Detecting Opinionated Claims in Online Discussions

Sara Rosenthal
Department of Computer Science
Columbia University
New York, NY 10027, USA
Email: sara@cs.columbia.edu

Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY 10027, USA
Email: kathy@cs.columbia.edu

***Abstract*—This paper explores the automatic detection of sentences that are opinionated claims, in which the author expresses a belief. We use a machine learning based approach, investigating the impact of features such as sentiment and the output of a system that determines committed belief. We train and test our approach on social media, where people often try to convince others of the validity of their opinions. We experiment with two different types of data, drawn from LiveJournal weblogs and Wikipedia discussion forums. Our experiments show that sentiment analysis is more important in LiveJournal, while committed belief is more helpful for Wikipedia. In both corpora, n-grams and part-of-speech features also account for significantly better accuracy. We discuss the ramifications behind these differences.**

I. Introduction

We explore the automatic detection of claims, statements in which the author presents an opinion that he thinks should be adopted by others. Such claims are common in social media, such as weblogs and discussion forums, where participants often aim to convince others of the validity of their own viewpoint. For example, the following sentences are claims found in Wikipedia discussion forums and LiveJournal blogs respectively: "I consider deletion of the link to be an act of vandalism." and "I think this is seriously one of the funniest things SNL has ever done!". Automatically detecting claims is useful for identifying Disputed Claims (claims that are not trustworthy [1], [2]), as well as analyzing discourse for social acts such as argumentation: claim followed by justification, agreements (two claims that agree on the same topic), and disagreements (two claims disagreeing on the same topic) [3], [4], [5], [6].

Given that we aim to identify opinionated and personal views where the author is committed to their opinion we hypothesize that sentiment detection [7] and committed belief [8] 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. We present a machine learning approach to detecting claims where we investigate the impact of the two main features, sentiment and committed belief. We also measure traditional lexical features as well as words and abbreviations used in social media. We use annotated data drawn from LiveJournal weblogs, and Wikipedia discussion forums. Part of our goal is to determine characteristics of these different types of social media. Our experiments show that we are able to detect claims significantly better than a majority and question mark baseline and that sentiment and parts of speech are very useful indicators of claims in LiveJournal while committed belief and n-grams are useful indicators of claims in Wikipedia.

More formally, we use *claim* to refer to assertions by a speaker who is attempting to convince others that his opinion is true. 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."*

In this paper, we investigate claims that express an opinionated belief; our goal is to identify when a speaker asserts a belief that s/he would like to convince to others that is true. A claim expresses belief if it is a personal view that others can disagree with. The sentence "*This, as I said earlier, is a complex issue .*" is an example of an opinionated claim because it is a personal view of the author. Table I shows examples of claims in each corpus. To establish a claim as truth, a speaker may often choose to justify the claim [3] to strengthen its impact. We do not explore justification in this paper.

It is often difficult to distinguish when a claim expresses a belief as opposed to a request for action or a statement of a fact. The sentences '*I have a job at Walmart.*" and "*So If you wish , go ahead and change it back .*" are simple examples of a statement of fact and request for action, respectively. On the other hand, in the sentence "*Lots of articles get a few instances of vandalism a day .*" it is unclear whether the claim is a fact or a belief.

In the sentence "*Would be good if you could say what those reasons are .*" 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, "*You*

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

TABLE I
EXAMPLES OF CLAIMS FROM EACH CORPUS

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 as good as possible 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 book is POV or the writer has only a slender grasp of relevant issues , material can be wrong .

say that “ ... for similar reasons I do not think the dead soldier is , either . ””. 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 the following sections, we first present related work, defining what we mean by sentiment analysis and committed belief. We then present our corpora, the methods we used, and our experiments and evaluation.

II. Related Work

As far as we are aware, there is very little work on identifying claims of the type we describe here. Exceptions are two recent companion papers, where Bender et al (2011) [9] and Marin et al (2011), [10] discuss the annotation and detection of authority claims on the sentence level. Authority claims can be credentials (a person's education, training, or history), experiential (based on the witness of an event), institutional (position within an organization), forum (policy norms), and external (outside authority, e.g. references). They only run experiments on detecting 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 hand-picked words. This research is parallel to our work. 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?” [9] which is a question and not belief.

In other related work, Kwon et al (2007) [11] , 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). They're 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: words, bigrams, and subjectivity, but 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 and have an accuracy of 55% using boosting. Their approach could benefit from our claim detection system to narrow down the potential main claims.

