the phrase which provided a 4% increase compared to our best results during the prior year (77.6% to 81.5%) and a rank of 10/20 among the constrained systems, which used no external data. Our results on the 2014 test set was 76.54% for a rank of 14/20. We do not do well in detecting the polarity of phrases in sarcastic tweets. This is consistent with the other teams as sarcastic tweets tend to have their polarity flipped. The improvements to our system provided a 1% boost in the SMS data with a rank of 15/20. Finally, in the LiveJournal dataset we had an F-Score of 78.19% for a rank of 12/20.

4.6.2 TAC KBP: Sentiment Slot Filling

When text is presented to its audience as objective, expressions of sentiment are often elusive, obscured by evasive language or displaced from one entity to another. Criticism, for example, is often implicit, and its source difficult to locate:

> As the government wraps up its Troubled Asset Relief Program, the company that

> received the most from the fund, the American International Group, is offering an exit plan with no clear sense of whether the taxpayers will end up with a gain or a loss (*NYT*, January 10, 2010).

Here, does the reporter express opinion or fact? If it is sentiment, the target is unclear: perhaps the unaccountable financial firm, its government caretaker, or both. In many cases, it can be difficult for even a human reader to determine who, exactly, the opinion comes from, and toward what, exactly, it is targeted.

Detecting sentiment toward an entity can be useful for influence detection to infer that the people in the conversation share opinion on the same things. In the future we would like to incorporate sentiment towards a topic into our influence system. This section briefly describes our participation in the TAC KBP Sentiment Slot Filling Task [Rosenthal *et al.*, 2013; Rosenthal *et al.*, 2014a] and the challenges in detecting sentiment towards a topic.

The Sentiment Slot Filling task evaluation[5] consisted of a list of queries regarding the sentiment of an entity. Each query was in the form {**positive/negative** sentiment **from/towards** Entity **X**}. The goal of the task was to find entities within the corpus (TAC 2013) or document (TAC 2014) that are appropriately related to the entity in the query. For example, consider the following query:

> **Negative** sentiment **from Allan West**

In the following text containing the query entity, potential target entities are bolded:

> **Klein** said **he** doesn't regret his votes for the health care and stimulus bills, measures that *West* and other **Republicans** used against **Democrats** nationwide.

Additional examples of queries that were provided during the evaluation and training period are shown in the first column of Table 4.10.

Given a query, our system first finds all candidate entities occurring within the same paragraph as the query (paragraph information is provided by the corpus). It then weeds out any entities that do not have a valid towards/from relationship. Afterwards, the system determines whether the sentiment surrounding the candidate entity matches the sentiment provided in the query. All duplicate mentions

[5] http://www.nist.gov/tac/2013/,http://www.nist.gov/tac/2014/

Query (entity, sentiment)	**Text** (correct phrase labeled, entities bolded)	Valid Slotfillers	Invalid Slotfillers
Pelosi, pos-from	Indeed [**liberals** credit *Pelosi*] with pressuring **Obama** when **he** was inclined to cave.	liberals	"Pelosi" is not an object in relation to "Obama" and "he"; they are not evaluated for sentiment
Barber, pos-towards	[**Talib's** pick in the fourth quarter was about as clutch of a play that I've seen around here in a long, long time, *Ronde Barber*] said of his **fellow cornerback**.	Talib	Though "fellow cornerback" refers to Talib, this mention would be eliminated because "said of" is not subjective
Israel, neg-towards	"This assault proved once again, clearly, that the current [*government of Israel* does not want peace in the region," **Erdogan**] told reporters in **Chile**.	Erdogan	Text between "Erdogan" and "Chile" is not subjective
Benedict, neg-from	But U.S. [**victims** of clerical abuse were not impressed by **Benedict's**] selections, saying some of the **bishops** themselves had "troubling" records on confronting abuse.	victims	"Benedict" is not an object in relation to "bishops"

Table 4.10: Examples of queries, expressions of sentiment, and valid and invalid slot fillers

that co-refer to the same entity (such as "Klein" and "he", which refer to U.S. Representative Ron Klein) are winnowed into a single slot filler based on the confidence reported by the sentiment polarity analysis, and the most representative name is reported. We also experiment with using only the mention of an Entity that appears closest to the query. This can be useful as mentions occurring later tend to be less likely to be relevant based on distance from the query. Finally, we also have several heuristics used to adjust our confidence of an answer:

- Entity is a noun: $+.1$
- Entities have Subject/Object Relationship: $+.2$
- Entities have Object/Subject Relationship: $-.2$

Error	Occurrence
Invalid Entity	14
Incorrect / Lack of appropriate toward/from Relationship	25
Incorrect Sentiment	6
Sentiment From a Place to the Query	8

Table 4.11: The most common errors in a subset of 8 queries consisting of 44 answers

- Prior Sentiment (stored in Knowledge Base and based on sentiment of entity in entire corpus) is the same as the Query Sentiment: $+.1$
- Author is the Entity and the word I is in the justification: $+.1$
- Author is the Entity and the word I is not in the justification: $-.1$

The impact of these filters to the confidence score was computed empirically based on observing the results during training time. Additional experiments would be useful in the future. Query examples and candidate answers are shown in Table 4.10. The data was provided by the evaluation and consists of newswire, and discussion forums (where the author can hold the opinion). Our best submission performed poorly with an F-score of 10.2%. No submission to TAC 2014 performed better than 20%.

The common cause of error was due to choosing entities poorly. For example, given the query "**positive** sentiment **towards Federer**", our system chose "French" as a candidate entity with the justification "Federer said the rivals chatted briefly Wednesday, and Nadal congratulated him for winning the French". There are several problems with this answer. First of all, places are generally not good entities. This is especially the case when trying to find sentiment *towards* a person as a place can not have an opinion about a person. The second issue with this answer is that SERIF [Ramshaw and Weischedel, 2005], BBNs coreference output provided by TAC, did a poor job of extracting the entity as the entity should be the "French open". On a more positive note, the sentiment of the justification phrase was correctly determined to positive. Another issue is choosing entities that are not related to the query entity. For example, given the query "**positive** sentiment **towards Mayor Bloomberg**", our system chose "Bush" as the candidate entity with the justification "The mayor said that Obama deserved praise for working out a deal with Republican leaders to retain Bush". The

	excellent	amazing	terrific	worthless	rotten	awful	united	vice	grow
AFINN	0.8	0.9	0.9	0.3	N/A	0.2	0.6	N/A	N/A
ANEW	0.92	N/A	0.90	N/A	0.16	N/A	N/A	N/A	N/A
crrugent	0.82	0.78	0.77	0.11	0.18	0.16	N/A	0.35	0.66
DAL	0.57	0.63	N/A	N/A	N/A	0	0.69	0.32	0.81
General-Inquirer	1	1	1	0	0	0	N/A	0	0.5
MPQA	1	1	1	0	0	0.25	0.25	0	0.75
NRC # Sentiment	0.86	0.76	0.85	0.19	0.19	0.22	0.50	0.50	0.50
Sentiment140	0.67	0.62	0.64	0.43	0.38	0.32	0.51	0.58	0.55
SentiStrength	0.75	0.75	0.88	0.25	0.25	0.13	0.5	0.5	0.5
SentiWordNet	1	0.75	0.65	0.25	0.13	0.04	0.5	0.63	0.5

Table 4.12: A list of normalized scores between 0-1 for words in publicly available sentiment dictionaries.

error here is that there is no subject/object relationship between Bush and Bloomberg.

We analyzed two queries of each type (8 in total) for a combined total off 44 answers. The types of errors found are shown in Table 4.11. It is clear that the majority of errors are due to invalid entities and entity relationships indicating that improvement on top of SERIF and dependency parsing are necessary. The majority of the sentiment errors shown were related to one query where the justification was objective text describing the history of a person. Finally, we also computed the number of errors due to choosing a place as the candidate entity. It may be possible to avoid this type of error by excluding all places as answers to pos- and neg-towards person queries as a place will usually not have an opinion towards a person.

4.7 Discussion

Our results in the public evaluations show that there is room for improvement in our method. In this section we explore what we believe is the main cause of our poor performance in comparison to the top teams in the SemEval tasks; our lexicon of choice, the DAL.

We have found that there are times where the DAL has questionable scores. For example, the word *evil*, which one would consider to be a negative and unpleasant word has a pleasantness score

	Dictionary	Concept	Polarity Range	Size
Social Media	NRC # Sentiment	Sentiment values for hashtagged words and bigrams in Twitter	-10 - 10	371K
	Sentiment140	Sentiment values for words and bigrams using emoticons in tweets	-5 - 5	740K
	SentiStrength	basic sentiment analyzer using short texts from MySpace. Positive and negative scores are given for each word	-1 - -5 & 1 - 5	2546
	AVAYA smiley	A dictionary with smiley values. It uses capitalization and length of words to modify polarity.	-1 - 1	11740
	AFINN	A list of words with sentiment attached making use of slang/obscene words often used in microblogs	-5 - 5	2477
NewsWire	SentiWordNet	Assigns each word 3 scores: positivity, objectivity and negativity. Uses tenary classifiers to determine the polarity	0 - 1	117K
	MPQA	classifies each phrase as neutral or polar and then determines the contextual polarity.	pos, neg, both, neu	8000
	General-Inquirer	Tags any word as having any of 182 characteristics such as positive, negative, power, weak etc.	pos, neg	11788
	DAL	scores for pleasantness, activeness, and imagery of word	1-3	8742
	ANEW	scores for pleasure, arousal, and dominance of word	1-9	1034
	crrugent	Extends ANEW with ratings for valence, arousal and domination	1-9	13915

Table 4.13: A description of publicly available sentiment dictionaries geared towards newswire and social media.

of 1.875/3. In addition, the AE space score which combines activeness and pleasantness does not always appropriately map to polarity. For instance, the word grow is a pleasant and active word, but it is often used in a neutral setting (e.g. “The tree will grow this year”.) where it is unclear whether it portrays positive or negative meaning without further context.

In the future, we would like to explore using different or additional dictionaries to the DAL and SentiWordNet, specifically focusing on ones geared towards social media. Towards this goal, we have performed a comprehensive analysis of the scores of several publicly available dictionaries as shown in Table 4.12. The scores in this table were normalized from their original dictionary score to be between 0-1. A score closer to 0 is negative and a score closer to 1 is positive. The scores

vary widely. For example, the normalized score for "excellent" ranges from .57 to 1.There are three main aspects that should be taken into account when choosing a dictionary: coverage, domain, and reliability in score. We describe this information for the dictionaries in Table 4.13. The coverage of the dictionaries varies from small to large; for example, the ANEW dictionary [Bradley and Lang, 1999] has poor coverage as evident by its small size. These dictionaries include those geared towards newswire text [Warriner *et al.*, 2013; Bradley and Lang, 1999; Wiebe *et al.*, 2005; Whissel, 1989; Baccianella *et al.*, 2010; Stone *et al.*, 1966] and social media text [Mohammad *et al.*, 2013; Thelwall *et al.*, 2010; Becker *et al.*, 2013; Nielsen, 2011]. Finally, the reliability of the scores vary based on how fine-tuned they are with scores on a scale of -10 to 10 [Mohammad *et al.*, 2013] and different scores for each of the three polarities [Baccianella *et al.*, 2010].

4.8 Conclusion

In this chapter we described a method for performing subjectivity and polarity detection on the phrase level. Our method is based on prior work [Agarwal *et al.*, 2009] which we geared towards social media by including several lexicons (e.g. emoticons) and style features (e.g. word lengthening). We compared our method across several genres, and show that data from different sources can be useful in predicting sentiment, thus providing motivation for applying domain adaptation to opinion detection in online genres in the future. We also show that although using social media features help a little bit, they do not provide a significant improvement across all domains. This indicates that data is more important than the genre specific features. Since the MPQA is a very large dataset, it does well in detecting sentiment in online genres. We suspect that this is only relevant to a general social media system. A system that focuses on a single source (e.g. Twitter), will benefit significantly from social media features related to that source (e.g. hashtags).

We have also discussed the importance of lexicon choice when developing a sentiment system. Lexicons differ in domain, coverage, and range of scores, all of which can affect the performance of sentiment system. We began our experiments by using the DAL as our lexicon. However, our recent analysis has discovered that the DAL often has an odd choice of scores for its words. In the future, we would like to experiment with changing the lexicon we use to see statistically how the different lexicons facilitate our approach as well as whether using several lexicons at once provides

an improvement compared to using one lexicon.

In addition to developing a system that detects sentiment we also participated in several public evaluations. This was never the main focus of our work, which was reflected in our performance in comparison to other participating teams. However, we did find that our system tended to be competitive. In addition, participating in public evaluations allowed us to find weaknesses in our system, such as the choice of lexicon as well as the impact of genre specific features (e.g. using hashtags in Twitter). Through our participation in evaluations we also explored sentiment towards a topic which is an important task albeit not a primary goal of this thesis. In the future we would like to continue exploring this research area.

To conclude, sentiment detection is a useful component in detecting influence which we have motivated through two weapons of influence: reciprocation and liking. Sentiment detection is a subcomponent which is used in our agreement and claim components. We have generated a subjectivity ($\{$subjective,objective$\}$ and polarity $\{$positive,negative$\}$ model using balanced data from all the corpora (Twitter, LiveJournal, Wikipedia, MPQA). The subjectivity model is used in claim detection as our claims must be opinionated, but the polarity is not important. In agreement detection, we use the polarity model because the type of sentiment is important. If two people have sentiment of the same polarity (e.g. both like Diet Coke), that can indicate agreement. If two people have sentiment of the opposite polarity (e.g. one likes Diet Coke and one hates diet Coke) that can indicate disagreement. Further details regarding the use and contribution of sentiment in each component will be discussed in the Persuasion (Chapter 6) and Agreement chapters (Chapter 5).

Chapter 5

Agreement

> *Computer programming is the single best professional opportunity in the world. We need more Americans in the field. Let's go!*
>
> Steve Ballmer. Former CEO, Microsoft, *code.org*

> *I think that great programming is not all that dissimilar to great art. Once you start thinking in concepts of programming it makes you a better person...as does learning a foreign language, as does learning math, as does learning how to read.*
>
> Jack Dorsey. Cofounder, Twitter, *code.org*

Any time people have a discussion, whether it be to solve a problem, discuss politics, products, or more casually, gossip, they will express their opinions. As a conversation evolves, the participants of the discussion will agree or disagree with the views of others. The ability to automatically detect agreement and disagreement (henceforth referred to as (dis)agreement) in the discussion is useful for understanding how conflicts arise, how they are resolved, and the role of each person in the conversation. One such example of agreement is shown above between Steve Balmer and Jack Dorsey both stating how great computer programming is. Furthermore, detecting (dis)agreement has been found to be useful for other tasks, such as detecting subgroups [Hassan *et al.*, 2012], stance

[Lin *et al.*, 2006; Thomas *et al.*, 2006], power [Danescu-Niculescu-Mizil *et al.*, 2012; Biran *et al.*, 2012], and interactions [Mukherjee and Liu, 2013].

Several weapons of influence [Cialdini, 2007] indicate that agreement can be useful for detecting influencers. The obvious connection is **reciprocation**; if someone says something positive, or speaks in a nice manner, a person will be more likely to agree with them in return and if someone says something negative, or using a rude tone, a person will be likely to disagree with them. Furthermore, once someone agrees with someone else they are likely to agree again. This follows the weapon of **commitment and consistency**. This is evident from a study done by Freedman and Fraser [1966] that people in a California neighborhood were more likely (76% vs 17%) to allow a "Drive Carefully" sign be displayed on their lawn if they had previously signed a petition that favored "keeping California beautiful" which most people signed. In addition, **scarcity**, in the form of a restriction, is likely to cause a negative reaction which will cause others to react [Brehm and Brehm, 1981] by defying and disagreeing with the restriction and the person who placed it. There are several examples of agreement, a_i, visible in the influencer example in Chapter 1, Table 2.2.

In this thesis, we explore a rich suite of features to detect (dis)agreement between two posts, the *quote* and the *response* (Q-R pairs [Walker *et al.*, 2012]), in online discussions. We analyze the impact of features including meta-thread structure, lexical and stylistic features, Linguistic Inquiry Word Count categories, sentiment, sentence similarity and accommodation. Our research indicates that conversational structure, as indicated by meta-thread information as well as accommodation between participants, plays an important role. *Accommodation* [Giles *et al.*, 1991], is a phenomenon where conversational participants adopt the conversational characteristics of the other participants as conversation progresses. Our approach represents accommodation as a complex interplay of semantic and syntactic shared information between the Q-R posts. Both meta-thread structure and accommodation use information drawn from both the quote and response; these features provide significant improvements over information from the response alone.

We detect (dis)agreement in a supervised machine learning setting using 3-way classification (agreement/disagreement/none) between Q-R posts in several datasets annotated for agreement, whereas most prior work uses 2-way classification. In many online discussions, none (i.e., the lack of (dis)agreement) is the majority category so leaving it out makes it impossible to accurately classify the majority of the sentences in an online discussion with a binary classification model.

Example of disagreement in an ABCD discussion indicated by different sides (Against and For).
Abortion is WRONG! God created that person for a reason. If your not ready to raise a kid then put it up for adoption so it can be with a good family. Dont murder it! Its wrong. If you can have sex then you should be ready for the consequences tht come with it! **Side:** ***Against***
So, those, who were raped through the multiple varieties of means, are expected to birth this child although it was coerced rape. I don't think so. Therefore, taking a woman's right to choice is wrong regardless what a church or the government suggests. **Side:** ***For***
Example of agreement in an ABCD discussion indicated by the same side (Against).
HELL NO! ... IF YOU WERE GROWN ENOUGH TO SPREAD YOUR FUCKING LEGS THEN YOU ARE GROWN ENOUGH TO TAKE THE RESPONSIBILITY OF YOUR FUCKING KID! ... YOU KNOW THERE ARE OTHER OPTIONS!!!! KILLING A INNOCENT BABY ISN'T GONNA JUST GO AWAY LIKE A RAINY DAY!!!! YOU WILL HAVE TO LIVE WITH THE GUILT FOREVER!!!!!!! **Side:** ***Against***
————————————> That is soo true living with the guilt forever know you murder you child it would have been even better if the murder hadn't been born. **Side:** ***Against***
Example of no (dis)agreement in an ABCD discussion between the original post and a response.
Coke or Pepsi?
They taste the same no big difference between them for me

Table 5.1: Examples of Agreement, Disagreement, and None in ABCD discussions

We also present a new naturally occurring agreement corpus, Agreement by Create Debaters (ABCD), derived from a discussion forum website, createdebate.com, where the participants are required to provide which side of the debate they are on. This enabled us to easily gather over 10,000 discussions in which there are over 200,000 posts containing (dis)agreement or the lack of, *25 times larger* than any pre-existing agreement dataset. We show that this large dataset can be used to successfully detect (dis)agreement in other forums (e.g. 4forums.com and Wikipedia Talk Pages) where the labels cannot be mined, thereby avoiding the time consuming and difficult annotation process.

In the following sections, we first discuss related work in spoken conversations and discussion forums. We then turn to describe our new dataset, ABCD, as well as two other manually annotated corpora, Internet Argument Corpus (IAC), and Agreement in Wikipedia Talk Pages (AWTP). We explain the features used in our system and describe our experiments and results. We conclude with

System	Features	Data	Classification	Method	Results
Hillard *et al* [2003]	p, l	ICSI	3-way	Decision Trees, unsupervised	79% A
Galley *et al* [2004]	s, d, l	ICSI	3-way	Bayesian Networks, HMM	86.9% A
Hahn *et al* [2006]	l	ISCI	3-way	semi-supervised, SVM	87.1% A
Germesin & Wilson [2009]	l, p, s, d, o	AMI	3-way	CRF	80.3% A/56.1% F
Wang *et al* [2011a; 2011b]	l, p, s, d	Broadcast News	2-way	CRF	58.8% F

Table 5.2: Related work in Agreement Detection in Spoken Dialog. Classification is either 2-way (Agreement/Disagreement) or 3-way (Agreement/Disagreement/None). Results in F-score (F) or Accuracy (A). The features used are lexical (l), prosodic (p), structural (s), duration (d), and opinion (o).

a discussion containing an error analysis of the hard cases of (dis)agreement detection.

5.1 Related Work

Early prior work on detecting (dis)agreement has focused on spoken dialogue [Galley *et al.*, 2004; Hillard *et al.*, 2003; Hahn *et al.*, 2006] the majority of which detect (dis)agreement on spurts[1] using the ICSI meeting corpus [Janin *et al.*, 2003]. Galley *et al* [2004] detected adjacency pairs and used Bayesian Networks to outperform the earlier work of Hillard *et al* [2003]. Hahn *et al* [2006] used unlabeled data to achieve comparable results to Galley *et al* [2004]. Germesin and Wilson [2009] detect (dis)agreement on dialog acts in the AMI meeting corpus [Mccowan *et al.*, 2005] using CRFs to the third order. and Wang *et al* [2011a; 2011b] detect (dis)agreement in broadcast conversation in English and Arabic on the utterance level using CRF. They avoid detecting adjacency pairs by assuming that a person can only respond to the person who spoke prior to them. Prior work in spoken dialog has motivated some of our features (e.g., lists of agreement and disagreement terms, sentiment and n-grams). However, it is difficult to directly compare the results of our approach to prior work

[1] A spurt is a period of speech that has no pauses $>$ .5 seconds

System	Features	Data	Size	Classification	Method	Results
Yin *et al* [2012]	l,o,d	Political Forum, USMB posts	818, 170	2-way	Logistic Regression	77% F, 65% F
Abbot *et al* [2011a]	l,s,o	IAC segments	8242	2-way	Jripper	68.2% A
Misra & Walker [2013]	l,d,o	IAC segments	2677	2-way	J48	60.1% A
Mukherjee & Liu [2012]	l	volconvo.com posts	1889	2-way	JTE	85.5% F
Opitz & Zirn [2013]	l	AAWD sentences	2302	2-way	SVM	72.8% F
Wang & Cardie [2014]	l,s,o	AAWD utterances, IAC segments	3663, 7253	3-way	CRF	59.7% F, 63.6% F

Table 5.3: Related work in Agreement Detection in Online Discussions. Classification is either 2-way (Agreement/Disagreement) or 3-way (Agreement/Disagreement/None). Results in F-score (F) or Accuracy (A). The features used are lexical (l), structural (s), durational (d), and opinion (o).

due to the different genre (discussion forums vs spoken conversations) and the size of the datasets. A comprehensive list comparing the different methods is shown in Table 5.2

Recent work has turned to (dis)agreement detection in online discussions [Yin *et al.*, 2012; Abbott *et al.*, 2011a; Misra and Walker, 2013; Mukherjee and Liu, 2012]. The prior work performs 2-way classification between agreement and disagreement using features that are lexical (e.g. n-grams), basic meta-thread structure (e.g. post length), social media features (e.g. emoticons), and polarity using dictionaries (e.g. SentiWordNet). Yin *et al* [2012], detect local and global (dis)agreement in discussion forums where people debate topics. Their focus is global (dis)agreement, which occurs between a post and the root post of the discussion. They manually annotated posts from US Message Board [2] (818 posts) and Political Forum [3] (170 posts) for global agreement. This approach ignores off-topic posts in the discussion which can indicate incorrect labeling These discussions differ from ours in that agreement is more common than disagreement. and the small size makes it difficult to determine how consistent their results would be in unseen datasets. Abbott *et al* [2011a], look at (dis)agreement using 2,800 annotated pairs from the Internet Argument Corpus (IAC) [Walker *et al.*,

[2] usmessageboard.com

[3] politicalforum.com

2012]. They performed two-way classification on a balanced test set. Their work was extended to topic independent classification by Misra and Walker [2013]. Since it is the largest previously used corpus, we use the IAC corpus in our experiments. Lastly, Mukherjee and Liu [2012], developed an SVM+Joint Topic Model classifier to detect (dis)agreement using 2,000 posts. They studied accommodation across (dis)agreement by classifying over 300,000 posts and explore the difference in accommodation across LIWC categories. They did not use accommodation to detect (dis)agreement, but they found that it is more common in agreement for most categories, except for a few style dimensions (e.g. negation) where it is reversed. This paper highly motivates our inclusion of accommodation for (dis)agreement detection.

In other work, Opitz and Zirn [2013] detect (dis)agreement on sentences using the Authority and Alignments in Wikipedia Discussions (AAWD) corpus [Bender *et al.*, 2011b] which is different than the AWTP corpus used in this thesis. In the future we would like to explore whether we could incorporate this corpus into ours. Wang and Cardie [2014] also detect (dis)agreement on the sentence level using AAWD and between Q-R pairs using the IAC. They achieve significant improvements to prior work using Isotonic CRF on all the sentences or Q-R pairs in a post. It is, however, hard to interpret their results as they do not mention whether they used a held out test set, and its size, or if they used cross-validation. Although they use 3-way classification it is impossible to compare to our results using the IAC because we perform (dis)agreement detection on the post level while they perform it on the pair level. Furthermore, they don't follow the labeling of prior work and consider far less annotations to be neutral. A comprehensive list comparing the prior work in (dis)agreement in online discussions is shown in Table 5.3.

Finally, Gözde Kaymaz [2013] automatically generates (dis)agreement labels in Twitter on over 112,000 tweets using retweets and a list of (dis)agreement terms to determine the label. This dataset is most comparable to ours in size, but differs significantly from our data due to the Twitter character limit for tweets and use of retweets. He does not detect (dis)agreement but rather uses it to find opinion leaders via the social network.

Our approach differs from prior work in that it explores (dis)agreement detection on a large, naturally occurring dataset where the annotations are derived from participant information. We explore new features representing aspects of conversational structure (e.g. sentence similarity) and the more difficult 3-way classification task of detecting agreement/disagreement/none.

5.2 Data

In this work we focus on direct (dis)agreement between quote-response (Q-R) posts in the three datasets described in the following subsections. Across all datasets we only include discussions of depth > 2 to ensure a response chain of at least three people and thus, a thread. We also excluded extremely large discussions (more than 500 posts) to improve processing speed. We only consider entire posts in Q-R pairs. Our datasets are Agreement by Create Debaters (ABCD), the Internet Argument Corpus (IAC) [Abbott *et al.*, 2011a], and Agreement in Wikipedia Edit Discussions (AWTP). Information regarding accessing the datasets is available in Appendix B.

5.2.1 Agreement by Create Debaters (ABCD)

Create Debate is a website where people can start a debate on a topic by asking a question. On this site, a debate can be:

- **open-ended**: there is no side
- **for-or-against**: two sided
- **multiple-sides**: three or more sides

In this thesis, we only focus on debates of the for-or-against nature where there are two sides. For example, we use a debate discussing whether people are for or against abortion[4] in our examples throughout this Chapter. Prior work [Abu-Jbara *et al.*, 2012] has used the side label of this corpus to detect the subgroups in the discussion. We annotate the corpus as follows: the side label determines whether a post (the *Response*) is in agreement with the post prior to it (the *Quote*[5]). If the two labels are the same, then they agree. If the two labels are different, they disagree. When the author is the same for both posts, there is no (dis)agreement as the second post is just a continuation of the first. Finally, the first post and its direct responses do not agree with anyone; the first post does not have a side as it is generally a question asking whether people are for, or against the topic of the debate. Examples of (dis)agreement and none are shown in Table A.4. We call this corpus Agreement by Create Debaters or ABCD.

[4] `www.createdebate.com/debate/show/Abortion_9`

[5] Note that we refer to the earlier post as the *Quote*. It however is not actually quoted. We use this notation to match prior work in the IAC corpus [Walker *et al.*, 2012]

Dataset	Thread Count	Post Count	Agree	Disagree	None
ABCD	9981	185479	38195	60991	86293
IAC	1220	5940	428	1236	4276
AWTP	50	822	38	148	636

Table 5.4: Statistics for full datasets

Our dataset includes over 10,000 discussions which include 200,000 posts on a variety of topics. Additional statistics for ABCD are shown in Table 5.4. There are far more disagreements than agreements as people tend to be argumentative when they are debating a topic.

5.2.2 Internet Argument Corpus (IAC)

The second dataset we use is the IAC [Walker *et al.*, 2012]. The IAC consists of posts gathered from `4forums.com` discussions that were annotated on Mechanical Turk. The Turkers were provided with a Q-R pair and had to indicate the level of (dis)agreement using a scale of $[-5, 5]$ where -5 indicated high disagreement, 0 no (dis)agreement, and 5 high agreement. As in prior work with this corpus [Abbott *et al.*, 2011a; Misra and Walker, 2013], we converted the scalar values to (dis)agreement with $[-5, -2]$ as disagreement, $[-1, 1]$ as none, and $[2, 5]$ as agreement[6]. In this dataset it is possible for multiple annotations to occur in a single post. We combine the annotation to the post level as follows. We ignored the none annotations unless there was no (dis)agreement. This means that the only way a post can be annotated as none is if all Q-R pairs in that post were annotated as none. In all other cases, we use the average (dis)agreement score as the final score for the post. 10% of the posts had more than one annotation label. The number of annotations per class is shown in Table 5.4. Not all Q-R posts in a thread were annotated for agreement as is evident by the ratio of threads to post annotations.

5.2.3 Agreement in Wikipedia Talk Pages (AWTP)

Our last corpus is 50 Wikipedia talk pages (used to discuss edits) containing 822 posts (see full statistics in Table 5.4) that were manually annotated as the AWTP corpus as described in Andreas *et*

[6]Wang and Cardie [2014] deviate from this and consider [-5,-1] to be disagreement, [0] to be none, and [1,5] to be agreement

al [2012] and discussed in detail in Appendix A, Section A.3 of this thesis. Although smaller than the IAC, the advantage to this dataset is that each thread was annotated in its entirety. We use this dataset to determine how well the automatically annotated data performs on other online discussions. As in the create debate discussions, disagreement is more common than agreement due to the nature of the discussion. These annotations were on the sentence level where multiple sentences can be part of a single annotation. In 99% of the Q-R posts, there was just one pair of sentences that were annotated with a (dis)agreement label and we used that annotation for the post. When there was one more than one pair, we used the majority annotation. The post was labeled with none only when all sentences within the post had the none label. AWTP was annotated by three different people. Inter-Annotator Agreement (IAA) using the sentence pairs was very high because most annotations were none. Therefore, we computed IAA by randomly sampling an equivalent amount of sentences pairs per label from two of the annotators (A1 & A2) and had the third annotator (A3[7]) annotate all of those sentence pairs. Cohen's κ for A1,A3 was .90 and for A2,A3 was .70 indicating high IAA.

5.3 Method

We model our data by posts. Each data point (the *Response*) is a single post and its label indicates whether it agrees, disagrees, or none, to the post it is responding to (the *Quote*). The following sections discuss the features used to train our model. Each feature is computed within the entire post. In addition, in all applicable features, we also indicate if the feature occurs in the first sentence of the post as our analysis showed that (dis)agreement tends occur within the first sentence of the response. Due to our decision to use Mallet [McCallum, 2002] for classification, all features are binary.

5.3.1 Meta-Thread Structure

We generate several features related to the meta-structure of the threaded discussion:

- **The post is the root of the discussion**: This is useful because the root of the discussion tends to be a question (e.g., "Are you for or against abortion") and thus, does not express (dis)agreement.

[7]We unfortunately could not perform pairwise IAA because annotators A1 and A2 were no longer available

- **The reply was by the same author**: The second post is just a continuation of the first.
- **The distance, or depth, of the post from the beginning of the discussion**: Anyone that replied to the root (Depth of 1) has no (dis)agreement because the root is a question and therefore has no side. The average depth per thread is 4.9 in ABCD, 12.7 in IAC and 6.2 in AWTP.
- **The number of sentences in the response**: People who disagree tend to write more than those who agree.

In the future we would like to add the duration of time between posts as used in prior work [Abbott *et al.*, 2011a; Misra and Walker, 2013].

5.3.2 Lexical Features

Lexical features are generated for each post. We use n-gram features (1-3 words) and also generate 1-4 gram part-of-speech (POS) tag features [Toutanova *et al.*, 2003] for each word in the post. We include all unigram POS tags and perform Chi-Squared feature selection on everything else. The top 15 n-grams per class (normalized by class size) in the ABCD and IAC training data are shown in Table 5.5. It is clear that the ABCD terms are more general, while the IAC terms have more to do with certain topics, probably due to its smaller size.

In addition, we also manually generated small lists of negation terms (e.g. not, nothing; 11 terms in total), agreement terms (e.g. agree, concur; 14 terms in total), and disagreement terms (e.g. disagree, differ; 14 terms in total) and generate a binary feature for each list indicating that the post has one of the terms from the respective list of words. The full list of terms is shown in Table 5.6. Finally, we also include a feature indicating whether there is a sentence that ends in a question as when someone asks a question, it may be followed by (dis)agreement, but it probably won't be in (dis)agreement with the post preceding it.

5.3.3 Lexical Stylistic Features

We include lexical stylistic features that fall into two groups, **general**: ones that are common across online and traditional genres, and **social media**: ones that are far more common in online genres. Examples of general style features are exclamation points and ellipses. Examples of social media

ABCD			IAC		
Agreement	Disagreement	None	Agreement	Disagreement	None
yeah	my	think	yes	the	these
i agree with	that is	good	agree	to	abortion
thank	me	better	for a	that	2
when i	well	i would	i agree	a	answer
thanks	said	always	they will	in	but it
i could	point	feel	often	if	help
funny	i have	long	is why	this	and then
totally	did not	will be	in their	as	3
hey	debate	new	we should	what	of what
haha	here	she	agree with	but	lives
damn	i am not	best	yeah	so	home
i get	though	great	had to	no	wonder
know i	that i	course	yes i	would	argue
i have been	are you	day	to give	of the	kids
i still	i will	i had	if there	do not	time to

Table 5.5: List of top 15 agreement , disagreement, and none n-grams in the ABCD and IAC training datasets.

style features are emoticons and word lengthening (e.g. sweeeet). A comprehensive list of lexical stylistic features can be found in the beginning of this part of the thesis in Section 3.1.

The occurrence of lexical style features in the ABCD and IAC training sets are shown in Figure 5.1. Emoticons and repeated exclamation points are indicative of agreement in the IAC training data and repeated question marks are indicative of disagreement in the IAC training data. Emoticons and acronyms are more indicative of agreement and question marks and all caps are indicative of disagreement in the ABCD training data. Clearly, emoticons and question marks have similar trends in the training data across both genres.

Negation	Agreement	Disagreement
not	agree	disagree
no	admit	deny
nobody	comply	disapprove
nothing	concede	dispute
neither	concur	dissention
nowhere	grant	contradiction
never	recognize	oppose
hardly	consent	refuse
scarcely	accede	reject
barely	okay	repudiate
n't	ok	resist
	permit	protest
	assent	decline
	acknowledge	differ

Table 5.6: Lists of negation, agreement , and disagreement terms used as features.

5.3.4 Linguistic Inquiry Word Count

The Linguistic Inquiry Word Count (LIWC) [Tausczik and Pennebaker, 2010] aims to capture the way people talk by categorizing words into a variety of categories such as negative emotion, past tense, money, and health. It has been used previously in agreement [Abbott *et al.*, 2011a]. The 2007 LIWC dictionary contains 4487 words with each word belonging in one or more categories. We use all the categories as features to indicate whether the response has a word in the category. In general, in the training data, most categories are more indicative of the most popular disagreement class. However, assent is indicative of agreement in the ABCD and IAC training data. The friend category is indicative of none in both training datasets.

5.3.5 Sentiment

By definition, (dis)agreement indicates whether someone has the same, or different, opinion than the original speaker. A sentence tagged with subjectivity can help differentiate between (dis)agreement and the lack thereof, while polarity can help differentiate between agreement and disagreement. We

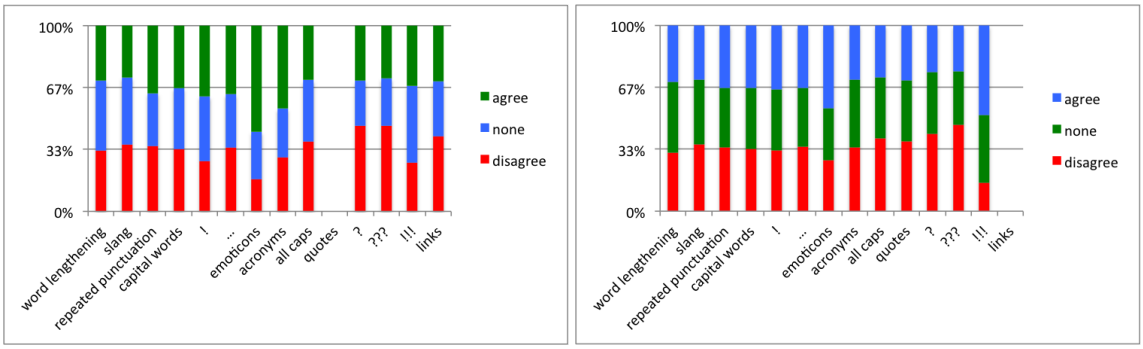


Figure 5.1: Occurrence of lexical style features in the training corpus for ABCD and IAC

use our phrase-based sentiment detection system described in Chapter 4 that has been optimized for lexical style to tag the sentences with opinion and polarity. For example, it produces the following tagged sentence "[That is soo true]/*Obj* [living with the guilt forever]/*neg* [know you murder you child]/*neg*..." We use the tagged sentence to generate several opinion-related features. We generate bag of words for all words in the polarity phrases, labeling each word as to which class it belongs to based on the polarity of the phrase (positive (w_+), or negative (w_-)). We also have binary features indicating the prominence of opinion and polarity (opinion, positive or negative) by using the majority occurrence within the post.

5.3.6 Sentence Similarity

A useful indicator for determining whether people are (dis)agreeing or not is if they are talking about the same topic. We use sentence similarity [Guo and Diab, 2012] to determine the similarity between the Q-R posts. For example the disagreement posts in Table A.4 are similar because of the statements *"LIVE WITH THE GUILT FOREVER!!!!!!!"* and *"living with the guilt forever"*. We use the output of the system to indicate whether there are two similar sentences above some threshold and whether all the sentences are similar to one another.

Furthermore, we also look at similar Q-R phrases in conjunction with sentiment. We generate phrases using the Stanford parser [Socher *et al.*, 2013] by adding reasonably sized branches of the parse tree as phrases. (we found breaking the tree at 15 branches to produce meaningful phrases). We then find the similarity [Guo and Diab, 2012] and opinion of the phrases using the system described Chapter 4. We hypothesize that this could be indicative of (dis)agreement because if similar phrases

have the same polarity (e.g. both are positive) that could indicate agreement and if they have different polarity (e.g. one is positive and one is negative) that could indicate disagreement. We also extract the unique words in the similar phrases as features. We hypothesize that this could help indicate disagreement, for example, if the word "not" was mentioned in one of the phrases, e.g. *"I do not see anything wrong with abortion =/"* vs *"I do see something wrong with abortion ..."*. We also include unique negation terms using the negation list described in the Lexical Feature section and features to indicate whether there is a similar phrase and if its opinion in the Q-R posts are of the the same polarity (agreement) or different polarity (disagreement).

5.3.7 Accommodation

It has been found that when people speak to each other, they tend to take on the speaking habits and mannerisms of the person they are talking to [Giles *et al.*, 1991]. This phenomenon is known as *accommodation*. Mukherjee and Liu [2012] found that accommodation does differ among people who (dis)agree, which strongly motivates the inclusion of accommodation in (dis)agreement detection[8]. We partly capture this via sentence similarity which explores whether they share the same words (as well as whether their words share semantic meaning). We also explore whether Q-R posts use the same syntax (via POS n-grams), copy lexical style, and use the same category of words (via LIWC). We use the features as described in prior sections but only include ones that exist in both the quote and response. The top accommodation features per class (normalized by class size) are shown in the ABCD and IAC training data are shown in Table 5.7. The top features were computed based on occurrence and popularity within the class in comparison to the other classes. For example, capitalized words were a common feature in agreement in ABCD and also occurred more frequently in agreement than in disagreement or none. Much of accommodation in the training data occurs through the response adopting part-of-speech grams of the quote (i.e. syntactic structure). In the Create Debate training data, accommodating towards several lexical styles is indicative of agreement. Interestingly, people who lack (dis)agreement do not seem to accommodate towards the quote. In the IAC training data, similarly to in n-grams, we find that some of the accommodation (e.g. sexual) appears to be topic related, and more domain specific. In the future it would be interesting to look other shared features for accommodation, such as similar emoticon usage.

[8] Accommodation was not used in classifying (dis)agreement in Mukherjee and Liu [2012.