Since detection of subjectivity and committed belief are major components of our work, we present related work on these two topics as well.

A. Subjectivity

There has been a large amount of work on sentiment detection, for example [12], [13], [14], [15], [16], [17], but most has been carried out on edited text. Previous approaches to sentiment detection in weblogs and forums tend to classify the sentiment of the entire document (blog post, or discussion forum) [18], [19], [20] instead of sentiment at the sentence level as we do. Here, we extend previous approaches [7] and gear it towards online corpora, exploiting social media characteristics.

Agarwal et al (2009) [7] used the Dictionary of Affect and Language (DAL), extended with WordNet to determine the polarity of phrases. They also used contextual features such as the top n-grams and POS tags to improve their results. We modify their work to look at subjectivity and will briefly discuss how we extended their approach in an upcoming section.

B. 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., [21], [22]). Committed belief is an integral part of claim detection because a statement 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 the sentence described earlier as fact, ‘*I have a job at Walmart.*”, “have” is committed belief, but the sentence is not an opinion. Prabhakaran et al (2010) [8] created a system that automatically tags words in a sentence for three types of belief [23]: committed (“I know”), non-committed (“I may”), and not applicable (“I wish”) which vary in intensity from strong to weak respectively. 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.

III. Corpora

Our corpus consists of two datasets: 285 LiveJournal blogposts and 51 Wikipedia discussion forums. Each dataset consists of 2,000 sentences that are between 30-120 characters. The statistics of each corpus are described in

TABLE II
STATISTICS FOR EACH CORPUS; LIVEJOURNAL AND WIKIPEDIA

Corpus	Claims	Not Claims	Subjective Phrases	Objective Phrases	Vocabulary Size
LiveJournal	1197 (60%)	791 (40%)	3035 (39%)	4709 (61%)	4747
Wikipedia	1282 (64%)	715 (36%)	1319 (37%)	4496 (63%)	4342

Table II. LiveJournal is a virtual community on the web where bloggers frequently post entries about their personal lives. LiveJournal tends to be very informal. For example, there is slang and ellipses in sentence 1 in Table I and it is written in lower case.

A wikipedia article is a webpage that can be edited by anyone to provide encyclopedia entries on any topic. However, there are usually a committed group of individuals that work together to edit the entries. These individuals discuss and debate with each other how to edit existing pages. This discussion occurs on the Wikipedia discussion forum and it is rich in opinionated claims; the Wikipedia dataset has 4% more claims than the LiveJournal dataset.

Our datasets were annotated for claim by two annotators. The annotators were told that a claim is a statement that is a belief that can be justified. The annotators were given a list of 2,000 sentences for each corpus. Our goal is to determine if a sentence is a claim on its own. Therefore, we did not provide the context of the sentence (we 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 2000 annotations for each corpus. Inter-annotator agreement prior to resolving disagreements was 75.4%, Cohen's $\kappa = 50.0$, on a subset of 663 LiveJournal sentences and 79.2%, Cohen's $\kappa = 55.7$, on a subset of 997 Wikipedia sentences.

TABLE IV
THE AVERAGE NUMBER OF SUBJECTIVE (SUB.) AND OBJECTIVE (OBJ.) PHRASES IN A SENTENCE THAT IS AND IS NOT A CLAIM.

	Claim		Not Claim	
	Sub.	Obj.	Sub.	Obj.
LiveJournal	1.8	2.6	1.5	2.0
Wikipedia	1.6	2.4	1.4	2.0

Each of our datasets were independently annotated for sentiment on Mechanical Turk by 3-5 Mechanical Turk workers at 3-5 cents a hit. We spent around $500 in total. Each worker marked the start and end of each subjective words/phrases in the sentence. The annotations of each worker were combined using intersection. A word had to appear in 2/3 of the annotations in order to be considered subjective. Table IV indicates the average number of subjective and objective phrases in sentences that were annotated as claim and not claim. Sentences that are 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 claims.

IV. METHODS

We use a supervised machine learning approach. We hypothesized that sentiment analysis should have a major impact in identifying claims since claims are an expression of opinions. However, the force of 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.

We pre-process the sentences to add POS tags using the CRF tagger[2] and chunk the sentences using the CRF chunker[3] . The chunker uses three labels, 'B' (beginning of chunk), 'I' (in chunk), and 'O' (out of chunk). The 'O' label tends to be applied to punctuation which one typically wants to ignore. However, in this context, punctation can be important (e.g. exclamation points, and emoticons). Therefore, we append words/phrases and punctuation tagged as O to the prior B-I chunk. Prior to tagging and chunking, we expand contractions and convert emoticons into a corresponding key to ensure that they would stay intact in the tagging and chunking steps. The emoticons were returned to the sentence afterwards. We also tagged our corpus for belief [8] as described below. Our claim system uses all or a subset of the methods discussed in this section as features in system evaluation.

A. Sentiment

Our sentiment system is a modification to the original work of Agarwal et al (2009) [7]. Similarly to their system, we automatically detect the sentiment of phrases using the DAL. We also expand it to use social media features such as emoticons, acronyms, and misspellings to adapt it towards our corpora. Since we are interested in whether an opinion exists, polarity is not important. Thus, we modified the system to detect subjective vs. objective phrases. In contrast to the original approach, we only perform feature selection on the 100 most frequent n-grams and do not use WordNet to expand the DAL.

[2]Xuan-Hieu Phan, CRFTagger: CRF English Phrase Tagger, http://crftagger.sourceforge.net/, 2006

[3]Xuan-Hieu Phan, CRFChunker: CRF English Phrase Chunker, http://crfchunker.sourceforge.net/, 2006

TABLE III
A list of the most common Part-of-Speech, Belief, and n-gram features. Each list contains features from the claim and not claim class. [1]The n-gram feature 'lllinkkk' refers to a url.

	LiveJournal			Wikipedia			LiveJournal + Wikipedia		
	POS	Belief	n-grams	POS	Belief	n-grams	POS	Belief	n-grams
1	VBZ	good	i do	VBZ	added	–	CD	please	added
2	VBP	hard	is	RB	agree	a	JJ	added	added by
3	RB	love	is not	JJ	disagree	agree	JJS	agree	are you
4	PRP	call	it	CD	needs	be	PRP	disagree	–
5	JJ	comment	it is	VBD	think	facts	RB	hard	bad
6	US	i	not	RBS	Preceding	however	VBD	irrelevant	be
7	CD	right	pretty	WRB	please	i think	VBP	right	but
8	MD	knows	that	VBP	irrelevant	is	VBZ	stay	facts
9	WDT	great	call	JJS	keep	it is	MD	think	however
10	VBD	means	do you	$	seem	much	$	Preceding	i
11	VBN	means	pretty	#	admit	needs	NNP	meet	i think
12		saying	lllinkkk[1]		characterized	not	UH	i	is
13		like	things		reflect	pov		needs	is not
14		better	i do not		accepted	sabbath		nice	it
15		checking	very		ask	think		do	it is

We ran several experiments on the system which found that using more n-grams and WordNet did not improve the results in online corpora, but increased the runtime significantly.

We train the sentiment system on the same sentences as our claim corpus and use the gold sentiment annotations during training of the claim system. 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, and lastly, the count for each of the subjective/objective patterns of the sentence using 1-3 chunks. For example, the sentence "Some posts seem to serve no purpose but to make people pissed ." was chunked and tagged as "[Some posts]/o [seem to serve]/o [no purpose]/s [but]/o [to make]/o [people pissed]/s". It has sentiment, 1/3 of the chunks are subjective and it contains the subjective/objective patterns 'o', 's', 'o o', 's o', 'o s', 'o o s', 'o s o', and 's o o'.

B. Committed Belief

We use the system described in Prabhakaran et al (2010) [8] to extract words in the sentence that express belief. The system detects how the writer intends the reader to interpret a stated proposition by tagging the sentences with three types of belief [23]: 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 and used all three of the tags to indicate whether the sentence conveyed belief.

We use the output of the system in a bag of words approach by counting the occurrence of each of the belief words and performing feature selection on them using the Chi Square test in Weka. The 15 most common belief features for each corpus are listed in Table III. In the future we would like to annotate our sentences for committed belief to train it in our genre and to better analyze how strong of an indicator of opinionated claims it is.

C. Question

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.

D. Lexical Features

We include Part-of-Speech tags and n-gram features (1-3 words). We counted the occurrence of all POS and performed feature selection on them using the Chi Square test in Weka. 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 top 250 n-grams using the Chi Square test in Weka. The most common n-grams and POS tags for each corpus are listed in Table III.