ABCD			IAC		
Agreement	Disagreement	None	Agreement	Disagreement	None
capitalized words^{LS}	function words^{LIWC}	nnp^{POS}	prp vb^{POS}	function words^{LIWC}	nn^{POS}
1st person single^{LIWC}	vb^{POS}	friends^{LIWC}	prp vbp^{POS}	common verbs^{LIWC}	dt^{POS}
assent^{LIWC}	nn^{POS}	vbz jjr^{POS}	assent^{LIWC}	capitalized words^{LS}	in^{POS}
!^{LS}	common verbs^{LIWC}		vb dt nn^{POS}	total PR+DT^{LIWC}	jj^{POS}
. . .^{LS}	cognitive processes^{LIWC}		. . .^{LS}	cognitive processes^{LIWC}	dt nn^{POS}
emoticons^{LS}	total PR+DT^{LIWC}		prp nn in^{POS}	auxiliary verbs^{LIWC}	nns^{POS}
!!!^{LS}	auxiliary verbs^{LIWC}			prepositions^{LIWC}	nnp^{POS}
nn nnp vbp^{POS}	present tense^{LIWC}			present tense^{LIWC}	RPT punctuation^{LS}
vb to jj^{POS}	impersonal PR+DT^{LIWC}			punctuation^{LS}	nn in^{POS}
dt jjs jj^{POS}	rp^{POS}			relativity^{LIWC}	sexual^{LIWC}
vbn vbg dt^{POS}	prp^{POS}			social processes^{LIWC}	to^{POS}
nn rb uh^{POS}	social processes^{LIWC}			personal PR+DT^{LIWC}	3rd person single^{LIWC}
vbd vb prp^{POS}	prepositions ^{LIWC}			articles^{LIWC}	nn nn^{POS}
	rp^{POS}			punctuation^{LS}	sadness^{LIWC}
	prp^{POS}			in^{POS}	to vb^{POS}

Table 5.7: List of top 15 agreement , disagreement, and none accommodation features in the ABCD and IAC training dataset. LS refers to lexical style features, POS refers to part-of-speech features, and LIWC refers to Linguistic Inquiry Word Count features. LIWC pronouns are abbreviated by the Penn Treebank notation as PR+DT.

5.4 Experiments

All of our experiments were run using Mallet [McCallum, 2002]. We experimented with Naive Bayes, Maximum Entropy (i.e. Logistic Regression), and J48 Decision Trees and found that Maximum Entropy consistently outperformed or there was no statistically significant difference to the other classifiers; we only show the results for Maximum Entropy here. We show our results in terms of None, Agreement, and Disagreement F-Score as well as macro-average F-score for all three classes. The ABCD and IAC datasets were split into 80% train, 10% development, and 10% test. We use the entire AWTP dataset as a test set because of its small size. All results are shown using a balanced training set by downsampling and the full test set. It is important to use a balanced dataset for training

Experiment	None	Agreement	Disagreement	Average
majority	63.2	0.0	0.0	21.1
n-gram	45.7	35.6	41.3	40.9
Lexical+Lexical-Style in R	58.7[1]	42.2	51.6	50.8
Thread Structure	100	45.8	62.0	69.2
Sentence Similarity (SS)	64.9	14.8	42.0	40.6
Accommodation w/o SS	71.6	41.0	57.1	57.9
Accommodation	74.0	45.1	59.1	59.4
Thread+Accommodation	99.6	57.8	68.2	75.2
All	99.6	58.0	73.1	76.9
Best	100	58.5	73.0	77.6

Table 5.8: The effect of conversational structure in the ABCD corpus. The results shown are None, Agreement, Disagreement and Average F-score. Lexical+Lexical-Style indicates n-grams, POS, lexical-style, and LIWC features.The best experiment includes the features found to be most useful during development. All results are statistically significant over the n-gram baseline and all results except for one[1] are significant over the majority baseline.

because the ratio of agreement/disagreement/none differs in each dataset. We tuned the features using the development set and ran an exhaustive experiment to determine which features provided the best results and use that best group of features as an additional experiment in the test sets.

In order to show the impact of our large dataset, we experimented with increasing the size of the training set by starting with 25 posts from each class and increased the size until the full dataset is reached (e.g. 25, 50, 100, ...). We also show a more detailed analysis of the various features using the full datasets.

We compute statistical significance using the Approximate Randomization test [Noreen, 1989; Yeh, 2000], a suitable significance metric for F-score. The approximate randomization test examines all n cases where the two results differed and randomly shuffles the results 2^n times or 2^{20} times if $n > 20$. If the difference between the shuffled results are worse than the difference between the actual results a significant amount of times. Significance is computed as $((nc + 1)/(nt + 1)$ where nc is the number of trials that were worse than the actual results and nt is the number of trials) with

the threshold of significance at $< .05$.

We compare our experiments to two baselines. The first is the majority class, which is none. Although none is more common, it is important to note that we would actually prefer to achieve higher f-score in the other classes as our goal is to detect (dis)agreement. The second baseline is n-grams, the commonly used baseline in prior work.

5.4.1 Agreement by Create Debaters (ABCD)

Our initial experiments were performed on the large ABCD dataset of almost 10,000 discussions described in the Data Section. We experimented with balancing and unbalancing the training dataset and the balanced datasets consistently outperformed the unbalanced datasets. Therefore, we only used balanced datasets in the training set for the rest of the experiments. Table 5.8 shows how accommodation and meta-thread structure are very useful for detecting (dis)agreement. In fact, using n-grams, POS, LIWC, and lexical style features in just the response yields an average F-score of 50.8% whereas using POS, LIWC and lexical style that exist in both the quote and response as well as sentence similarity yields a significant improvement of 8.6 points or 16.9% to an average F-score of 59.4%, indicating that conversational structure is very indicative of (dis)agreement. Using all features and the best features (computed using the development set) provide a statistically significant improvement at $\leq .05$ over both baselines. Our best results include all features except for polarity with an average F-Score of 77.6%. As one would expect, Figure 5.2 shows that as the training size increases the results on the ABCD test set improve.

5.4.2 Internet Argument Corpus (IAC)

In contrast to most prior work using the IAC we detect (dis)agreement on the *post* level as a 3-way classification task: agreement, disagreement, none. Detecting (dis)agreement without including none posts is unrealistic in a threaded discussion where the majority of posts will be neither agreement or disagreement. Additionally, we do not balance the test set as do Abbott *et al* [2011a] and Walker *et al* [2013], but rather use all annotated posts to maintain a realistic agreement/disagreement/none ratio.

In this section we analyze how useful a large out-of-domain self labeled dataset is in comparison to gold in-domain labeled data by comparing results from the small manually annotated in-domain IAC corpus and the large self labeled ABCD corpus. In contrast to the ABCD experiments, we

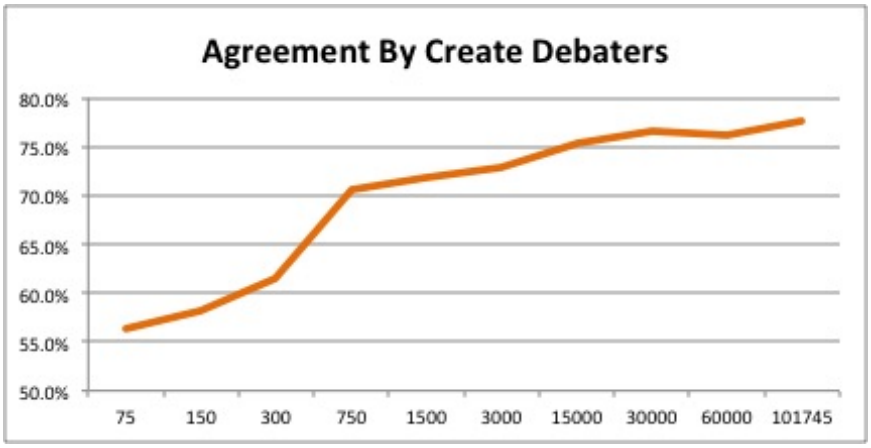
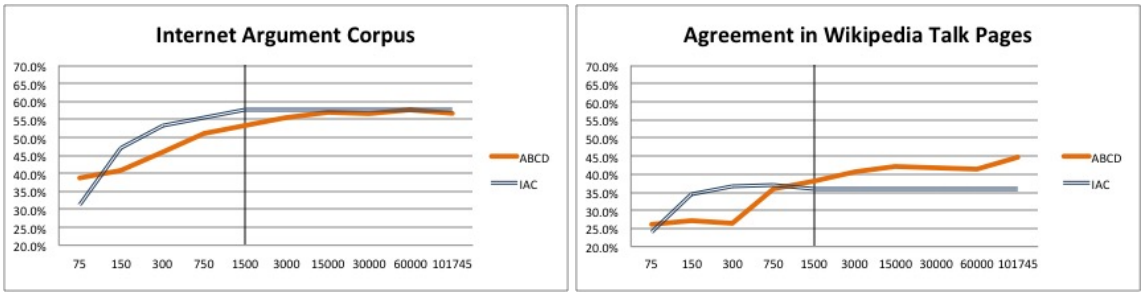


Figure 5.2: Average F-score as the ABCD training size increases when testing on the ABCD.



(a) (b)

Figure 5.3: Avg. F-score as the training size increases. The vertical line is the size of the IAC training set. The F-score succeeding the vertical line is the score at the peak size, included for contrast.

did not find accommodation to be significantly useful when training and testing using the IAC. We believe this is due to the large amount of none posts in the dataset (71.9%) where one does not expect accommodation to occur. However, in examining the average F-score for (dis)agreement, without none, we found that accommodation provides a 2.7 point or 11% improvement over only using features from the response. This improvement is masked by a 1.2 reduction in the none class where accommodation is not useful. The best IAC features differ depending on the training set and were computed using the IAC development set. Using the IAC training set, meta-thread structure, the LIWC, sentence similarity, and lexical style were most important. Using the ABCD corpus, the best features on the IAC development set were meta-thread structure, polarity, sentence similarity, the

Features	IAC				ABCD			
	None	Agree	Disagree	Avg	None	Agree	Disagree	Avg
majority	85.1	0.0	0.0	28.4	85.1	0.0	0.0	28.4
n-gram	58.6	11.7	27.8	32.7	46.7	7.8	36.6	30.3
Lexical+Lex. Style in R	54.1	12.0^{α}	29.7^{α}	31.9	43.9	13.6^{α}	30.1^{α}	29.2
Thread Structure	87.4^{β}	$25.3^{\alpha\beta}$	$50.0^{\alpha\beta}$	54.2^{β}	87.3^{β}	$26.4^{\alpha\beta}$	$53.8^{\alpha\beta}$	55.8^{β}
Sentence Similarity (SS)	25.4	12.0	0.0	12.5	83.7	0.0	7.9 30.5	
Accommodation w/o SS	52.7	14.5	32.3	33.2	48.4	12.5	34.2	31.7
Accommodation	52.9	13.9^{α}	32.4^{α}	33.1	51.7	14.7^{α}	34.3^{α}	33.6
Thread+Accommodation	87.5^{β}	$26.5^{\alpha\beta}$	48.9^{β}	$54.3^{\alpha\beta}$	87.2^{β}	$28.0^{\alpha\beta}$	$55.5^{\alpha\beta}$	56.9^{β}
All	83.5^{β}	$28.8^{\alpha\beta}$	$50.4^{\alpha\beta}$	54.2^{β}	87.3^{β}	$27.0^{\alpha\beta}$	41.2^{α}	51.8
Best	87.4^{β}	$31.5^{\alpha\beta}$	$54.4^{\alpha\beta}$	57.8^{β}	87.3^{β}	$25.5^{\alpha\beta}$	$57.3^{\alpha\beta}$	56.7^{β}

Table 5.9: The effect of conversational structure in the IAC test set using the IAC and ABCD as training data. The results shown are None, Agreement, Disagreement and Average (Avg) F-score. Lexical+Lexical-Style indicates n-grams, POS, lexical-style, and LIWC features. The best experiment includes the features found to be most useful during development and differs per dataset. Results highlighted to indicate statistical significance over majority$^{\alpha}$ and n-gram$^{\beta}$ baselines.

LIWC, and the negation/agreement/disagreement terms and question lexical features. We found it especially interesting that polarity and lexical features were useful on the ABCD while lexical style was useful for the IAC indicating clear variations in content across genres. Using the best features per corpus found from tuning towards the development sets (e.g. training and tuning on ABCD) provide a statistically significant improvement at $\leq .05$ over the n-gram baseline. The best and all (dis)agreement results provide a statistically significant improvement over the majority baseline. More detailed results are shown in Table 5.9. Finally, Figure 5.3a shows how increasing the size of the automatic ABCD training set improves the results compared to the manually annotated training set using the best feature set. Interestingly, there is little variation between the use of both datasets using the best features. We believe this is because thread structure is the most useful feature due to the large occurrence of none posts.

Features	IAC				ABCD			
	N	A	D	Avg	N	A	D	Avg
majority	87.2	0.0	0.0	29.1	87.2	0.0	0.0	29.1
n-gram	68.1	12.7	21.3	34.1	36.5	11.6	32	26.7
Lexical+Lex. Style in R	64.1	12.1^{α}	22.7^{α}	33.0	54.0^{β}	$27.7^{\alpha\beta}$	$36.2^{\alpha\beta}$	39.3^{β}
Thread Structure	58.0	12.4^{α}	23.7^{α}	31.4	63.6^{β}	15.0^{α}	33.4^{α}	37.3
Sentence Similarity (SS)	32.1	8.7	1.3	14.0	72.9	6.4	26.0	35.1
Accommodation w/o SS	53.4	11.1	29.6	31.4	45.5	18.3	40.1	34.6
Accommodation	52.4	12.4^{α}	$30.7^{\alpha\beta}$	31.8	50.7^{β}	$17.5^{\alpha\beta}$	$40.1^{\alpha\beta}$	36.1^{β}
Thread+Accommodation	55.0	14.9^{α}	$37.2^{\alpha\beta}$	35.7	62.9^{β}	$21.3^{\alpha\beta}$	$52.2^{\alpha\beta}$	43.9^{β}
All	64.2	15.5^{α}	$36.4^{\alpha\beta}$	38.7	61.9^{β}	$25.8^{\alpha\beta}$	$43.5^{\alpha\beta}$	43.7^{β}
Best	59.3	14.4^{α}	$34.5^{\alpha\beta}$	36.1	63.6^{β}	$23.3^{\alpha\beta}$	$46.8^{\alpha\beta}$	44.4^{β}

Table 5.10: The effect of conversational structure in the AWTP test set using the IAC and ABCD as training data. The results shown are None (N), Agreement (A), Disagreement (D) and Average (Avg) F-score. Lexical+Lexical-Style indicates n-grams, POS, lexical-style, and LIWC features. The best experiment includes the features found to be most useful during development (but not necessarily at test time) and differs per dataset. Results highlighted to indicate statistical significance over majority$^{\alpha}$ and n-gram$^{\beta}$ baselines.

5.4.3 Agreement in Wikipedia Talk Pages (AWTP)

Our last set of experiments were performed on the AWTP which was annotated in-house (Appendix A, Section A.3). The advantage to the AWTP corpus is that the annotators were given the entire thread during annotation time, and annotated all (dis)agreement, whether between Q-R pairs or not. In contrast, the IAC annotators were not provided with the entire thread. It was annotated only between Q-R pairs and even all Q-R pairs in a thread were not annotated. This means that each AWTP thread can be used for (dis)agreement detection in its entirety. Having fully annotated threads preserves the ratio of agreement/disagreement/none pairs better (the IAC has posts that are missing annotations).

Due to its small size of only 50 discussions we do not have a training set for AWTP. Instead, we experiment with predicting (dis)agreement in AWTP using the large naturally occurring ABCD dataset and the gold out-of-domain IAC dataset as training data. Despite the IAC's advantage of having gold labels, we found that using the ABCD as training consistently outperforms using the

IAC as training on out-of-domain data, excluding when using just n-grams. In contrast to the other datasets, meta-thread structure and accommodation individually perform worse than using similar features found in the response alone. We believe this is because meta-thread structure is not strictly enforced in Wikipedia Talk Pages, providing an inaccurate representation of who is responding to who. Using all and the best features found during development (e.g. via training and tuning on ABCD) provide a statistically significant improvement at $\leq .05$ over the n-gram baseline for ABCD. The all and best (dis)agreement results provide a statistically significant improvement over the majority baseline for training on ABCD and IAC. More detailed results are shown in Table 5.8. We ran identical experiments to those performed on the IAC by increasing the training size of the ABCD corpus and IAC corpus to show their effects on the test set as shown in Figure 5.3b. The IAC dataset performs worse than using the ABCD dataset once the size of the ABCD training set exceeds the size of the IAC training set. This is further indication that automatic labeling is useful.

5.5 Discussion

We performed an error analysis on 50 ABCD posts and 50 IAC posts from the development sets that were predicted incorrectly by the classifier to determine the kind of errors our system was making. In the ABCD posts we focused on agreement posts that were predicted incorrectly as our performance was worst in this class. Our analysis indicated that in most cases, 72.7% of the time, the error was due to the incorrect label; it should have been disagreement or none and not agreement as suggested by the side of the post. This is unsurprising as the label is determined using the side chosen by the post author. However, what is more surprising is that this was the common cause of error in the IAC dataset as well, occurring 58.3% of the time. This is because the IAA using Cohen's κ among Amazon Turk workers for the IAC is low, averaging to .47 [Walker *et al.*, 2012] across all topics. In addition, detecting agreement is hard as is evident in the incorrectly labeled examples in Table 5.11. Other errors were in posts where the agreement was a response, (5 in ABCD, 8 in IAC), an elaboration, (5 in ABCD, 2 in IAC), there was no (dis)agreement, (0 in ABCD, 12 in IAC), and a conjunction indicating the post contained agreement and disagreement. (1 in ABCD, 1 in IAC). The high number of IAC errors where there was no (dis)agreement, (i.e. label was none,) is due to overfitting to the ABCD corpus where the none class is easy to determine. Finally, we

Dataset	Quote	Response	Description
ABCD	The same thing people use all words for; to convey information.	to convey information. Give me an example of when you are fully capable of saying this without offending someone.	The first sentence sounds like agreement but the second sentence is argumentative
ABCD	Belief is mainly opinionated, I do not know that God is real but I still do not fucking believe in him!	Congratulations do you want a medal?	Disagreement. It is sarcasm, or it may be None
IAC	Nowhere does it say, that she kept a gun in the bathroom emoticon_xkill	And nowhere does it say she went to her bedroom and retrieved a gun.	Agreement. It is an elaboration. Further context would help.
IAC	Oh my goodness. This is a trick called semantics. I guess you got sucked in. Yes, abortion is used as contraception unfortunately.	Yea I know that, but there is no "abortion" pill that will terminate a fetus.	Agreement (*Yea I know that*) and Disagreement (*but there is ...*)

Table 5.11: Hard examples of (dis)agreement in ABCD and IAC

noticed in the response and elaboration errors that the first sentence tended to be more indicative of agreement than the rest of the post. This caused us to add the first features which boosted overall performance. To gain true insight into our model and gauge the impact of mislabeling, the labels of a small set of 60 ABCD threads (908 posts) were manually annotated by one external annotator to correct (dis)agreement errors resulting in 99 label changes. We allowed a post to be both agreement and disagreement and avoided changing labels to none as it is not a self-labeling option. We found that this did not provide a significant change in F-score.

As is evident from our experiments, exploiting meta-thread structure and accommodation provide significant improvements. We also explored whether additional context would help by exploring the entire thread structure using general CRF. However, our experiments found that using CRF did not provide a significant improvement compared to using Maximum Entropy in the ABCD and AWTP corpora. This may be explained by our error analysis, which showed that in only 2/50 ABCD posts (4%) and 9/50 IAC posts (18%) further context beyond the Q-R posts would possibly help make it clearer whether it was agreement or disagreement. On the other hand, it is possible that CRF would

help in the IAC, but it is not possible to experiment with this without annotating the threads in their entirety.

5.6 Conclusion

In this chapter we describe a system that detects (dis)agreement in threaded discussions. We have shown that by exploiting conversational structure our system achieves significant improvements compared to using lexical features alone. In particular, our approach demonstrates the importance of meta-thread features, and accommodation between participants of an online discussion reflected in the semantic, syntactic and stylistic similarity between their posts.

Furthermore, one of our main contributions to agreement detection is that we use naturally occurring labels derived from Create Debate, to achieve improvements in detecting (dis)agreement compared to using smaller manually labeled datasets of the IAC and AWTP. Our ABCD dataset is 30 times larger than prior datasets used in agreement detection. Information regarding accessing the datasets is available in Appendix B. We achieve an average F-score of 77.6% using only automatically labeled ABCD data for training and testing. Using the automatically labeled ABCD dataset to train our model with the manually annotated datasets from the IAC and AWTP as testing achieves comparable or better results compared to using the in-domain data with an average F-score of 56.7% on the IAC, and insignificant difference in F-score compared to using the IAC as training and 44.4% on the AWTP and 8.3 point improvement compared to using the IAC as training. There results are promising for domains where no annotated data exists; the dataset can be used to avoid performing a time consuming and costly annotation effort. In the future we would like to take further advantage of existing manually annotated datasets by using domain adaptation to combine the datasets. In addition, our error analysis indicated that a significant amount of errors were due to mislabeling. We would like to explore improving results by using the system to automatically correct such errors in held-out training data and then using the corrected data to retrain the model.

Including agreement in influence detection is motivated by three weapons of influence: reciprocation, commitment and consistency, and scarcity. **Reciprocation** indicates that a person feels obligated to return favors. Therefore if someone agrees with them, they will feel the need to agree back. Similarly, once someone (dis)agrees with someone, for the sake of **commitment and consis-**

tency they will feel obligated to continue (dis)agreeing. Finally, **scarcity**, in the form of restrictions will cause people to defy and disagree with the restrictions being placed. We use the output from the agreement system to generate features for detecting influencers such as the number of times a person was (dis)agreed with and the number of times people (dis)agreed with the person. Further details regarding the features and their usefulness in influence detection is described in Part III of this Thesis.

Chapter 6

Persuasion

> *C is quirky, flawed, and an enormous success*
>
> Dennis M. Ritchie, *The Development of the C Language*

> *Design and programming are human activities; forget that and all is lost.*
>
> Bjarne Stroustrup, *The C++ Programming Language*

> *We don't get a chance to do that many things, and every one should be really excellent. Because this is our life. Life is brief, and then you die, you know? So this is what we've chosen to do with our life.*
>
> Steve Jobs, *Fortune Magazine, 2008*

In this chapter we explore persuasion as a component for detecting influence. There are two types of persuasion: persuade to believe, and persuade to act. Persuade to believe is an attempt to change someone's opinion such as Steve Job's quote at the beginning of this chapter. In contrast, persuade to act is an attempt to convince someone to do something. For example, "It's going to snow, you should go buy the bread and milk." is an example of a persuade to act. In this work we only focus on attempts of persuading someone of a belief.

An attempt to persuade is defined as a set of contributions made by a single participant which may be made anywhere within the thread, and which are all concerned with stating and supporting a single opinionated claim. The subject of the opinionated claim does not matter: an opinion may seem trivial, but the argument could still have the structure of a persuasion. Three types of support can follow a opinionated claim: argumentation, grounding, and reiteration. *Argumentation* is a justification. *Grounding*, is an appeal to an external source, knowledge or authority to support the opinionated claim. Grounding is motivated by both the **liking** and **authority** weapons of influence, which discuss that familiarity and association to the well-known cause influence. *Reiteration* is a restatement or paraphrase of the original opinionated claim. Reiteration is motivated by **commitment and consistency** as a person feels the need to repeat their arguments to be consistent. Finally, in this work, we only explore persuasive attempts to change someone's belief that consist of an opinionated claim followed by at least once instance of argumentation. We did not implement systems for reiteration as it is very hard to detect. The argumentation system also takes care of grounding.

In this chapter we will describe related work, the data, and our systems for detecting opinionated claims, argumentation and combining them to form an attempt to persuade. Examples of attempt to persuade ({opinionated claim, argumentation} pairs) are shown in Table 2.2: $\{pc_1, pa_1\}$ and $\{pc_3, pa_3\}$.

6.1 Related Work

Detecting attempts to persuade falls under the broader area of argumentation mining [Palau and Moens, 2009]. Argumentation mining is defined as the automatic detection of identifying and structuring arguments. Argumentation mining has become a popular topic of interest with recent papers related to dataset creation [Aharoni *et al.*, 2014] and classification [Stab *et al.*, 2014; Ghosh *et al.*, 2014; Schneider, 2014].

Several authors [Anand *et al.*, 2011; Young *et al.*, 2011; Freedman *et al.*, 2011] have explored attempts to persuade similar to our approach in blogs, microtext, and Google groups. These approaches differ from ours in that they look at attempt to persuade that are beliefs (persuasion of an opinion) and acts (persuasion to do something), without differentiating between them. We focus on detecting attempt to persuade that aim to change someone's belief. These papers do not describe a method for

detecting attempts to persuade, but rather how to annotate documents for persuasion and how to use persuasion in a larger system.

As far as we are aware, there is very little work on identifying opinionated claims of the type we describe here. Exceptions are two recent companion papers, where Bender *et al* [2011c] and Marin *et al* [2011], discuss the annotation and detection of authority claims on the sentence level. Authority claims are those in which a person attempts to bolster their credibility in the discussion. They describe several types of authority claims but only run experiments on detecting forum claims. Forum claims are based on "policy, norms, or contextual rules of behavior in the interaction". They detect forum claims using lexical features such as n-grams and Part-of-Speech (POS) and a few other features such as sentence length, capital words, and the number of URLs. Their best result was 63% using a manually tuned list of words. This research is orthogonal to our work as a claim can either be an opinionated belief, an authority claim, or neither. For example, one sentence they provide as authority is "Do any of these meet wikipedia's [[WP:RS — Reliable Sources]] criteria?" [Bender *et al.*, 2011c] which is a question and not belief.

In other related work, Kwon *et al* [2007], identify and classify the main subjective claims in order to understand the entire document in the public's comments about the Environmental Protection Agency (EPA). Their goal is to identify not just whether a sentence is a claim, but if it is the main claim of the writer and classify its stance (support/oppose/propose). They use several similar features to ours: words, bigrams, and subjectivity. They differ from our approach in that they take the entire document into account as opposed to just the sentence by looking at its position and topic. They have an accuracy of 55% using boosting.

Automatically detecting opinionated claims is useful for identifying Disputed Claims (claims that are not trustworthy [Ennals *et al.*, 2010; Adler *et al.*, 2010]), as well as analyzing discourse for social acts such as argumentation (claim followed by justification or a supporting claim) as used in this work, support agreements (two claims that agree on the same topic), and disagreements (two claims disagreeing on the same topic) [Biran and Rambow, 2011b; Schneider *et al.*, 2012; Abbott *et al.*, 2011b; Bender *et al.*, 2011a; Park and Cardie, 2014].

Corpus	Opinionated Claims	Not Opinionated Claims	Subjective Phrases	Objective Phrases	Vocabulary Size
LiveJournal	1410 (56%)	1120 (44%)	3035 (39%)*	4709 (61%)*	4747*
Wikipedia	1282 (64%)	715 (36%)	1319 (37%)	4496 (63%)	4342

Table 6.1: Statistics for each corpus; LiveJournal and Wikipedia. *Subjective phrases, objective phrases and vocabulary size for LiveJournal are based on the portion of the corpus annotated for sentiment.

6.2 Data

In total, the persuasions dataset consists of 309 LiveJournal blogs and 118 Wikipedia Talk Pages. Portions of this dataset were annotated for opinionated claims, argumentation, and opinion.

309 blogs from LiveJournal and 118 discussions from Wikipedia Talk Pages were annotated for attempt to persuade by labeling a pair of sentences as an opinionated claim followed by support in the form of *argumentation*, *grounding*, and *reiteration*. The annotators were asked to first search for opinionated claims in the entry, and then for each opinionated claim that was found, search for argumentations, reiterations, and groundings supporting that opinionated claim. If at least one case of support was found, the annotator was to consider it to be an attempt to persuade. Inter-annotator agreement was performed on a subset of the LiveJournal blogs with a κ of .69 and an F-measure of .75. Although we have annotated the data for different forms of attempt to persuade, we only use the argumentation which accounts for 92% of the persuasion attempts found in the data. Other forms of persuasion attempts are excluded.

A subset of the persuasion corpus was annotated for opinionated claims, regardless of whether it was followed by argumentation. The opinionated claims dataset consists of 161 LiveJournal blogposts and 71 Wikipedia talk pages, with 20 threads from each dataset reserved for training. Each training dataset consists of at least 2,000 sentences that are between 30-120 characters. The statistics of each corpus are described in Table 6.1. The discussion in Wikipedia Talk pages is very argumentative, and therefore rich in opinionated claims; the Wikipedia dataset has 8% more opinionated claims than the LiveJournal dataset. The datasets were annotated for claim by two annotators. The annotators were told that an opinionated claim is an opinionated statement that is a belief that can be justified. The

	Opinionated Claim		Not Opinionated Claim	
	Subjective	Objective	Subjective	Objective
LiveJournal	1.8	2.6	1.5	2.0
Wikipedia	1.6	2.4	1.4	2.0

Table 6.2: The average number of subjective and objective phrases in a sentence that is and is not an opinionated claim.

annotators were given a list of at least 2,000 sentences for each corpus. Our goal is to determine if a sentence is an opinionated claim on its own. Therefore, we did not provide the context of the sentence (we also found that the context was usually not necessary). Upon completion of the annotation, the annotators compared their answers and resolved all disagreements to provide a gold set of annotations for each corpus. Inter-annotator agreement prior to resolving disagreements was .75, with a Cohen's κ of .50 on a subset of 663 LiveJournal sentences. Within a subset of 997 Wikipedia Talk Page sentence the IAA was .79, with a Cohen's κ of .56.

In addition to our training datasets, we also had approximately 10 sentences taken from 20 unseen documents resulting in a test set of around 200 sentences annotated for opinionated claim in both genres. The majority of the sentences in both the training and test sets are identical to the LiveJournal and Wikipedia sentences used for training by the sentiment system described in Section 4. The LiveJournal opinionated claims corpus contains an additional 500 training sentences annotated for claim that are not part of the sentiment corpus and were not annotated for sentiment. Table 6.2 indicates the average number of subjective and objective phrases in sentences that were annotated as opinionated claim and not opinionated claim. Sentences that are opinionated claims tend to have more subjective and objective phrases. All sentences have more objective phrases than subjective phrases. Only 32.5% of Wikipedia sentences and 25.9% of LiveJournal sentences that were completely objective were marked as opinionated claims.

6.3 Claim Detection

To be persuasive, a person must try to convince others of the validity of their opinions. In order to detect this phenomenon, we must be able to detect when someone makes an opinionated claim.

This is known as *Claim Detection* and it explores the detection of sentences that are opinionated in which the author expresses a belief. This is consistent with the definition of *claim* drawn from Oxford Dictionaries Online[1]: *"An assertion of the truth of something, typically one that is disputed or in doubt."* An example of an opinionated claim is evident in the chapter quote by Dennis Ritchie where he states his opinion and belief about the C language.

Given that we aim to identify opinionated and personal views where the author is committed to their opinion we hypothesize that sentiment detection (Chapter 4) and committed belief [Prabhakaran *et al.*, 2010] could be useful. A sentence has sentiment if it conveys an opinion, and a sentence has committed belief if the writer indicates that he believes the proposition. Opinion and belief are both necessary for the detection of opinionated claims. It is possible to have an opinionated term without having an opinionated claim. For example, *happy*, in the phrase *"Happy Birthday"* is a positive opinion term, but the phrase is not a opinionated claim. It is also possible to have a sentence conveying belief without opinion. For example, *is*, in the sentence *"It is raining outside"* indicates that author believes that it is raining, but this is not an opinion.

We present a supervised machine learning approach to detecting opinionated claims where we investigate the impact of the two main features, sentiment and committed belief. We also measure traditional lexical features as well as lexical-stylistic features such as those common in social media (Chapter 3.1). Our approach is trained and tested on two online genres: LiveJournal and Wikipedia. Examples of opinionated claims from each corpus are shown in Table 6.3. This work (originally discussed in [Rosenthal and McKeown, 2012], and extended in this thesis) explores how claim detection differs in different online genres and whether cross-genre classification can be successful.

As discussed, in order to be persuasive, a person must convince others of the validity of their opinions by stating them in a manner that indicates they believe in them. The two main components of claim detection (opinion and belief) can be motivated by several weapons of influence. The **liking** and **reciprocation** influence weapons motivate the use of opinion detection (see Section 4). Committed belief is motivated by **commitment and consistency**; as a person needs to stand by the beliefs they make in their claim in order for them to hold value. There are several examples of opinionated claims within the context of influence, pc_i, in the influence example in Table 2.2.

In the rest of this section, we describe our method, (including opinion and committed belief),

[1]http://oxforddictionaries.com/definition/claim?region=us

LiveJournal	1	oh yeah, you mentioned the race... that is so un-thanksgivingish!
	2	A good photographer can do awesome work with a polaroid or ‘phonecam.
	3	/*hugs/* I feel like I have completely lost track of a lot of stuff lately.
Wikipedia	4	The goal is to make Wikipedia and, more specifically, this article as good as possible.
	5	This was part of his childhood, and should be mentioned in the article.
	6	If the writer has only a slender grasp of relevant issues, material can be wrong.

Table 6.3: Examples of opinionated claims from the LiveJournal and Wikipedia corpus

experiments, and conclude with a discussion of difficult examples and error analysis.

6.3.1 Method

We use a supervised machine learning approach. We hypothesized that sentiment analysis should have a major impact in identifying opinionated claims since such claims are an expression of opinions. In addition, the statement must be an expression of the author’s belief; as noted earlier, it is not enough to have a small amount of subjective material in the sentence. Thus, we also explored the impact of committed belief as it would enable us to determine when the author believes in the expressed proposition. Traditional lexical features as well as POS tags could also play a role, as well as words and abbreviations typically found in social media.

6.3.1.1 Opinion

We detect the subjectivity of each sentence using the sentiment system described in Chapter 4. We experimented with using the gold sentiment annotations and system output sentiment annotations during training of the opinionated claim system. We found that system output performs better since system output is used at test time. We use the output of the sentiment system to compute three features. First, we use it to determine whether sentiment exists in the sentence, second to determine the ratio of the sentence that is subjective. For example, the sentence in Chapter 4, Table 4.2 on has sentiment, and 1/3 of the chunks are subjective. Finally, we also include the important opinionated words by applying χ^2 feature selection on all the words in the subjective phrases. The top opinion words are shown in Table 6.4. It is possible for some words that do not appear to be opinionated to occur (e.g. the). This is because the word may occur frequently in an opinionated phrase. In

LiveJournal			Wikipedia			LiveJournal + Wikipedia		
opinion	belief	n-grams	opinion	belief	n-grams	opinion	belief	n-grams
not	think	i	more	think	is	that	think	is
that	feel	is	metal	agree	i think	not	agree	i
but	could	that	heavy	has	it is	is	has	not
like	love	not	blues	is	not	more	feel	it is
too	done	but	better	want	very	like	make	it
is	know	be	actually	should	think	too	find	that
so	mean	is not	paragraph	needs	this is	very	being	be
because	good	it	present	make	facts	good	good	but
a	fine	like	no		it	better	love	very
very	sure	a	what		be	but	sure	a
much		of	certainly		a	think	makes	is not
think		it is	really		sabbath	things	is	i think
good		the	non-ie		however	a	hope	think
the		they	gets		much	say	because	much
things		very	doing		pov	bad	nice	of

Table 6.4: A list of the most common opinion, belief, and n-gram features from the balanced datasets after applying feature selection. Each list contains features from the opinionated claim class.

addition, as described in Chapter 4 page 30 (Opinion Detection), on occasion the scores assigned by the Dictionary of Affect and Language (DAL) are questionable. For example, both heavy and metal, referring to ”heavy metal”, the topic of one of the Wikipedia Talk page discussions, are considered to be slightly negative words by the DAL.

6.3.1.2 Committed Belief

When reading a text, such as a newspaper article, a human reader can usually determine if the author believes a specific proposition is true. This is the problem of determining *committed belief*; it falls into the general area of determining the cognitive state of the speaker (e.g., [Rao and Georgeff, 1998; Stolcke *et al.*, 2000]). Committed belief is an integral part of claim detection because a statement

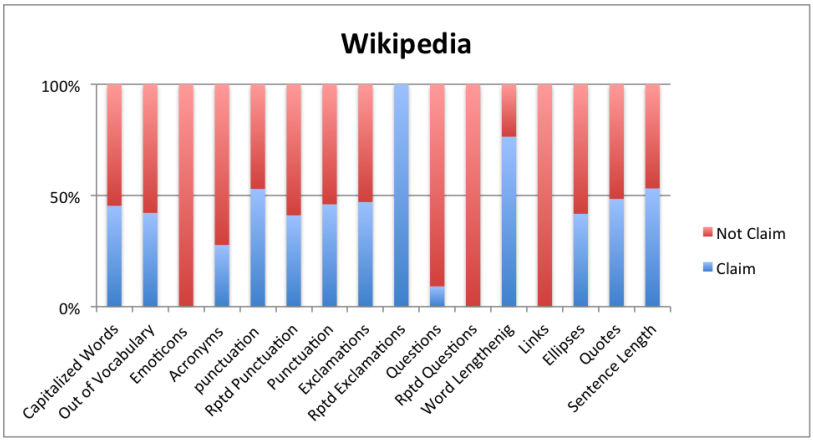


Figure 6.1: Percentage of the lexical stylistic features that are indicative of opinionated claims in the Wikipedia Talk Page corpus.

can not be an opinionated claim without committed belief. However, a sentence can have committed belief and still not be an opinionated claim. For example, in Sentence 1 of Table 6.7, the word *"have"* indicates that the sentence is a committed belief, but the sentence is not an opinion.

Prabhakaran *et al* [2010] created a system that automatically tags the verbs in a sentence for three types of belief [Diab *et al.*, 2009]: committed ("I know"), non-committed ("I may"), and not applicable ("I wish") which vary in intensity from strong to weak respectively. We are interested in all these forms of belief (or the lack of) and used all three of the tags to indicate whether the sentence conveyed belief. Non-committed and not applicable tags should be indicative of a lack of claim. Their system is trained on a diverse corpus of 10,000 words that includes newswire, e-mails, and blogs. The system uses lexical features (e.g. POS, is-number) and syntactic features (e.g. is-predicate, lemma, root of parse) and the best system achieves an accuracy of 64%. We use their system to provide features for determining claims.

We use the output of the belief tagger [Prabhakaran *et al.*, 2010] to generate two types of features. The first feature counts the occurrence of committed and non-committed belief normalized by the length of the sentence. The second set of features extracts the words in the sentence that express belief. We are interested in both committed and non-committed belief and used the occurrence of both these tags to indicate whether the sentence conveyed belief. We use the head verb of the phrase tagged with all three forms of belief in a bag of words approach by counting the occurrence of each

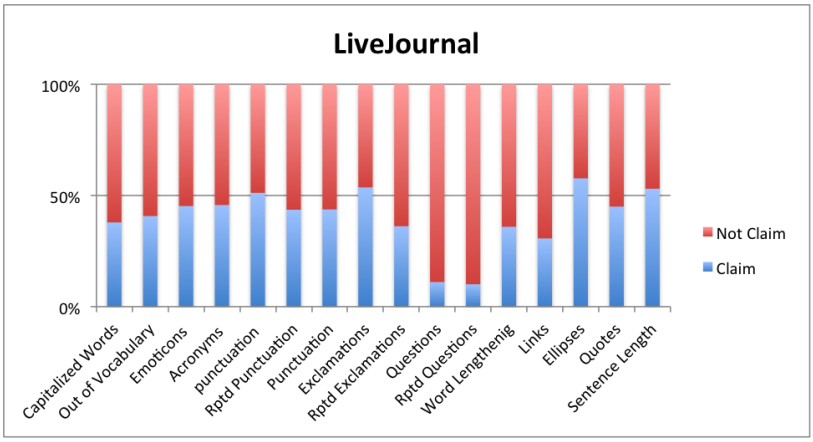


Figure 6.2: Percentage of the lexical stylistic features that are indicative of opinionated claims in the LiveJournal corpus.

of the verbs and performing feature selection on them using the χ^2 test in Weka. The top belief related verb words are shown in Table 6.4. We hypothesize the belief and non-committed belief words can be further indication of an opinionated claim whereas words tagged as being non-belief can be indicative of the sentence not being an opinionated claim.

6.3.1.3 Lexical

Our lexical features include questions, lexical style, n-grams, and Part-of-Speech (POS) tags. Sentences that are questions are often not claims. For example, *"Can you help me fix it?"* Therefore, we use whether the sentence ends in a question mark (?) as a binary feature. We use the question feature as a baseline.

We include Part-of-Speech tags using the CRF tagger [Phan, 2006b] and count the normalized occurrence of all POS tags. We include the top 250 n-gram features (1-3 words) per class. We experimented with taking the top 0, 100, 250, and 500 n-grams. We chose 250 n-grams as more caused the system to perform worse. We performed feature selection on the n-grams using the χ^2 test in Weka. The top n-grams for each class are shown in Table 6.4. We also use the lexical stylistic features, which include features geared towards social media, as described in Chapter 3.1. Across all the datasets, the lexical-stylistic features on average account for less than one word in a sentence, with LiveJournal having more than Wikipedia. Figures 6.2 and 6.1 display the frequency of several

	Balanced			Unbalanced		
Experiment	LiveJournal	Wikipedia	LJ + Wiki	LiveJournal	Wikipedia	LJ + Wiki
majority	N/A	N/A	N/A	49.5	52.5	51.0
question	50.5	62.5	56.5	50.5	62.5	56.5
lexical-style	55.5	66.0	58.9	50.5	63.0	59.7
sentiment	60.4	55.5	60.0	58.4	53	56.2
belief	55.5	60.5	57.7	49.5	54.5	52.2
n-grams	57.4	71.5	**65.9**	55.0	69.5	62.2
pos	56.4	70.5	**65.7**	56.9	68.0	**64.2**
all	57.4	**74.5**	**64.7**	**60.4**	68.5	**64.2**
best	**62.4**	**77**	**70.4**	**62.4**	**73.5**	**68.2**

Table 6.5: Results in accuracy for experiments using a test set on balanced and unbalanced training sets. We use two baselines: The majority class, and the question feature. The features used are question, lexical-style, sentiment, belief, n-grams, and pos. Results in bold are statistically significant over both baselines at $p \leq .05$. The features for the best results shown in the last row differ per experiment as described in the text.

lexical-stylistic features in sentences that are indicative of opinionated claims vs. not opinionated claims across the two genres. Question marks are clearly indicative of a sentence not being a claim. Interestingly, in Wikipedia, repeated exclamation points only occur in opinionated claims, and in LiveJournal repeated exclamation points tend to point to sentences that are not opinionated claims.

6.3.2 Experiments and Results

We ran cross-validation experiments to tune the features and then ran experiments on each test corpus in Weka using Logistic Regression. We experimented with other classifiers as well (e.g. Naive Bayes and SVM) and found that Logistic Regression consistently outperformed or did the same as the other classifiers. The results on each test set is shown in Table 6.5 and 6.6. We compare our results against a majority baseline and the sentence ending in a question mark baseline and compute statistical significance using McNemars test at $p \leq .05$.

The experiments showed that opinion detection is more important than committed belief in LiveJournal (60.4% and 55.5% respectively) , while committed belief is more helpful than opinion

	Balanced		Unbalanced	
Experiment	L-W	W-L	L-W	W-L
majority	N/A	N/A	52.5	49.5
question	62.5	50.5	62.5	50.5
lexical-style	64.5	52.0	65.5	51.5
sentiment	62.0	48.5	63.5	49.5
belief	59.5	55.0	53.5	50.0
n-grams	70.0	49.0	70.5	51.0
pos	**73.5**	60.4	**71.5**	56.9
all	68.0	48.5	**73.5**	53.5
best	**79.0**	60.4	**77.5**	56.9

Table 6.6: Accuracy for using each corpora for training and testing respectively. We experimented with training on LiveJournal and testing on Wikipedia (L-W) and training on Wikipedia and testing on LiveJournal (W-L) with balanced and unbalanced training datasets. The features used are question, lexical-style , sentiment, belief, n-grams, and pos. Results in bold are statistically significant over both baselines at $p \leq .05$. The features for the best results shown in the last row differ per experiment as described in the text

detection for Wikipedia (60.5% and 55.5% respectively). The best feature set in LiveJournal is using opinion and POS tags for both balanced and unbalanced training datasets. The best feature set in Wikipedia is the question, lexical-style, n-gram, and POS features for the balanced dataset and belief, question, opinion, and POS for the unbalanced dataset. Although belief is more useful in Wikipedia it is still not as powerful as the lexical features. These experiments indicate that it is easier to correctly predict opinionated claims in Wikipedia. The last columns in Table 6.5 shows the results for combining LiveJournal and Wikipedia sentences into one corpus. Similarly to the single corpus classification, we find that POS and n-grams are very useful. The best balanced system includes the question, opinion, lexical-style, and POS features and the best unbalanced system includes belief, question, opinion, and POS features. The combined dataset performs better than the LiveJournal experiment but worse than the Wikipedia experiment.

Although necessary for opinionated claims we found that opinion and belief are not the most useful group of features. We hypothesize that this is because both opinion and belief are too common

in sentences that are not opinionated claims as well. In fact, we analyzed a subset of approximately 150 training sentences and found that 90% of the sentences have some sentiment. On the other hand, sentences that are completely objective should be predicted to be not opinionated claims. However, there are difficult cases where this is not the case. For example, the sentiment in the following sentence is very subtle, if it exists: ":| @ dogs being in clothes ... especially dresses .". The emoticon ":|" may make it an opinion but it is not clear. Another example is the sentence "Other metro riders seem to think so" where the opinion is very subtle and it is not even clear that this sentence should be an opinionated claim. Committed belief is not as common as opinion; We counted the number of sentences where at least some committed belief occurred, and it occurs in 29% and 30.5% of the LiveJournal and Wikipedia sentences respectively. For example, the opinionated claim "and, I love this city, it is time to remember that" and a sentence that is not an opinionated claim: "Day three in office and the Global Gag Rule (A.K.A " The Mexico City Policy ") is gone !" both have committed belief. Committed belief does occur more often among opinionated claims: 34.9% to 23% in LiveJournal and 42.5% to 18.5% in Wikipedia for claim and non-claim sentences respectively. This indicates that it should add some value. These percentages reflect our findings as committed belief is more useful in Wikipedia where there is clearly a larger tendency to find committed belief in claims.

In addition to our experiments in each genre, we also ran cross-domain experiments to determine how similar opinionated claims are in Wikipedia and LiveJournal. The cross-domain classification experiments (see Table 6.6) did perform well. In fact, the best system when training on LiveJournal and testing on Wikipedia performed slightly (though not significantly) better than the best system when training and testing on Wikipedia. We experimented with applying domain adaptation (as described in Chapter 10, Section 10.5), but it did not improve the results for these datasets. However, the cross-domain results are encouraging for applying domain adaptation to claim detection in other online genres where little training data exists.

6.3.3 Discussion

Not all sentences can clearly be defined as opinionated claims. It is often difficult to distinguish when a claim expresses a belief as opposed to a request for action or a statement is a fact but not does contain opinion. Sentences 1 and 2 in Table 6.7 are simple examples of a statement of fact and

	Sentence	Description
1	I have a job at Walmart.	fact; not an opinionated claim
2	So If you wish , go ahead and change it back	request; not an opinionated claim
3	Lots of articles get a few instances of vandalism a day	fact or opinion depending on context; may be an opinionated claim
4	Would be good if you could say what those reasons are .	request or belief; may be an opinionated claim
5	You say that " ... for similar reasons I do not think the dead soldier is , either . "	fact; sentence within quotes is an opinionated claim

Table 6.7: Examples of sentences that are clearly not opinionated claims and sentences that are difficult to distinguish.

request for action, respectively. However, in sentence 3, it is unclear whether the claim is a fact or a belief. If the statement was made by a moderator, and followed with a percentage of vandalized articles it would be a fact. On the other hand, it is an opinion if it was said without validation. We do not distinguish between fact and non-opinionated claims. In sentence 4 it is unclear if the statement is a belief or a request for action. In this case it is a request for action, because while the author is stating an opinion, it is an opinion about the request for action, and not an overall belief; a subjective word does not necessarily imply belief. Sentences that include quotes can also be confusing as shown in sentence 5. This sentence is a fact, because the speaker is stating a fact about what someone else said, even though the quote is a belief. In this work, our goal is to detect opinionated claims and we do not differentiate between non-opinionated claims, facts, and requests.

In this regard, the cause for error in the claim detection model often reflected sentences that could be interpreted as a claim or a fact. Table 6.8 shows a list of sentences that were incorrectly classified in each corpus. For example, are the first and fourth sentences in Table 6.8 a claim the person is making about them self, or just a fact? Similarly, is sentence 7 a claim about missing a TV show or are they upset that they missed it? It is also difficult to tell whether sentence 8 is a personal opinion or fact due to the uncertainty in the sentence. Another cause for error was questions that did not have a question mark such as sentence 3, or in contrast sentence 2 which was actually an opinionated claim even though it was a question. Finally, short sentences tend to to not be claims and there is less likelihood of sentiment and belief being useful in short sentences due to less occurrence. This is the

#	source	sentence	opinionated claim
1	Wikipedia	@Somnabot: I exist, and willfully am arguing for this.	Y
2		-ProChoice:ProLife:- Are pictures of protesters really necessary in the article?	Y
3		Do we all agree that the pro-life image should be changed to Image: Pro-life protest.	N
4		(reset indent) Sorry I only violate the NPA when I think the discussion is effectively over.	N
5	LiveJournal	And I was talking about my friends stupid .	Y
6		I thought I was going to die.	Y
7		I missed Kings Of Leon on Conan last night.	N
8		Well, maybe it is something about winter Actually, it was in early autumn...	N

Table 6.8: Examples of sentences that were incorrectly classified by the system.

probable cause for error in sentences 6 and 7.

6.3.4 Conclusion

Our research reveals that sentiment analysis and detection of committed belief play a role in the detection of opinionated claims. However, we discovered that n-grams and POS tags have a stronger impact on accuracy. Furthermore, we found that sentiment was more important for LiveJournal, while committed belief was more important for Wikipedia discussion forums. These results are supported by the fact that sentiment is 5% more common in LiveJournal than in Wikipedia. LiveJournal opinionated claims tend to focus on the emotions as well as likes and dislikes of the poster. In contrast, in Wikipedia, authors are truly arguing for the changes that they want to make. Their posts have more to do with their opinions about the appropriate edits to make and thus, emotions and likes are not central. Finally, our cross-domain classification experiments are encouraging for applying domain adaptation to future datasets where little gold data exists.

6.4 Argumentation

Participants in a conversation need to justify, or argue, the claims they make, to convince the reader the claim is true and/or relevant to the discourse. For example, in the sentence "I'd post an update with the new date immediately in case anyone makes plans between now and when you post the reminder.", the claim "I'd post an update with the new date immediately" is followed by the argumentation "in case anyone makes plans between now and when you post the reminder". Another example is evident in the chapter quote where Bjarne Stroustrup argues (albeit sarcastically) that all will be lost if design and programming are not human activities.

Social science experiments have found that argumentation is indicative of influence. For instance, an experiment [Langer, 1989; Cialdini, 2007] examining requests to cut a waiting line found that if the person followed their request with a reason, people were more likely to concede. In fact, the study showed that people will concede 93% of the time if the requester simply used the word because, even if the reason that followed does not add any new information. There are several examples of argumentation, pa_i, visible in the influencer example in Table 2.2.

In previous work [Biran and Rambow, 2011a; Biran and Rambow, 2011b], my colleagues, Or Biran and Owen Rambow, automatically detected argumentation. We use this work, along with claim detection, in our Attempt to Persuade system. We do not make any improvements to their system.

Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] refers to several relations identified in discourse. In this work, Biran and Rambow [Biran and Rambow, 2011a; Biran and Rambow, 2011b], explore argumentation which does not refer to a single discourse relation (e.g. "JUSTIFY" in RST), but instead a discourse contribution of a large number of relations. They exploit the RST relations in the RST Penn Treebank [Carlson *et al.*, 2003] to extract lists of of indicators of relations and co-occurring content word pairs for each of the indicators (using Wikipedia). The system uses the gold standard opinionated claims found by the annotators to determine how well the system performs in detecting argumentation.

The system is applied to pairs of sentences where the first sentence is known to be an opinionated claim and the second sentence may be an argumentation of the first. The basic features generated for each pair of sentences is sentence length, and whether the argumentation comes before or after the opinionated claim. In addition, the RST relations are used to generate n-gram features by using the spans of text within each relation. Biran and Rambow experimented with several ways of using the

n-grams to generate word pairs between the two sentences such as whether the word pair exists in addition to constraints on the order of the word pair.

The experiments were performed using cross-validation and using a test set on LiveJournal weblogs in Weka using the Naive Bayes classifier. The best F-measure was 50.8% using cross-validation and 47.41% using a test set. The experiments found that discourse relations are indeed useful and that the location of word pairs is important.

6.5 Conclusion

Thus far, we have described our opinionated claim and argumentation components. Our research reveals that sentiment analysis and detection of committed belief play a role in the detection of opinionated claims. However, we discovered that n-grams and POS tags have a stronger impact on accuracy. Research in argumentation detection found that discourse relations are indeed useful and that the location of word pairs is important.

Persuasion is simply defined as an opinionated claim sentence followed by an argumentation sentence, and is found by looking at whether an argumentation closely follows each opinionated claim. All such <opinionated claim,argumentation> pairs are considered an attempt to persuade. Thus, the persuasion pairs are a subset of the argumentation and claim sentences. In the future, we would like to explore additional methods for finding more difficult cases of persuasion, such as those using reiteration and those where the argumentation is within the same sentence as or precedes the opinionated claim.

In Part III of the thesis we use Claim, Argumentation, and Persuasion as three components in detecting influencers. Motivation for including these three components in influence detection is through the reciprocation, liking, commitment and consistency, and authority weapon's of influence. In the beginning of this chapter we defined persuasion as a claim followed by support in the form of grounding, argumentation, and reiteration. Claim is motivated by **reciprocation** and **liking** because the claim must be an opinion. Recall reciprocation is motivation for opinion detection because expressing positive and/or negative sentiment will cause similar sentiment to be returned. In addition, giving praise to a person will cause liking. Claim detection is also motivated by **commitment and consistency** because the claim must be a belief that the person is committed to. Grounding is

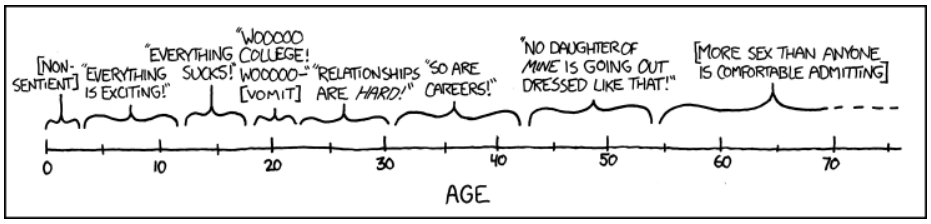
motivated by both the **liking** and **authority** weapons of influence, which discuss that familiarity and association to the well-known cause influence. Reiteration is motivated by **commitment and consistency** as a person feels the need to repeat their arguments to be consistent.

We run each discussion through our claim and argumentation systems which tags sentences as claims and argumentation respectively. We then find persuasion using the output from both systems. The output from each of these components is used to generate features related to occurrence (e.g. the author has a claim), n-grams (the words in the sentences that were tagged as argumentation), and whether it occurs in the first sentence of the post. We suspect that a claim in the beginning of the post is stronger. Further details regarding the features and their usefulness in influence detection is described in Part III of this Thesis.

Chapter 7

Author Traits

“



”

Randall Munroe, *Ages*[1]*, XKCD*

The psychological phenomenon of **social proof** entails that a person will be influenced by others in their surroundings. The social proof phenomenon is apparent when a person attempts to reflect the behavior of others. This occurs because people tend to assume, whether true or false, that others know more about the situation at hand. In fact, social proof, has such a strong influence on people that it has been evident in copycat suicides. Known as the *“Werther effect”*, a suicide tends to cause similar people to commit suicide [Phillips and Carstensen, 1988]. Furthermore, social proof is most effective when a person perceives the people in their surroundings to be similar to them [Cialdini, 2007]. This tendency is known as *homophily*. One manner in which people can be similar is through shared author traits such as demographics as well as political affiliation. Thus, social proof strongly motivates the inclusion of author traits in influence detection.

[1] http://xkcd.com/907/

In addition to social proof, other weapons of influence also motivate including author traits in influence detection: liking and scarcity. First, the association to a person through age, location, or gender will cause **liking**. One such example of this is evident in people rooting for a local sports team. Second, **scarcity** is also motivation for including author traits in influence detection. When an opportunity becomes less available or restricted it causes a *psychological reactance* [Brehm, 1989] to the loss of freedom. This becomes evident across groups of people that are affected by the restriction. For example, teenagers have been found to be most affected by *scarcity* [Driscoll *et al.*, 1972]; perhaps because they are looking for their individuality and are therefore most susceptible to restrictions. Teenagers will act out and defy restrictions. It has also caused reaction within gender groups. For instance, smoking ads have targeted women to be independent and unbound by chauvinistic restraints. This perception implies that women have been restricted in the past by not being independent. It has unfortunately been a successful ad campaign; there has been a rise in the percentage of women who smoke worldwide and it is expected to continue [2].

In this chapter, we present an author trait classifier based on prior state-of-the art methods and apply it towards the detection of four author traits. The author traits we explore are age (year of birth), gender (male/female), and religion (Christian/ Jewish/ Muslim/ Atheist), as well as political party (Republican/Democrat).We use the author trait classifier to build a model to classify each author trait. We train the classifier using automatically labeled or prior existing datasets in each demographic. Our models achieve accuracy comparable to or better than prior work in each author trait. The author trait classifiers are used to automatically label the author traits of each person, and ultimately detect influence (Chapter 10). In particular, we also describe, in depth, the impact of Social Proof in detecting influence in Wikipedia Talk Pages in Chapter 12.

In the rest of this chapter, we first discuss related work in demographic and political affiliation detection. We then describe our author trait classifier and the datasets used to train the models. Information regarding accessing the datasets is available in Appendix B. Afterwards, we describe our method and results.

[2] http://www.who.int/bulletin/volumes/89/3/10-079905/en/

System	Features	Data	Classification	Results
MacKinnon [2007]	age of friends	blogs	exact	5 M
Burger [2006]	lexical, blog style, profile info	blogs	$\leqslant 18$	N/A
Schler *et al* [2006]	lexical and lexical style	blogs	10s/20s/30s	76.2% A
Goswami *et al* [2009]	lexical, slang, sentence length	blogs	10s/20s/30s	80.4% A
Tam & Martell [2009]	n-grams	chat	teens/other	95.5% F
Rao *et al* [2010]	lexical, lexical style, network	Twitter	$\leqslant 30$	74.1% A
Nguyen [2011]	lexical and gender	blogs, forums	exact	5.8 M

Table 7.1: Results in related work of age detection. Results are shown in either Accuracy (A) or F-score (F) for binary classifiers and Mean Absolute Error (M) for regression. Results are averaged for classifiers that are 3-way or more. Lexical features refers to those that are textual in nature such as n-gram and part-of-speech

7.1 Related Work

Prior work in demographic detection has used traditional features such as n-grams (1-3 words), Part-of-Speech (POS) tags (e.g. is the word a noun or verb), and stylistic features (e.g. [Schler *et al.*, 2006; Rao *et al.*, 2010; Mukherjee and Liu, 2010]), as well as genre specific features such as hashtags and the social network in Twitter [Nguyen and Lim, 2014; Burger *et al.*, 2011; Conover *et al.*, 2011; Zamal *et al.*, 2012] and friends and interests in LiveJournal [Burger and Henderson, 2006]. In this work we aim to make our author trait detector as general as possible and therefore only use features available in all online discussion forums by excluding genre specific features. Thus, our system performs as well or better than the results of prior work that exclude genre specific features. It does not, however, perform better than systems using genre-specific features as reported in the following section (e.g. gender results in Twitter [Burger *et al.*, 2011; Bamman *et al.*, 2012]).

7.1.1 Age

Most prior work in age detection has explored 2-3 way classification by dividing the people into categories by age in blogs [Schler *et al.*, 2006; Goswami *et al.*, 2009] and tweets [Rao *et al.*, 2010]. In previous work, Mackinnon [2007], predict age with very high accuracy using only the social network of a person. This, however, is not always available. Initial research on predicting age

without using the ages of friends focuses on identifying important candidate features, including blogging characteristics (e.g., time of post), text features (e.g., length of post), and profile information (e.g., interests) [Burger and Henderson, 2006]. They aimed at binary prediction of age, classifying LiveJournal bloggers as either over or under 18, but were unable to automatically predict age with more accuracy than a baseline model that always chose the majority class. Prior work by Schler *et al* [2006] has examined metadata such as gender and age of bloggers in blogger.com. In contrast to our work, they examine bloggers based on their age at the time of the experiment, whether in the 10's, 20's or 30's age bracket. They identify interesting changes in content and style features across categories, in which they include blogging words (e.g., "LOL"). They can distinguish between bloggers in the 10's and in the 30's with relatively high accuracy (above 96%) but many 30s are misclassified as 20s, which results in a overall accuracy of 76.2%. Their work shows that ease of classification is dependent in part on what division is made between age groups and in turn motivates our decision to study whether the creation of social media technologies can be used to find the dividing line(s). Goswami *et al* [2009] add to Schler *et al*'s approach using the same data and have a 4% increase in accuracy. However, the paper is lacking details and it is entirely unclear how they were able to do this with fewer features than Schler *et al*. In other work, Tam and Martell [2009] attempt to detect age in the NPS chat corpus between teens and other ages. They use an SVM classifier with only n-grams as features. They achieve $> 90\%$ accuracy when classifying teens vs 30s, 40s, 50s, and all adults and achieve at best 76% when using 3 character gram features in classifying teens vs 20s. Age prediction has also been explored in Twitter [Rao *et al.*, 2010] using network structure in addition to contextual features. Recent work has explored predicting the exact age using regression in blogs [Nguyen *et al.*, 2011] and Dutch tweets [Nguyen *et al.*, 2013a]. A comparison of all prior work is shown in Table 7.1. It is important to note that although the results are shown it is difficult to directly compare them among all the systems as the datasets and classification setup vary.

7.1.2 Gender

Prior work in gender detection has detected gender using binary classification in blogs [Schler *et al.*, 2006; Mukherjee and Liu, 2010; Goswami *et al.*, 2009; Nowson and Oberlander, 2006; Herring and Paolillo, 2006; Yan and Yan, 2006] and Twitter [Rao *et al.*, 2010; Burger *et al.*, 2011; Bamman *et al.*, 2012]. Schler *et al* [2006] take the same approach that they used to predict age, and

System	Features	Data	Results
Schler *et al* [2006]	lexical and lexical style	blogs	80.1% A
Goswami *et al* [2009]	lexical, slang, sentence length	blogs	89.3% A
Herring [2006]	word forms (e.g. POS)	blogs	N/A
Yan *et al* [2006]	lexical and lexical style	blogs	60% F
Nowson *et al* [2006]	dictionary and n-grams	blogs	91.5% F
Mukherjee & Liu [2010]	lexical	blogs	88.6% A
Rao *et al*[2010]	lexical, lexical style, and network structure	Twitter	72.3% A
Burger *et al* [2011]	n-grams and profile info	Twitter	91.8% A
Bamman *et al* [2012]	lexical, lexical style, clustering	Twitter	88.0% A

Table 7.2: Results in related work of gender detection. Results are shown in either Accuracy (A) or F-score (F). Lexical features refers to those that are textual in nature such as n-gram and part-of-speech

obtain an accuracy of 80.1%. Herring *et al* [2006] found that the typical gender related features were based on genre and independent of author gender. Yan *et al* [2006] used text categorization and stylistic web features, such as emoticons, to identify gender and achieved 60% F-measure. Nowson *et al* [2006] employed dictionary and n-gram based content analysis and achieved 91.5% accuracy using an SVM classifier. Mukherjee and Liu [2010] detect gender in a small set of blogs using contextual features such as n-grams, POS, and collocations. Similar approaches have been applied in Twitter [Rao *et al.*, 2010; Burger *et al.*, 2011; Bamman *et al.*, 2012]. In addition, to the prior features discusses these methods have explored using genre related features such as hashtags and retweets [Rao *et al.*, 2010; Burger *et al.*, 2011] or using followers to employ clustering techniques [Bamman *et al.*, 2012]. A comparison of all prior work is shown in Table 7.2. It is important to note that although the results are shown it is difficult to directly compare them among all the systems as the datasets and classification setup vary.

7.1.3 Politics

Most prior work in predicting political orientation or ideologies has focused on predicting political views as left-wing vs right-wing in Twitter [Conover *et al.*, 2011; Cohen and Ruths, 2013; Rao *et al.*, 2010] or debates [Iyyer *et al.*, 2014; Gottipati *et al.*, 2013]. These prior works combine Liberals with

System	Features	Data	Classification	Results
Conover *et al* [2011]	lexical, hashtags, network structure	Twitter	L/R	94.9%
Cohen & Ruths [2013]	lexical, hashtags	Twitter	L/R	87%
Rao *et al* [2010]	lexical, lexical style, network structure	Twitter	D/R	N/A
Iyyer *et al* [2014]	lexical, RNN	debates	L/R	69.3%
Gottipati *et al* [2013]	political stances, clustering	debates	L/R	88.9%

Table 7.3: Results in related work of political detection. Results are shown in accuracy. Lexical features refers to those that are textual in nature such as n-gram and part-of-speech. Classification is between either Left vs Right wing (L/R) or Democrat vs Republican (D/R).

System	Features	Data	Classification	Results
Nguyen and Lim [2014]	lexical, network structure	Twitter	Christian vs Muslim	89.5% F
Koppel *et al* [2009a]	lexical	Twitter	Islamic Ideology	73.0% A

Table 7.4: Results in related work of religion detection. Results are shown in either Accuracy (A) or F-score (F). Lexical features refers to those that are textual in nature such as n-gram and part-of-speech.

Democrats and Conservatives with Republicans or classified just Republican vs Democrat [Rao *et al.*, 2010]. A comparison of all prior work is shown in Table 7.3. It is important to note that although the results are shown it is difficult to directly compare them among all the systems as the datasets and classification setup vary.

7.1.4 Religion

There is little work on predicting religion with the only known prior work found to be on the prediction of Christian vs Muslim Twitter users from Singapore [Nguyen and Lim, 2014] and work on classifying documents by Islamic ideology (e.g Muslim Brotherhood) and organization (e.g. Hamas) [Koppel *et al.*, 2009a]. A comparison of the prior work is shown in Table 7.4. It is important to note that although the results are shown it is difficult to directly compare them among all the systems as the datasets and classification setup vary.

7.1.5 Other

Finally, there is also related work in author profiling or attribution, most notably the PAN evaluation tasks, where several demographics, among other author attributes such as those related to personality, are explored at once [Koppel *et al.*, 2009b; Argamon *et al.*, 2009; Rangel *et al.*, 2013; Rangel *et al.*, 2014]. In this related work, age and gender are explored using methods as described above. However, politics and religion are not classified.

In contrast to prior work we develop models for all four author traits using the same system. We exclude genre specific features (e.g. hashtags in Twitter) and domain specific features such as (e.g. ideological stance) since we are interested in a more general system that can be ported to other genres and domains. It is important to note that this does cause a reduction in performance within Twitter where hashtags are very useful features. We build models to predict age and gender using similar methods to prior work. However, we use a much larger dataset. We also build two models to predict age: binary classification and linear regression. In contrast to most prior work we use the year of birth as the label, since the age of a person changes. In political detection, we chose to classify political party as Republican or Democrat. In comparison to prior work in religion which was between two religions, we perform a 4-way classification of Christians, Muslims, Jews, and Atheists.

7.2 Data

Our author trait data comes from two different types of online sources; weblogs (livejournal.com, blogger.com) for age and gender and microblogs (Twitter) for politics and religion. Information regarding accessing the datasets is available in Appendix B.

7.2.1 Age and Gender

We use the publicly available blogger.com authorship corpus [Schler *et al.*, 2006] and our LiveJournal age corpus [Rosenthal and McKeown, 2011] to detect age and gender. The Blogger corpus is self-labeled for age and gender (It is also labeled for industry and astrological sign where available. We do not use these labels here) while the LiveJournal corpus provides the date of birth for each poster. For uniformity, we converted the blogger age in the authorship corpus to the date of birth based

author trait	source	label	size
age	blogger.com	year of birth	19098
	livejournal.com	year of birth	21467
gender	blogger.com	Male	9552
		Female	9546
	livejournal.com	Male	4249
		Female	3287
political party	twitter.com	Republican	1247
		Democrat	1200
religion	twitter.com	Christianity	5207
		Islam	1901
		Atheism	1815
		Judaism	1486

Table 7.5: The size (in users) of each author trait corpus

on the time of download (2004). For example, a 22 year old in 2004 was born in 1982. We then automatically generated gender labels for the LiveJournal corpus internally. We generate gender labels by looking at the first name of the blogger if it was provided. We used the Social Security Administration lists[3] to determine the appropriate gender based on the popularity of the name for each gender. If the name is predominantly male or female at a 2:1 ratio we assign it that gender. Otherwise, we exclude the blogger from the gender corpus. The size of the age and gender corpora are shown in Table 7.5.

7.2.2 Politics and Religion

There are several websites that either automatically generate labels (tweepz.com), or allow users to self-label (twellow.com and wefollow.com) their Twitter account into categories. Previous work [Zamal *et al.*, 2012] has used the labels from wefollow.com to automatically download Twitter users related to desired categories. We follow this approach to download Twitter users based on political party (Republican/Democrat), and religion (Christian, Jewish, Muslim, Atheist). We scraped

[3]http://www.ssa.gov/oact/babynames/limits.html

each of the websites using a python script that used the Mechanize [4] headless browser and regular expressions to find the appropriate content on each website. After downloading the list of users we performed some post-processing to exclude non-English speakers based on the language in their bio . We excluded any users whose bios contained many (40%) foreign characters and non-english words. Additionally, we discarded users that appeared in more than one category within a single author trait (e.g. a person cannot be labeled as Republican *and* Democrat).

We then used the Twitter API to download the last 100 tweets of each user on November 4th, 2014. Downloading on this date was desirable because it ensured that the data was rich in political information because it was election day in the US. We only include the most popular political parties and views. Our political party tweets consist of Republican and Democrat. We downloaded tweets pertaining to the four most popular religions in the United States[5]: Christianity, Judaism, Islam, and Atheism. The full data statistics are provided in Table 7.5.

7.3 Method

We present a supervised method that draws on prior work in the area as discussed in the prior section. We experimented with different classifiers in Weka including Naive Bayes, SVM, and Logistic Regression and found that SVM (referred to as SMO in Weka) always performs the same or better than the other methods. We use this classifier to build several models which detect each author trait by training and testing on the relevant data (e.g. the classifier is trained using the age data to build a model to predict age). The only exception is that we use Linear Regression to predict the exact age of each user using year of birth. We apply χ^2 feature selection to all groups of features to reduce the feature set to the most useful features. We generate the features by looking at the past 100 tweets or 25 blogs per user per corpus. We also limit the text to 1000 words per user to improve processing time. We include three type of features: lexical, lexical-stylistic, and online behavior.

LIWC		Syntax Bigrams	
Old	Young	Old	Young
function words	assent	i have	i dont
prepositions	swear words	i am	i u
relativity	non-fluencies	i and	well i
cognitive processes	fillers	that have	i love
common verbs		you have	i cant
total pronouns		i i	i n
articles		is what	i cuz
space		i do	i didnt
auxiliary verbs		i think	dun i
social processes		it has	i dunno

Table 7.6: Age Features: The top 10 LIWC categories and syntax bigrams for younger and older people.

7.3.1 Lexical Features

We include three kinds of lexical features: n-grams, part-of-speech (POS), and collocations which have all been found to be useful in prior work [Schler *et al.*, 2006; Rao *et al.*, 2010; Mukherjee and Liu, 2010; Rosenthal and McKeown, 2011]. We keep the top 1000 features of each type. Collocations are syntax bigrams that take the subject/object (S/O) relationship or verb/object (V/O) relationship of terms into account. We achieve this using Xtract, the collocation method developed by Frank Smadja [1993]. We ran our own implementation of Xtract on the most recent 100 blog posts or tweets per user. In the Twitter datasets we run Xtract on all the text. Due to the large size of the blog corpora, we limit it to the 2,000 most recent words per user. We include the syntax bigrams (e.g. voting democrat), POS bigrams (e.g. we VB) and collocation pos bigrams (e.g. vote NN) generated from Xtract as features.

The top 10 collocation bigram features within each author trait are shown in Tables 7.6,7.7,7.8, and 7.9. We compute the top 10 features by examining the difference in occurrence of each feature

[4]https://pypi.python.org/pypi/mechanize/

[5]www.census.gov/compendia/statab/cats/population/religi-on.html

LIWC		Syntax Bigrams	
Male	Female	Male	Female
articles	personal pronouns	it has	i i
prepositions	total pronouns	urllink this	i and
space	first person singular	united states	i have
work	total function words	you get	i am
achievement	social processes	check it	i love
relativity	common verbs	this post	i in
leisure	cognitive processes	john kerry	i think
quantifiers	present tense	it does	i do
money	3rd person singular	george bush	my in
tentativeness	adverbs	first post	do you

Table 7.7: Gender Features: The top 10 LIWC categories and syntax bigrams for males and females.

normalized by the total occurrence per trait. Both young and old people write in the first person, but older people talk more about *"having"* and younger people talk more about *"not having"* (e.g. *"don't"*, *"can't"*). Women frequently write in the first person (e.g. *"i"*) and men talk about politics (e.g *"George Bush"*, *"John Kerry"*). Republicans use the word *"republican"* frequently and also *"god bless"*. Democrats frequently talk about *"President Obama"* as well as political issues such as *"climate change"* and *"minimum wage*. Atheists talk most often in the first person (e.g. *"i"*). Christians use *"you"* frequently, as well as *"thank"* and *"god"*. Jews talk a lot about religious holidays (e.g. *"rosh hashana"*, *"yom kippur"*. Interestingly, Muslims, use Twitter specific words, such as *"follow"* frequently. They also mention God (*"Allah"*) frequently.

7.3.2 Lexical-Stylistic Features

We include two types of lexical-style features: general and social media. General features can be found in any genre, such as the number of capital words, exclamation points, and question marks. Social Media features are those common in online discussions such as word lengthening (e.g. looooooong), emoticons, and acronyms. These features were described in more detail in the beginning of this part of the thesis in Section 3.1. Figure 7.1 shows the difference among the

LIWC		Syntax Bigrams	
Republican	Democrat	Republican	Democrat
articles	biological processes	happy birthday	president obama
social processes	assent	republican party	mitch mcconnell
religion	sexual processes	god bless	climate change
		county gop	minimum wage
		republican the	scott brown
		new ad	marriage equality
		harry reid	lib dem
		kay hagan	democratic underground
		republican women	supreme court
		county party	koch brothers

Table 7.8: Political Party Features: The top 10 LIWC categories and syntax bigrams for Republicans and Democrats.

significant features per author trait. Younger people tend to use the social media features more: slang, emoticons, acronyms, and word lengthening. Older people use more quotes and links. Women use more emoticons and repeated punctuation. Men use more quotes. In general, there is very little significant difference in lexical style based on political party. However, Democrats do use more emoticons than Republicans. Finally, in general, Muslims use more lexical stylistic features than the other religions. Particularly, Muslims often use slang, emoticons, and word lengthening. Jews and Christians use the most exclamation points, and Atheists use the most question marks.

Finally, in our recent experiments, we also include LIWC categories [Tausczik and Pennebaker, 2010] as features as in prior work [Schler *et al.*, 2006]. Some examples of these categories are positive or negative emotion, technology words, and money words. The top 10 LIWC features within each author trait are shown in Tables 7.6,7.7,7.8, and 7.10. We compute the top 10 features by examining the difference in occurrence of each feature normalized by the total occurrence per trait. Older people write more coherent sentences as evident by function words, prepositions, and pronouns. Younger people use more swear words and non-fluencies. Men use articles and prepositions often, and talk about work and money. Women use pronouns often and talk in the first person. Republicans talk more about religion while Democrats talk more about biological and sexual processes. Christians

Atheists	Christians	Jews	Muslims
i think	have you	join us	one person
am i	you do	jewish community	followed person
what is	thank you	shabbat shalom	people followed
i do	god bless	thank you	followed me
it like	what is	rosh hashana	me unfollowed
you like	am i	yom kippur	unfollowers person
say i	pope francis	new york	yang dan
you get	love you	see you	may allah
have it	what do	terror attack	happy birthday
i wish	check it	new blog	get statistics

Table 7.9: Religion Features: The top 10 syntax bigrams for Atheists, Christians, Jews, and Muslims.

talk about religion the most. Jews talk about space and time. Muslims talk about humans and family. Atheists use function words and pronouns most often. Although, not in the top 10 they also use more swear words than others.

7.3.3 Online Behavior

Online behavior refers to features such as comments and friends that are often generated from a profile page. However, these features are not always available. In our earlier experiments in age detection [Rosenthal and McKeown, 2011] we included all such features available in LiveJournal. The full list of features are shown in Table 7.11.

In our more recent work, we want to avoid being genre dependent. Therefore, we exclude all features that don't occur in all datasets (e.g. comments, friends, interests, and hashtags). There is one online behavior feature that is found in all discussions. That is a time-stamp indicating when the person posted. We used this to generate two features, the mode time (e.g. 10 PM GMT) and mode day (e.g. Sunday). The mode time, refers to the most common hour of posting from 00-24 based on GMT time. We didn't compute time based on the time zone because city/state is often not included. In our experiments we found the time does not provide a significant increase in accuracy. Therefore, we discarded it as a feature in our author trait experiments. We did, however, include it in the model

Atheists	Christians	Jews	Muslims
function words	function words	relativity	assent
cognitive processes	prepositions	prepositions	humans
pronouns	social processes	function words	numbers
common verbs	relativity	space	family
auxiliary verbs	religion	time	
present tense	cognitive processes	articles	
first person singular	common verbs	work	
personal pronouns	positive emotion	inclusion	
impersonal pronouns	pronouns	leisure	
adverbs	inclusion	first person plural	

Table 7.10: Religion Features: The top 10 LIWC categories for Atheists, Christians, Jews, and Muslims.

Feature	Explanation	Example
# of Friends	Number of friends the blogger has	45
# of Posts	Number of downloadable posts (0-25)	23
# of Lifetime Posts	Number of posts written in total	821
Time	Mode hour and day the blogger posts	11/Monday
Comments	Average number of comments per post	2.64

Table 7.11: List of online behavior features

used to predict author traits for influencers for completeness.

7.4 Experiments and Results

We trained our classifier on each author trait. The classifier was tuned using cross-validation and all results are shown on a held-out test set of 10% of the data. All datasets were kept unbalanced. The gender, religion, political party, and binary age author traits were classified using Support Vector Machines (SVM). The results are shown in Table 7.12. It is clear that n-grams alone is the most useful feature across all author traits. In the following sections we discuss the results for each author trait. In particular, we focus on several experiments that were the early focus of the dissertation work

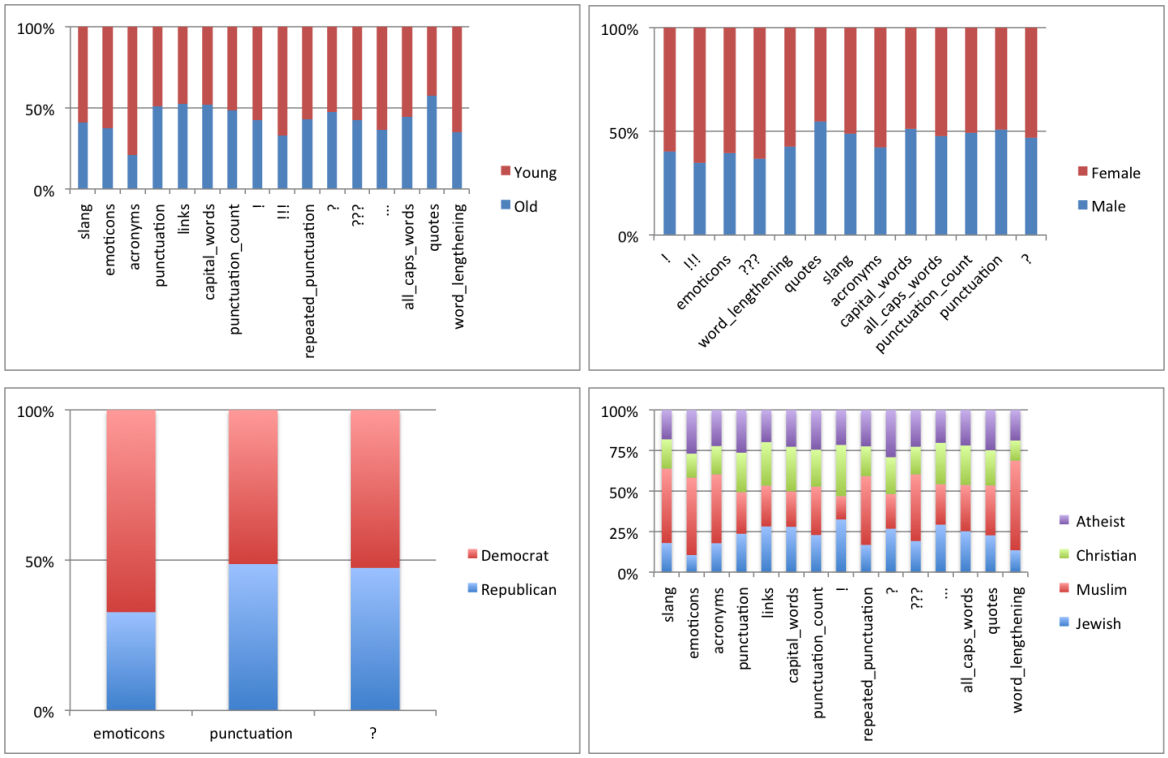


Figure 7.1: Significant lexical style features per author trait.

in age detection. These experiments laid the initial groundwork for all of author trait detection.

7.4.1 Age

We classified age using two models. First, we tried to predict the exact year using Linear Regression; we achieved a Mean Absolute Error (MAE) of 5.1 years and a .55 correlation (r) which is slightly better than the results in prior work [Nguyen *et al.*, 2011] when avoiding blog-specific features.

The next approach we took was performing binary classification using 1982 as the splitting point. This year was found to be significant in our earliest work on age detection [Rosenthal and McKeown, 2011]. In this work we used logistic regression over 10 runs of 10-fold cross-validation using the t-test to compute statistical significance.