E. Social Media Features

TABLE V
List of Social Media Features and corresponding examples

Feature	**Example**
Capital Words	WHAT
Slang	dunno
Emoticons	:)
Acronyms	LOL
Punctuation	.
Repeated Punctuation	#$@.
Punctuation Count	5
Exclamation Points	!
Repeated Exclamations	!!!!
Question Marks	?
Repeated Questions	???
Ellipses	...

In order to take into account the nature of online corpora, we compute statistics on emoticons, slang and

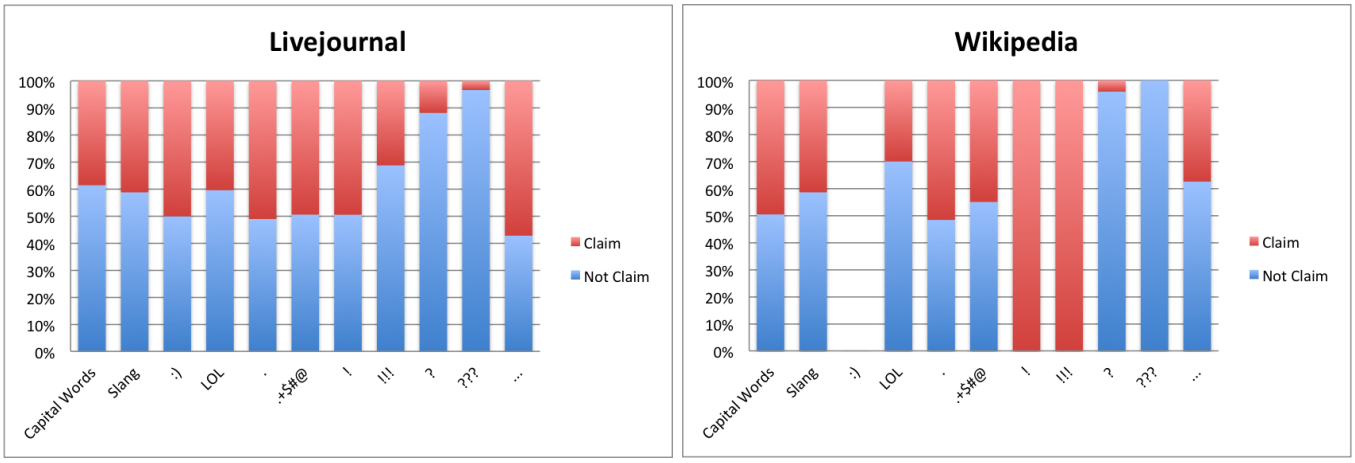

Fig. 1. Percentage of social media features that are claims in each corpus.

other features which we call "social media features" as described in Table V. We count the occurrence of several social media features in each sentence. Some of the features are more specific, such as exclamation points and ellipses, while others cover the more general case. A marker can only be counted as one feature (e.g. '!' is an exclamation point, not punctuation). On the other hand, the punctuation count feature takes into account the total number of features (e.g. '...' and '?' is 4). The count values are normalized by the length of the sentence. Across all the datasets, the social media features on average account for less than one word in a sentence, with LiveJournal having more than Wikipedia. Figure 1 displays the frequency of social media features in sentences that are claims vs. not claims across the two genres. Question marks are clearly indicative of a sentence not being a claim. Interestingly, in wikipedia, exclamation points only occur in claims, and in LiveJournal repeated exclamation points tend to point to sentences that are not claims.

TABLE VI
ACCURACY FOR CROSS-VALIDATION ON BALANCED DATASETS. WE USE TWO BASELINES: THE MAJORITY CLASS, AND THE QUESTION FEATURE. THE FEATURES USED ARE QUESTION (?), SOCIAL MEDIA (SM), SENTIMENT, BELIEF, N-GRAMS, AND POS. THE BEST RESULTS ARE DISPLAYED IN BOLD

Experiment	LJ	Wiki
Majority	50	50
?	55.5	57.3
?+sm	58.1	59.4
?+sentiment	63	57.6
?+belief	59	63
?+n-grams	63	68.1
?+pos	65.7	66.2
?+n-grams+pos	65.4	**71.1**
?+sentiment + belief	63	59.2
?+belief+pos	65.4	68
?+sm+pos+n-grams	65	**71.1**
?+sentiment+n-grams+pos	**66.4**	67.4
?+sentiment+belief+n-grams+pos	**66.2**	66.3
?+ belief+sm+pos+n-grams	65	**71.4**
All	65.6	66.1

V. Experiments and Results

We ran all of our experiments in Weka using Logistic Regression. Each experiment uses some permutation of the methods described in section IV. While several other permutations were run as well, we only show the groups that are most informative and have the most impact on results. In each cross-validation experiment we did 10 runs of 10-fold cross-validation. Due to our limited amount of data, we used the same sentences to train the opinion classifier and claim classifier. For each fold of the cross-validation, we first ran that fold on the opinion classifier and then used it in the claim classifier with the train and test portions the same for each.

A. Single Corpus Classification

We ran each experiment using a balanced and unbalanced dataset (See Table II for size); the balanced dataset is the size of the smaller class multiplied by 2) in each corpus. Our results are shown in Table VI and Table VII with the best results highlighted in bold (there is no statistically significant difference between all the results that are highlighted). We use the majority class and the question feature as our baselines. All the features provided a statistically significant improvement over the baselines with $p = .001$ with the exception of the 'question+belief' experiment on unbalanced LiveJournal data (which performed worse). The most useful feature in the LiveJournal corpus was POS tags and the most useful feature in Wikipedia were n-grams. Sentiment has more of an impact in LiveJournal and belief has more of an impact in Wikipedia. Social media features provide similar improvements to both datasets. Interestingly, even though some of the more complex features help in their own right, the performance of POS and n-grams together does better than almost all other experiments. The one exception is the 'question+sentiment+n-grams+pos' experiment which does statistically significantly better ($p = .01$)

TABLE VII
ACCURACY FOR CROSS-VALIDATION ON UNBALANCED DATASETS. WE USE TWO BASELINES: THE MAJORITY CLASS, AND THE QUESTION FEATURE. THE FEATURES USED ARE QUESTION (?), SOCIAL MEDIA (SM), SENTIMENT, BELIEF, N-GRAMS, AND POS. THE BEST RESULTS ARE DISPLAYED IN BOLD

Experiment	LJ	Wiki
Majority	60.2	64.2
?	64.1	69.3
?+sm	64.5	69.7
?+sentiment	65.5	69.4
?+belief	63	69.6
?+n-grams	65.6	73.3
?+pos	**67.6**	70.6
?+n-grams+pos	**67.3**	**74.5**
?+sentiment + belief	65.9	69.5
?+belief+pos	**67.6**	72.1
?+sm+pos+n-grams	**67.4**	**74.7**
?+sentiment+n-grams+pos	**67.4**	73.9
?+sentiment+belief+n-grams+pos	66.9	73.1
?+ belief+sm+pos+n-grams	**67.3**	73.4
All	66.9	**74.7**

TABLE VIII
ACCURACY FOR USING BALANCED TRAINING DATASETS ON A TEST SET. WE USE TWO BASELINES: THE MAJORITY CLASS, AND THE QUESTION FEATURE. THE FEATURES USED ARE QUESTION (?), SOCIAL MEDIA (SM), SENTIMENT, BELIEF, N-GRAMS, AND POS. THE BEST RESULTS ARE DISPLAYED IN BOLD

Experiment	LJ	Wiki
Majority	N/A	N/A
?	51.5	60
?+sm	51.5	68.5
?+sentiment	53	60.5
?+belief	55.9	65
?+n-grams	53.5	72
?+pos	57.9	75
?+n-grams+pos	57.9	76
?+sentiment + belief	54.5	62
?+belief+pos	**62.4**	**77.5**
?+sm+pos+n-grams	55	75.5
?+sentiment+n-grams+pos	56.4	69
?+sentiment+belief+n-grams+pos	56.9	65
?+ belief+sm+pos+n-grams	58.4	74
All	57.4	68

TABLE IX
ACCURACY FOR USING UNBALANCED TRAINING DATASETS ON A TEST SET. WE USE TWO BASELINES: THE MAJORITY CLASS, AND THE QUESTION FEATURE. THE FEATURES USED ARE QUESTION (?), SOCIAL MEDIA (SM), SENTIMENT, BELIEF, N-GRAMS, AND POS. THE BEST RESULTS ARE DISPLAYED IN BOLD