We approached age prediction as attempting to identify a shift in writing style over a 14 year time span from birth years 1975-1988:

For each year X = 1975-1988:

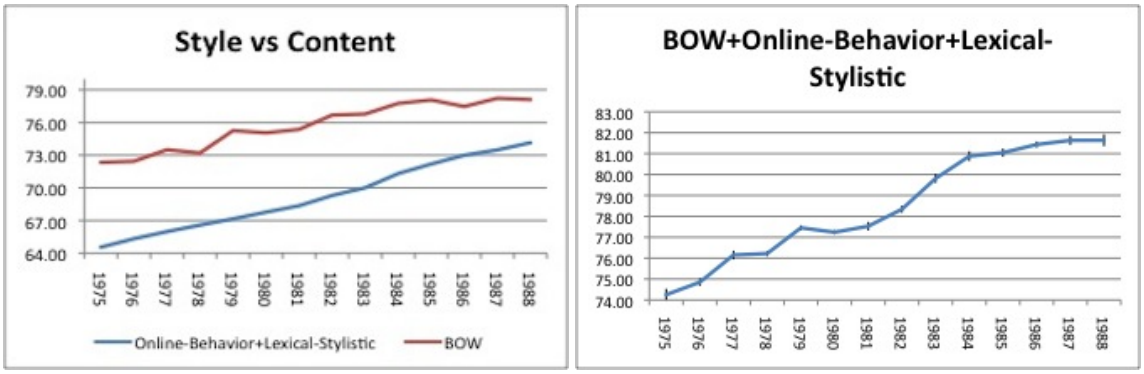
Author Trait	Age	Gender	Political Party	Religion
Majority	57.1	51.9	51.3	50.0
n-grams	78.6	75.1	69.9	79.3
Lexical-Style	52.9	62.2^{α}	56.5^{α}	54.5
LIWC	70.6^{α}	70.5^{α}	52.4	62.7^{α}
POS	72.4^{α}	69.0^{α}	65.9^{α}	74.7^{α}
collocations + POS collocations	76.7^{α}	73.1^{α}	71.5	74.3^{α}
All	$79.6^{\alpha\beta}$	$76.4^{\alpha\beta}$	$75.2^{\alpha\beta}$	78.3^{α}

Table 7.12: The author trait results of SVM classification shown using accuracy. Significance is shown in comparison to majority$^{\alpha}$ and n-gram$^{\beta}$ baselines at $p \leq .05$

- get 1500 blogs ($\sim$33,000 posts) balanced across years BEFORE X
- get 1500 blogs ($\sim$33,000 posts) balanced across years IN/AFTER X
- Perform binary classification between blogs BEFORE X and IN/AFTER X

The experiment focuses on the range of birth years of bloggers from 1975-1888 to identify at what point in time, if any, shift(s) in writing style occurred among college-aged students in generation Y. We were motivated to examine these years due to the emergence of social media technologies during that time. Furthermore, research by Pew Internet [Zickuhr, 2010] has found that this generation (defined as 1977-1992 in their research) uses social networking, blogs, and instant messaging more than their elders. The experiment is balanced to ensure that each birth year is evenly represented. We balance the data by choosing a blogger consecutively from each birth year in the category, repeating these sweeps through the category until we have obtained 1500 blogs. We chose to use 1500 blogs from each group because of processing power, time constraints, and the amount of blogs needed to reasonably sample the age group at each split. Due to the extensive running time, we only examined variations of a combination of online-behavior, lexical-stylistic, and BOW features.

Figure 7.2a shows that content helps more than style, but style helps more as age decreases. However, as shown in Figure 7.2b, style and content combined provided the best results. We found 5 years to have significant improvement over all prior years for p $\leq$ .0005: 1977, 1979, and 1982-1984.



(a) Style vs Content (b) Style and Content

Figure 7.2: Accuracy from 1975-1988 for Style vs Content (a) and Style + Content (b). Style is Online-Behavior+Lexical-Stylistic features. Content is Bag-of-Words features (BOW).

Generation Y is considered the social media generation, so we decided to examine how the creation and/or popularity of social media technologies compared to the years that had a change in writing style. We looked at many popular social media technologies such as weblogs, messaging, and social networking sites. Figure 7.3 compares the significant birth years 1977,1979, and 1982-1984 against when each technology was created or became popular [Urmann, 2009] among college aged students. We find that all the technologies had an effect on one or more of those years. AIM and weblogs coincide with the earlier shifts at 1977 and 1979, SMS messaging coincide with both the earlier and later shifts at 1979 and 1982, and the social networking sites, MySpace and Facebook coincide with the later shifts of 1982-1984. On the other hand, web forums and Twitter each coincide with only one outlying year which suggests that either they had less of an impact on writing style or, in the case of Twitter, the change has not yet been transferred to other writing forms.

In our recent work we chose to split the dataset using 1982, the middle most significant year. The newer system, which includes more data ([Schler *et al.*, 2006]) and more features has an accuracy of 79.9% on a held-out test set compared to 78.2% using cross-validation. The LIWC, n-grams, POS, and the syntax collocations features are all useful for detecting age. The lexical-style features do not perform well compared to the majority baseline.

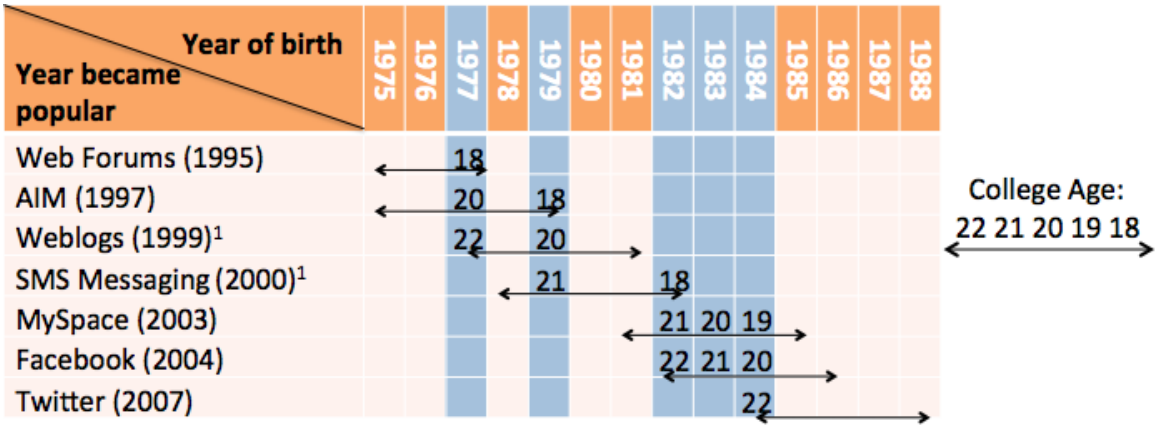


Figure 7.3: The impact of social media technologies: The arrows correspond to the years that generation Yers were college aged students. The highlighted years represent the significant years in our experiments. [1]Year it became popular (the technology was available prior to this date)

7.4.2 Gender

Our accuracy of 76.4% (Table 7.12) on gender detection is slightly worse than leading methods [Schler *et al.*, 2006; Mukherjee and Liu, 2010]. However, we think this is due to prior work using cross-validation as opposed to a held-out test set. In fact, our cross-validation results were 82.5%, slightly better than Schler *et al* [2006]. It is more difficult to compare to the work of Mukherjee and Liu [2010] as the datasets are different and much smaller in size. Mukherjee and Liu have a collection of blogs from several websites (e.g. technorati.com and blogger.com) and only 3100 posts. In contrast we generate our model with blogs from livejournal.com and blogger.com [Schler *et al.*, 2006] and over 25,000 blogs labeled with gender. All the feature groups help in predicting gender. Lexical-style is more useful in gender than any other author trait.

7.4.3 Politics

Prior work in detecting politics on tweets tends to combine Republican and conservative to "right-wing" and Democrat and liberal to "left-wing" and uses Twitter-specific features such as political orientation of friends to achieve high accuracy making it difficult to compare against them. Our system using all features achieves an accuracy of 75.2% in detecting political party (Table 7.12)

which is comparable or better than the results in prior work where Twitter-specific features are excluded. Collocations are the most useful feature in political party detection. The LIWC categories have the least impact.

7.4.4 Religion

We achieve an accuracy of 78.3% in 4-way classification of religion (Table 7.12) using all the features. Using just n-grams actually performs better with 79.3% accuracy. LIWC categories, POS, and collocations all have a positive impact on religion detection. The only prior work we have found in religion detection is two-way classification of Muslim vs Christian, making it difficult to compare against their results.

7.5 Conclusion

We present an author trait detection system which predicts four different author traits: age, gender, religion, and political party. We show that style and content are useful in author trait detection, but that n-grams is the most powerful feature in differentiating author traits. One key advantage of our system in comparison to prior work is that we present a single classifier using the same group of features that is used to train models on *several* author traits. In addition, we purposely designed our system to use features that should be available in all online genres, making our system is robust. Our prediction of all author traits is binary (e.g. male vs female for gender) and achieve competitive or better results to similar prior work (i.e excluding genre specific features) in all author traits. In addition, we also use linear regression to predict the exact age of the author.

In the future we would like to use the different author traits to help improve predicting each of the individual author trait results. For example, we could use the predicted age and gender to improve the model for predicting political party. We would also like to explore additional author traits such as race, geography, and education. We also think it could be useful to enter our system in the PAN evaluation task.

The social proof, liking, and scarcity weapons of influence motivate the inclusion of author trait detection in detecting influencers. **Social proof** entails that a person will be influenced by others in their surroundings, particularly those who are similar to them. In addition, association to a person

through author traits such as age and gender will cause **liking**. Finally, **scarcity**, in the form of restrictions can affect demographic groups (e.g. assigning teenagers a curfew causes them to want to stay out late). In Part III of the thesis we use the predicted author traits of each author to generate features for influence detection. We include two main group of features. The first group of features is simply based on the value of each author trait (e.g. male/female). These can be a useful features as they can indicate homophily to other authors in the corpus. The second group of features is related to social proof by determining if the author is similar to the other authors in the particular thread. We determine this by examining if they follow the majority author trait in the thread (e.g. The author is Democrat and most participants are Democrat). Additional features and further analysis of the features in detecting influencers along with the impact of author traits on influence detection can be found in Part III. Finally, in Chapter 12 of this thesis, we will take a closer look at social proof and its impact on influence detection. We will show that influencers tend to be aligned with the majority author traits of the other participants in the conversation and that homophily is important. This indicates that author traits, and thus social proof, is indeed a useful measure for detecting influence.

Chapter 8

Direct Features

Our last two components are features that are used directly within the influencer system. They are credibility features and dialog pattern features. The credibility features aim to capture how credible the author is using various writing styles within each post of the author. The dialog features aim to capture the impact of the thread structure on the authors. Since these features are not stand alone systems, but only used within in influence detection, we describe all experiments with them in the experiment section on influence detection in Chapter 11.

8.1 Credibility

> *Never trust a computer you can't throw out a window*
>
> Steve Wosniak, *Reality Check by Guy Kawaski*

People are more likely to be influenced by those who have established credibility. This is evident mainly in the **Authority** weapon of influence as those that are in authority tend to be considered more credible. In addition, being credible may cause a person to be **liked**. We provide several features directly related to determining the credibility or lack thereof of each person in a conversation:

Grounding providing sources to back up claims.

Name Mentions How often a participant refers to other participants as well as indicating honorifics (e.g. President)

Type	Example
Link	Supporting Evidence: Getting Away With Torture (**LINK**)
Quotes	**"**he should investigate the possible tens of thousands of innocent Iraqi civilians that were murder during the second Iraq war, all on Bush's hands.**"** I must disagree.
Stastistics	**35,000** attendees does not a "fringe" constitute or represent.
Author Name	Hence they say nothing .RT **@professorkck** Media fear claims of liberal bias.
Honorific	**President** Barack Obama is obliged to order a criminal investigation under the Convention against Torture to which the US is a party.
Out of Vocabulary	And we all know the **governent** doesn't rat itself out.
Inquisitiveness	I'd like to replace that section in this article. Any objections**?**

Table 8.1: Examples of each type of credibility.

Out of Vocabulary The amount of informal and mispelled speech by the participant

Inquisitiveness The percentage of the participant's sentences that are questions.

Each credibility indicator is motivated by at least one weapon of influence [Cialdini, 2007] as described in the following sections.

8.1.1 Grounding

A person can make themselves appear more credible by providing sources to back up the their claims. We compute this by keeping track of the number of links, quotes and numbers of each author. A person can provide a url or direct quote as a source. Numbers can indicate useful statistical information. An example of each type of grounding is shown in Table 8.1.

8.1.2 Name Mentions

There are several weapons of Influence related to name detection: **Social Proof, Liking, and Authority**. First, people are often disregarded when they are in need of aid because of social norms.

Ms	Honorable	Master	Rev	Reverend	Ambassador	Doctor
Mrs	President	Pres	Gov	Governer	Coach	Officer
Mr	Representative	Dr	Treasurer	Secretary	Corporal	Sargent
Commander	Lieutenant	Colonel	General	Sir	Madam	PhD
Prof	Attorney	Professor	Superintendent	Senator	Administrative	Captain
Queen	King	Prince	Princess	Atty	Monsignor	

Table 8.2: A list of the honorifics used in the credibility component.

People tend to look at what others are doing and follow the crowd. It has also been found that an item becomes more desirable if it is associated with famous people. This is called the art of *name-dropping*, which is a common tactic applied by sales people (e.g. using a famous actor in a commercial). Lastly, people tend to be influenced by a person of higher authority. A prime example, would be a moderator in a discussion forum. In addition, this authority does not necessarily have to be true, but if someone is thought to have power, such as being a doctor, they are listened to more blindly. All these examples are motivation for name detection.

We add name detection through two different methods. In the first method we look for the mention of author names within the discussion using an exact matching. In the future we would like to explore using a named entity tagger and coreference resolution to increase coverage. In the second method we look for honorifics such as President and Doctor. A comprehensive list of honorifics is shown in Table 8.2. An example of each type of name mention is shown in Table 8.1.

8.1.3 Out of Vocabulary

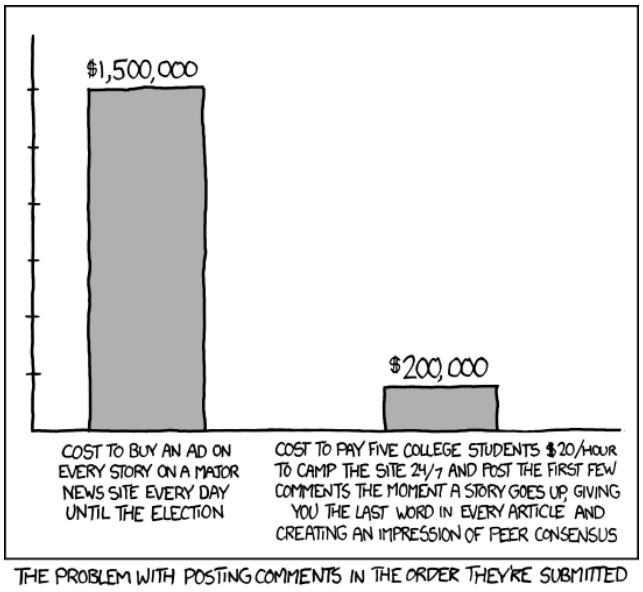
We refer to misspellings as out of vocabulary by computing the occurrence of non-dictionary terms of each author. Misspellings can indicate a lack of credibility.

8.1.4 Inquisitiveness

Inquisitiveness can indicate a lack of influence as a person who asks questions will tend to lack **credibility**. We compute three forms of inquisitiveness [Swayamdipta and Rambow, 2012]; does the participant ask the most questions, does the participant ask any questions, and the percentage of posts with questions. An example of inquisitiveness is shown in Table 8.1.

8.2 Dialog Patterns

“



”

Randall Munroe, *First Post*[1]*, XKCD*

Thus far we have discussed several components that are contextual in nature. In addition, the structural style of the conversation can also be very indicative of influence. For example, as the XKCD comic quoted at the beginning of this section shows, politicians can cheaply pay students to write the first comments under an article to achieve the appearance of peer consensus, a form of social proof.

We have explored several dialog patterns that can be indicative of influence [Biran *et al.*, 2012][2] The tree structure found in threaded documents is exploited to find these patterns. The Dialog Patterns component contains the following patterns:

[1] http://xkcd.com/1019/

[2] Excluding interval, these features were initially defined by my colleagues, Or Biran and Swabha Swayamdipta under the guidance of Owen Rambow

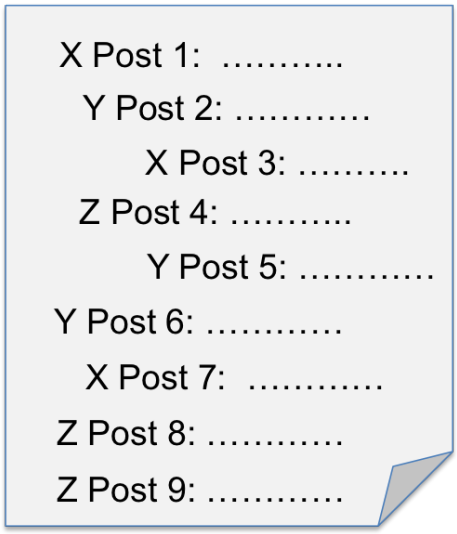


Figure 8.1: Example of a hypothetical thread structure in a discussion.

Initiative The participant is or is not the first poster of the thread.

Irrelevance The percentage of the participant's posts that are not replied to by anyone.

Incitation The length of the longest branch of posts which follows one of the participant's posts.

Investment The participant's percentage of all posts in the thread.

Interjection The point in the thread at which the participant enters the discussion.

Interval The period of time it takes to receive a response.

Each of these dialog patterns is motivated by at least one weapon of influence [Cialdini, 2007] as described in the following sections. An example of a hypothetical thread structure is shown in Figure 8.1.

8.2.1 Initiative

Social proof dictates people follow what others do. This would imply that the first person to do the action has influenced the others. Furthermore, people tend to follow those with **authority**; in certain online corpora, such as weblogs, there is an owner of the particular document. That owner tends to have the authority on the page. In contrast to initiative we also include whether the person is the last person to post in the conversation as this too can be indicative of influence. Author *X* has the initiative in Figure 8.1 and author *Z* is the last person to post in the discussion.

8.2.2 Irrelevance

It is human nature to automatically respond, or **reciprocate** [Cialdini, 2007], to what others have provided. This principal is so strong that it often becomes an *obligation*; this is particularly evident in returning a favor or gift. If a person is not reciprocated, but made to feel irrelevant or ignored, would imply a strong lack of influence. Irrelevance is also indicative of a lack of **social proof**; a person that is ignored is not causing people to follow them. We compute irrelevance as the percentage of a persons posts that are left without a response. For example, author *X* is ignored 2/3 times in Figure 8.1. We also compute whether the person has consecutive posts (meaning they respond to themselves) which may indicate uncertainty and a lack of influence. Similarly, we also compute whether the participant has alternating posts indicating they are active in the dialogue.

8.2.3 Incitation

As mentioned in irrelevance, it is human nature to **reciprocate**. Perhaps the opposite of being irrelevant, is to have people respond to you often. This is in part captured by incitation. In addition, incitation also follows from **social proof**; once one person responds to a post others will respond too causing a long chain. For example, the longest thread stemming from author *X* consists of three posts (posts 1-3) in Figure 8.1. We also compute whether an author has alternating posts which can indicate further responsiveness.

8.2.4 Investment

Investment too, is a byproduct of **social proof**. A person is more likely to participate in a conversation if other people are doing so. Investment also implies that a person is **committed** to the conversation, and can also indicate **consistency** to their beliefs and views. We compute investment as the percentage of the conversation a person has participated in and whether the participant has the most, and longest posts. For example, author *X* wrote 3/9 posts in Figure 8.1. We also indicate if they only participated in the conversation one time or more than once as the former tends to indicate a lack of influence and the latter is more indicative of influence.

8.2.5 Interjection

Another dialog pattern that captures **social proof** is interjection; the point at which a person joins the conversation indicates how quick they are to follow others and how quickly others follow them. For example, author *Z* joins the conversation at 4/9 in Figure 8.1.

8.2.6 Interval

We include a dialog pattern that looks at the average response time and active time, or **Interval**, between posts for each participant. The intuition behind this dialog pattern can be thought of as the opposite of irrelevance; a person who is influential is likely to be responded to faster than someone who is not influential. Robert B. Cialdini [2007] described this phenomenon under **Scarcity**. A person will be afraid that they will be ignored or are missing out if they don't respond immediately (similarly to a person being afraid of missing out on a "good" sale).

8.3 Conclusion

In this chapter we described two system components: credibility and dialog patterns. Although considered components, each of the described features are computed per author directly within the discussion and are *not* stand-alone systems. Credibility refers to features that can indicate whether the author has or lacks authority within the conversation and is extracted per participant from the context in their posts. Dialog patterns refers to features based on the thread structure of the discussion.

In this chapter we described and motivated each of the features in the dialog and credibility components used for influence detection. Including credibility in detecting influencers is mainly motivated by **authority** as those who are in authority tend to be more credible. Including dialog patterns in influence detection is motivated by several weapons of influence. The main weapon is **social proof**. This is because the structure of thread indicates many ways that people can follow what others do. For example, the actions of the first person in the thread, when a person joins a conversation, and how much they speak can all be indicative of how much they will be followed and how influential they will be.

The impact of these components in influence detection will be described in further detail in Part III of this thesis. In particular, we will show that dialog patterns are a very successful component in

detecting influence. Most notably, we find that initiative tends to be very indicative of influence. This is intuitive as the person who initiates the discussion is either the owner of the blog (e.g. LiveJournal) or initiating the debate or edit in a discussion forum (e.g. Wikipedia and Create Debate).

Part III

Influence across Genres

Chapter 9

Introduction

Thus far, in Part II of this thesis, we have discussed the system components that can be useful for detecting the influencers in a discussion. Our components are: Claims, Argumentation, Persuasion, Agreement, Author Traits, and Dialog Structure and Credibility. Each component has been motivated using social science through the weapons of influence [Cialdini, 2007].

In this portion of the thesis we describe our supervised method for using the system components to detect situational influence across online genres. We detect influencers across five online data sources that have been annotated for influencers: Wikipedia Talk Pages, LiveJournal weblogs, Political Forum discussions, Create Debate debate discussions, and Twitter microblog conversations. First, we will describe a rich suite of features that were generated using each of the system components. Then, we describe our experiments and results including using domain adaptation to exploit the data from multiple online genres.

We conclude this portion of the thesis with a detailed analysis of a single weapon of influence, social proof, and its impact in detecting influencers in Wikipedia Talk Pages. This provides a single example of the usefulness of providing comprehensive components in the detection of influence. In the future, we would like to explore in further detail the impact of other weapons of influence within multiple online genres.

P1 by Richard001 Two recently added images include a dead dog, which I find random and in somewhat bad taste, and an image of the pope, along with caption making POV assertions about his entrance to the 'afterlife'. Can we please have discussion before changing the image.
P2 by Bleh999 The Dog is clearly the best illustration of death put forth so far imho, the pope one is indeed POV. Why do you think the dog is bad taste?
P3 by Richard001 I've added small thumbs of each of the images we have had recently. As for the dead dog, it's not particularly appealing to me. There are plenty of other images we could use as well, the important thing is to discuss the changes.
P4 by Ksyrie I am for your remarks–
P5 by Bleh999 Note the image was only used thumbnail size in the article so if you were so offended why did you click so to see the larger version?
P6 by Richard001 I don't think it's a valid point that it's small - would it be okay to have a thumbnail of a detailed photograph of two people enthusiastically copulating on the sexual intercourse article if you couldn't see their genitalia in the thumbnail? I strongly suggest using an image from long enough ago that it won't cause any edit disputes - the Mongol invasions for example.
P8 by Ksyrie I dount if any available pic far away too 1000 years ago,a slaughter scene from Sparta War?–
P9 by Bleh999 Once again you are wrong, take a look at wikipedia commons there are graphic images of sexual acts available
P10 by Fusion7 I am rather upset with the current image Image:Civil War graves.JPG This article should be mainly about the science of death, and its ties with human culture secondary. However, this image does the opposite. It focuses not on death itself, but our traditions that revolve around its occurrence.
P11 by Fusion7 I agree with your comments

Table 9.1: Influence Example: A portion of a Wikipedia Talk Page discussion regarding an Image for Death displaying Richard001 as the influencer. Replies are indicated by indentation (for example, P2 is a response to P1).

Chapter 10

Method

> *(Computer Science) developed the machine that assisted the power of the brain rather than muscle*
>
> Grace Hopper

Our approach to influence detection uses each of the system components described in Part II of this thesis. Our components are: Claims, Argumentation, Persuasion, Agreement, Author Traits, and Dialog Structure and Credibility. Each component has been motivated using social science through the weapons of influence [Cialdini, 2007]. We will describe them briefly, and in more extensive detail in the following sections. We first describe the influence definition used during annotation and then describe the datasets. They are: LiveJournal, Wikipedia Talk Pages, Create Debate, Political Forum, and Twitter. Then, we describe the features generated from the output of each system component. Finally, we explain our use of domain adaptation to improve results across multiple online genres. This method is used to predict the influencers in each discussion. Our experiments and results follow in Chapter 11.

10.1 Introduction to Features

We briefly describe the features generated from the system components here. Further details regarding motivation of the features for each component, including examples, can be found later in this chapter.

10.1.1 Persuasion, Claim, Argumentation

The first set of features relate to persuasion. In this work we focus on attempts to persuade someone to change their opinion. Causing someone to change their mind can be a strong indication of being influential. An attempt to persuade consists of two parts: making a claim and closely following it within an argumentation. A sentence is considered to be a claim if it is opinionated and the author expresses a belief. An argumentation is a justification to a claim. We use a blackbox component to automatically label the sentences that are claims argumentations and persuasions (further detail regarding these blackbox components can be found in Chapter 6) in a discussion.

This output is used to generate the following features for claim, argumentation, and attempts to persuade. Henceforth we refer to persuasion as being applicable for features generated for attempt to persuade and its subparts, claim and argumentation.

We include features to indicate persuasiveness by examining whether the participant **is persuasive**, **total occurrence** of persuasion, **all/number of persuasive posts**, and the **post with the most** attempts to persuade. These features are normalized by post size and/or sentence size in the thread where applicable.

In addition, we also include a feature to indicate the number of times the first sentence in the post was an argumentation or a claim normalized by the number of posts by the participant. Starting a post with an argument or claim can indicate that it is stronger. Finally, we also include Bag-of-Words from the text of the argumentative and claim sentences. This represents meaningful content that the participant has written.

10.1.2 Agreement

The agreement system (Chapter 5) labels each post in the discussion as to whether it agrees or disagrees with the post it is responding to. It is also possible for a person to neither agree nor disagree with those posts. Henceforth we refer to these three possibilities as (dis)agreement. (Dis)agreement is an important indicator of influencers for several reasons. A person who agrees with others is indicating they are following someone else's opinions. This can imply that they are less likely to be influential. On the other hand, disagreeing can indicate influence because the person has their own conflicting opinions they are trying to push. A lack of (dis)agreement can indicate that the post is not relevant and not influential.

Each of the features is computed per discussion by examining the agreement and disagreement labels of all posts per participant in the discussion. We have features to indicate how often the person is (dis)agreed with (**from**) and how often they (dis)agree with others (**to**). We also have features to indicate what the first to and from post are as a **person's initial position** (**from**) and **first impression** (**to**) in the discussion can have a greater impact than later posts.

10.1.3 Author Traits

The author trait system (Chapter 7) classifies each author according to their gender, age, political party, and religion. These author traits are used to generate three groups of features: **single** binary features, **majority** features based on the thread, and **combination** author trait features (e.g. majority gender *and* age).

10.1.4 Dialog Patterns, Credibility

The dialog patterns (Chapter 8.2) and credibility (Chapter 8.1) components are a group of features used directly within influence detection. Briefly, the dialog features are: Initiative, Irrelevance, Incitation, Investment Interjection, Repetition, and Time. The credibility features are: Inquisitiveness, Honorifics, Name Mentions, Links and Quotes, and Statistics.

10.2 Definition

In this thesis we focus on detecting **situational influence**. Situational influence refers to finding the person who is most influential within a single discussion. In contrast, global influence refers to how influence spreads throughout a community (commonly found in social networks). Situational influence is important because the person who is influential can change depending on the topic at hand. In the introduction to this thesis we motivated influence detection and our components through social science using Robert B. Cialdini's "Weapons of Influence" [Cialdini, 2007]. The annotators were given the following general definition: An influencer is someone who has credibility in the group, persists in attempting to convince others, and introduces topics/ideas that others pick up on or

support [1]. In addition, the annotators were also provided with a detailed explanation of who is *not* an influencer as well as sample threads. Further details regarding annotation and the full annotation manual are described in Appendix A.

10.3 Data

We have conversations from five different sources that have been annotated for Influence: LiveJournal, Wikipedia Talk Pages, Political Forum, Create Debate, and Twitter.

LiveJournal is a virtual community in which people write about their personal experiences in a weblog and can typically be considered as a public diary.

Wikipedia Talk Pages are the talk pages associated with each Wikipedia Page where proposed edits for the page are discussed.

Create Debate is a discussion forum whose format is a debate style where people can post on either a *for* or *against* side of an argument.

Political Forum is a discussion forum website used to discuss political topics.

Twitter is the most popular micro-blogging site on the web where people write short blurbs between 0-140 characters. A threading structure can be generated using tweets via retweets (tweeting something written by someone else) and mentions (mentioning another user).

A more detailed explanation of each of the sources is described in Chapter 1, Section 2. The LiveJournal dataset is a subset of the annotations gathered for age detection (Chapter 7, Section 7.2) in 2010. We obtained the Wikipedia talk pages by examining Wikipedia data statistic pages such as the most edited talk pages[2]. These pages are more likely to have considerable content and controversy making them more likely to obtain influencers. The Create Debate, Political Forum, and Twitter discussions were gathered by crawling the website for discussions related to general politics as well as politics related to the republican primary using a list of keywords (see Table 10.1) during the time

[1] This definition was compiled as a group effort of multiple universities and government under the IARPA Socio-cultural Content in Language (SCIL) program `http://www.iarpa.gov/index.php/research-programs/scil`

[2] `http://en.wikipedia.org/wiki/Wikipedia:Most-edited_talk_pages`

General Politics	republican, democrat, "us senate" OR u.s. senate OR "united states" senate, house or representatives, cabinet member, speaker of the house, obama, biden, michelle obama, democratic national convention or DNC, u.s. president OR "us president" OR "united states" president
Republican Primary	romney, santorum, gingrich, republican primary, gop, paul ryan, jon huntsman, michelle bachmann, herman cain, chris christie, tim pawlenty, republican national convention OR rnc, reince priebus, ann romney

Table 10.1: List of keywords (separated by commas) used to search for threads and tweets that are relevant to politics

	General Politics			Republican Primary		
Corpus	# Threads	# Users	# Posts	# Threads	# Users	# Posts
Political Forum	5078	2080	413607	5094	2039	410671
Create Debate	2077	3635	51754	940	2727	28540
Twitter	N/A	3154069	10071273	N/A	1716552	10,881,171

Table 10.2: List of statistics for each of the unannotated political datasets

period of November 2011-March 2012. A summary of the statistics of all the downloaded political data is shown in Table 10.2.

A portion of documents across all online genres were annotated for situational influence by several annotators using the definition described in Section 10.2. Seven annotators annotated the Wikipedia and LiveJournal threads with an average IAA using Cohen's κ of .53 among all annotators on a portion of 12 Wikipedia Talk Pages. The political threads were annotated by four annotators with an average IAA using Cohen's κ of .57 among all annotators computed on six threads, two from each genre. This indicates that although it is a doable task, even among annotators it was hard to achieve a common consensus regarding the influencer in the discussion. In the future we would like to mitigate the differences by adopting a ranking scheme instead of asking the annotators to provide who they think are the influencers. In this scheme we would ask the annotators to list the participants in the discussion from most influential to least influential. Our hope is that this would result in less variation because although the annotators may not agree on the top influencer, they would agree out of order on the top n influencers.

Corpus	Number of				Average per Thread		Posts /	Threads w/o
	Threads	Authors	Posts	Influencers	Posts	Authors	Author	Influence
Wikipedia	509	3130	8749	456	17.2	6.2	2.8	76
LiveJournal	219	1157	2026	223	9.3	5.3	1.8	0
Political Forum	101	1147	3202	91	31.7	11.4	2.8	17
Create Debate	106	1737	6494	94	61.3	16.4	3.7	18
Twitter	99	1602	1919	101	19.4	16.2	1.2	4

Table 10.3: List of statistics for each of the annotated influence datasets

Our Wikipedia and LiveJournal datasets are the largest with 509 and 219 annotated discussion threads respectively. The Political Forum, Create Debate, and Twitter datasets are smaller in size, with around 100 discussions threads each. Further statistics for each of the annotated datasets is shown in Table 10.3. The Create Debate discussions are the most lengthy, have the most authors, and most posts per thread. In contrast, LiveJournal has the least content, and number of authors per thread. Although the Twitter discussions are long, the average number of posts per author is only one. Table 10.4 displays the occurrence of the contextual components in the training data across the multiple online genres. It is clear that Twitter has the least contextual information. This is expected due its size limit of 144 characters per post. On the other hand, Create Debate is the richest in contextual information. This is mainly due to the large size of each discussion. Wikipedia, Create Debate, and Political Forum are all more argumentative in nature. This is evident by the high occurrence of arguments, claims, persuasion, and disagreement. In contrast, Twitter and LiveJournal are less argumentative and have more agreement than disagreement.

Typically there is one influencer in each discussion, and on rare occasion two (e.g. 14/509 or 2.8% in Wikipedia). In addition there were some threads where no influencer was present. Since our goal is detecting influence, we excluded those discussions from the experiments.

An example of an influencer in a Wikipedia Talk Page discussion regarding an Image for Death is shown in Table 9.1 where Richard001 is considered to be the influencer. Richard001initiates the thread, and maintains participation throughout. He gets support for his ideas about changing pictures, and debates with dissenters. Opposition to his opinions are mild as well. An additional example of influence in a Wikipedia Talk Page discussion can be found at the beginning of this thesis (in

Corpus	Claim	Argument	Persuade	Agree	Disagree	Influencers
Wikipedia	17.1	17.9	11.4	7.0	4.5	0.9
LiveJournal	5.2	5.4	2.5	2.7	0.4	1.0
Political Forum	25.6	30.7	18.7	0.2	24.4	1.0
Create Debate	110.3	100.3	84.7	19.4	29.0	0.9
Twitter	1.2	0.9	0.2	3.3	2.4	1.0

Table 10.4: Occurrence of contextual components across the training datasets of the online genres normalized by the number of discussions

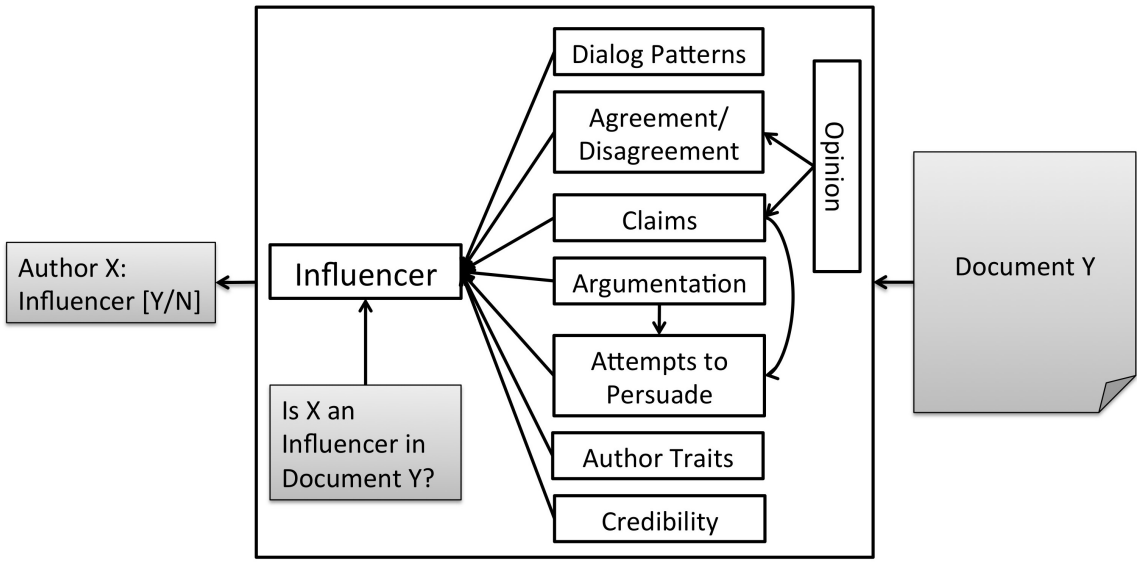


Figure 10.1: The complete influencer pipeline.

Chapter 1,Table 2.2). Influencers in the other online sources can be found in the data chapter of this thesis 1.5.

10.4 Features

In this section we describe the features generated from the system components described in Part II. They are, Claim, Argumentation, Persuasion, Agreement, Author Traits, Credibility, and Dialog. Figure 10.1 shows the pipeline of the Influencer system. The pipeline is given a document as input. Each document Y has X_n authors. For each author X_i in document Y the system outputs *Yes*

indicating the author *is* an influencer or *No* indicating the author *is not* an influencer. The system consists of 1 subcomponent, 7 components, and the influence component. Opinion is a subcomponent used in agreement and claim detection. A claim and argumentation make up an attempt to persuade. All components feed in to the influence component. In the following sections we describe the features generated per system component that are used within the influence component along with examples pertaining to the Wikipedia Talk Page discussion regarding an Image for Death as shown in Table 9.1.

10.4.1 Claim

Our first set of features are produced from the claim component (Chapter 6, Section 6.3). Each of the features is computed per discussion by examining the sentences tagged as claims by the claim component per participant:

Has Claim: True if the participant made at least one claim. If a person does not make any claims it is unlikely that they will be influential.

Claim Count: The number of claims made by the participant. We have two features: normalized by the total number of claims in the thread and normalized by the number of sentences in the thread. The former is motivated by examining how successful the participant is in making claims in this thread in comparison to the other authors. The latter examines how successful the participant is in making claims in comparison to participants in other discussions by taking the length of the discussion in to account.

All of Posts have Claims: All the posts of the participant have at least one claim. The motivation of this feature is to determine how committed the author is in relation to all his posts.

Post with most Claims: The number of claims in the post with the most claims normalized by the number of sentences in the post. This feature explores how strong the strongest post (in terms of making claims) of the author is. A single strong post can be very important in terms of being influential.

Claim First in Post: The number of times the first sentence in the post was a claim normalized by the number of posts by the author. Starting a post with a claim can indicate that it is a stronger claim.

Claim Behavior	Participants			
	Bleh999	Fusion7	Ksyrie	Richard001
Has Claim	T	T	F	T
Claim Count normalized by # of claims	2/16	2/16	0/16	12/16
Claim Count normalized by # of sentences	2/56	2/56	0/56	12/56
All Posts have a Claim	F	T	F	F
Claims in Post with most Claims	1.0	1.0	0.0	0.5
Claim First in Post	2/3	2/2	0/2	4/8

Table 10.5: The claim feature values for each of the participants in the Wikipedia Talk Page discussion thread regarding an image for death shown in Table 9.1.

Claim Content: Bag-of-Words from the text of the sentences that are claims. This can be considered meaningful content that the participant has written.

Table 10.4 shows the difference in claims across the multiple online genres. Claims are common in most genres, most particularly in Create Debate. Table 10.5 lists the values of the claim features for each of the participants in the Wikipedia Talk Page discussion about an image for death shown in Table 9.1. Figure 10.2 shows how useful each feature is across each genre in the training datasets. Making claims is a positive indication of influence across all genres, and having many claims is an even stronger indication of influence. The least useful features are having claims in the first sentence of the post and having all posts have a claim, probably because both of us these features are not very common in general. Claims are the least useful in Create Debate perhaps due to the very argumentative nature of the dataset. As Table 10.4 showed we find that claims are very common in Create Debate indicating they may be less useful for discriminating between influencers and non-influencers.

10.4.2 Argumentation

Our next set of features are produced from the argumentation component (Chapter 6, Section 6.4 and are identical to those in the claim component, but for argumentation. Each of the features is computed per discussion by examining the content labeled as argumentative by the argumentation component per participant:

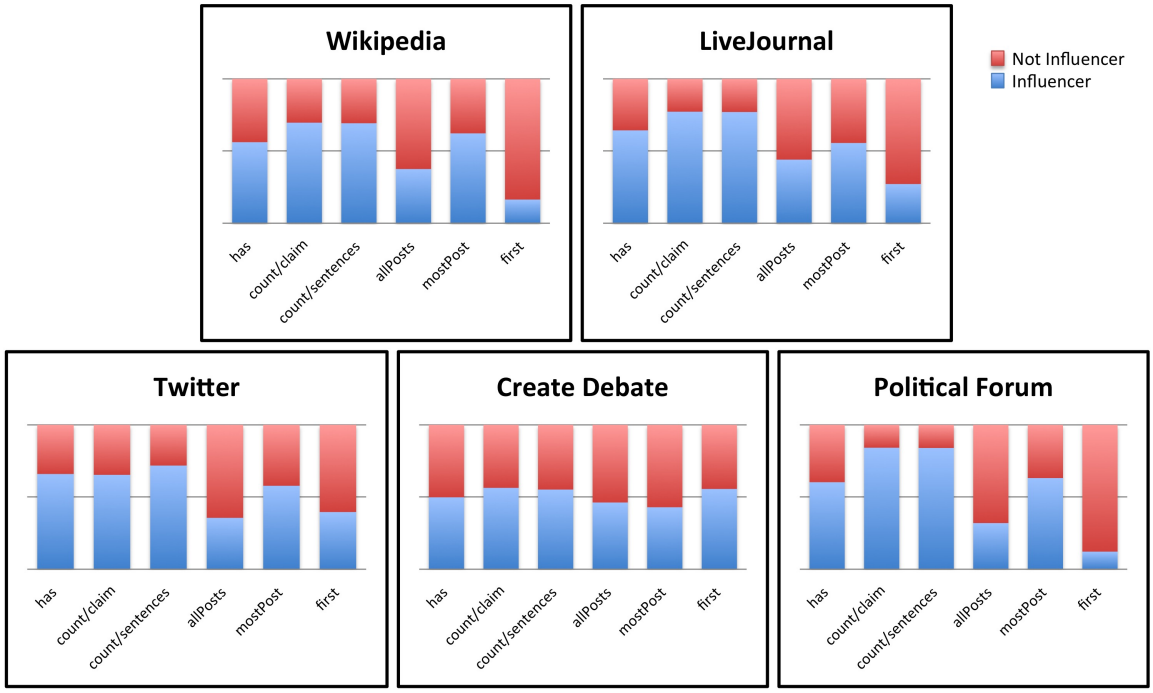


Figure 10.2: The ratio of claim features towards influencers in each of the training datasets.