Experiment	LJ	Wiki
Majority	49.5	52
?	51.5	60
?+sm	52	63
?+sentiment	54.5	60
?+belief	52.5	61.5
?+n-grams	54.5	76
?+pos	56.9	75
?+n-grams+pos	57.9	**82**
?+sentiment+belief	**58.4**	60.5
?+belief+pos	57.9	70
?+sm+pos+n-grams	55.9	78.5
?+sentiment+n-grams+pos	54	66.5
?+sentiment+belief+n-grams+pos	52.5	67
?+ belief+sm+pos+n-grams	55.4	74.5
All	54	69.5

TABLE X
ACCURACY FOR BALANCED AND UNBALANCED DATASETS USING LIVEJOURNAL AND WIKIPEDIA AS ONE CORPUS. WE USE TWO BASELINES: THE MAJORITY CLASS, AND THE QUESTION FEATURE. THE FEATURES USED ARE QUESTION (?), SOCIAL MEDIA (SM), SENTIMENT, BELIEF, N-GRAMS, AND POS. THE BEST RESULTS ARE DISPLAYED IN BOLD

Experiment	Bal.	Unbal.
Majority	50	62.2
?	56.4	66.7
?+sm	58	67.5
?+sentiment	63.6	67.7
?+belief	61.4	66.9
?+n-grams	65.7	70.3
?+pos	66.6	70.9
?+n-grams+pos	**68.9**	**72.6**
?+sentiment+belief	62.6	69.1
?+belief+pos	67.4	71.2
?+sm+pos+n-grams	**68.8**	**72.7**
?+sentiment+n-grams+pos	68.2	72
?+sentiment+belief+n-grams+pos	68.3	71.6
?+ belief+sm+pos+n-grams	**68.9**	72.1
All	68.1	71.8

than the 'pos' and 'n-gram+pos' experiment.

In addition to our cross-validation experiments, we also ran our experiments on a held out test set in both LiveJournal and Wikipedia that were only annotated for claim. Approximately 10 sentences were taken from 20 unseen documents resulting in a test set of around 200 sentences in both genres. The results using balanced and unbalanced training datasets are shown in Table VIII and Table IX. Similarly to the cross-validation experiments, we find that the most useful feature in the LiveJournal corpus was POS tags and the most useful feature in Wikipedia were n-grams and POS tags. In contrast to the cross-validation experiments we find that committed belief is more useful in both LiveJournal and Wikipedia in balanced datasets. Sentiment may not have as big an impact in these experiments because our phrases are based solely off of the chunks and not the sentiment annotations which can group multiple chunks together.

B. Cross-Domain Classification

The impact of the lexical and social media features differ in LiveJournal and Wikipedia as is evident in Table III. Are these corpora too different to allow classification across the two corpora? In this section, we show our experiments for combining the two corpora as well as using each corpus for training and testing respectively to illustrate how it affects performance.

Table X shows the results for combining LiveJournal and Wikipedia sentences into one corpus. Similarly to the single corpus classification, we find that POS and n-

TABLE XI
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 BALANCED DATASETS. THE FEATURES USED ARE QUESTION (?), SOCIAL MEDIA (SM), SENTIMENT, BELIEF, N-GRAMS, AND POS. THE BEST RESULTS ARE DISPLAYED IN BOLD

	Balanced		Unbalanced	
Experiment	L-W	W-L	L-W	W-L
majority	50	50	64.2	60.2
?	64.1	69.3	64.1	69.3
?+sm	62.3	67.2	64.7	69.9
?+sentiment	66.7	69.9	65.5	69.3
?+belief	65.6	71.6	66.1	73.7
?+n-grams	66.2	71.0	70.5	77.4
?+pos	66.9	67.9	70.4	72.4
?+sentiment+belief	69.1	72.4	69.0	73.5
?+belief+pos	68.9	71.5	72.4	77.4
?+sm+pos+n-grams	69.9	73.6	73.7	79.4
?+sentiment+n-grams+pos	72.2	75.0	70.7	77.6
?+sentiment+belief + n-grams+pos	**74.3**	**76.0**	72.1	80.0
?+belief+sm+pos + n-grams	71.6	**76.0**	**75.6**	**82.2**
All	74.0	**76.1**	72.8	80.8

grams are very useful. The combined dataset performs better than the LiveJournal experiment but worse than the Wikipedia experiment.