Has Argumentation: True if the participant was argumentative at least once. If a person does not make any arguments it is less likely that he will be influential.

Argumentation Count: The number of arguments made by the participant. As in claims, we have two features: normalized by the total number of arguments in the thread and normalized by the number of sentences in the thread. The former is motivated by examining how successful the participant is in making arguments in this thread in comparison to the other authors. The latter examines how successful the participant is in making arguments in comparison to participants in other discussions by taking the length of the discussion in to account.

Number of Argumentative Posts: The number of posts where the participant made an argument normalized by the number of posts by the participant. The motivation of this feature is to determine how convincing the author is in relation to all his posts.

Most Argumentative Post: The number of arguments in the most argumentative post normalized by the number of sentences in the post. This feature explores how strong the strongest post (in

Argumentative Behavior	Participants			
	Bleh999	Fusion7	Kysrie	Richard001
Has Argumentation	T	T	F	T
Argumentation Count normalized by # of argumentations	2/11	1/11	0/11	8/11
Argumentation Count normalized by # of sentences	2/56	1/56	0/56	8/56
Number of Argumentative Posts	2/3	1/2	0/2	7/8
Most Argumentative Post	1.0	.166	0.0	0.5
Argument First in Post	0/3	0/2	0/0	0/8

Table 10.6: The argumentation feature values for each of the participants in the Wikipedia Talk Page discussion thread on Death shown in Table 9.1.

terms of making arguments) of the author is. A single strong post can be very important in terms of being influential.

Argument First in Post: The number of times the first sentence in the post was argumentative normalized by the number of posts by the participant. Starting a post with an argument can indicate that it is stronger.

Argumentative Content: Bag-of-Words from the text of the argumentative sentences. This can be considered meaningful content that the participant has written.

Table 10.4 shows the difference in argumentation across the multiple online genres. Argumentation is common in most genres, most particularly in Create Debate. Table 10.6 lists the values of the argumentative features for each of the participants in the Wikipedia Talk Page discussion regarding an image for death shown in Table 9.1. Figure 10.3 shows how useful each argumentation feature is across each genre in the training datasets. Making arguments is indicative of influence across all genres, with the strongest impact in LiveJournal and Political Forum. Argumentation rarely occurs in Twitter, and in fact never occurs in the first sentence. However, when it does occur it is indicative of influence. Interestingly, starting a post with an argument is indicative of a lack of influence in Create Debate, indicating that starting off with an argumentative tone causes other participants to not want to listen to them.

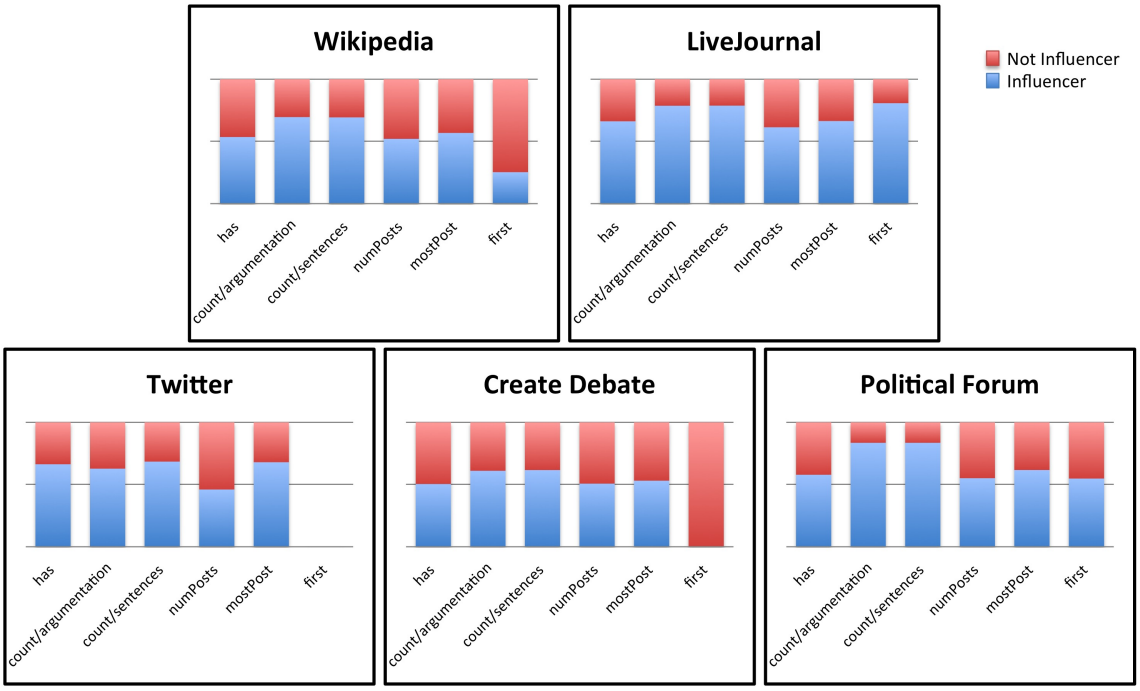


Figure 10.3: The ratio of argumentation features towards influencers in each of the training datasets.

10.4.3 Persuasion

This set of features is produced from the persuasion component. Recall from Chapter 6 that an attempt to persuade is a claim followed by an argumentation. This indicates that persuasion is a subset of the output from the claim and argumentation components. Each of the features is computed per discussion by examining the content labeled as an attempt to persuade by the persuasive component per participant:

Has Persuasion: True if the participant made an attempt to persuade at least once. If a person does not make any persuasive attempts it is less likely that he will be influential.

Persuasion Count: The number of persuasive attempts made by the participant. As in claims and argumentation, we have two features: normalized by the total number of persuasive attempts in the thread and normalized by the number of sentences in the thread. The former is motivated by examining how successful the participant is in making persuasive attempts in this thread in comparison to the other authors. The latter examines how successful the participant is in

Persuasive Behavior	Participants			
	Bleh999	Fusion7	Kysrie	Richard001
Has Persuasion	T	T	F	T
Persuasion Count normalized by # of persuasions	1/9	1/9	0/9	7/9
Persuasion Count normalized by # of sentences	1/56	1/56	0/56	7/56
All Posts have Persuasion	F	F	F	F
Most Persuasive Post	1	.166	0	.5

Table 10.7: The feature values for each of the participants in the Wikipedia Talk Page discussion thread shown in Table 9.1.

making persuasive attempts in comparison to participants in other discussions by taking the length of the discussion in to account.

All Persuasive Posts: All the posts have at least one persuasive attempt. The motivation of this feature is to determine if the author is always persuasive.

Most Persuasive Post: The number of attempts to persuade in the most persuasive post normalized by the number of sentences in the post. This feature explores how strong the strongest post (in terms of making persuasive attempts) of the author is. A single strong post can be very important in terms of being influential.

Table 10.4 shows the difference in persuasion across the multiple online genres. Persuasion is rare in both LiveJournal and Twitter, the least argumentative corpora. Table 10.7 lists the values of the persuasion features for each of the participants in the Wikipedia Talk Page discussion regarding an image for death shown in Table 9.1. Figure 10.4 shows the distribution of persuasion features for influencers across the training datasets of all genres. Being persuasive is a positive indicator of influence as most features show. However, persuasion occurs infrequently in most datasets making it unlikely for all the posts by a person to have persuasion. In other words, the "all persuasive posts" feature is misleading as it indicates that having persuasion in all posts is indicative of not being influential, when in fact this feature occurs infrequently making it less useful.

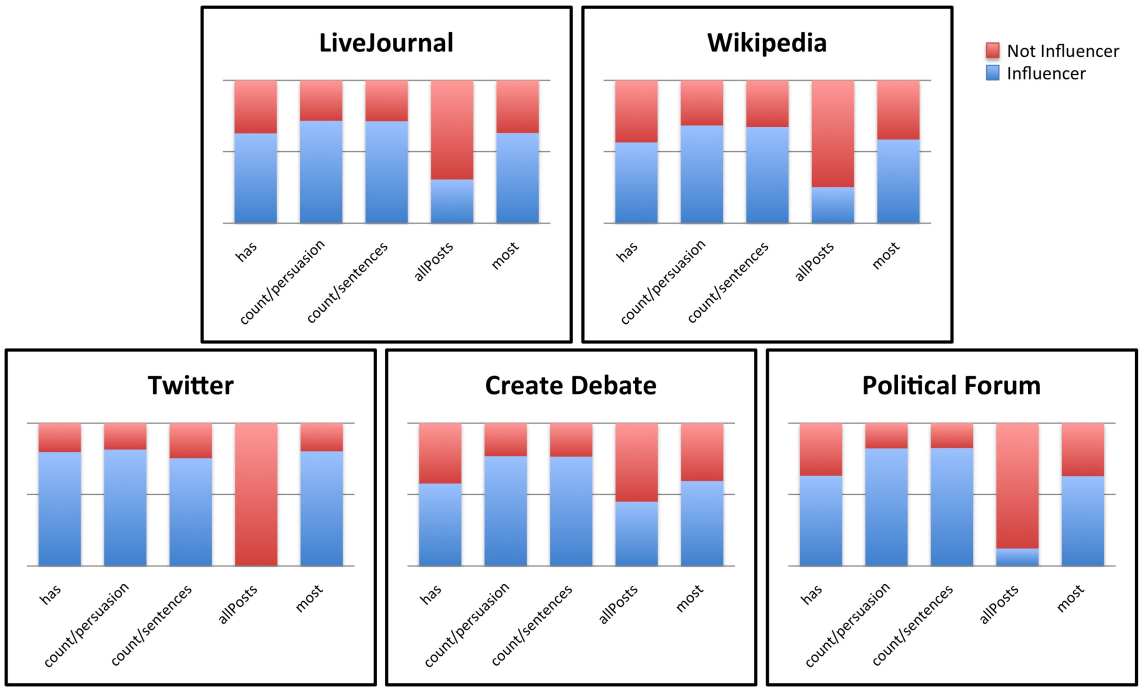


Figure 10.4: The ratio of persuasion features towards influencers in each of the training datasets.

10.4.4 Agreement

This set of features are produced from the agreement component (Chapter 5). Each of the features is computed per discussion by examining the (dis)agreement labels of all posts per participant in the discussion. It is also possible for a person to neither agree or disagree with those posts (the *none* label as described in Chapter 5).

Agreement/Disagreement/None from: Three features indicating the number of times the participant agrees, disagrees, or neither with the participant they are responding to normalized by the total number of posts. A person who agrees with others is indicating they are following someone else's opinions. This can imply that they are less likely to be influential. On the other hand, disagreeing can indicate influence because the person has their own conflicting opinions they are trying to push. A lack of (dis)agreement can indicate that the post is not relevant and not influential.

Agreement/Disagreement/None to: Three features indicating the number of times other partici-

Agreement Behavior	Participants			
	Bleh999	Fusion7	Kysrie	Richard001
Agreement From	2/15	1/15	2/15	4/15
Disagreement From	1/15	1/15	0/15	0/15
None From	0/15	0/15	0/15	4/15
Agreement To	4/15	1/15	0/15	4/15
Disagreement To	1/15	0/15	0/15	1/15
None To	0/15	0/15	0/15	1/15
First Agreement From	T	F	T	F
First Disagreement From	F	T	F	F
First None From	F	F	F	T
First Agreement To	T	T	F	T
First Disagreement To	F	F	F	F
First None To	F	F	F	F

Table 10.8: The agreement feature values for each of the participants in the Wikipedia Talk Page discussion thread shown in Table 9.1.

pants responding to this participant agree, disagree, or neither normalized by the total number of posts. Agreement indicates a person's ideas are being followed. This can imply that the person is influential. Disagreement indicates a person's ideas are compelling and controversial which can have a negative impact on influence. As in the prior feature, a lack of (dis)agreement can indicate that the post is not relevant and the person is not influential.

First Agreement/Disagreement/None from: True if the participants first post agrees, disagrees, or neither with the person he is responding to. A person's initial position in the discussion can have a greater impact than later posts.

First Agreement/Disagreement/None to: True if the first person to respond to the participant agrees, disagrees, or neither with them. A person's first impression in the discussion can have a greater impact than later posts.

Table 10.4 shows the difference in agreement and disagreement across the multiple online genres. Disagreement is more common in Political Forum and Create Debate whereas agreement is more

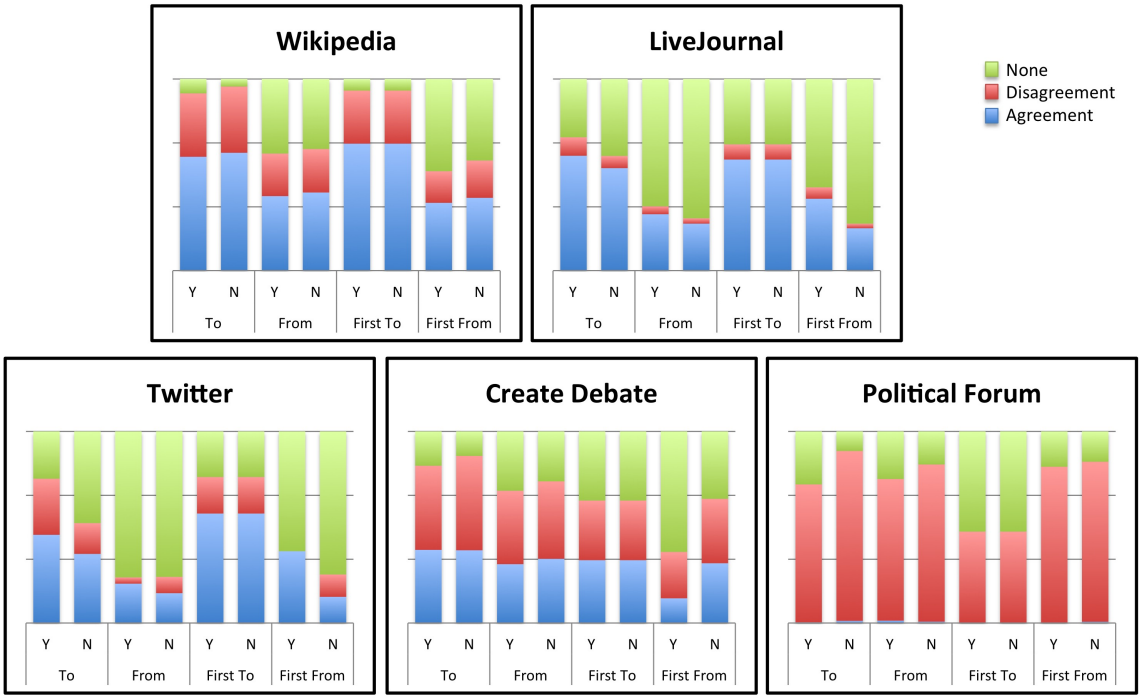


Figure 10.5: The ratio of agreement, disagreement, and none features towards influencers (Y) and non-influencers (N) in each of the training datasets.

common in LiveJournal, Wikipedia, and Twitter. Table 10.8 lists the values of the agreement features for each of the participants in the Wikipedia Talk Page discussion regarding an image for death shown in Table 9.1. Figure 10.5 shows the distribution of agreement features in detecting influence in the training datasets across genres. In Twitter and LiveJournal, influencers have more agreement. In Political Forum, agreement is very rare relative to disagreement, and disagreement is indicative of not being an influencer. In Create Debate, influencers have slightly less disagreement than non-influencers. Finally, in Wikipedia, influencers have slightly more none (the lack of (dis)agreement) than non-influencers.

10.4.5 Author Trait

The output of the author trait component (Chapter 7) is used to generate three groups of features related to each of the five author traits: age, gender, religion, and political party. The groups of features are single features, majority features, and combination features.

Author Trait Behavior	Bleh999	Fusion7	Ksyrie	Richard001
Gender	Male	Male	Male	Male
Majority Gender	Yes	Yes	Yes	Yes
Age	Old	Young	Old	Old
Year-of-Birth	1976	1988	1985	1904
Majority Age	Yes	No	Yes	Yes
Religion	Atheist	Jewish	Atheist	Atheist
Majority Religion	Yes	No	Yes	Yes
Political Party	Democrat	Republican	Democrat	Democrat
Majority Political Party	Yes	No	Yes	Yes
Majority All	Yes	No	Yes	Yes
Majority None	No	No	No	No

Table 10.9: The author traits for each of the participants in the discussion on an image for Death as shown in Table 9.1.

Single: Each author trait is represented as a single feature, marked as true for the predicted trait. Gender is male or female. Age is older or younger than 1982 (See Chapter 7, Section 7.4 for explanation of chosen year). Religion is Atheist, Christian, Jewish, or Muslim. Political Party is Republican or Democrat. Finally, we also have a feature indicating the exact year of birth that was predicted for each author (e.g. 1983).

Majority: Social proof indicates that people will be influenced by those that are like them. We measure this per author trait by determining if a person is predicted to be in the majority within the discussion and have a majority feature corresponding to each author trait. For example, if the majority of the people in a discussion are predicted to be republican, we expect that the influencer is likely to be predicted to be republican as well.

Majority All/None: We also include features to indicate whether the participant is in the majority in *all* author traits or in *no* author traits.

Combination: We also explore whether combining author traits is beneficial. There have been many studies which show that certain tendencies towards an issue are based on many author traits. In particular, this applies to combining demographics and politics. For example, women

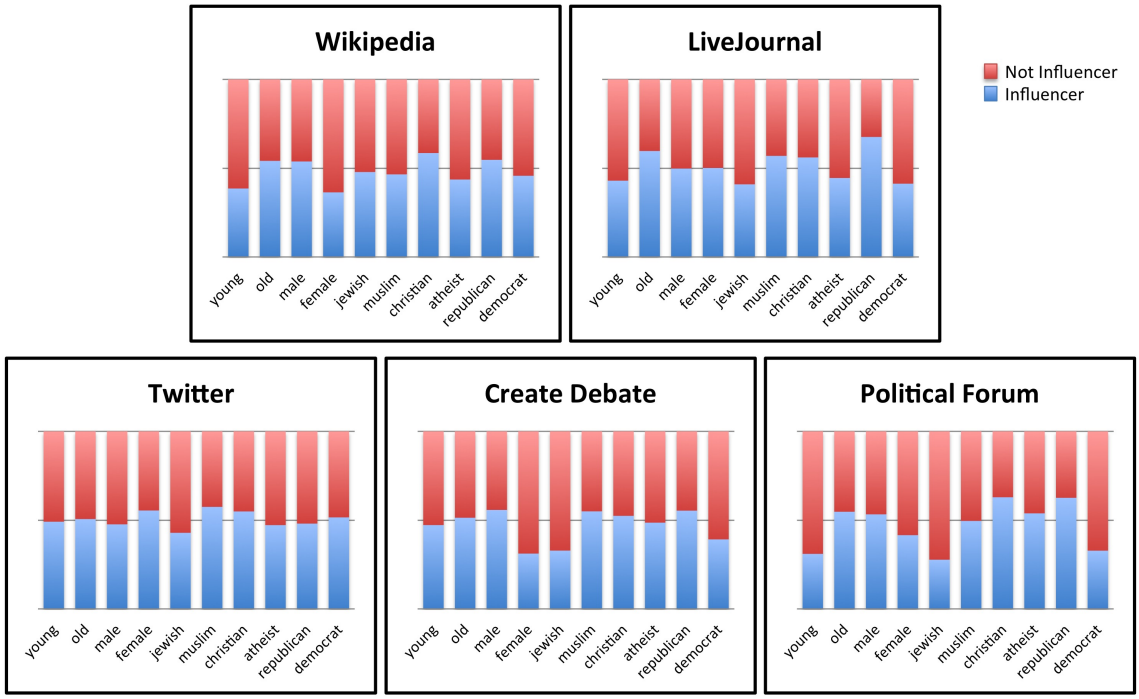


Figure 10.6: The ratio of single author trait features towards influencers in each of the training datasets.

tend to vote for democrats[3] and Christians tend to vote for republicans[4]. Another example is that younger people tend to be less religious than older people[5]. This has motivated us to combine author traits as additional features. We achieve this by combining every two author traits, d_i, d_j, into a binary feature $<d_i, d_j>$ and majority feature $<dm_i, dm_j>$. For example, <gender,political party> and <gender_m,political party_m>.

Table 10.9 lists the values of the single and majority author trait features for each of the participants in the Wikipedia Talk Page discussion regarding an image for death shown in Table 9.1. The combination features can be computed by combining the results for the single features. Figures 10.6 and 10.7 show the occurrence of the majority and binary features within each of the training datasets

[3]http://www.pewresearch.org/fact-tank/2014/11/05/as-gop-celebrates-win-no-sign-of-narrowing-gender-age-gaps/

[4]http://www.pewforum.org/2014/11/05/how-the-faithful-voted-2014-preliminary-analysis/

[5]http://www.pewforum.org/2010/02/17/religion-among-the-millennials/

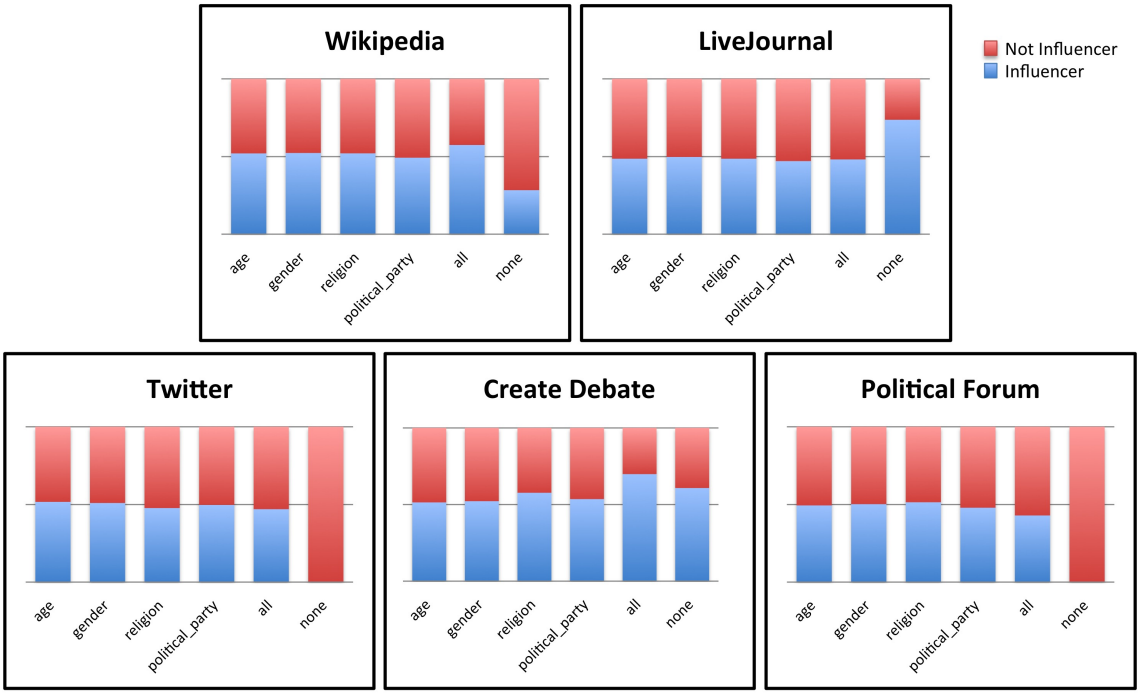


Figure 10.7: The ratio of majority author trait features towards influencers in each of the training datasets.

based on influence. Some features appear to be indicative across several genres. For example, men and older people are almost always (except for Twitter) more influential than women and younger people. In contrast, the effect of being in the majority, i.e. social proof, differs for influence across genres. It clearly is a stronger indication of influence in Wikipedia and Create Debate; the two datasets containing the most text. It unfortunately does not seem to have a big impact in the other datasets. We believe this may be due to the amount of text available per participant, particularly in Twitter. This may indicate that it was harder to correctly predict the author traits in these genres. In Chapter 12 we will provide further discussion regarding author traits and the impact of social proof in Wikipedia Talk Pages.

10.4.6 Credibility

The credibility component is a group of features used directly within influence detection. Briefly, the features are:

Credibility Behavior	Participants			
	Bleh999	Fusion7	Kysrie	Richard001
Quotes	0	0	0	0
Links	0	0	0	0
Numbers	0.33	0	0.5	0.25
Honorifics	0	0.5	0	0
Participant Mentions	0	0	0	0
Participant Mentioned	0	0	0	0
Misspellings	0.66	0.50	1.0	.625
Asked Question	T	F	T	T
Most Questions	F	F	F	T
Percent Questions	.33	0.0	.5	.375

Table 10.10: The credibility feature values for each of the participants in the Wikipedia Talk Page discussion thread shown in Table 9.1.

Grounding: providing sources to back up claims through quotes, links, and numbers.

Name Mentions: How often a participant is referred to by other participants (mentions) and how often others refer to the participant (mentioned)

Honorifics: how often a person uses honorifics in their posts (e.g. President)

Slang: The amount of informal speech by the participant

Inquisitiveness: The percentage of the participant's sentences that are questions, did the participant ask a question, and the person who had the most questions.

Further details and motivation regarding each feature can be found in Chapter 8.1 on the Credibility component. All features are normalized by the number of posts by the author. Table 10.10 lists the values of the credibility features for each of the participants in the Wikipedia Talk Page discussion regarding an image for death shown in Table 9.1. Figure 10.8 shows the distribution of the credibility features towards influencers in the training datasets across genres. These features vary across genre. In some cases no url links were found due to preprocessing in those genres that discarded them from the text. Being mentioned by others is indicative of influence in LiveJournal,

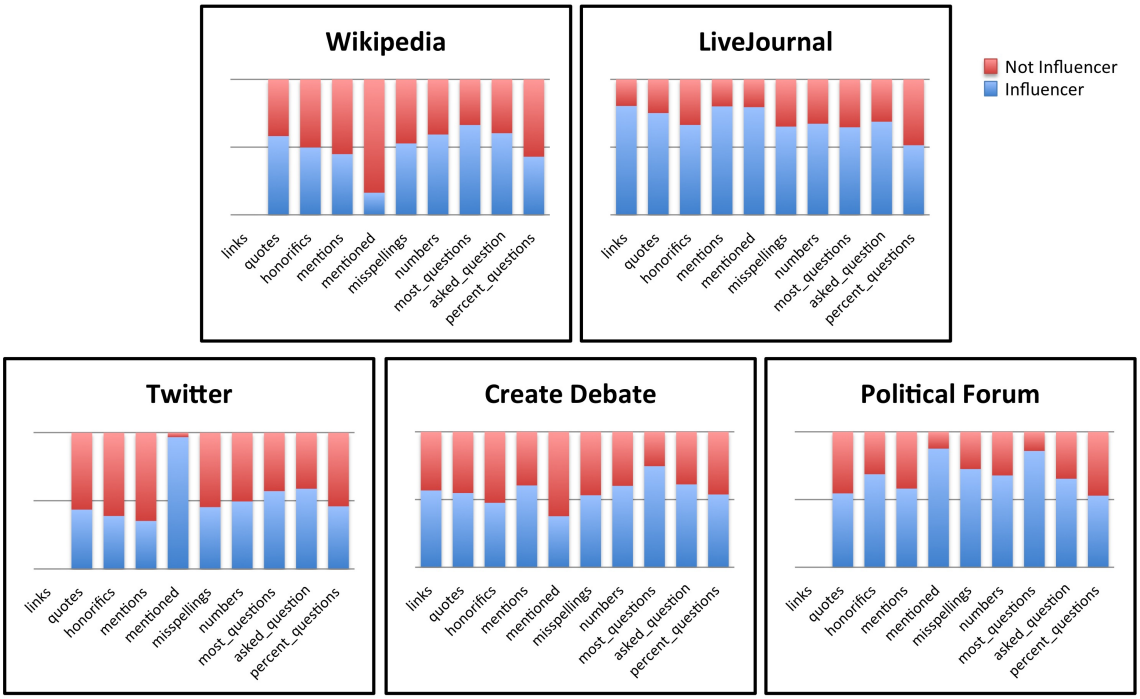


Figure 10.8: The ratio of credibility features towards influencers in each of the training datasets.

Twitter, and Political Forum. It is particularly useful in Twitter. This is intuitive due to the common use of retweets (RT) in Twitter, and the influencer usually being retweeted the most. In LiveJournal, asking questions is correlated with influencers the least. In the other genres asking the most questions is indicative of influence. On the other hand, having a high percentage of questions is indicative of a lack of influence.

10.4.7 Dialog Patterns

The dialog component is a group of features used directly within influence detection. Briefly, the features are:

Initiative: The participant is or is not the first poster of the thread

Last: The participant is the last poster in the thread

Irrelevance: The percentage of the participant's posts that were not replied to by anyone

Dialog Behavior	Participants			
	Bleh999	Fusion7	Kysrie	Richard001
Initiative	F	F	F	T
Last	F	F	F	T
Investment	3/15	2/15	2/15	8/15
Irrelevance	0	.5	1	.5
Incitation	3	1	0	14
Interjection	2/15	11/15	5/15	0/15
Max Posts	F	F	F	T
Longest Post	F	F	F	T
Consecutive Posts	F	F	F	T
Repetition	T	T	T	T
Response Time	391.33	48449.5	178	15813
Active Time	15527	9529	188	77802
Once	F	F	F	F

Table 10.11: The dialog feature values for each of the participants in the Wikipedia Talk Page discussion thread shown in Table 9.1.

Incitation: The length of the longest branch of posts which follows one of the participant's posts

Investment: The participant's percentage of all posts in the thread

Interjection: The point in the thread at which the participant enters the discussion

Max Posts: True if the participant has the most posts

Longest Post: True if the participant has the longest post

Consecutive Post: True if the participant has consecutive posts

Repetition: True if the participant has two or more posts

Response Time: The average period of time, in minutes, it takes to receive a response

Active Time: The amount of time, in minutes, the participant was active in the conversation (last_post_time - first_post_time)

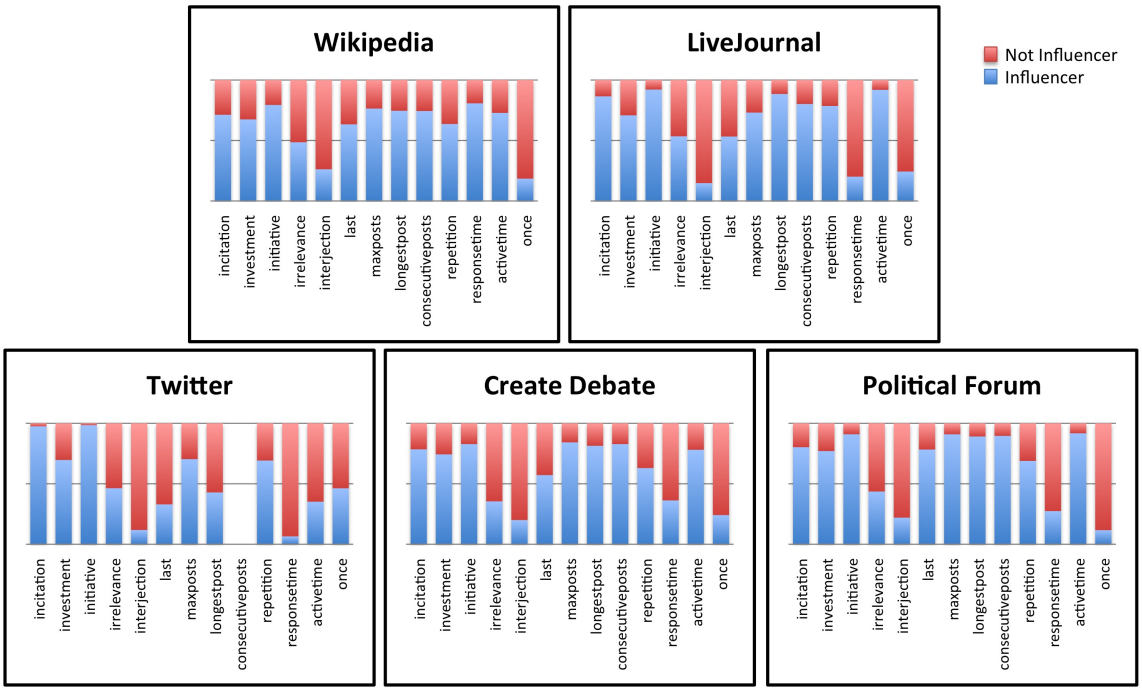


Figure 10.9: The ratio of dialog features towards influencers in each of the training datasets. Interjection, incitation, irrelevance, investment, response time, and active time are all computing using the sum of the feature values per class and should be interpreted accordingly.

Once: The participant is part of the conversation only once

Further details regarding each feature can be found in Chapter 8.2. Table 10.11 lists the values of the dialog features for each of the participants in the Wikipedia Talk Page discussion regarding an image for death shown in Table 9.1. Figure 10.9 shows the distribution of the dialog features towards influencers in the training datasets across genres. It is important to note that for numerical values the sum of the feature is used to generate the ratio. This indicates that interjection should be interpreted as joining early is more indicative of being an influencer. Similarly, influencers tend to have the longest branches following their posts (incitation). It is interesting to note that although the ratios may vary across genres, the trends for the dialog features appear to be more similar than any other group of features maintaining similar ratios across the majority of features. For example, Initiative, investment, incitation, and joining the conversation early (interjection) are positive indicators of

influence across all genres. Posting only once, being ignored, and taking a long time to respond are all indicative of a lack of influence across all genres.

10.5 Domain Adaptation

Domain adaptation has become a popular method for improving results in recent years. For example, it has been used to improve results in Named Entity Recognition and POS tagging [Daumé, 2007], parsing [Plank and van Noord, 2011], and machine translation [Axelrod *et al.*, 2011]. Different supervised approaches to domain adaptation have adopted augmenting the feature space [Daumé, 2007; Daumé *et al.*, 2010], considering the target domain a mixture of source domains [Plank and van Noord, 2011], and using perplexity and cross-entropy [Axelrod *et al.*, 2011; Moore and Lewis, 2010].

Domain adaptation is particularly useful in situations where crowdsourcing and distant supervision are not viable options. These are annotation efforts that are complex, time-consuming, and costly. Annotating discussions for influence is time consuming and costly. The annotator must be trained and must read each discussion in its entirety to make an appropriate decision. It took the annotators 25 minutes on average to choose an influencer for each discussion. This indicates that our entire dataset of 1035 discussions took around 430 hours to be annotated.

In this work, we chose to use a supervised domain adaptation technique that augments the feature space among the target and source domains [Daumé, 2007]. We chose this method due to its great performance, even though it is quite simple to implement. Daumé [2007] augments the source and target feature space to include three versions of each feature; a target version, source version and general version: $< t, s, g >$. The general version is always visible and the target and source version are visible depending on the origin of the feature. A target feature will become $< 1, 0, 1 >$ and a source feature will become $< 0, 1, 1 >$. For example, given the target language, Political Forum (p), and the source languages of Wikipedia (w) ,LiveJournal (l), Create Debate (c), and Twitter (t) all features would be converted to $< p, w, l, c, t, g >$ for each of the genres and the general version. For example, the initiative feature of $< 1 >$ (indicating the person did start the conversation) would become $< 1, 0, 0, 0, 0, 1 >$ for a Political Forum author, $< 0, 1, 0, 0, 0, 1 >$ for a Wikipedia author, $< 0, 0, 1, 0, 0, 1 >$ for a LiveJournal Author and so on. This method is relatively simple to implement,

performs very well, and can be used in conjunction with any learning algorithm. More recent work by Finkel and Manning [2009] describe a similar method that uses priors. Daumé also modified his original method to use semi-supervised learning [Daumé *et al.*, 2010] when unlabeled target data is available as well.

10.6 Conclusion

In this chapter we have described our datasets, definition of influencers, and seven groups of features based on our system components: Claim, Argumentation, Persuasion, Agreement, Author Traits, Credibility, and Dialog. These features have been tuned for supervised learning using the Wikipedia Talk Page development set. We have shown the usefulness of each group of features across all genres in the training sets. It is clear that, in general, the useful features vary across genres indicating that some components may be more useful in some genres than in others. For example, the Credibility mentioned feature is a far more useful indication of influencers in Twitter and Political Forum. Another example is that having a high ratio of claims is indicative of influencers in LiveJournal and Political Forum. Additionally, agreement to others is a positive indication of being an influencer in Twitter. The main exception to this is the Dialog component which tends to have similar trends across all genres for influencers. This indicates that the Dialog component should be useful for domain adaptation which we will show to be true in the following chapter.

In the future we would like to explore adding additional components and features. Particularly, features related to topic have been found to be useful in detecting influencers and power [Nguyen *et al.*, 2013b; Prabhakaran and Rambow, 2014]. We opted to exclude it thus far since most of our discussions tend to be on a single topic. Another component that may be useful is accommodation. Thus far we have opted to include it, with success, in agreement detection. Motivation for including accommodation in influence detection is the combination of reciprocation and social proof weapon's of influence. People feel obligated to return favors and people tend to follow those who are similar to them. Both of these weapon's can be apparent through accommodating towards another person in the discussion.

We also describe domain adaptation; a method we use to take advantage of the multiple online genres we have annotated. In the following chapter we will describe experiments that use domain

adaptation and our system components to achieve improved F-score in detecting influencers across several online genres. We will show that the useful components vary across genre. In the future we would like to explore applying domain adaptation to our system components individually where multiple online genres tend to be available as well (e.g. opinion and agreement detection).

Chapter 11

Experiments and Results

> *Language is a process of free creation; its laws and principles are fixed, but the manner in which the principles of generation are used is free and infinitely varied. Even the interpretation and use of words involves a process of free creation.*
>
> Noam Chomsky, *Language and Freedom (1970)*

Robert B. Cialdini [2007] has stated that the principles of influence that work best differ depending on the conditions of each situation. This chapter describes the experiments and results in predicting influencers in multiple online genres: Wikipedia, LiveJournal, Create Debate, Political Forum, and Twitter. The task of the system is to decide for each participant in a thread whether or not he or she is an influencer in that particular thread.These datasets include discussions where no influencer was present. In the majority of our experiments we exclude threads for which no influencer was found from our experiments, narrowing our task to finding the influencers when they exist which is our main interest. At the end of this chapter we show the difference in performance when the discussions without influencers are included in the test set. For each participant X in each thread Y, the system answers the following question: *Is X an influencer in Y*?

Since we know that there is at least one influencer in each discussion we use the confidence of the classifier to determine the most likely influencer by choosing the person with the highest confidence of being an influencer as the influencer. We consider this approach to be *ranking*. In some cases, using ranking to predict the influencer outperforms the equivalent system without ranking. In the

	Target					Domain Adaptation				
	W	L	T	P	C	W	L	T	P	C
Argue	✓	✓		✓	✓	✓				✓
Claim	✓	✓		✓	✓	✓			✓	✓
Persuade	✓	✓		✓	✓				✓	✓
Agree	✓	✓	✓	✓	✓	✓		✓	✓	
Dialog	✓	✓	✓	✓	✓		✓	✓	✓	✓
Author Traits							✓		✓	✓
Credibility		✓	✓	✓		✓			✓	✓

Table 11.1: Components that are useful for predicting influencers using target training data and provide a positive improvement for domain adaptation within each genre.

future, we would like to adjust the annotation method to rank all of the people in the discussion based on influence instead of just choosing the influencer(s). This would allow us to employ ranking methods such as RankSVM [Herbrich *et al.*, 2000]. We believe this will reduce disagreement among annotators and supply a more accurate picture of influence within a discussion.

We compare our results to two baselines - predicting everyone as an influencer, and a classifier that uses the number of words a person wrote as its only feature. Early work in social science has established that the person who speaks the most is usually the influencer [Bales *et al.*, 1951; Scherer, 1979; Brook and Ng, 1986; Ng *et al.*, 1993; Ng *et al.*, 1995; Reid and Ng, 2000; Bales, 1969]. Therefore, the number of words a person wrote is a strong baseline. All following experiments include the number of words as a feature. Within each genre, we show the results using the features produced from each system component individually, all components (including with and without ranking), and the best components which differ per genre. In addition, we compare using only the Target genre (e.g training and testing on Wikipedia), the Target and Source genre (e.g training on all genres and testing on Wikipedia), and Domain Adaptation (e.g. Training on all genres, applying domain adaptation to take the source/target genre into account, and testing on Wikipedia).

We experimented with a number of different classification methods, including bayesian, rule-based models, and logistic regression. We found that SVM produced the best results. Rather than balancing the training set using downsampling, we balance the class weights of the influencer

examples based on their occurrence in the training data. This ensures that the classifier knows we are more interested in finding influencers without incurring a considerable loss in data. We compute statistical significance using the Approximate Randomization test [Noreen, 1989; Yeh, 2000], a suitable significance metric for F-score. The approximate randomization test examines all n cases where the two results differed and randomly shuffles the results 2^n times or 2^{20} times if $n > 20$. If the difference between the shuffled results are worse than the difference between the actual results a significant amount of times $((nc + 1)/(nt + 1)$ where nc is the number of trials that were worse than the actual results and nt is the number of trials) the results are considered to be significant. All significance is shown at $\leq .05$.

A summary of which components provide an improvement per genre in comparison to both baselines is shown on the left side of Table 11.1. It is clear that the dialog and agreement components are the most useful across all genres. Although part of the best systems in some genres, the author trait component by itself does not provide a positive improvement over the number of words baseline in any genre. The components which have a positive impact when applying domain adaptation are shown on the right side of Table 11.1. The Create Debate and Political Forum genres achieve the biggest impact from applying domain adaptation with 6/7 components providing a positive improvement. More detailed results per genre are provided in the following sections.

11.1 Wikipedia

The complete Wikipedia results are shown in Table 11.2. The dialog, agreement, and argumentation components consistently perform the best as shown in rows 3, 5, and 8. The goal of making an edit to a Wikipedia page means that the dialog components such as speaking a lot and not being ignored are important. In addition, others must agree with proposed edits for them to be of value. Argumentation is important too because a person must be able to clearly argue their opinions with confidence. The second and third to last rows show experiments using all the components with and without ranking. As is evident, there is very little difference in results when using ranking. Using all of the components performs worse than the best system, and some single components on their own clearly indicating that the most useful components vary among genres. The last row in Table 11.2 shows the best Wikipedia results per experiment. The best target experiments for Wikipedia use the

System	Target			Target and Source			Domain Adaptation		
	P	R	F	P	R	F	P	R	F
All Yes	16.7%	100.0%	28.7%	16.7%	100.0%	28.7%	16.7%	100.0%	28.7%
Num Words	34.3%	51.1%	41.0%	30.9%	53.2%	39.1%	30.9%	53.2%	39.1%
Agreement	34.0%	70.2%	$45.8\%^{\alpha}$	32.7%	74.5%	$45.5\%^{\alpha}$	33.0%	76.6%	$46.2\%^{\alpha}$
Argumentation	39.0%	53.2%	$45.2\%^{\alpha}$	36.1%	74.5%	$48.6\%^{\alpha\beta}$	34.6%	76.6%	$47.7\%^{\alpha}$
Claim	28.7%	70.2%	$40.7\%^{\alpha}$	28.2%	74.5%	$40.9\%^{\alpha}$	27.9%	80.8%	$41.5\%^{\alpha}$
Credibility	27.8%	53.2%	$36.5\%^{\alpha}$	21.2%	53.2%	30.3%	21.9%	55.3%	31.3%
Author Trait	29.9%	48.9%	$37.2\%^{\alpha}$	21.1%	59.6%	31.1%	19.6%	74.5%	31.0%
Dialog	39.1%	72.3%	$50.7\%^{\alpha}$	34.2%	83.0%	$48.5\%^{\alpha}$	36.0%	87.2%	$50.9\%^{\alpha\beta}$
Persuasion	29.1%	68.1%	$40.7\%^{\alpha}$	29.1%	68.1%	$40.8\%^{\alpha}$	29.1%	68.1%	40.8%
All	34.3%	70.2%	$46.2\%^{\alpha}$	32.1%	76.6%	$45.3\%^{\alpha}$	41.9%	66.0%	$51.2\%^{\alpha\beta\gamma\delta}$
All Ranking	34.0%	70.2%	$45.8\%^{\alpha}$	32.1%	76.6%	$45.3\%^{\alpha}$	40.8%	66.0%	$50.4\%^{\alpha\beta\gamma\delta}$
Best	41.6%	78.7%	$\mathbf{54.4\%^{\alpha\beta}}$	37.4%	78.7%	$\mathbf{50.6\%^{\alpha\beta}}$	46.6%	72.3%	$\mathbf{56.7\%^{\alpha\beta}}$

Table 11.2: **Influence Detection Results on the Wikipedia Test Set**: The results are shown when training on Wikipedia (Target), all Genres (Target and Source), and applying Domain Adaptation. Performance is in terms of Precision (P), Recall (R), and F-measure (F) using the baselines (everyone is an influencer, and number of words), all features (full), individual features one at a time, and the best feature combination for each data set which differs for each genre. Significance is shown in comparison to $^{\alpha}$ all-yes baseline, $^{\beta}$ num-words baseline, $^{\gamma}$domain adaptation to target, $^{\delta}$ domain adaptation to target and source, and $^{\epsilon}$ target and source to target.

features from the Argumentation and Dialog components and uses ranking with an F-score of 54.4%. We expect these components are most useful because of the nature of Wikipedia Talk Pages as we have already described. The best target and source experiments for Wikipedia use the features from the Claim, Persuasion, Dialog, and Argumentation components with an F-Score of 50.6%. We have motivated the importance of Dialog in Wikipedia. The rest of the components heavily rely on context, an important element in Wikipedia Talk Page conversations where the conversations are lengthy. Finally, the best domain adaptation experiments for Wikipedia use the Claim, Persuasion, Dialog, Credibility, and Argumentation components with an F-score of 56.7%. Here the classifier is able to learn the importance of these components within Wikipedia while taking advantage of additional data.

For example, dialog features are based on patterns so the additional data makes the kind of patterns more clearer and gives more impact to the ones that are best in Wikipedia and in general. Similarly, claims, argumentation, and persuasion occur frequently in Wikipedia, additional data provides further information regarding the context in which they occur. Among all the best experiments, using domain adaptation performs best with an F-Score of 56.7%. However, the difference in comparison to the target result is insignificant. This is unsurprising as Wikipedia has the most annotated data. The results indicate that a small amount of additional out-of-domain data does not provide significant improvements in results when a large amount of in-domain data exists.

11.2 LiveJournal

The complete LiveJournal results are shown in Table 11.3. The dialog and agreement components consistently perform the best as shown in rows 3 and 8. This is due to the nature of LiveJournal as a weblog often used as a personal diary. Dialog is important because the owner of the blog tends to be the influencer. Similarly, people tend to agree with the blogger because it is their personal space. The second and third to last rows show experiments using all the components with and without ranking. As is evident, there is very little improvement in LiveJournal when using ranking. It is also interesting to note that using all the components and domain adaptation does worse than just the target experiments. This indicates that the components that do not adapt well have a negative impact on results. The last row in Table 11.3 shows the best LiveJournal results per experiment. The best target experiments for LiveJournal use the features from the Dialog component for an F-score of 78.3%. Dialog features are the most useful because the blogger, who initiates the discussion, tends to be the influencer. The best target and source experiments for LiveJournal use the Dialog, Author Trait, and Agreement components for an F-Score of 81.6%. This indicates that including additional data is useful, particularly for some of the contextual features that are not as common in LiveJournal but still may have an impact such as agreement. Finally, the best domain adaptation experiments for LiveJournal uses the Dialog component with an F-Score of 79.2% which we have shown to be the most useful feature because of the tendency for the blogger to be the influencer. Among all the best experiments, using the target and source data performs best with an F-Score of 81.6%. However, the difference among the three results is insignificant. This is unsurprising as

System	Target			Target and Source			Domain Adaptation		
	P	R	F	P	R	F	P	R	F
All Yes	19.4%	100.0%	32.5%	19.4%	100.0%	32.5%	19.4%	100.0%	32.5%
Num Words	42.5%	85.0%	56.7%	45.5%	25.0%	32.3%	45.5%	25.0%	32.3%
Agreement	63.0%	85.0%	$72.3\%^{\alpha\beta}$	54.5%	90.0%	$67.9\%^{\alpha\beta}$	56.3%	90.0%	$69.2\%^{\alpha\beta}$
Argumentation	48.6%	85.0%	$61.8\%^{\alpha\beta}$	43.6%	85.0%	$57.6\%^{\alpha\beta}$	50.0%	70.0%	$58.3\%^{\alpha\beta}$
Claim	43.9%	90.0%	$59.0\%^{\alpha}$	42.9%	75.0%	$54.5\%^{\alpha}$	48.3%	70.0%	$57.1\%^{\alpha\beta}$
Credibility	45.0%	90.0%	$60.0\%^{\alpha}$	43.8%	70.0%	$53.9\%^{\alpha}$	34.0%	80.0%	47.8%
Author Trait	37.8%	85.0%	$52.3\%^{\alpha}$	28.6%	50.0%	36.4%	22.5%	90.0%	36.0%
Dialog	69.2%	90.0%	$\mathbf{78.3\%^{\alpha\beta}}$	65.5%	95.0%	$77.6\%^{\alpha\beta}$	67.9%	95.0%	$\mathbf{79.2\%^{\alpha\beta}}$
Persuasion	47.1%	80.0%	$59.3\%^{\alpha}$	50.0%	60.0%	$54.5\%^{\alpha}$	42.3%	55.0%	$47.8\%^{\alpha}$
All	57.7%	75.0%	$65.2\%^{\alpha}$	64.5	100.0%	$78.4\%^{\alpha\beta}$	51.7%	70.0%	57.1%
All Ranking	55.2%	80.0%	$65.3\%^{\alpha}$	64.5%	100.0%	$78.4\%^{\alpha\beta}$	50.0%	70.0%	58.3%
Best	69.2%	90.0%	$\mathbf{78.3\%^{\alpha\beta}}$	69.0%	100.0%	$\mathbf{81.6\%^{\alpha\beta}}$	67.9%	95.0%	$\mathbf{79.2\%^{\alpha\beta}}$

Table 11.3: **Influence Detection Results on the LiveJournal Test Set**: The results are shown when training on LiveJournal (Target), All Genres (Target and Source), and applying Domain Adaptation. Performance is in terms of Precision (P), Recall (R), and F-measure (F) using the baselines (everyone is an influencer, and number of words), all features (full), individual features one at a time, and the best feature combination for each data set which differs for each genre. Significance is shown in comparison to $^{\alpha}$ all-yes baseline, $^{\beta}$ num-words baseline, $^{\gamma}$domain adaptation to target, $^{\delta}$ domain adaptation to target and source, and $^{\epsilon}$ target and source to target.

LiveJournal has a considerable amount of annotated data and the results are high with just target data. Interestingly, knowing whether the data comes from LiveJournal has a negative impact because the dialog component is so useful in LiveJournal.

11.3 Political Forum

The complete Political Forum results are shown in Table 11.4. The agreement and argumentation components consistently perform the best as shown in rows 3 and 4. This portrays the characteristics of the political forum website where people go specifically to discuss political topics that they are

System	Target			Target and Source			Domain Adaptation		
	P	R	F	P	R	F	P	R	F
All Yes	9.1%	100.0%	16.7%	9.1%	100.0%	16.7%	9.1%	100.0%	16.7%
Num Words	27.8%	57.7%	37.5%	29.3%	65.4%	40.5%	29.3%	65.4%	40.5%
Agreement	37.3%	84.6%	$51.8\%^{\alpha\beta}$	43.8%	80.8%	$56.8\%^{\alpha\beta}$	42.9%	80.8%	$56.0\%^{\alpha\beta}$
Argumentation	44.4%	76.9%	$56.3\%^{\alpha}$	42.6%	76.9%	$54.8\%^{\alpha\beta}$	51.7%	57.7%	$54.5\%^{\alpha\beta}$
Claim	33.9%	73.1%	$46.3\%^{\alpha}$	29.6%	80.8%	$43.3\%^{\alpha}$	41.5%	65.4%	$50.7\%^{\alpha}$
Credibility	27.8%	57.7%	$37.5\%^{\alpha}$	21.0%	65.3%	$31.8\%^{\alpha}$	36.1%	50.0%	$41.9\%^{\alpha}$
Author Trait	16.3%	53.8%	$25.0\%^{\alpha}$	15.4%	61.5%	$24.6\%^{\alpha}$	18.3%	42.3%	$25.6\%^{\alpha}$
Dialog	30.1%	84.6%	$44.4\%^{\alpha}$	40.8%	76.9%	$53.3\%^{\alpha\beta\epsilon}$	36.1%	84.6%	$50.6\%^{\alpha\gamma}$
Persuasion	35.7%	76.9%	$48.8\%^{\alpha}$	22.8%	80.8%	$35.6\%^{\alpha}$	40.4%	80.8%	$53.8\%^{\alpha\beta\delta}$
All	50.0%	34.6%	$40.9\%^{\alpha}$	39.6%	80.8%	$53.2\%^{\alpha\beta}$	40.0%	38.5%	$39.2\%^{\alpha}$
All Ranking	42.9%	46.2%	$44.4\%^{\alpha}$	39.6%	80.8%	$53.2\%^{\alpha\beta}$	38.2%	50.0%	$43.3\%^{\alpha}$
Best	62.1%	76.9%	$\mathbf{69.0\%^{\alpha\beta}}$	52.4%	84.6%	$\mathbf{64.7\%^{\alpha\beta}}$	71.4%	76.9%	$\mathbf{74.1\%^{\alpha\beta}}$

Table 11.4: **Influence Detection Results on the Political Forum Test Set**: The results are shown when training on Political Forum (Target), All Genres (Target and Source), and applying Domain Adaptation. Performance is in terms of Precision (P), Recall (R), and F-measure (F) using the baselines (everyone is an influencer, and number of words), all features (full), individual features one at a time, and the best feature combination for each data set which differs for each genre. Significance is shown in comparison to $^{\alpha}$ all-yes baseline, $^{\beta}$ num-words baseline, $^{\gamma}$domain adaptation to target, $^{\delta}$ domain adaptation to target and source, and $^{\epsilon}$ target and source to target.

passionate about. The second and third to last rows show experiments using all the components with and without ranking. Ranking helps in both the target and domain adaptation experiment. Using all of the components performs worse than the best system and some single components on their own clearly indicating that the most useful components vary among genres. The best target experiments for Political Forum use the features from the Persuasion, Argumentation, Agreement, and Credibility components to ensure an influencer is chosen in each discussion with an F-score of 69.0%. This is indicative of the importance of context in Political Forum. The best target and source experiments for Political Forum use the features from the Claim, Argumentation, and Agreement components with an

F-score of 64.7%. These results are further indication of the importance of the contextual components. Finally, the best domain adaptation experiments for Political Forum are almost identical to the target experiment using the Persuasion, Argumentation, and Agreement components and ranking to ensure an influencer is chosen in each discussion with an F-score of 74.1%. This may explain why domain adaptation is so useful, the components with a positive impact adapt well from the other domains. Using domain adaptation performs best with an F-Score of 74.1%, a 5.1 improvement compared to only using the target genre.

11.4 Create Debate

The complete Create Debate results are shown in Table 11.5. The agreement, argumentation, and claim components in rows 3-5 have the most impact indicating the importance of context in Create Debate. This is evident in the style of create debate where agreement is very important to indicate what side of the debate a person is on. In addition, to be convincing in a debate a person must make claims and arguments. Our results in Create Debate are lower than the results in any other genres. This is due to the size of the conversation. Each create debate discussion has a lot of participants, but still only one to two influencers. This is evident in the low all-yes baseline (10.8% F-score) which indicates how rare influencers are in Create Debate. The second and third to last rows show experiments using all the components with and without ranking. Ranking does not provide an improvement but there is no significant difference between the results. Using all of the components performs worse than the best system and some single components on their own clearly indicating that the most useful components vary among genres. The best target experiments for Create Debate use all features except for Persuasion with an F-score of 30.6%. It is not entirely clear why persuasion is the only one that is not useful. It does occur often in Create Debate, but perhaps it does not provide additional information on top of claims and argumentation (recall persuasion is a subset of the two; not all claim and argumentation sentences will be classified as persuasion). The best target and source experiments for Create Debate use the Persuasion, Agreement, and Credibility components with an F-score of 33.7%. This is indication that context is important in Create Debate. Finally, the best domain adaptation experiments for Create Debate is the Dialog and Agreement components with an F-score of 34.1%. These components are useful in the target experiments and they adapt well

System	Target			Target and Source			Domain Adaptation		
	P	R	F	P	R	F	P	R	F
All Yes	5.7%	100.0%	10.8%	5.7%	100.0%	10.8%	5.7%	100.0%	10.8%
Num Words	15.2%	42.9%	22.4%	13.1%	71.4%	22.1%	13.1%	71.4%	22.1%
Agreement	18.7%	60.7%	$28.6\%^{\alpha}$	18.8%	32.1%	$23.7\%^{\alpha}$	19.2%	35.7%	$25.0\%^{\alpha}$
Argumentation	19.3%	57.1%	$28.8\%^{\alpha}$	14.3%	21.4%	$17.1\%^{\alpha}$	24.5%	46.4%	$32.1\%^{\alpha\beta\delta}$
Claim	17.8%	67.9%	$28.1\%^{\alpha}$	17.3%	67.9%	$27.5\%^{\alpha}$	20.5%	57.1%	$30.2\%^{\alpha\beta}$
Credibility	10.0%	78.6%	$17.7\%^{\alpha}$	10.6%	71.4%	$18.4\%^{\alpha\epsilon}$	19.5%	28.6%	$23.2\%^{\alpha}$
Author Trait	9.6%	46.4%	$15.9\%^{\alpha}$	8.0%	78.6%	$14.5\%^{\alpha}$	21.6%	39.3%	$27.8\%^{\alpha\gamma\delta}$
Dialog	15.3%	75.0%	$25.5\%^{\alpha}$	17.3%	32.1%	$22.5\%^{\alpha}$	22.0%	39.3%	$28.2\%^{\alpha}$
Persuasion	16.4%	75.0%	$26.9\%^{\alpha}$	11.7%	71.4%	$20.1\%^{\alpha}$	21.2%	39.3%	$27.5\%^{\alpha}$
All	20.4%	35.7%	$26.0\%^{\alpha}$	16.7%	35.7%	$22.7\%^{\alpha}$	15.2%	17.9%	16.4%
All Ranking	19.6%	35.7%	$25.3\%^{\alpha}$	16.7%	35.7%	$22.7\%^{\alpha}$	12.2%	17.9%	14.5%
Best	22.8%	46.4%	$\mathbf{30.6\%^{\alpha}}$	22.7%	53.6%	$\mathbf{33.7\%^{\alpha\beta}}$	25.9%	50.0%	$\mathbf{34.1\%^{\alpha\beta}}$

Table 11.5: **Influence Detection Results on the Create Debate Test Set**: The results are shown when training on Create Debate (Target), All Genres (Target and Source), and applying Domain Adaptation. Performance is in terms of Precision (P), Recall (R), and F-measure (F) using the baselines (everyone is an influencer, and number of words), all features (full), individual features one at a time, and the best feature combination for each data set which differs for each genre. Significance is shown in comparison to $^{\alpha}$ all-yes baseline, $^{\beta}$ num-words baseline, $^{\gamma}$domain adaptation to target, $^{\delta}$ domain adaptation to target and source, and $^{\epsilon}$ target and source to target.

across datasets. Among all the best experiments, using the Domain Adaptation performs best with an F-Score of 34.1%. This is not a significant improvement over the best Target result. However, while the best target result is not significant over the number of number of words baseline, the domain adaptation experiment is.

11.5 Twitter

The complete Twitter results are shown in Table 11.6. The best Twitter components are consistently agreement and dialog in rows 3 and 8 of Table 11.6. The dialog component is very useful in Twitter

System	Target			Target and Source			Domain Adaptation		
	P	R	F	P	R	F	P	R	F
%All Yes	7.8%	100.0%	14.4%	7.8%	100.0%	14.4%	7.8%	100.0%	14.4%
Num Words	11.9%	26.7%	16.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Agreement	61.7%	96.7%	$75.3\%^{\alpha\beta}$	80.6%	96.7%	$87.9\%^{\alpha\beta\epsilon}$	67.4%	96.7%	$79.5\%^{\alpha\beta}$
Argumentation	11.1%	23.3%	15.1%	50.0%	13.3%	21.1%	0.0%	0.0%	0.0%
Claim	12.1%	23.3%	15.9%	33.3%	16.7%	22.2%	50.0%	6.7%	11.8%
Credibility	61.4%	90.0%	$73.0\%^{\alpha\beta}$	18.8%	70.0%	$29.6\%^{\beta}$	61.4%	90.0%	$73.0\%^{\alpha\beta}$
Author Trait	8.5%	33.3%	13.5%	9.3%	40.0%	$15.1\%^{\beta}$	9.7%	10.0%	9.8%
Dialog	55.6%	100.0%	$71.4\%^{\alpha\beta}$	77.8%	93.3%	$84.8\%^{\alpha\beta\epsilon}$	76.3%	96.7%	$85.3\%^{\alpha\beta\gamma}$
Persuasion	11.9%	26.7%	16.5%	33.3%	3.3%	6.1%	33.3%	3.3%	6.1%
All	61.9%	86.7%	$72.2\%^{\alpha\beta}$	77.8%	93.3%	$84.8\%^{\alpha\beta\epsilon}$	58.1%	83.3%	$68.5\%^{\alpha\beta}$
All Ranking	62.8%	90.0%	$74.0\%^{\alpha\beta}$	77.8%	93.3%	$84.8\%^{\alpha\beta\epsilon}$	57.8%	86.7%	$69.3\%^{\alpha\beta}$
Best	67.5%	90.0%	$\mathbf{77.9\%^{\alpha\beta}}$	88.2%	100.0%	$\mathbf{93.8\%^{\alpha\beta\epsilon}}$	80.0%	93.3%	$\mathbf{86.2\%^{\alpha\beta}}$

Table 11.6: **Influence Detection Results on the Twitter Test Set**: The results are shown when training on Twitter (Target), All Genres (Target and Source), and applying Domain Adaptation. Performance is in terms of Precision (P), Recall (R), and% F-measure (F) using the baselines (everyone is an influencer, and number of words), all features (full), individual features one at a time, and the best feature combination for each data set which differs for each genre. Significance is shown in comparison to $^{\alpha}$ all-yes baseline, $^{\beta}$ num-words baseline, $^{\gamma}$domain adaptation to target, $^{\delta}$ domain adaptation to target and source, and $^{\epsilon}$ target and source to target.

because the first person is usually the influencer. In addition, due to retweets they are also responded to the most. Agreement is very useful because retweets are a form of agreement very common in Twitter. Similarly, Credibility is very useful because being mentioned by name is common via retweets. The second and third to last rows show experiments using all the components with and without ranking. Ranking provides a small, but insignificant, improvement between the two results. Using all of the components performs worse than the best system and some single components on their own clearly indicating that the most useful components vary among genres. The best target experiments for Twitter use the features from the Dialog and Credibility components to ensure an influencer is chosen in each discussion for an F-Score of 77.9%. The best target and source

experiments for Twitter use the features from the Claim, Argumentation, and Agreement components. All three of these components improved when going from using only target data to all data. This indicates that the initial data helped. Finally, the best domain adaptation experiments for Twitter is the Claim, Agreement, and Credibility components and ranking to ensure an influencer is chosen in each discussion. As in the target experiments the most useful of these components is agreement. Among all the best experiments, using the Target and Source data performs best with a recall of 100% an F-Score of 93.8% significantly outperforming the target results. Agreement is very useful in all Twitter experiments because of retweets (RT), which are essentially the same post as the one being retweeted. Using additional genres as training data is very useful, particularly due to the small size of the dataset. We expect the significant improvement using all data without applying domain adaptation is due to the lack of context in Twitter (because of the 144 character limit of tweets). We expect Twitter gains from the genres with additional context and not being informed of the genre at training time (as occurs in domain adaptation) helps give more weight to out-of-domain genres.

11.6 Discussions Without Influencers

In this section we show the impact of excluding discussions without influencers from our datasets. In these experiments we run the same models that have been trained only on discussions with influencers, and testing on discussions with and without influencers. Table 11.7 shows a comparison between the best systems with and without influencer in the test sets across all genres and the target, target and source, and domain adaptation experiments. In LiveJournal and Twitter there is very little difference because there were very few (Twitter) or no threads (LiveJournal) without influencers. The best system does vary in the domain adaptation experiment in Twitter where just the Dialog component provides the best system when including all discussions. We expect this is due to the structural nature of these features. It appears that they can adapt better to a discussion where no influencers are present. The biggest difference occurs in Political Forum where the variation in the target and domain adaptation experiments is around 10 points. The best system did not change in all of the experiments. This indicates that the system felt the need to always predict an influencer in the discussions where no influencer was annotated. In Create Debate the best experiments vary by around 5 points. The only experiment that differed is the target experiment where when excluding all discussions without

Genre	System	Target			Target and Source			Domain Adaptation		
		P	R	F	P	R	F	P	R	F
Wikipedia	Influencer only	41.6	78.7	54.4	37.4	78.7	50.6	46.6	72.3	56.7
	All discussions	37.0	78.7	50.3	32.5	83.0	46.7	41.5	72.3	52.7
LiveJournal	Influencer only	69.2	90.0	78.3	69.0	100.0	81.6	67.9	95.0	79.2
	All discussions	69.2	90.0	78.3	69.0	100.0	81.6	67.9	95.0	79.2
Political Forum	Influencer only	62.1	76.9	69.0	52.4	84.6	64.7	71.4	76.9	74.1
	All discussions	50.0	76.9	60.6	46.8	84.6	60.3	55.6	76.9	64.5
Create Debate	Influencer only	22.8	46.4	30.6	22.7	53.6	33.7	25.9	50.0	34.1
	All discussions	16.5	71.4	26.9	21.5	50.0	30.1	20.6	50.0	29.2
Twitter	Influencer only	67.5	90.0	77.9	88.2	100.0	93.8	80.0	93.3	86.2
	All discussions	62.5	100.0	76.9	85.7	100.0	92.3	74.4	96.7	84.1

Table 11.7: **Influence Detection Results on Discussions without Influencers**: A comparison of the best systems when including only discussions with influencers (Influencer Only) and all discussions in the test set. The discussions with influencers results is equivalent to the best results shown in the prior tables. The results are shown when training on each genre (Target), All Genres (Target and Source), and applying Domain Adaptation. Performance is in percentage (%) in terms of Precision (P), Recall (R), and F-measure (F).

influencers the system used all components except for Persuasion. In contrast, the best system including all discussions used the Persuasion and Claim components. Finally, in Wikipedia, the best experiments vary by around 5 points as in Create Debate. The only experiment that differed is the target and source experiment. The best system when excluding all discussions without influencers included the Claim, Persuasion, Dialog, and Argumentation components. The best system when including all discussions used the Dialog, Demographic, and Agreement components. In the future we feel it would be useful to have a system that determined the likelihood of an influencer being present in the discussion *prior* to running the influence detection system.

11.7 Conclusion

Across all genres, our best systems significantly beat two baselines: predicting everyone as an influencer, and the number of words a person wrote. The number of words a person wrote is a strong baseline because the person who contributes to the conversation the most is usually the influencer. The components that are useful individually, and as part of the best system, differ per genre. Overall each component is useful in multiple genres. Dialog structure, claims, and agreement are all important in Wikipedia. These components point to the desire to make edits in Wikipedia. It is important to be clear and have people agree with you to have influence for your changes to be acknowledged. In LiveJournal, dialog structure and agreement are most important. This is because the owner of the blog is often the influencer and people tend to agree with them since it is their blog. Agreement and argumentation are most important in Political Forum. In Political Forum, people discuss political topics and can become very argumentative. Of course, having people agree with your arguments makes it more likely for you to be an influencer. In Create Debate, the agreement, argumentation, and claim components are all important. This is indicative of the debate style. All of these components are important to being convincing in a debate. Finally, in Twitter, dialog structure and agreement are most important. The first person to tweet is usually the influencer, and being retweeted make agreement important. Although author traits are not useful individually, they are part of the best system in all genres except for Political Forum. Similarly, persuasion is part of the best system in Twitter, Political Forum, and Wikipedia. Credibility is part of the best system in Wikipedia. A summarization of the components that are part of the best system in each genre and experiment can be found in Table 11.8.

The usefulness of all the components validates our features and further drives home the point Robert Cialdini has made about his weapons of influence: “All the weapons of influence discussed in this book work better under some conditions than under others. If we are to defend ourselves adequately against any such weapon, it is vital that we know its optimal operating conditions in order to recognize when we are most vulnerable to its influence.” [2007]. Knowing that different components work in different situations can be very useful. This information could be used to not only say *who* is influential, but *how* to be influential. For example, showing a person which components they use to be influential could also help them improve how they influence others, and help them adjust to different settings.

	Target					Target & Source					Domain Adaptation				
	W	L	P	C	T	W	L	P	C	T	W	L	P	C	T
Argue	✓		✓	✓		✓		✓		✓	✓		✓		
Claim				✓		✓		✓		✓	✓				✓
Persuade			✓			✓			✓		✓		✓		
Agree			✓	✓			✓	✓	✓	✓			✓	✓	✓
Dialog	✓	✓		✓	✓	✓	✓				✓	✓		✓	
Author Traits				✓			✓								
Credibility			✓	✓	✓				✓		✓				✓
Ranking	✓												✓		✓

Table 11.8: Components that are useful in the best systems for predicting influencers using target training data, all training data, and domain adaptation within each genre.

As our results show, domain adaptation provides significantly better results in the best system in Political Forum. In Twitter, using the Target and Source data provides a significant improvement. Finally, although using domain adaptation does not provide a significant improvement over the best target results in Create Debate, it does beat the number of words baseline. The best target results do not. The components that are successful vary across genres, even when applying domain adaptation. In fact, using all the components often performs worse with domain adaptation. The size, formatting, and writing style all have an impact on the results:

Size: Domain adaptation does not have a significant impact in Wikipedia and LiveJournal because these genres have the most annotated data.

Formatting Style: Across the majority of genres, Dialog Patterns achieve the biggest increase when applying domain adaptation. This is intuitive as it is the feature that is not dependent on context, but on formatting. In general, the formatting structure is similar across all genres and most features apply even in the cases where the format is slightly different (e.g. Create Debate is a two sided thread structure). For example, there is always a person who posts first, posts the most, is ignored etc... In Twitter, although domain adaptation helps, using all of the genres without domain adaptation provides an even more significant improvement. We expect this is because of the word limitation in Twitter. The other genres provide further contextual information enabling other components to be

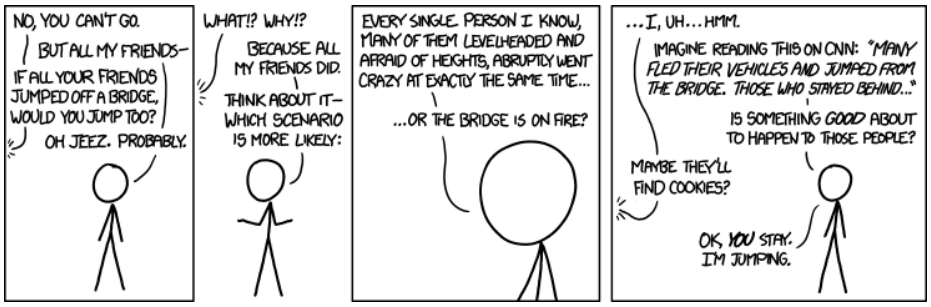
beneficial.

Writing Style: Political forum benefited the most from domain adaptation. This could suggest that perhaps political forum is most similar in context to Wikipedia, our largest dataset. One thing is certain, each author participates a the same amount in Wikipedia and Political Forum discussions as shown in the prior Chapter in Table 10.3. The ratio of the occurrence of claim, argumentation, and persuasion is also most similar among Wikipedia and Political Forum as shown in Table 10.4 in the prior chapter. Although the ratio of agreement and disagreement differs, the combination (e.g. (dis)agreement) is also most similar among the two genres.

In conclusion, we use several datasets and components for detecting influence in online genres. We show that different components work best in each situation. We also apply domain adaptation to these datasets to take advantage of data across genre. We show that style and size have an impact on domain adaptation results. In the future, we would like to explore a semi-supervised approach to exploit unlabeled data. We would also like to experiment with annotating the discussions by ranking all the participants in the discussion from most to least influential. We believe this would help reduce disagreement among annotators.

Chapter 12

Social Proof



Randall Munroe, *Bridge*[1]*, XKCD*

In the previous chapter, we showed how all of the components contribute in detecting influence. Here, we focus on a single component (author trait detection) and its motivation in social science to deeply recognize its impact in influence detection. The psychological phenomenon of *social proof* dictates that people will be influenced by others in their surroundings. Furthermore, social proof is most effective when a person perceives the people in their surroundings to be similar to them [Cialdini, 2007]. This tendency is known as *homophily*. One manner in which people can be similar is through shared author traits such as age (year of birth), gender (male/female), and religion (Christian/ Jewish/ Muslim/ Atheist), as well as political party (Republican/Democrat).

[1]https://xkcd.com/1170/

In this chapter, we explore the impact of social proof via author traits in detecting the most influential people in Wikipedia Talk Page discussions. We use the author trait detection system described in Chapter 7 to classify individuals in the Wikipedia Talk Page discussions along four author traits: age, gender, religion, and political party. Our experiments show that, although not always significant, there are social proof tendencies among participants and that the topic of the discussion plays a role in determining which author traits are useful. For example, religion is more indicative of influence in discussions that are religious in nature such as a discussion about the Catholic Church.

In the rest of this chapter we show the impact of author traits on influence detection via the author trait single, majority, and combination features. In addition, we also use topic features that are only available in Wikipedia as described in the following section.

12.1 Method

Our method involves four groups of features. The first three, as described in Chapter 10, are: single features related to each author trait, features indicating if the author trait is the majority in the discussion, and a combination of author traits. We also include features related to the issue being discussed in the Wikipedia Page, which is not readily available in other online genres. We will describe the features in greater detail in relation to the Wikipedia Talk Page dataset in the rest of this section.

In addition, as in Chapter 10, we include a classifier that uses only the number of words the participant has written as a feature as a baseline. This feature is important because in addition to indicating the likelihood of someone being influential (if someone barely participates in the discussion it reduces their chances of being influential), the odds of the predicted author trait being correct decreases if the provided text is minimal.

12.1.1 Single Features

We explore the occurrence of influence based on each author trait as an indication of what type of people are more influential. Each author trait is represented as a binary feature during classification. The breakdown of each feature by influence in the training set is shown in Figure 12.1. As indicated,

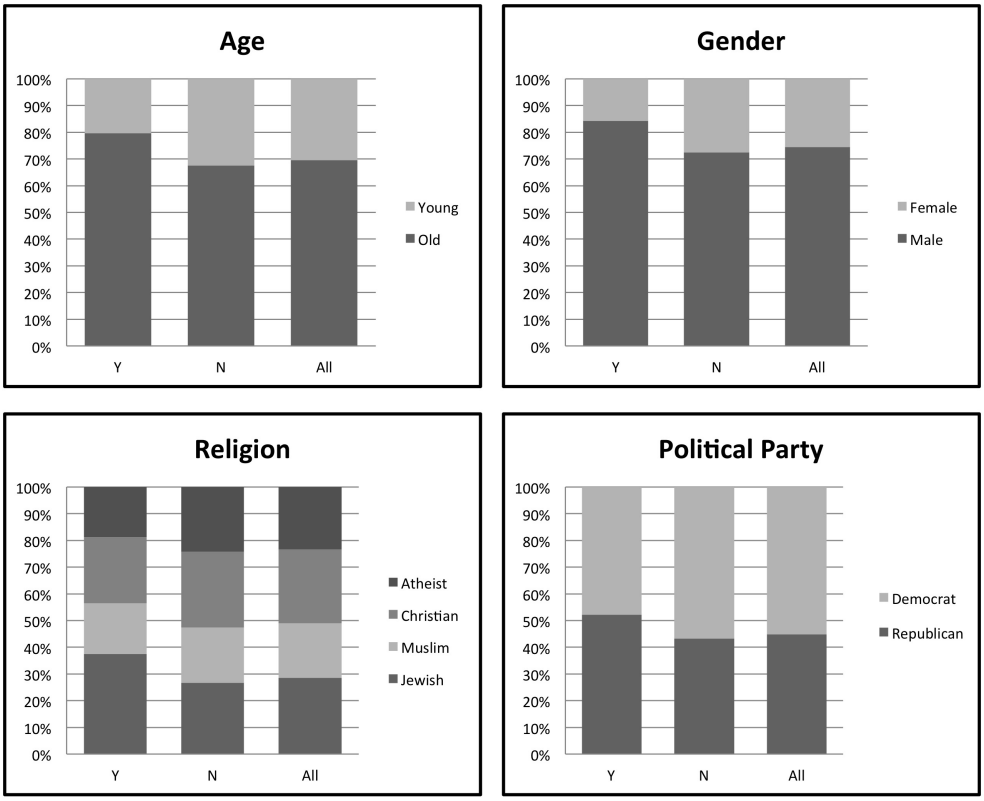


Figure 12.1: The breakdown the author trait single features by influence (Y/N) and overall (All) in the Wikipedia Talk Page training set.

there tend to be more old people in Wikipedia, but there is also a clear indication that older people are more influential. We have similar findings with the male gender, the republican political party, and the Jews and Atheists in religion. Interestingly, the majority of Wikipedia contributors are liberal, but conservatives are more influential. We suspect that the tendency towards a certain author trait may be dependent on the topic of the Wikipedia article as discussed in the following section. For example, political party may play a more important role in a discussion regarding abortion and religion may play a more important role in a discussion regarding Israel. Finally, we also have a feature indicating the exact year of birth that was predicted for each author (e.g. 1983).

Topic	G	A	R	P
Abortion	35	8	9	3
Catholic Church	27	9	553	7
George W. Bush	4	0	8	68
Israel	4	18	623	12
Michael Jackson	2	42	4	0

Table 12.1: A list of topics and the occurrence of issues associated with them in **A**ge, **G**ender, **R**eligion, and **P**olitics. An occurrence > 5 indicates it is an issue relevant to that topic.

Author Trait	Keywords
gender	male, female, gender, women, men
age	young, old, teenager, generation, 2000s, 1990s, 1980s, 1970s, 1960s, 1950s
religion	religion, religious, jewish, judaism, christian, catholic, islam, muslim, muslims, jews, israel, atheist, atheism, church
politics	politics, political, republican, democrat, conservative, liberal

Table 12.2: List of labels and synonyms used to assign author trait topics to Wikipedia articles

12.1.1.1 Topic Features

The topic in a discussion can indicate what kind of issues will be addressed. This in turn can indicate a stronger presence of different author traits. For example, a Wikipedia article about the Catholic Church will be more likely to be edited by religious people than a Wikipedia article about the pop star Michael Jackson. This in turn can indicate the author trait tendencies of the people in the Talk Page. In order to analyze the impact of topic on influence and author traits we automatically inferred the author traits that were likely to be related to the Wikipedia article.

We implemented this by counting the occurrence of the labels and related synonyms of each author trait within the Wikipedia article. For example, male and female are gender labels. A full list of labels and synonyms is shown in Table 12.2. This alone was sufficient for our task since we want high precision and care less about recall. It is important to stress, that we did not do this in the Wikipedia Talk Pages but rather in the *actual* Wikipedia article. If an author trait term occurred more

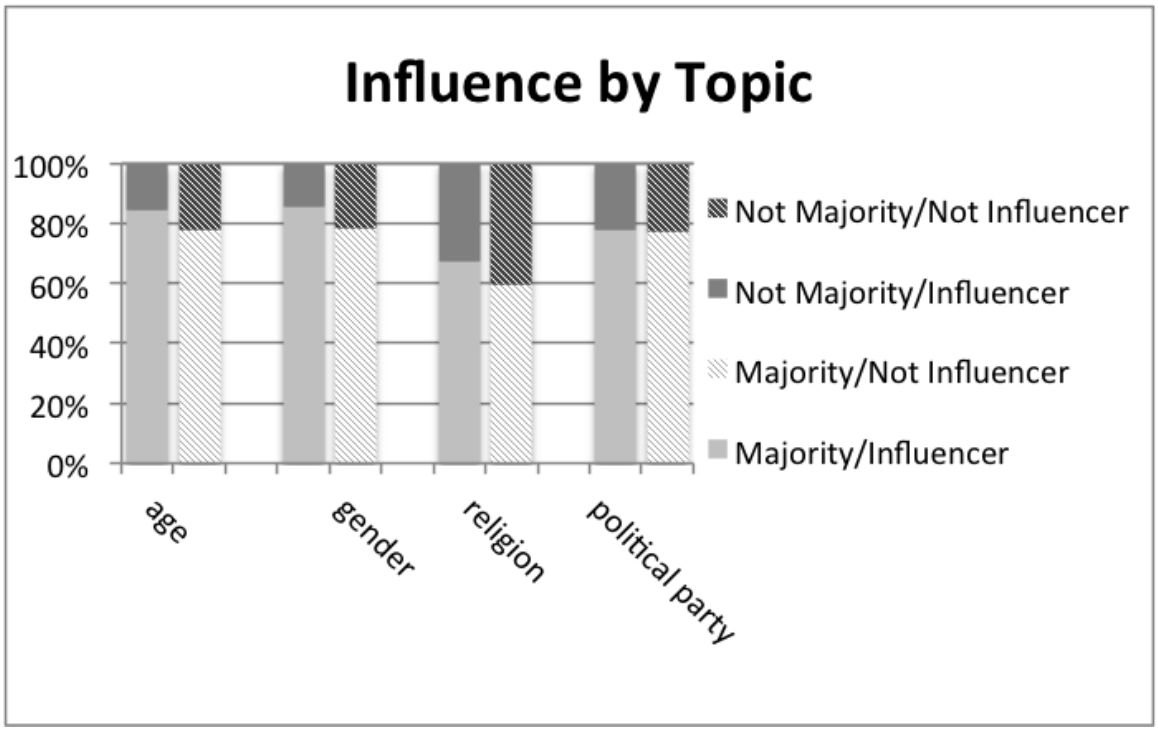


Figure 12.2: The breakdown of the users being in the majority within their document for each author trait with topic being taken into account.

than five times[2] it was considered to be an issue related to that author trait to ensure the occurrence was more than happenstance. Table 12.1 lists an example of topics and the occurrence of issues within the Wikipedia article.

Using this method, there were 38 age, 42 gender, 66 religious, and 58 political articles, with most articles overlapping on more than one author trait issue. There are a total of 99 topics with one or multiple discussions from the Talk Page associated to the topic.

We use each issue as a feature which is set to true if that topic is associated with the article and false if it is not. For example, the gender, age, and religion issues would be set to true for Abortion Talk Pages.

12.1.2 Majority Features

Social proof indicates that people will be influenced by those that are like them. We measure this per author trait by determining if a person is predicted to be in the majority within the discussion and have a majority feature corresponding to each author trait. For example, if the majority of the people in a discussion are predicted to be republican, we expect that the influencer is likely to be predicted to be republican as well. Furthermore, we expect this to be most evident when the discussion is relevant to the particular author trait. For example, a discussion on abortion would be relevant to religion, politics, and gender. Thus, we only include discussions per related topic in our figure. Figure 12.2 illustrates that influencers are in the majority more than non-influencers when the issue is relevant in the Wikipedia article. As Figure 12.2 shows, in general all people tend to be in the majority author trait in a discussion, but there is a stronger tendency towards being in the majority when a person is an influencer. The results displayed take the topic of the document into account in that only documents applicable to each author trait are shown in the chart. In other words, the gender bars in the graph include a subset of 42 documents where the topic is gender. A topic can fall under multiple author traits in which case it will be included in multiple bars on the graph. For example, discussions on abortion are only included in the bars on religion, politics, and gender. We also include features to indicate whether the participant is in the majority in *all* author traits or in *no* author traits.

In order to determine whether the majority features should be useful, in addition to using the single features, we needed to verify whether there were enough cases where the overall minority author trait was still the majority author trait within a reasonable amount of discussions. We find that in the training data, in 84.1% of the discussions the majority is older people and in 88.5% of the discussions the majority is male. These percentages are in line with the trends found in the single features as shown in Figure 12.1. However, there still are many discussions where the majority is female (11.5%) or younger people (15.9%). In contrast to our findings in the single features shown in Figure 12.1, where overall there were slightly more Republicans than Democrats, we found that in 55.8% of the discussions in the training data the majority is Democrat whereas slightly more editors are Republican. In terms of religion we found that in 41.5%, 16.6%, 18.8%, and 23.2% of the discussions the majority is Jewish, Muslim, Christian, and Atheist respectively. Although

[2]The split of terms among documents is such that documents have no terms whatsoever most often and fewer than 6 terms related to an issue 48.5% times whereas 51.5% of the issues have 6 or more terms.

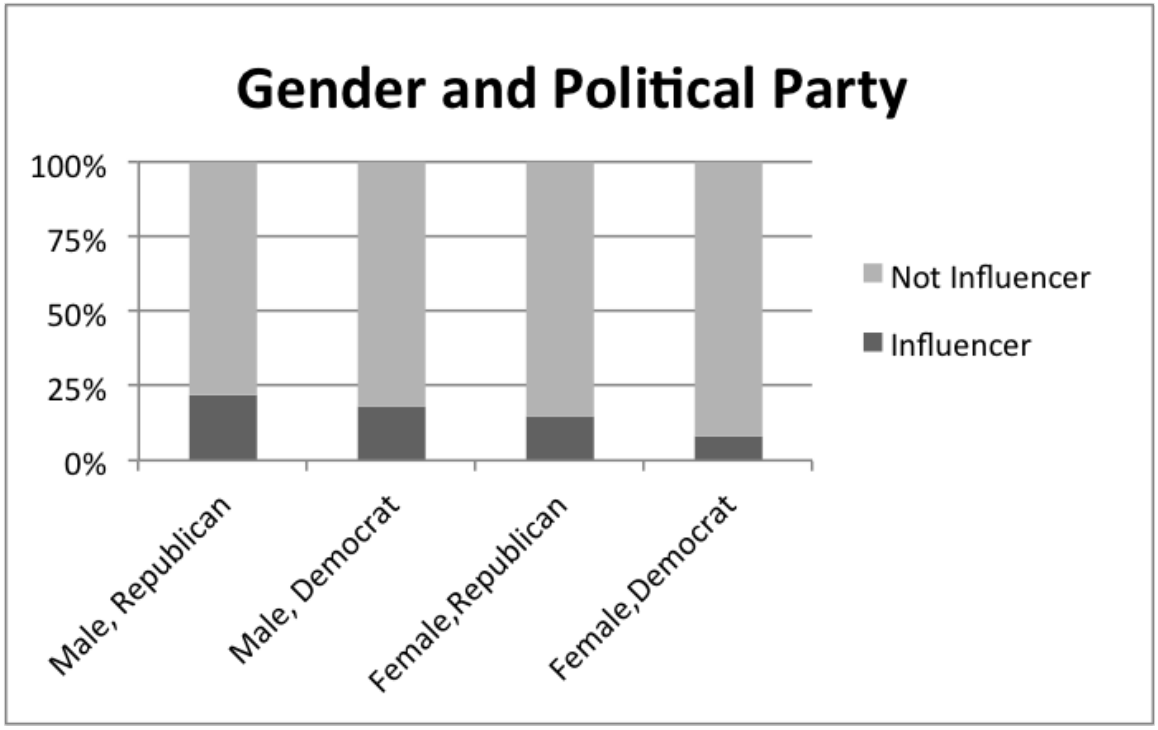


Figure 12.3: The breakdown of influencers and non-influencers in the training data based on the single combination feature of gender and political party.

Christianity is the most commonly predicted religion overall (see Figure 12.1), we expect that in the discussions the majority is Judaism due to the many articles that are controversial to the state of Israel (e.g. regarding Gaza and the Israeli Defense Force). This indicates that, in particular, using the majority religion feature should have a positive impact on predicting influencer in addition to the single religion features.

12.1.3 Combination Features

In addition to looking at a single author trait of a person at a time, we also explore whether combining author traits is beneficial. There have been many studies which show that certain tendencies towards an issue are based on many author traits. Examples of when this is applicable are described in Chapter 10. We achieve this by combining every two author traits, d_i, d_j, into a single feature $<d_i, d_j>$ and majority feature $<dm_i, dm_j>$. For example, <gender,political party> and <gender_m,political party_m>. As one example, we find that indeed in our dataset women are 62.5%

more likely to be Democrat. However, we find that women that are Republican are more likely to be influential than women who are Democrat as shown in the breakdown of the <gender,political party> feature in the training data as shown in Figure 12.3. We also find that a person is more likely to be an influencer if they are in the majority for both political party and gender.

12.2 Experiments and Results

As in Chapter 11, all results were predicted using the SVM classifier in Weka built with logistic models to provide confidence values. All SVM parameters were optimized using the development set. Rather than balancing the training set using downsampling, we balance the class weights of the influencer examples based on their occurrence in the training data. This ensures that the classifier knows we are more interested in finding influencers without incurring a considerable loss in data. All results reported are shown using F-score because influence is rare. Statistical significance is computed using the Approximate Randomization test [Noreen, 1989; Yeh, 2000]. We compare our experiments to two baselines, picking everyone as an influencer (all-yes baseline), and the number of words a person wrote in the discussion (num-words baseline).

Since we know that there is at least one influencer in each discussion we use the confidence of the classifier to determine the most likely influencer by choosing the person with the highest confidence of being an influencer as the influencer. Using ranking to predict the influencer outperforms the equivalent system without ranking in some cases.

All results following the baselines include the number of words and topic features unless otherwise mentioned. We show experiments on the development and test set. We felt it was important to include both results because we could not achieve a significant improvement in comparison to the number of words baseline on the test set. This indicates that further work still needs to be done in this analysis.

In the development set, the system using just the best majority features gives 5.0 points improvement in F-score compared to using just the number of words in a sentence (Table 12.3, row 3). These features are gender, religion, and political party, and never-the-majority features. Ranking was also useful in this system. In row 4 of Table 12.3, we show that the best system using just single features achieves a 5.5 points improvement in F-score compared to using just the number of

	Development				Test			
Experiment	Confusion Matrix	P%	R%	F%	Confusion Matrix	P%	R%	F%
all-influencer	**41** 0 174 0	19.1	**100.0**	32.0	**47** 0 235 0	16.7	**100.0**	28.7
num words	23 36 18 138	39.0	56.1	45.0	24 46 23 189	34.3	51.0	41.0
majority best	27 40 14 134	40.3	65.9	50.0^{R}	28 54 19 181	34.1	59.6	43.4^{R}
single best	26 36 15 138	41.9	63.4	50.5	26 45 21 190	36.6	55.3	44.1
majority+single best	26 35 15 139	42.6	63.4	51.0	20 51 18 184	36.3	**61.7**	45.7^{R}
best w/o topic	26 38 15 136	40.6	63.4	49.5	27 51 20 184	34.6	57.5	43.2^{R}
best	**28** 37 13 137	**43.1**	**68.3**	$\mathbf{52.8}^{\alpha}$	**29** 50 18 185	**36.7**	**61.7**	$\mathbf{46.0}^{R}$

Table 12.3: The results of all groups of features on influence detection using author traits. The confusion matrix is filled as [TP FN] in the first row and [FP TN] in the second row. ^{R} indicates that ranking was used in the results. The best results are highlighted in bold. All results are significant in comparison to the all-yes baseline. $^{\alpha}$ significant in comparison to the num-words baseline.

words in the sentence. This system uses the three age features (exact, binary, and distance), religion, and political party. The success of the single features indicates that homophily is important. In other words, having author traits similar to most other people in Wikipedia is very indicative of being influential. The best system using single and majority features combined (Table 12.3, row 5) gave an improvement of 6.0 points in F-score overall. These features are the exact age and distance from mean age, religion, and political party single features, and the majority age, gender, religion, political party, and never-the-majority features. Finally, in the last row of Table 12.3, the best set of combination and majority features had a 7.8 points improvement in F-score using majority gender, never-the-majority, and exact age, binary age, and gender. In addition it included combination

features: majority <political party, gender>, and single <gender, binary age>, <political party, gender>, <political party, binary age>, and <religion, binary age>. This provides evidence that homophily and social proof are both important in predicting influencers. Finally, as a comparison, we show the best system without using the topic features. We show that excluding topic features causes a reduction in performance (Table 12.3, rows 6 & 7). All results on the development set were statistically significant at less than 0.01 compared to the all-influencers baseline. The best result on the development set of 52.8% F-score had a significance of < 0.05 compared to the number of words baseline.

In the test set, the system using just the best majority features gives 2.4 points improvement in F-score compared to using just the number of words in a sentence (Table 12.3, row 3) using all of the majority features. Ranking was also useful in this system. In row 4 of Table 12.3, we show that the best system using just single features achieves a 3.1 points improvement in F-score compared to using just the number of words in the sentence. This system uses gender, religion, and political party. The best system using single and majority features combined (Table 12.3, row 5) gave an improvement of 4.7 points in F-score overall. These features are the exact age and distance from mean age, and religion single features, and the majority, gender, religion, political party, always-the-majority, and never-the-majority features as well as using ranking. Finally, in the last row of Table 12.3, the best set of combination and majority features had a 5.0 points improvement in F-score using the same features as in the single and majority system in addition to combination features: majority <political party, gender>, and single <religion, gender> and uses ranking. As in the development set, this provides evidence that homophily and social proof are both important in predicting influencers. Finally, as a comparison, we show the best system without using the topic features. We show that excluding topic features causes a reduction in performance (Table 12.3, rows 6 & 7). All results on the test set were statistically significant at less than 0.01 compared to the all-influencers baseline, but not significant in comparison to the number of words baseline.

12.3 Discussion

Our goal in this chapter is not to produce the best system for influence detection, but rather to analyze the impact of social proof in influence detection. Our results show that social proof is important

in being influential. This is indicated by the usefulness of the majority features and a 5.0 boost in F-score using the best group of author trait features.

There are many cases where the predicted author trait may be incorrect. It is interesting to note that although the author trait of a person may be predicted incorrectly, certain tendencies are found in discussions on different issues. This indicates that topic is important. For example, the majority religion in most articles regarding the Catholic Church is predicted to be Christian. If a person is not predicted to be Christian (regardless of the truth of that prediction) it indicates that they are not staying on topic. This is indication that they are not the influencer. Similarly, one would expect that the influencer would usually be Christian.

We believe that the biggest drawback to our author trait predictions in the Wikipedia Talk Pages is due to the limited amount of text available for some people. Roughly half of the participants write less than 100 words within the discussion and we are unlikely to correctly predict their author traits. We included the number of words as a feature to help address this issue. The classifier should use this feature to learn that the author trait features are less reliable when the author has written less. We can improve the author trait predictions by increasing the amount of text for each author. This can be achieved by combining the text written by each person throughout the entire corpus since most authors appear in more than one article. We would like to explore doing this in the future to improve the author trait predictions. We would also like to have the discussions annotated for author traits to analyze the upper bound impact of author traits on influence prediction.

The author trait models are trained on different corpora than Wikipedia and as a result we do not know how accurate the author trait predictions on Wikipedia are. We do find that there are similar trends in our predictions in the Wikipedia training data in comparison to reported statistics of Wikipedia Editor demographics [3]. For example, in a 2013 study it was found that 83% of the Wikipedia editors were male. As shown in Figure 12.1, we find that approximately 75% of the users are predicted to be male. The reported demographics on age indicate that 47% of the users were born before 1980, 27% were born after 1989, and 26% were born between 1980 and 1989. This indicates that there are more old people than young people and that the 50% split occurs somewhere between 1980-1989. Similarly, we find that the majority of users are born before 1982 (See Figure 12.1),

[3]https://en.wikipedia.org/wiki/Wikipedia:Wikipedians#cite _note-UNU-M-6, https://meta.wikimedia.org/wiki/List_of_Wikimedians_ by_religion

indicating they are older and that 1982 is likely a good split for Wikipedia. Finally, the most popular religions of contributors on Wikipedia in 2012 are Christianity (35%), no religion (36%), Judaism (9%), and Islam (6%). In our predictions, we find that Christianity is the most common with Judaism following next. We expect the discrepancy with atheism is because atheism is a subset of no religion. Statistics on the political party of Wikipedia editors could not be found. The relationships between the trends in our training data and the most recent reported statistics are encouraging and indicative of positive labeling of author traits in our dataset. In the future, we would also like to have the discussions annotated for author traits to analyze the upper bound impact of author traits on influence prediction.

Finally, does being in the minority indicate that it will be harder to be influential? For example, as shown, men are more influential than women in this dataset (see Figure 12.1). Does this mean that women have no hope of being influential, particularly in a male dominant setting? On the surface, yes. Women may have to work harder to be influential in a male dominant setting. We, however, do not have to lose hope if we are in the minority! There are many traits and their importance varies across discussions. Gender may not play an important role in some discussions. For example, political party may be more important. In other words, if the majority of people in a political discussion are democrats it would be better to be a female democrat than a male republican. Social proof does, however, indicate that if a person has *nothing* in common with the other participants in the discussion being influential will be nearly impossible. The key then is to find something, no matter how small, that can help one relate to others in a discussion. This connection can then be exploited to become influential.

12.4 Conclusion

We show that there are certain tendencies of influencers to have particular author traits within the Wikipedia Talk Page dataset. These are particularly dependent on the issue being discussed. We also show that influencers tend to be aligned with the majority of the other participants in the conversation. This indicates that social proof is indeed a useful measure for detecting influence. We find that including such features gives at best a 5.0 improvement in F-score in the test set compared to using simply the number of words of each participant in the discussion for an F-score of 46.0%. This

result was significant in comparison to the all-yes baseline. However, this improvement was not significant in comparison to the number of words baseline. This indicates that there is still research to be explored in the effect of social proof on detecting influencers.

In the future, we would like to use the different author traits to help improve each of the individual author trait results. For example, using the predicted age and gender to improve the model for predicting political party. To improve our result in influence detection, we would like to use the content per author across the corpus for author trait prediction at once. When available, the increase in content would allow us to more accurately predict the correct author traits of a person. We would also like to annotate the influencer corpus for gold author trait labels to gain a stronger grasp of the importance of author traits in influence prediction. Finally, we would also like to explore the impact of social proof in the other online genres.

Part IV

Conclusions

Chapter 13

Conclusion

> *It isn't what you have or who you are or where you are or what you are doing that makes you happy or unhappy. It is what you think about it.*
>
> Dale Carnegie, *How to Win Friends and Influence People*

Detecting influence in online conversations is a new line of research that has not been explored by many people. It is something that was difficult to do on a large scale in the past, but with the recent boom in social media, written conversations are easier to attain, allowing this interest problem to be explored. In this thesis, we have created several system components that have been used to successfully detect influence in multiple online genres. In doing so, we have provided several valuable contributions to future research. In the rest of this chapter we discuss our contributions, limitations, and future work in more detail.

13.1 Contributions

In this thesis we have successfully detected influence in multiple online genres. This research has resulted in several valuable contributions. The main contributions are:

13.1.1 Motivation in Social Science

The first contribution of our research is to provide rationale for our system using social science. This is important, because in addition to showing that our system is successful, we also show

that it validates the earliest research in analyzing influencers. We showed this by motivating our system through Robert Cialdini's weapons of influence [Cialdini, 2007] which we described in the introduction in Chapter 1. Each system component in Part II was then motivated by one or more weapons of influence.

13.1.2 System Components

In addition to detecting influence, we also create several components that are stand alone systems. They are Opinion, Claims, Agreement, and Author Traits. All the systems use a supervised approach.

The opinion system detects subjectivity and polarity on the phrase level and is geared towards social media. We have created annotated data from Twitter, LiveJournal, and Wikipedia and compare it against the MPQA [Wiebe *et al.*, 2005]. We performed a cross-genre analysis by training and testing on alternating genres that indicate domain adaptation may be useful to achieve improvements in accuracy. In particular, the MPQA was able to successfully predict sentiment in sentences from social media sources. This is good because it indicates that we can use pre-existing datasets to improve results. Part of our contribution in this chapter is a comprehensive analysis of existing lexicons.

The claim system detects sentences that are opinionated claims that contain belief. It was trained on LiveJournal and Wikipedia sentences. As in opinion, we also provided a cross-genre analysis with encouraging results towards domain adaptation. We show that n-grams and POS are very useful for detecting claims, but sentiment and belief also have an impact.

Our agreement system detects agreement and disagreement between quote and response (Q-R) posts using a large self-labeled create debate corpus (ABCD). We also compare it against the manually labeled Internet Argument Corpus (IAC) [Walker *et al.*, 2012] and our Wikipedia corpus (AWTP). We show that a large out-of-domain self-labeled dataset can do as well or better than a gold dataset. In addition we show that accommodation between the quote and response is important for agreement detection.

Finally, our author trait system is a single method that was used to generate four different models to detect age, gender, political party, and religion. Our training data for age and gender comes from LiveJournal and the Blogger authorship corpus [Schler *et al.*, 2006]. Our political party and religion data comes from Twitter. All of the data was self-labeled, or automatically determined (e.g. gender

based on the person's name). We show that lexical-style (e.g. emoticons, word-lengthening) and lexical-content (e.g. n-grams, POS) are both valuable features.

13.1.3 Comprehensive Analysis of Influencers

We use our system components to create a rich and complex suite of features. Each system component is used to tag the entire discussion for that component and that output is then used to generate features. For example, the claim system tags all the sentences in the discussion as to whether or not they are claims. We use this information to generate features per author such as "did the person have at least one claim", "how many claims did the person have", and "how often was the first sentence in the post a claim". We have similar features for argumentation and persuasion. Some examples of agreement features are "how often did they (dis)agree with others" and "how often did others agree with them". Our author trait system produces the predicted age, gender, political party, and religion for each participant in the discussion. This information is then used to create binary features (e.g. gender of the participant), majority features (e.g. in the majority gender within the discussion), and combination features (e.g age and gender together).

All of the features are then used in a supervised method to detect whether each person in the discussion is or is not an influencer. We have shown that these systems are useful for detecting influence. Furthermore, using the system components to detect influence has a key advantage. In addition to predicting who the influencer is, we have shown that we can also explain *why* they are the influencer. One way that this could be useful is for individuals to improve their influence. For example, showing a person which components they do and don't use to be influential could help them improve how they influence others, and help them adjust to different settings where different approaches may appropriate.

In addition to our broader comprehensive analysis, we provided a detailed analysis of one weapon of influence, social proof, within Wikipedia Talk pages. We show that there are certain tendencies of influencers to have particular author traits within the Wikipedia Talk Page dataset. These are particularly dependent on the issue being discussed. We also show that influencers tend to be aligned with the majority of the other participants in the conversation. This indicates that social proof is indeed a useful measure for detecting influence.

13.1.4 Cross-Genre Analysis

One key advantage of our research over prior work is that we compare how influence differs across multiple online genres. The online genres we explore are comprehensive in the types of social media being explored. They are: LiveJournal (weblogs), Wikipedia Talk Pages (task oriented discussion forum), Political Forum (discussion forum), Create Debate (debate forum), and Twitter (microblog). In Chapter 11, we show that the useful components differ across genre. Overall each component is useful in multiple genres.

Dialog structure, claims, and agreement are all important in Wikipedia. These components point to the desire to make edits in Wikipedia. It is important to be clear and have people agree with you to have influence. It is also important for your changes to be acknowledged. In LiveJournal, dialog structure and agreement are most important. This is because the owner of the blog is often the influencer and people tend to agree with them since it is their blog. Agreement and argumentation are most important in Political Forum. In Political Forum, people discuss political topics which can become very argumentative. Of course, having people agree with your arguments makes it more likely for you to be an influencer. In Create Debate, the agreement, argumentation, and claim components are all important. This is indicative of the debate style. All of these components are important to being convincing in a debate. Finally, in Twitter, dialog structure and agreement are most important. The first person to tweet is usually the influencer, and being retweeted means that agreement is important. Although author traits are not useful individually, they are part of the best system in all genres except for Political Forum. Similarly, persuasion is part of the best system in Twitter, Political Forum, and Wikipedia. Credibility is part of the best system in Wikipedia.

The usefulness of all the components validates our features and further drives home the point Robert Cialdini has made about his weapons of influence: “All the weapons of influence discussed in this book work better under some conditions than under others. If we are to defend ourselves adequately against any such weapon, it is vital that we know its optimal operating conditions in order to recognize when we are most vulnerable to its influence.” [2007]. Knowing that different components work in different situations can be very useful. In addition, we show that domain adaptation and exploiting data from multiple online genres improves performance, particularly in genres with little annotated data. Our best result in Wikipedia is 60.6% F-score using just Wikipedia data, the genre with the most data. Our best result in LiveJournal is 81.6% using all data without

domain adaptation. Our best result in Political Forum is 74.1% using domain adaptation. Our best result in Create Debate is 34.1% using domain adaptation. Our best result in Twitter is 93.8% F-score using all data without domain adaptation. All of these results significantly beat our baselines of predicting everyone to be an influencer and our baseline using the number of words a person has written.

13.1.5 Annotation Manuals and Corpora

Through our research in detecting influence and its components, we have had the opportunity to develop several annotation tools and datasets as described in Chapter 2. In Appendix A, we provide rich annotation manuals as well as several annotation systems using Amazon's Mechanical Turk and stand alone web annotation systems. We have created several automatically labeled datasets as well as manually annotated datasets. We have developed datasets in multiple online genres for opinions, claims, argumentation, persuasion, agreement, author traits (age, gender, religion, and politics), and influence.

13.2 Limitations

In this section we describe the main limitations to our approach in detecting influence.

13.2.1 Data Collection

The biggest challenge in many tasks is gathering and annotating large suitable datasets. The emergence of social media technology has made suitable datasets available for influence detection. However, annotating influence is a complex and time consuming annotation process resulting in a costly effort. Despite this limitation we have been able to develop a rich resource consisting of over 1,000 discussions and 8,700 authors annotated for influence detection across five online genres. However, this resource is still limited in scope. Specifically, some of the data focuses on a variety of discussions, but approximately 1/3 focuses on political discussions. This research could benefit significantly from data in other domains. For example, product related discussions in particular would be useful due to the importance of finding influencers to improve advertising techniques. The most ideal situation would be to find a dataset that could be self-labeled. We have

found this to be beneficial in agreement detection where we were able to use the side a person chose in the debates in Create Debate to determine whether they agreed with the person they were responding to. It could be extremely helpful in influence detection as well. One possibility for doing this may be to use the *change my view* option in a subreddit discussion forum on Reddit (`https://www.reddit.com/r/changemyview`). Having a high change my view score can be indicative of influence.

Another drawback to our data is that we have only annotated which people are influencers per discussion. This is at most two people in the discussion. A more ideal form of annotating influencers in the future would be to rank the participants in the discussion from most influential to least influential. This would improve inter-annotator agreement (Cohen's κ is .53 and .57 due to the complexity of the task) and allow ranking algorithms to be used.

13.2.2 Situational Influence

In the beginning of this thesis we highlighted the two main types of influence: situational and global influence. Situational influence refers to influence by an individual as evident by their participation in the conversation and global influence refers to how influence spreads among a community. The focus of this research is on situational influence. It is important to explore situational influence because a person's influence can vary across discussions. For example, Barack Obama may be influential when it comes to politics, but he is not considered to be influential regarding tennis. Both situational and global influence are important and complementary. Global influence can be explored via social networks as well as through trends across documents over time. Global influence can be useful for advertising strategies as people who are largely connected can be targeted to deliver messages. Global influence is discussed as future work in the following section.

13.2.3 Influence and Causality

The strongest indication of a person being influential is successfully changing the mind of those they are trying to influence. This is known as *causality*. This is a fairly difficult task, mainly because of the challenge of finding suitable data to know that a change actually occurred. Particularly, data over a significant amount of time may be necessary to find change. Furthermore, even if it is clear that a person changed their mind, it is difficult to determine the cause. For example, they may have changed

their mind without being influenced. In this thesis we opted to disregard causality. Rather, we just focused on finding people who display the characteristics of someone who is influential. This is important and very valuable information in of itself. Simply knowing that someone can be influential is useful for targeting people in advertising, political strategies, and defense. Exploring influence and causality is an important future direction to our research which we will discuss in the next section.

13.2.4 Online Genres

In our analysis of influence detection in online genres we explored how influencers differ in five online genres: Wikipedia Talk Pages, LiveJournal weblogs, Political Forum discussions, Create Debate discussions, and Twitter microblogs. This is the most comprehensive analysis of influence detection to date, far surpassing the amount of genres analyzed in prior work (Nguyen *et al.* [2013b] explored two datasets, but did not compare them). However, despite this, there are always more online genres that could be explored. Particularly, we did not focus on discussions on news websites such as the NY times, or comments on YouTube videos. It is important to keep this limitation in mind when making assumptions about how our system would perform on online genres in general, while at the same time realizing that our comprehensive analysis does provide great depth into many types of online discussions.

13.3 Future Directions

In this section we discuss four future directions to our research. The first three are extensions of influence detection. They are: influence trends, influence and causality and influence and social networks. The final direction of future work is broader. It is author profiling. Influence detection can be considered one type of profile of a person. We would like to explore other personalities and profiles of people as well.

13.3.1 Influence Trends

The influencer system detects who the influencers are in an individual thread. We would like to extend the system to track trends over time. Tracking a person over time to examine the trends that occur is extremely difficult for several reasons. It would require a considerable amount of annotation

which is both timely and costly. Second, although the same people may post several times in a single corpus, it would be impossible to track the same person across online genres. Therefore, instead, we could track trends towards a particular *topic* or *entity* over time.

Our goal in such a system would be the ability to query the full corpus on any word or phrase. Each query will return all relevant documents and news articles within the time period requested along with the influencers in each of the online documents. This will give the system the ability to answer several questions related to influence such as:

Who are the influential people on the topic per day?
Which days were most influential?
Which online genre had the most influencers on the topic?
How did the influence of the topic spread across a genre?
Can the trend be related to a major news event?

The system will take advantage of the social network that occurs in each genre. However, it will not be built off of friend's lists but based on the responses and replies in the discussion threads per day. The system could also use news summaries and links that generated the summaries to analyze whether the trend started from a news article or an online genre and how the news articles effected it. This can be achieved through Newsblaster summaries [McKeown *et al.*, 2002] which collects articles daily from popular new sources such as the New York Times, Wall Street Journal, and The Washington Post.

13.3.2 Influence and Causality

As described in the prior section, the strongest indication of a person being influential is successfully changing the mind of those they are trying to influence. This is known as *causality*. Influence and causality have been explored within a social network. Bond *et al* [2012] ran an experiment that sent out political messages to 61 millions Facebook users. They found that these messages directly influenced the voting behavior of the users, the user's friends, and the friends of their friends. We would like to explore whether causality can be seen within influence in a conversation. This is a fairly difficult task, mainly because of the challenge of finding suitable data to know that a change actually

occurred. Particularly, data over a significant amount of time may be necessary to find change.

There are two datasets we would like to explore to see whether causality can be immediately evident without annotation. The first is using Wikipedia Talk Pages. We would like to explore whether we can take advantage of changes that were made to the Wikipedia page following the discussions on the talk page. A change in the Wikipedia Page could be indicative of influence and causality. This of course may not be so straight forward as edits could be unrelated or minor. For example, multiple edits to a page may occur in a small time frame making it difficult to determine which one is associated with the discussion. The other dataset we would like to explore for influence and causality is Create Debate. As a reminder, in the Create Debate discussions the contributors must indicate the side of the debate they are on (*for* or *against*). We hypothesize that if a person in the discussion changes sides at some point in the debate (e.g. from *for* to *against*) this could be indicative of being influenced. This is still difficult because the reason for the change could be unrelated or an error. It may only be possible to do this in a controlled environment where the participants could be asked if, and who, they were influenced by.

13.3.3 Social Network and Context

In the beginning of this thesis we described related work on influence in social networks and influence in conversations. These are known as situational and global influence. The focus of our thesis was on situational influence; detecting influence in a conversation. In contrast, prior work in influence in social networks has looked at how information is diffused from influential people through the social network. These two areas of research do not overlap, but rather complement each other. In the future, we would like to compare the differences and similarities among influential people in the social network and in the conversations. For example, are the people who are influential in the most conversations the ones who are the most influential in the social network? We would also like to develop a system that takes advantage of both forms of influence for improved performance.

13.3.4 Online Genres

As we discussed in the limitations section, there are always more genres that can be explored. In the future, it would be beneficial to develop an addition to the system that can attempt to classify discussions from unseen genres without performing additional annotations and achieving a

considerable loss of information due to the new genres. One method for achieving this may be to use domain adaptation to predict what type of discussions the new discussion is most similar too. This can be either one genre (e.g. LiveJournal), several genres, or a subset of multiple genres. Including this in influence detection will make our system more robust and general.

13.3.5 Author Profiling

There are many other personal characteristics that can be useful in understanding people and society, (e.g. Introversion, Bullying, and Selflessness). Detecting these characteristics about a person is difficult because they change depending on the context, and the personalities tend to be rare. I would like to explore these characteristics as a means of author profiling: how do these characteristics differ among large groups of people? How do they change over time and in different settings? And, how do they affect other people? These characteristics can also be explored across genres and regions. Specifically, translation or multilingual analysis can be used to explore these characteristics in discussions in multiple languages. I will strive to achieve this goal by exploring different discussion forums and web-logging sites, such as Twitter and Wikipedia Talk pages, over a long range of time, ideally several years, and across different topics, such as politics, science, and religion.

Automatically identifying personal character traits, or author profiling, is a prime example of a social science task that was not possible prior to the development of the Internet. In the past, such experiments were done in other fields, such as psychology, on a very small scale. The development of the Internet, and more specifically a means to communicate with others on a personal level, i.e. through weblogs and discussion forums, has given researchers the ability to explore these areas on a larger scale using data science tools to achieve these research goals. In the long term, I would like to explore other sociolinguistic experiments that have been performed on a smaller scale and re-implement these experiments on a larger scale using the abundant amount of data on the web. This will validate past experiments and provide the ability to create tools to conduct these experiments on a variety of data. In addition, I would also like to compare interdisciplinary social science experiments that have been explored across several fields, such as political science, journalism, and psychology. For example, the effect of influence on government action has been studied in political science and sociologists have investigated how influence affects different people. These different experiments can be explored to analyze how they differ, how they are the same, and what useful insights can be

exploited to enhance detection of such experiments in data science.

Part V

Bibliography

Bibliography

[Abbott *et al.*, 2011a] Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on LSM*, LSM '11, pages 2–11, Portland, Oregon, 2011. Association for Computational Linguistics.

[Abbott *et al.*, 2011b] Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 2–11, Portland, Oregon, June 2011. Association for Computational Linguistics.

[Abu-Jbara *et al.*, 2012] Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the ACL*, ACL '12, pages 399–409, Jeju Island, Korea, 2012. Association for Computational Linguistics.

[Adler *et al.*, 2010] B. Thomas Adler, Luca de Alfaro, and Ian Pye. Detecting Wikipedia Vandalism using WikiTrust - Lab Report for PAN at CLEF 2010. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[Agarwal *et al.*, 2009] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 24–32, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Agarwal *et al.*, 2011] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social*

Media (LSM 2011), pages 30–38, Portland, Oregon, June 2011. Association for Computational Linguistics.

[Aharoni *et al.*, 2014] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Anand *et al.*, 2011] Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, , and Philip Resnik. Believe me: We can do this!. annotating persuasive acts in blog text. In *Proceedings of the AAAI-2011 Workshop on Computational Models of Natural Argument*, 2011.

[Andreas *et al.*, 2012] Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[Aral and Walker, 2012] Sinan Aral and Dylan Walker. Identifying Influential and Susceptible Members of Social Networks. *Science*, 337(6092):337–341, July 2012.

[Aral *et al.*, 2009] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[Argamon *et al.*, 2009] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, February 2009.

[Axelrod *et al.*, 2011] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Baccianella *et al.*, 2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[Bakshy *et al.*, 2011] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[Bales *et al.*, 1951] R. F. Bales, Strodtbeck, Mills F. L., T. M., and M. Roseborough. Channels of communication in small groups. *American Sociological Review*, pages 16(4), 461–468, 1951.

[Bales, 1969] R.F. Bales. *Personality and interpersonal behavior*. Holt, Rinehart, and Winston, 1969.

[Bamman *et al.*, 2012] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender in twitter: Styles, stances, and social networks. *CoRR*, abs/1210.4567, 2012.

[Barbieri *et al.*, 2013] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Topic-aware social influence propagation models. *Knowledge and Information Systems*, 37(3):555–584, 2013.

[Barbosa and Feng, 2010] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *COLING (Posters)*, pages 36–44, 2010.

[Becker *et al.*, 2013] Lee Becker, George Erhart, David Skiba, and Valentine Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 333–340, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[Beineke *et al.*, 2004] Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the*

42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, pages 263–270, Barcelona, Spain, July 2004.

[Bender *et al.*, 2011a] E. M. Bender, J. T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, June 2011.

[Bender *et al.*, 2011b] Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on LSM*, LSM '11, pages 48–57, Portland, Oregon, 2011. Association for Computational Linguistics.

[Bender *et al.*, 2011c] Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon, June 2011. Association for Computational Linguistics.

[Bermingham and Smeaton, 2010] Adam Bermingham and Alan F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In Jimmy Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An, editors, *CIKM*, pages 1833–1836. ACM, 2010.

[Biran and Rambow, 2011a] Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168, Washington, DC, USA, 2011. IEEE Computer Society.

[Biran and Rambow, 2011b] Or Biran and Owen Rambow. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing*, 5(4):363–381, 2011.

[Biran *et al.*, 2012] Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. Detecting influencers in written online conversations. In *Proceedings of the Language in Social Media 2012 Workshop*, Montreal, June 2012.

[Bond *et al.*, 2012] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 09 2012.

[Bradley and Lang, 1999] M.M. Bradley and P.J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. C-1, The Center for Research in Psychophysiology, University of Florida., 1999.

[Brehm and Brehm, 1981] S.S. Brehm and J.W. Brehm. *Psychological reactance: a theory of freedom and control*. Academic Press, 1981.

[Brehm, 1989] Jack W. Brehm. Psychological reactance: Theory and applications. *Advances in Consumer Research*, 16(1):72 – 75, 1989.

[Brody and Diakopoulos, 2011] Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollll-lllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[Brook and Ng, 1986] M.E. Brook and S. H. Ng. Language and social influence in small conversational groups. *Journal of Language and Social Psychology*, pages 5(3), 201–210, 1986.

[Burger and Henderson, 2006] John D. Burger and John C. Henderson. An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 15–20. AAAI, 2006.

[Burger *et al.*, 2011] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Callison-Burch and Dredze, 2010] Chris Callison-Burch and Mark Dredze. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 1–12, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Carlson *et al.*, 2003] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers, 2003.

[Chesley *et al.*, 2006] Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 27–29, 2006.

[Cialdini *et al.*, 1976] Robert B. Cialdini, Richard J. Borden, Avril Thorne, Marcus R. Walker, Stephen Freeman, and Lloyd R. Sloan. Basking in Reflected Glory: Three (Football) Field Studies. *Journal of Personality and Social Psychology*, 34(3):366–375, 1976.

[Cialdini, 2007] Robert B. Cialdini. *Influence: The Psychology of Persuasion (Collins Business Essentials)*. Harper Paperbacks, revised edition, January 2007.

[Cohen and Ruths, 2013] Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! In *International AAAI Conference on Weblogs and Social Media*, 2013.

[Conover *et al.*, 2011] M Conover, B Gonçalves, J Ratkiewicz, A Flammini, and F Menczer. Predicting the political alignment of twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*, 2011.

[Danescu-Niculescu-Mizil *et al.*, 2012] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on WWW*, WWW '12, pages 699–708, NYC, USA, 2012. ACM.

[Daumé *et al.*, 2010] Hal Daumé, III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Daumé, 2007] Hal Daumé, III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[Diab *et al.*, 2009] Mona T. Diab, Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. Committed belief annotation and tagging. In *Linguistic Annotation Workshop*, pages 68–73, 2009.

[Dow *et al.*, 2013] P. Alex Dow, Lada A. Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013.

[Drachman *et al.*, 1978] David Drachman, Andre deCarufel, and Chester A Insko. The extra credit effect in interpersonal attraction. *Journal of Experimental Social Psychology*, 14(5):458 – 465, 1978.

[Driscoll *et al.*, 1972] Richard Driscoll, Keith E. Davis, and Milton E. Lipetz. Parental interference and romantic love: The romeo and juliet effect. *Journal of Personality and Social Psychology*, 24(1):1–10, 1972.

[Ennals *et al.*, 2010] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. Highlighting disputed claims on the web. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *WWW*, pages 341–350. ACM, 2010.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.

[Ferrara, 2012] Emilio Ferrara. A large-scale community structure analysis in facebook. *EPJ Data Science*, 1(1), 2012.

[Finkel and Manning, 2009] Jenny Rose Finkel and Christopher D. Manning. Hierarchical bayesian domain adaptation. In *Proceedings of the North American Association of Computational Linguistics (NAACL 2009)*, 2009.

[Freedman and Fraser, 1966] J. L. Freedman and S. C. Fraser. Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology*, 4(2):195–202, August 1966.

[Freedman *et al.*, 2011] Marjorie Freedman, Alex Baron, Vasin Punyakanok, and Ralph M. Weischedel. Language use: What can it tell us? In *ACL (Short Papers)*, pages 341–345. The Association for Computer Linguistics, 2011.

[Galley *et al.*, 2004] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 43nd Annual Meeting of the ACL*, page 669, Barcelona, Spain, 2004. Association for Computational Linguistics.

[Germesin and Wilson, 2009] Sebastian Germesin and Theresa Wilson. Agreement detection in multiparty conversation. In *ICMI*, pages 7–14. ACM, 2009.

[Ghosh *et al.*, 2014] Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Giles *et al.*, 1991] Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation*, pages 1–68. Cambridge University Press, 1991. Cambridge Books Online.

[Gladwell, 2002] Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, January 2002.

[Go *et al.*, 2009] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford, 2009.

[Godbole *et al.*, 2007] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[Goswami *et al.*, 2009] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In *International AAAI Conference on Weblogs and Social Media*, 2009.

[Gottipati *et al.*, 2013] Swapna Gottipati, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. Predicting user's political party using ideological stances. In Adam Jatowt, Ee-Peng Lim, Ying Ding, Asako Miura, Taro Tezuka, Gael Dias, Katsumi Tanaka, Andrew J. Flanagin, and Bing Tian Dai, editors, *SocInfo*, volume 8238 of *Lecture Notes in Computer Science*, pages 177–191. Springer, 2013.

[Goyal *et al.*, 2011] Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. A data-based approach to social influence maximization. *Proc. VLDB Endow.*, 5(1):73–84, September 2011.

[Guo and Diab, 2012] Weiwei Guo and Mona Diab. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 864–872, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[Hahn *et al.*, 2006] Sangyun Hahn, Richard Ladner, and Mari Ostendorf. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of HLT-NAACL*, pages 53–56, NYC, USA, June 2006. Association for Computational Linguistics.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update, 2009.

[Hassan *et al.*, 2012] Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the Joint Conference on EMNLP and CoNLL*, EMNLP-CoNLL '12, pages 59–70, Jeju Island, Korea, 2012. Association for Computational Linguistics.

[Herbrich *et al.*, 2000] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, page 115–132, January 2000.

[Herring and Paolillo, 2006] Susan C. Herring and John C. Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006.

[Hillard *et al.*, 2003] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT-NAACL*, Edmonton, Canada, 2003. ACL.

[Huang *et al.*, 2012] Junming Huang, Xue-Qi Cheng, Hua-Wei Shen, Tao Zhou, and Xiaolong Jin. Exploring social influence via posterior effect of word-of-mouth recommendations. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 573–582, New York, NY, USA, 2012. ACM.

[Iyyer *et al.*, 2014] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Association for Computational Linguistics*, 2014.

[Janin *et al.*, 2003] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The icsi meeting corpus, 2003.

[Jiang *et al.*, 2011] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Katz and Lazarsfeld, 1955] Elihu Katz and Paul F. Lazarsfeld. *Personal influence*. Free Press, Glencoe, IL, 1955. by Elihu Katz and Paul F. Lazarsfeld. With a foreword by Elmo Roper. "A report of the Bureau of Applied Social Research, Columbia University." Bibliography: p. 381-393.

[Kaymaz, 2013] Gozde Kaymaz. Detection of topic-based opinion leaders in microblogging environments. Master's thesis, Boğaziçi Univ., Istanbul, Turkey, 2013.

[Kim and Hovy, 2004] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

[Koppel *et al.*, 2009a] Moshe Koppel, Navot Akiva, Eli Alshech, and Kfir Bar. Automatically classifying documents by ideological and organizational affiliation. In *ISI*, pages 176–178. IEEE, 2009.

[Koppel *et al.*, 2009b] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *JASIST*, 60(1):9–26, 2009.

[Kwak *et al.*, 2010] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[Kwon *et al.*, 2007] Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. Identifying and classifying subjective claims. In *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*, dg.o '07, pages 76–81. Digital Government Society of North America, 2007.

[La Fond and Neville, 2010] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 601–610, New York, NY, USA, 2010. ACM.

[Langer, 1989] E. Langer. Minding Matters The consequences of mindlessness-mindfulness. *Advances in Experimental Social psychology*, 22:137–173, 1989.

[Lin *et al.*, 2006] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on CoNLL*, CoNLL-X '06, pages 109–116, NYC, USA, 2006. Association for Computational Linguistics.

[MacKinnon and Warren, 2007] Ian MacKinnon and Robert H. Warren. Age and geographic inferences of the livejournal social network. In Edoardo Airoldi, David M. Blei, Stephen E. Fienberg, Anna Goldenberg, Eric P. Xing, and Alice X. Zheng, editors, *Statistical Network Analysis: Models,*

Issues, and New Directions, volume 4503 of *Lecture Notes in Computer Science*, pages 176–178. Springer Berlin Heidelberg, 2007.

[Mann and Thompson, 1988] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[Marin *et al.*, 2011] Alex Marin, Bin Zhang, and Mari Ostendorf. Detecting forum authority claims in online discussions. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 39–47, Portland, Oregon, June 2011. Association for Computational Linguistics.

[McCallum, 2002] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet, 2002.

[Mccowan *et al.*, 2005] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus. In *Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research.*, 2005.

[McKeown *et al.*, 2002] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 280–285, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[Mei *et al.*, 2007] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of WWW*, pages 171–180, New York, NY, USA, 2007. ACM Press.

[Mejova and Srinivasan, 2012] Yelena Mejova and Padmini Srinivasan. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter. In *ICWSM*, 2012.

[Mishne, 2005] Gilad Mishne. Experiments with mood classification in blog posts. In *1st Workshop on Stylistic Analysis Of Text For Information Access*, 2005.

[Misra and Walker, 2013] Amita Misra and Marilyn Walker. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France, August 2013. Association for Computational Linguistics.

[Miura *et al.*, 2014] Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[Mohammad *et al.*, 2013] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[Moore and Lewis, 2010] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Mukherjee and Liu, 2010] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Mukherjee and Liu, 2012] Arjun Mukherjee and Bing Liu. Analysis of linguistic style accommodation in online debates. In *Proceedings of COLING 2012*, pages 1831–1846, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[Mukherjee and Liu, 2013] Arjun Mukherjee and Bing Liu. Discovering user interactions in ideological discussions. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 671–681, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[Myers *et al.*, 2012] Seth A. Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 33–41, New York, NY, USA, 2012. ACM.

[Nakov *et al.*, 2013] Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint*

*Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013.

[Nasukawa and Yi, 2003] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 70–77, New York, NY, USA, 2003. ACM.

[Ng *et al.*, 1993] S. H. Ng, D. Bell, and M. Brooke. Gaining turns and achieving high in influence ranking in small conversational groups. *British Journal of Social Psychology*, pages 32, 265–275, 1993.

[Ng *et al.*, 1995] S. H. Ng, M Brooke, and M. Dunne. Interruption and in influence in discussion groups. *Journal of Language and Social Psychology*, pages 14(4),369–381, 1995.

[Nguyen and Lim, 2014] Minh-Thap Nguyen and Ee-Peng Lim. On predicting religion labels in microblogging networks. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1211–1214, New York, NY, USA, 2014. ACM.

[Nguyen *et al.*, 2011] Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Nguyen *et al.*, 2013a] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. " how old do you think i am?" a study of language and age in twitter. In *ICWSM*, 2013.

[Nguyen *et al.*, 2013b] Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah Cai, Jennifer Midberry, and Yuanxin Wang. Modeling topic control to detect influence in conversations using nonparametric topic models. In *Machine Learning*, pages 1–41. Springer, 2013.

[Nielsen, 2011] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[Noreen, 1989] Eric W. Noreen. *Computer-Intensive Methods for Testing Hypotheses : An Introduction*. Wiley-Interscience, April 1989.

[Nowson and Oberlander, 2006] Scott Nowson and Jon Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI*. American Association for Artificial Intelligence (www.aaai.org), 2006.

[Opitz and Zirn, 2013] Bernd Opitz and Cecilia Zirn. Bootstrapping an unsupervised approach for classifying agreement and disagreement. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Linköping Univ. Electronic Press, 2013.

[Owoputi *et al.*, 2013] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of NAACL*, 2013.

[Pak and Paroubek, 2010] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[Palau and Moens, 2009] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA, 2009. ACM.

[Pang and Lee, 2004] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[Park and Cardie, 2014] Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Phan, 2006a] Xuan-Hieu Phan. Crfchunker: Crf english phrase chunker, 2006.

[Phan, 2006b] Xuan-Hieu Phan. Crftagger: Crf english tagger, 2006.

[Phillips and Carstensen, 1988] David P. Phillips and Lundie L. Carstensen. The effect of suicide stories on various demographic groups, 1968–1985. *Suicide and Life-Threatening Behavior*, 18(1):100–114, 1988.

[Plank and van Noord, 2011] Barbara Plank and Gertjan van Noord. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[Prabhakaran and Rambow, 2013] Vinodkumar Prabhakaran and Owen Rambow. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 216–224, 2013.

[Prabhakaran and Rambow, 2014] Vinodkumar Prabhakaran and Owen Rambow. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 339–344, 2014.

[Prabhakaran *et al.*, 2010] Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1014–1022, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Prabhakaran *et al.*, 2014a] Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1481–1486, 2014.

[Prabhakaran *et al.*, 2014b] Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014,*

October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1965–1976, 2014.

[Quercia *et al.*, 2011] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. In the mood for being influential on twitter. In *SocialCom/PASSAT*, pages 307–314. IEEE, 2011.

[Ramshaw and Weischedel, 2005] L.A. Ramshaw and R.M. Weischedel. Information extraction. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 5, pages v/969–v/972 Vol. 5, March 2005.

[Rangel *et al.*, 2013] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, Sheffield, UK, 2013.

[Rangel *et al.*, 2014] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, Sheffield, UK, 2014/09/18 2014.

[Rao and Georgeff, 1998] Anand S. Rao and Michael P. Georgeff. *Readings in Agents*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[Rao *et al.*, 2010] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[Read, 2005] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Reid and Ng, 2000] Scott A. Reid and Sik Hung Ng. Conversation as a resource for in influence: evidence for prototypical arguments and social identification processes. *European Journal of Social Psychology*, pages 30, 83–100, 2000.

[Rienks, 2007] Rutger Joeri Rienks. *Meetings in smart environments : implications of progressing technology*. PhD thesis, University of Twente, Enschede, the Netherlands, July 2007.

[Rosenthal and McKeown, 2011] Sara Rosenthal and Kathleen McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *proceedings of ACL-HLT*, 2011.

[Rosenthal and McKeown, 2012] Sara Rosenthal and Kathleen McKeown. Detecting opinionated claims in online discussions. In *Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing Special Session on Semantics and Sociolinguistics in Social Media*, ICSC '12. IEEE Computer Society, 2012.

[Rosenthal and McKeown, 2013] Sara Rosenthal and Kathy McKeown. Columbia nlp: Sentiment detection of subjective phrases in social media. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 478–482, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[Rosenthal *et al.*, 2013] Sara Rosenthal, Gregory J. Barber, and Kathleen McKeown. Columbia nlp: Sentiment slot filling. In *Text Analysis Conference Sentiment Slot Filling Workshop*, Gaithersburg, Maryland, USA, 2013. National Institute of Standards and Technology.

[Rosenthal *et al.*, 2014a] Sara Rosenthal, Suvarna Bothe, and Kathleen McKeown. Columbia nlp: Sentiment slot filling. In *Text Analysis Conference Sentiment Slot Filling Workshop*, Gaithersburg, Maryland, USA, 2014. National Institute of Standards and Technology.

[Rosenthal *et al.*, 2014b] Sara Rosenthal, Kathy McKeown, and Apoorv Agarwal. Columbia nlp: Sentiment detection of sentences and subjective phrases in social media. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 198–202, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[Rosenthal *et al.*, 2014c] Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August 2014.

[Rosenthal *et al.*, 2015] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015, Denver, Colorado, June 2015. Association for Computational Linguistics.

[Scherer, 1979] K. R. Scherer. Voice and speech correlates of perceived social influence in simulated juries. In *H. Giles and R. St Clair (Eds), Language and social psychology*, pages 88–120. Oxford: Blackwell, 1979.

[Schler *et al.*, 2006] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[Schneider *et al.*, 2012] Jodi Schneider, Tudor Groza, and Alexandre Passant. A review of argumentation for the social semantic web. *Semantic Web – Interoperability, Usability, Applicability*, 2012.

[Schneider, 2014] Jodi Schneider. Automated argumentation mining to the rescue? envisioning argumentation and decision-making support for debates in open online collaboration communities. In *Proceedings of the First Workshop on Argumentation Mining*, pages 59–63, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Smadja, 1993] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177, 1993.

[Snow *et al.*, 2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[Socher *et al.*, 2013] Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 455–465, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[Stab *et al.*, 2014] Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Frontiers and Connections between Argumentation Theory and Natural Language Processing*, page (to appear), July 2014.

[Stolcke *et al.*, 2000] Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26(3):339–373, September 2000.

[Stone *et al.*, 1966] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[Strzalkowski *et al.*, 2013] Tomek Strzalkowski, Samira Shaikh, Ting Liu, George Aaron Broadwell, Jennifer Stromer-Galley, Sarah M. Taylor, Veena Ravishankar, Umit Boz, and Xiaoai Ren. Influence and power in group interactions. In *SBP*, pages 19–27, 2013.

[Swayamdipta and Rambow, 2012] Swabha Swayamdipta and Owen Rambow. The pursuit of power and its manifestation in written dialog. In *ICSC*, pages 22–29. IEEE Computer Society, 2012.

[Tam and Martell, 2009] Jenny Tam and Craig H. Martell. Age detection in chat. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, ICSC '09, pages 33–39, Washington, DC, USA, 2009. IEEE Computer Society.

[Tausczik and Pennebaker, 2010] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Language and Social Pyschology*, 2010.

[Thelwall *et al.*, 2010] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text, 2010.

[Thomas *et al.*, 2006] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 327–335, 2006.

[Toutanova *et al.*, 2003] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, NAACL '03, pages 173–180, Edmonton, Canada, 2003. Association for Computational Linguistics.

[Travers *et al.*, 1969] Jeffrey Travers, Stanley Milgram, Jeffrey Travers, and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[Turney, 2002] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Urmann, 2009] David H. Urmann. The history of text messaging, 2009.

[Walker *et al.*, 2012] Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of the Eight International Conference on LREC (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[Wang and Cardie, 2014] Lu Wang and Claire Cardie. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Wang *et al.*, 2011a] Wen Wang, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. Identifying agreement/disagreement in conversational speech: A cross-lingual study. In *INTERSPEECH*, pages 3093–3096. ISCA, 2011.

[Wang *et al.*, 2011b] Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 374–378. Association for Computational Linguistics, 2011.

[Warriner *et al.*, 2013] AmyBeth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

[Watts and Dodds., 2007] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34:441–458, 2007.

[Whissel, 1989] C. M. Whissel. The dictionary of affect in language. In *R. Plutchik and H. Kellerman, editors, Emotion: theory research and experience*, volume 4, London, 1989. Acad. Press.

[Wiebe *et al.*, 2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, 2005.

[Wilson *et al.*, 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Wu *et al.*, 2011] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM.

[Yan and Yan, 2006] Xiang Yan and Ling Yan. Gender classification of weblog authors. In *AAAI Spring Symposium Series on Computation Approaches to Analyzing Weblogs*, pages 228–230, 2006.

[Yang *et al.*, 2007] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Emotion classification using web blog corpora. In *Web Intelligence*, pages 275–278, 2007.

[Yeh, 2000] Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[Yin *et al.*, 2012] Jie Yin, Paul Thomas, Nalin Narang, and Cecile Paris. Unifying local and global agreement and disagreement classification in online debates. In *Proceedings of the 3rd Workshop*

in WASSA, WASSA '12, pages 61–69, Jeju Island, Korea, 2012. Association for Computational Linguistics.

[Young *et al.*, 2011] Joel Young, Craig Martell, Pranav Anand, Pedro Ortiz, and IV Henry Gilbert. A microtext corpus for persuasion detection in dialog. In *AAAI*, 2011.

[Yu and Hatzivassiloglou, 2003] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Yu and Kübler, 2011] Ning Yu and Sandra Kübler. Filling the gap: semi-supervised learning for opinion detection across domains. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 200–209, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Zamal *et al.*, 2012] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[Zickuhr, 2010] Kathryn Zickuhr. Generations 2010. In *Pew Research Center*. Pew Research Center, 2010.

Part VI

Appendices

Appendix A

Annotation Manuals

This appendix describes the annotation manuals for the components where human annotation was used.

A.1 Influence

The annotators were provided with the following guidelines for annotating the influencers in a discussion developed by Or Biran with slight modifications. They were also given three discussion to practice annotating.

A.1.1 What are we looking for?

We are looking for participants who are influencers in online discussion threads. Currently, we will look at threads from the following sources: 1. Wikipedia Discussions (Talk pages) 2. LiveJournal We are not sure how appropriate either source is for finding influencers; it may be that there are very few influencers in the threads we give you.

A.1.2 Annotation Instructions

In the Excel spreadsheet named influencer.xls, add a single row for each influencer. The columns of the file should be: source, filename, participant, evidence, confidence. The (empty) spreadsheet is in your folder. You can use either MS Excel or Open Office to edit it. All decisions for all threads should go in the same spreadsheet.

The definitions of the columns are:

- **source** The source of the thread, e.g. LiveJournal
- **filename** The file name for the thread
- **participant** The name of the influencer as it appears in the thread
- **evidence** a short description of why you think the participant is an influencer
- **confidence** one of three values: HIGH, MEDIUM, or LOW. This is a measure of your confidence that the participant you chose really is an influencer based on our rather vague definition.

If there are no influencers in the thread, place ”none” as the participant. If there are multiple influencers, add multiple rows with the same filename value.

A.1.3 Who is an influencer?

Good question. We want to hear back from you about any insights you have. We do not have a very clear idea of what exactly it means to be an influencer, but in general, an influencer is someone who has credibility in the group and whose opinions or ideas are respected and in some way influence the discussion. We also have the bulleted definition, according to which an influencer:

1. Has credibility in the group
2. Persists in attempting to convince others, even if some disagreement occurs.
3. Introduces topics/ideas that others pick up on or support.
4. Must be a group participant but need not be active in the discussion(s) where others support/credit him. And also:
5. Influencer’s ideas or language may be adopted by others
6. others may explicitly recognize influencer’s authority. Who is not necessarily an influencer?

A.1.4 Who is not an Influencer?

We do have some idea of what an influencer is not. There are other types of power, other than influence, which a participant in a discussion may have, which are described below. IMPORTANT: a participant may be an influencer and also one of the other types we describe here; If someone has hierarchical power, for example, they may still have influence; but having hierarchical power does not automatically mean the participant is an influencer.

A.1.4.1 Hierarchical power:

person_1 appears to be above person_2 in a hierarchy.

For example:

1. person_1 appears to give an approval or a direct order
2. another person appears to be asking person_1 for approval
3. person_1 appears to have authority to make the final decision

In Wikipedia Discussions, for example, there may be moderators who have hierarchical power over other editors.

A.1.4.2 Situational power:

person_1 appears to have power (authority to direct and/or approve other people's actions) in the current situation or while a particular task is being performed. In Wikipedia Discussions, this may be a person who is very active in editing the particular article.

A.1.4.3 Power directing the communication:

person_1 actively attempts to achieve the goal of the discussion and directs communication towards that goal. In Wikipedia Discussions this may be the person who started the topic thread.

A.2 Sentiment

In this section we describe the annotation tool and guidelines for sentiment annotation [Nakov *et al.*, 2013; Rosenthal *et al.*, 2014c]. The instructions provided to the annotators, along with an example,

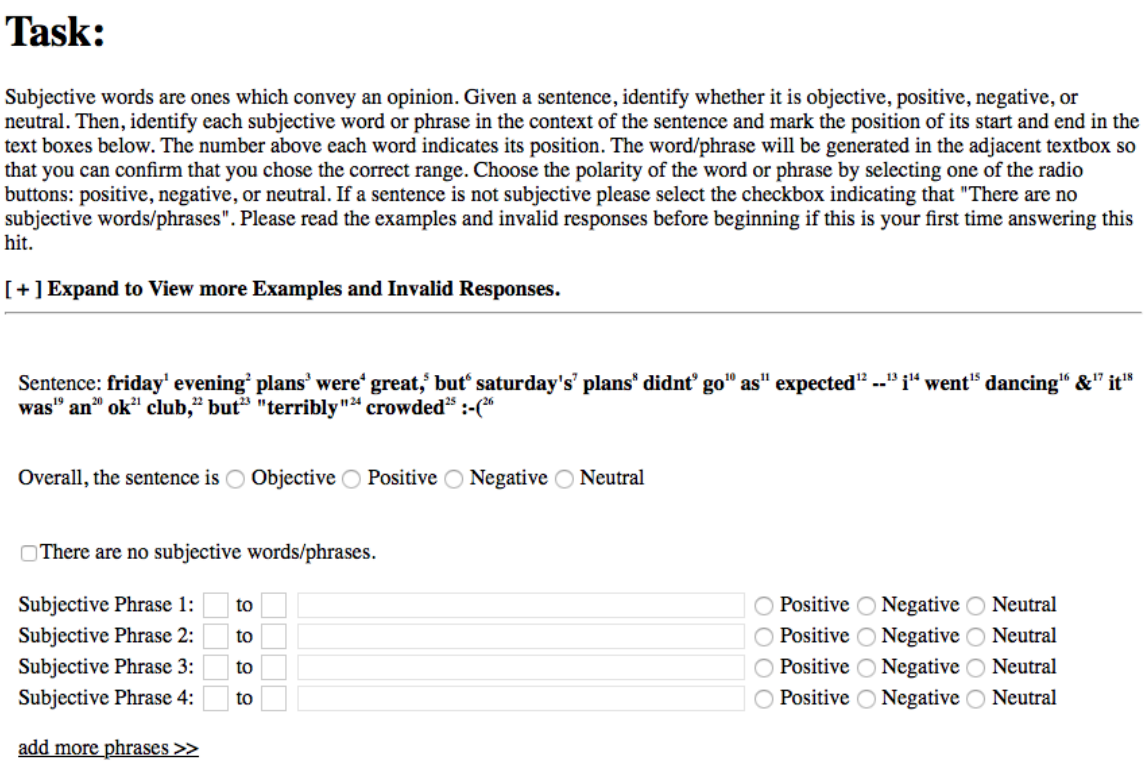


Figure A.1: Sentiment Annotation Screenshot

are shown in the screenshot in Figure A.1. We provided several additional examples to the annotators, shown in Table A.1.

In addition, we filtered spammers by considering the following kinds of annotations invalid:

- containing overlapping subjective phrases;
- subjective, but not having a subjective phrase;
- marking every single word as subjective;
- not having the overall sentiment marked.

A.2.1 Annotation Process

Our datasets were annotated for sentiment on Mechanical Turk. Each sentence was annotated by five Mechanical Turk workers (Turkers). In order to qualify for the hits, the Turker had to have an approval rate greater than 95% and have completed 50 approved hits. Each Turker was paid

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
”March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*,” according to Chen, also Taiwan’s chief WTO negotiator.
friday evening plans were great, but saturday’s plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: ”I’d like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

Table A.1: List of example sentences with annotations that were provided to the annotators. All subjective phrases are italicized. Positive phrases are in green, negative phrases are in red, and neutral phrases are in blue.

three-five cents per hit. The Turker had to mark all the subjective words/phrases in the sentence and say whether each subjective word/phrase was positive, negative or neutral. They also had to indicate the overall polarity of the sentence.

Figure A.1 shows the instructions and an example provided to the Turkers. The first five rows of Table A.2 show an example of the subjective words/phrases marked by each of the workers as positive, negative or neutral.

We combined the annotations of each of the workers using intersection as indicated in the last row of Table A.2. A word had to appear in 2/3 of the annotations in order to be considered subjective. Similarly, a word had to be labeled with a particular polarity (positive, negative, or neutral) 2/3 of the time in order to receive that label. We also experimented with combining annotations by computing the union of the sentences, and taking the sentence of the worker who annotated the most hits, but we found that these methods were not as accurate. The polarity of the entire sentence was determined

Worker 1	***I would love*** to watch Vampire Diaries ***:)*** and some Heroes! ***Great combination***	9/13
Worker 2	I would love to watch Vampire Diaries :) and some ***Heroes! Great*** combination	11/13
Worker 3	I ***would love*** to watch Vampire Diaries ***:)*** and some Heroes! ***Great*** combination	10/13
Worker 4	I would ***love*** to watch Vampire Diaries :) and some Heroes! ***Great*** combination	13/13
Worker 5	I would love to watch Vampire Diaries :) and some Heroes! ***Great*** combination	11/13
Intersection	I would ***love*** to watch Vampire Diaries :) and some Heroes! ***Great*** combination	

Table A.2: Example of a sentence annotated for subjectivity on Mechanical Turk. Words and phrases that were marked as subjective are italicized and highlighted in bold. The first five rows are annotations provided by Turkers, and the final row shows their intersection. The final column shows the accuracy for each annotation compared to the intersection.

based on the majority of the labels. If there was a tie, the sentence was discarded. In order to reduce the number of sentences lost, we combined objective and neutral labels, which Turkers tended to mix up.

A.3 Agreement

This section describes that annotation guidelines and tool for agreement on the sentence level as described in our previous work [Andreas *et al.*, 2012].

A.3.1 Annotation Guidelines

A thread consists of a set of posts organized in a tree. We use standard terminology to refer to the structure of this tree (so every post has a single "parent" to which it replies, and all nodes descend from a single "root"). Each post is marked with a timestamp and an author, a string (its "body").

Agreement annotation is performed on pairs of sentences $\{s, t\}$, where each sentence is a substring of the body of a post. s is referred to as the "antecedent sentence", and t as the "reaction sentence."

The antecedent sentence and reaction sentence occur in different posts written by different authors. Annotations are implicitly directed from reaction to antecedent; the reaction is always the sentence from the post with the later timestamp. Annotations between pairs of posts with the same

author are forbidden. Each pair is also annotated with a type.

Type

Each sentence pair can be of either type agreement or disagreement. Two sentences are in agreement when they provide evidence that their authors believe the same fact or opinion, and in disagreement otherwise.

Mode

Mode indicates the manner in which agreement or disagreement is expressed. Broadly, a pair of posts are in a "direct" relationship if one is an ancestor of the other, and indirect otherwise; they are in a "response" relationship if one explicitly acknowledges a claim made in the other, and "paraphrase" otherwise. More specifically:

Direct response: The reaction author explicitly states that they are in agreement or disagreement, e.g. by saying "I agree" or "No, that's not true." An agreement/disagreement is only a direct response if it is a direct reply to its closest ancestor, i.e. its parent. For example, in Table A.4, the reaction sentence *"I knoooooow."* in c_3 is a direct response to the antecedent sentence *"That jacket is gorgeous."* in c_2. In Table A.3, the reaction sentence *"Arcadian, I added only the lymphoma portion of the WHO classification."* in c_6 is a direct disagreement to the sentence *"Emmanuelm, aren't you the person who added those other categories on 6 July 2005?"* in c_5.

Direct paraphrase: The reaction author restates a claim made in an ancestor post. An agreement/disagreement is only a direct paraphrase if it is a direct rewording of its closest ancestor, i.e. its parent. For example, in Table A.4, the sentence *"It is a bit raggedy looking."* in c_4 is a direct paraphrase of the sentence *"but it's so, I don't know – raggedy looking!"* of its parent, c_1.

Indirect response: The reaction is a direct response to a claim, but the post does not descend from the source. This often occurs when the author pressed the "reply" button on a post other than the one they were attempting to respond to (this would be the case if, for example, c_3 descended from c_5 instead of c_2 above). Or, perhaps it is intended to answer more than one

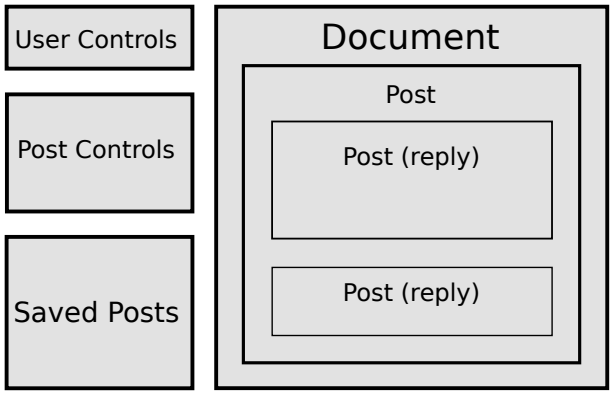


Figure A.2: Schematic of the annotation tool: The left side shows the controls used for navigation and the right displays the current thread.

previous post. The reaction of an indirect response should be the single sentence written closest in time to its antecedent.

Indirect paraphrase: The reaction restates a claim made in a post that is earlier in time, but not an ancestor of the post. The reaction of an indirect paraphrase annotation be the single sentence written closest in time to its antecedent. For example, in Table A.4, c_5 is an indirect paraphrase of c_2.

A.3.2 The Annotation Process

In recent years there has been an increasingly popular trend to use Amazon's Mechanical Turk to label data. Studies have shown [Snow *et al.*, 2008] that Mechanical Turk users are able to produce high quality data that is comparable to expert annotators for simple labeling tasks that can be completed in a few seconds such as affective text analysis and word sense disambiguation. However, others have shown [Callison-Burch and Dredze, 2010] that the annotations are considerably less reliable for tasks requiring a substantial amount of reading or involving complicated annotation schemes. Our annotation task was difficult for several reasons; observation of the entire thread was necessary to annotate each edge and the annotations themselves were fairly involved. Therefore, we decided to rely on two trained annotators rather than a large number of untrained annotators.

Both annotators were undergraduates, and neither had any previous experience with NLP annota-

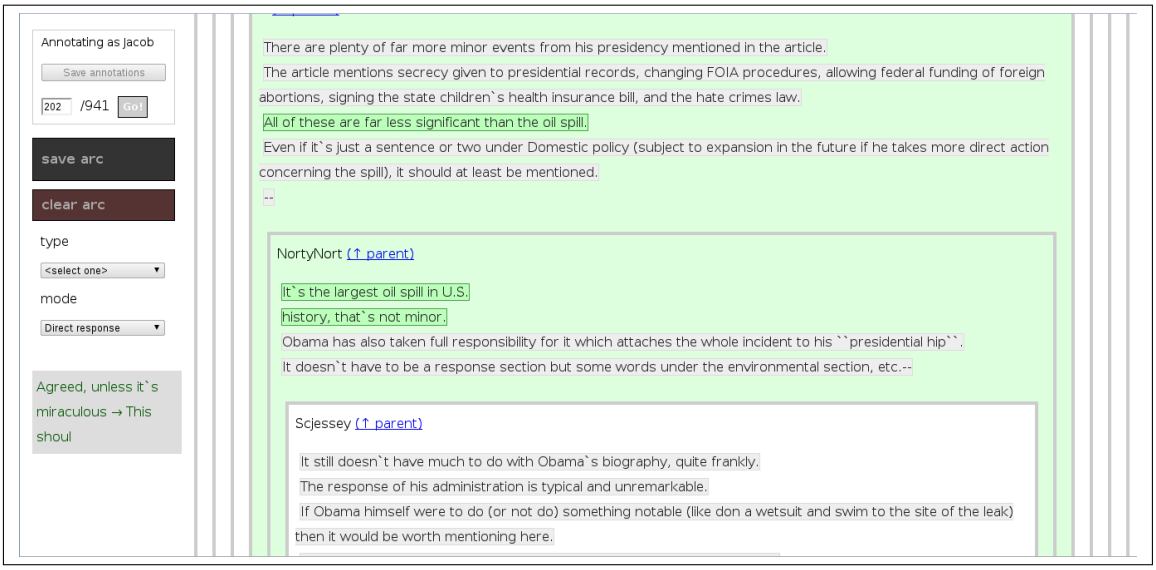


Figure A.3: Screenshot of the annotation tool in use.

tion tasks. They were trained to use the web-based annotation tool described in the following section for approximately one hour; they then annotated the remainder of the corpus on their own.

A.3.3 The Annotation Tool

The web-based annotation tool (Fig. A.2 and A.3) is used to provide a simple and easy way to annotate threads. The interface consists of two parts; the left hand-side which contains controls to navigate through threads and add agreements, and the right-hand side which displays the current thread.

The document is displayed using its thread structure indicated by both indentation and boxes nesting children under their parents as shown in Figure A.2. To clearly differentiate between possible sentences, each sentence is displayed on its own line.

Annotators begin by selecting the individual sentences from the antecedent and reaction which provide evidence of agreement or disagreement, and then mark type and mode using the post controls on the left-hand side. (While the response/paraphrase annotation must be encoded by hand, the direct/indirect distinction is inferred automatically from the document structure.) Each post pair that is added appears as a saved post on the left-hand side of the tool directly below the post controls. Saved posts can be removed if they were mistakenly added. Figure A.3 shows the annotation tool in

1. Did you find the tool easy to use?
2. What challenges did you encounter when using the tool?
3. Were the LiveJournal or Wikipedia discussions easier to annotate?
4. Were the LiveJournal or Wikipedia discussions faster to annotate?
5. Did you have any previous experience with annotation?
6. What was the learning curve associated with the task?
7. On average, how long did it take you to complete a single LiveJournal discussion? A Wikipedia discussion?
8. Was it easier to find agreements or disagreements?
9. Was it easier to find direct or indirect agreements/disagreements?
10. What is your general opinion about the task?
11. Is there anything else that you would like to let us know about with regard to this annotation task?

Figure A.4: Annotator questionnaire

use.

The system automatically prevents users from annotating the forbidden cases mentioned in Section A.3.1, such as the antecedent and reaction sentences being written by the same author. It also automatically determines the antecedent and reaction of the annotation based on the timestamps of the two posts involved.

The annotation tool outputs a CSV file (Table A.5) encoding the annotator ID, the document ID and the post structure as a JSON array with entries for each annotated arc.

A.4 User Studies

After completing their portion of the task, annotators were asked to fill out a brief questionnaire describing their experience (Fig. A.4). They reported that the annotation tool was “easy to use” and “effective”, that the annotation task was “interesting”, and that there were “no real challenges” in annotating. They reported that between the two genres, the LiveJournal entries were both

conceptually easier to annotate and required less time, primarily because the posts were shorter in length. Annotators reported that LiveJournal entries required an average 2 to 10 minutes to annotate, while Wikipedia discussions required 10 to 20 minutes. Annotators were divided in their opinions on whether agreements or disagreements, and direct or indirect were easier to identify indicating that it is a matter of personal preference.

These responses have given us confidence that the annotation tool succeeded in its purpose (of simplifying the data collection process for this corpus), and that it will be easy to further expand the corpus if we require additional data.

$\mathbf{c}_1$ There seems to be a much better list at the National Cancer Institute than the one we've got. It ties much better to the actual publication (the same 11 sections, in the same order). I'd like to replace that section in this article. Any objections?
$\mathbf{c}_2$ Not a problem. Perhaps we can also insert the relative incidence as published in this month's wiki Blood journal
$\mathbf{c}_3$ I've made the update. I've included template links to a source that supports looking up information by ICD-O code.
$\mathbf{c}_4$ Can Arcadian tell me why he/she included the leukemia classification to this lymphoma page? It is not even listed in the Wikipedia leukemia page! I vote for dividing the WHO classification into 4 parts in 4 distinct pages: leukemia, lymphoma, histocytic and mastocytic neoplasms. Let me know what you think before I delete them.
$\mathbf{c}_5$ Emmanuelm, aren't you the person who added those other categories on 6 July 2005?
$\mathbf{c}_6$ Arcadian, I added only the lymphoma portion of the WHO classification. You added the leukemias on Dec 29th. Would you mind moving the leukemia portion to the leukemia page
$\mathbf{c}_7$ Oh, and please note that I would be very comfortable with a "cross-coverage" of lymphocytic leukemias in both pages. My comment is really about myeloid, histiocytic and mast cell neoplasms who share no real relationship with lymphomas.
$\mathbf{c}_8$ To simplify the discussion, I have restored that section to your version. You may make any further edits, and I will have no objection.

Table A.3: Examples of agreement and disagreement in a Wikipedia discussion forum. Direct Response: $c_2 \rightarrow c_1$, $c_6 \rightarrow c_5$, $c_8 \rightarrow c_7$,c_6

$\mathbf{c}_1$ I want this jacket. Because 100% silk is so practical, especially in a household with cats. but it's so, I don't know – raggedy looking! That's awesome!
$\mathbf{c}_2$ That jacket is gorgeous. Impractical, way too expensive for the look, and pretty much gorgeous. guh.
$\mathbf{c}_3$ I knooooooow, and you're not helping. :)
$\mathbf{c}_4$ Monday! WHEE! It is a bit raggedy looking. I think it's because of the ties.
$\mathbf{c}_5$ Wow, that jacket looks really nice... I wish I could afford it!

Table A.4: Examples of a agreement in a LiveJournal weblog. Direct Response: $c_2 \rightarrow c_1$, $c_5 \rightarrow c_1$, Direct Paraphrase: $c_4 \rightarrow c_1$, Indirect Paraphrase: $c_5 \rightarrow c_2$

Field	Value
Document ID	5
Annotator	John Doe
Antecedent ID	13
Reaction ID	11
Antecedent	It does seem heavily censored
Reaction	Um, I can't help but notice that this article seems heavily censored.
Type	agreement
Mode	indirect paraphrase

Table A.5: Sample annotator output

Appendix B

Corpora

This appendix describes the corpora and code associated with this thesis. All are publicly available at `www.cs.columbia.edu/~sara/data.php` or upon request via e-mail unless other wise noted.

B.1 Influence

The influence system is written in Java. It consists of modules for each components. Modules can easily be added or removed. The code has the capabilities to perform all experiments in this thesis, including domain adaptation and cross genre experiments. It requires Weka and jar files for all the components as described below.

The influencer data comes from five sources: LiveJournal, Wikipedia, Create Debate, Political Forum, and Twitter. The discussions are provided in the LiveJournal xml format. Each of the datasets has a csv file with the list of influencers associated with it. The data from each source is broken into training, development (optional), and testing. The data is available upon request.

B.2 Sentiment

Our sentiment system is written in Java. It requires the DAL [Whissel, 1989], Weka [Hall *et al.*, 2009], CRF POS Tagger and Chunker [Phan, 2006a; Phan, 2006b], WordNet [Fellbaum, 1998], SentiWordNet [Baccianella *et al.*, 2010] (optional), and Wiktionary (optional) to be installed.

The sentiment data is annotated on the phrase level, and in most cases sentence level. It comes from four sources:

MPQA [Wiebe *et al.*, 2005] `http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/`

Twitter Semeval Twitter data: `http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools`

LiveJournal Part of semeval data: `http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools`

Wikipedia Available via download of claim data: `http://www.cs.columbia.edu/~sara/download/claims.tar.gz`

B.3 Agreement

The agreement system is written in Java. It requires Mallet [McCallum, 2002], the CRF POS Tagger [Phan, 2006b], Sentence Similarity [Guo and Diab, 2012] (optional), and the sentiment system.

The agreement data is from three sources:

ABCD The Agreement by Create Debaters Corpus is available upon request in the LiveJournal XML format. Annotations are between Q-R pairs .

IAC [Walker *et al.*, 2012] The Internet Argument Corpus is available in its original format at `https://nlds.soe.ucsc.edu/iac`. Available upon request with post level annotations in LiveJournal XML format.

AWTP The Agreement in Wikipedia Talk Pages Corpus is available on the sentence level at `http://www.cs.columbia.edu/~sara/download/agreement_annotations.tar.gz`. It is available upon request in the LiveJournal XML format.

B.4 Claim

The claim system requires Weka [Hall *et al.*, 2009]. the sentiment system, and committed belief [Prabhakaran *et al.*, 2010; Diab *et al.*, 2009] (optional).

The claim dataset consists of sentences annotated for claim in LiveJournal and Wikipedia. It also is accompanied with the sentiment labels for the majority of the sentences. It is available for download at `http://www.cs.columbia.edu/~sara/download/claims.tar.gz`.

B.5 Argumentation

The argumentation code and data can be requested via e-mail from Or Biran.

B.6 Author Traits

Our author trait system is written in Java. It requires Weka [Hall *et al.*, 2009], Xtract [Smadja, 1993], and Stanford Core NLP [Klein and Manning, 2003].

Our author trait data comes from three sources:

LiveJournal Weblogs labeled with age and gender. Available upon request in the LiveJournal XML format.

Blog Authorship Corpus Weblogs labeled with age and gender. Downloadable at `http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm`. Available upon request in the LiveJournal XML format.

Twitter Twitter users labeled with religion and political party. User Ids are available upon request. Code to download and convert to the LiveJournal XML format are available upon request.

B.7 Xtract

Xtract is the collocation tool developed by Frank Smadja [1993] that we have used in Author Trait Detection. We have reimplemented it in Java. It is available upon request.

Appendix C

Glossary of Terms

accuracy A statistical measure that indicates how well a binary classification test correctly identifies or excludes a condition.

argumentation An argumentation is a justification to a claim.

claim A claim is opinionated text in which the author expresses a belief.

collocation A collocation is sequence of words or terms that co-occur more often than would be expected by chance.

comment In weblogs, any responses to the entry is a comment.

corpus A corpus is a collection of written texts.

cross-genre analysis Cross-Genre analysis is an analysis of multiple online genres.

dataset See **corpus**

dialog Dialog refers to the conversation between participants in online discussions.

(dis)agreement Agreement and disagreement together are referred to as (dis)agreement. They refer to having the same opinion (agreement) or different opinion (disagreement) from another.

domain adaptation Domain adaptation refers to exploiting and adapting information from other domains, or genres, to improve results.

entry In weblogs, the post by the owner of the blog is called an entry.

f-score F-score is the harmonic mean of precision and recall.

lexicon A lexicon is a dictionary. This is mainly used in the opinion detection chapter of this thesis (Chapter 4).

online genre A online genre is a specific type of social media from a single website.

part of speech (POS) Part of Speech is a category of lexical terms which have similar grammatical properties. Words can be tagged for different parts of speech using a part of speech tagger.

persuasion Persuasion is a claim followed by support in the form of argumentation, reiteration, and grounding.

post A post is a single entry or comment in a discussion thread.

precision Precision indicates of the data predicted to be true, how many are correct.

sentiment Sentiment is an opinion that can refer to subjectivity or polarity. Subjectivity is subjective or objective. Polarity can be positive, negative, or neutral.

social media Social Media refers to Internet-based applications, such as virtual communities and networks, that allow the creation and exchange of user-generated content.

social science Social Science is academic disciplines concerned with the study of the social life of human groups and individuals including anthropology, economics, geography, history, political science, psychology, social studies, and sociology.

system components The system components are our core components used in influence detection. They are: Opinion, Claims, Argumentation, Persuasion, Agreement, Author Traits, Credibility, and Dialog Patterns.

recall Recall is how many true cases are found during classification of a dataset.

quote-response (Q-R) pair A quote-response pair is a pair of posts (or a portion of them) where the response is a direct reply to a quote

thread, discussion Thread and discussion refer to a online conversation with a threaded structure which indicates the reply chain.

weapons of influence The weapons of influence [Cialdini, 2007] are different means of influencing people. They are: Reciprocation, Commitment and Consistency, Social Proof, Liking, Authority, and Scarcity.