In our final experiment we explore how well the individual datasets can predict claims in the other. There is no cross-validation in this experiment. Our experiments show that the different datasets do perform well. The best Wikipedia system predicts LiveJournal sentences correctly 76.1% and 82.2% in balanced and unbalanced datasets respectively. The best LiveJournal system predicts Wikipedia sentences correctly 74.3% and 75.6% in balanced and unbalanced datasets respectively. The full set of results is shown in Table XI. Using Wikipedia sentences as training to predict LiveJournal sentences consistently outperforms using LiveJournal sentences as training to predict Wikipedia sentences.

VI. CONCLUSION AND FUTURE WORK

Our research does reveal that sentiment analysis and detection of committed belief play an important role in detection of claims. However, we also discovered that n-grams and POS tags have a strong impact on accuracy. Furthermore, we found that sentiment and POS tags were more important for LiveJournal, while committed belief and n-grams were more important for Wikipedia discussion forums.

These results are supported by the fact that sentiment is 5% more common in LiveJournal than in Wikipedia. The kinds of claims in LiveJournal tend to focus much more on the individual writing the post and thus, pronouns are common. LiveJournal 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.

Reflecting more broadly, while social media have different characteristics from grammatical and single author genres such as news, our results highlight the fact that sub-genres within social media do not share the same characteristics. Thus, it is important to investigate the characteristics of different kinds of social media.

Nonetheless, further work is needed to experiment with the different ways in which sentiment and committed belief can be used. For example, we could use opinionated words as features instead of the aggregate sentiment features that we experimented with. We also plan to explore the addition of new features to see if accuracy can be improved.

ACKNOWLEDGEMENT

The authors would like to thank Alex Liu for generously donating his time to help annotate all the sentences for claim.

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the U.S. Army Research Lab. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

REFERENCES

[1] R. Ennals, B. Trushkowsky, and J. M. Agosta, "Highlighting disputed claims on the web." in *WWW*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, Eds. ACM, 2010, pp. 341–350. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2010.html#EnnalsTA10

[2] B. T. Adler, L. de Alfaro, and I. Pye, "Detecting wikipedia vandalism using wikitrust," 2010.

[3] O. Biran and O. Rambow, "Identifying justifications in written dialogs," in *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ser. ICSC '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 162–168. [Online]. Available: http://dx.doi.org/10.1109/ICSC.2011.41

[4] J. Schneider, T. Groza, and A. Passant, "A review of argumentation for the social semantic web," *Semantic Web ÌâåÐ Interoperability, Usability, Applicability*, 2012.

[5] R. Abbott, M. Walker, P. Anand, J. E. Fox Tree, R. Bowmani, and J. 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)*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 2–11. [Online]. Available: http://www.aclweb.org/anthology/W11-0702

[6] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. 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)*, no. June, 2011, pp. 48–57. [Online]. Available: http://aclweb.org/anthology-new/W/W11/W11-07.pdf\#page=58

[7] A. Agarwal, F. Biadsy, and K. 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*, ser. EACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 24–32. [Online]. Available: http://portal.acm.org/citation.cfm?id=1609067.1609069

[8] V. Prabhakaran, O. Rambow, and M. T. Diab, "Automatic committed belief tagging," in *COLING (Posters)*, 2010, pp. 1014–1022.

[9] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. 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)*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 48–57. [Online]. Available: http://www.aclweb.org/anthology/W11-0707

[10] A. Marin, B. Zhang, and M. Ostendorf, "Detecting forum authority claims in online discussions," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 39–47. [Online]. Available: http://www.aclweb.org/anthology/W11-0706

[11] N. Kwon, L. Zhou, E. Hovy, and S. W. Shulman, "Identifying and classifying subjective claims," in *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*, ser. dg.o '07. Digital Government Society of North America, 2007, pp. 76–81. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248460.1248473

[12] T. Wilson, J. Wiebe, and P. Hoffman, "Recognizing contextual polarity in phrase level sentiment analysis," in *Proceedings of ACL*, 2005.

[13] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language." in *Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.*, 2005.

[14] P. Turney., "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews." in *Proceedings of ACL*, 2002.

[15] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts." in *Proceedings of ACL*, 2004.

[16] P. Beineke, T. Hastie, and S. Vaithyanathan, "The sentimental factor: Improving review classification via human provided information." in *Proceedings of ACL.*, 2004.

[17] S. M. Kim and E. Hovy., "Determining the sentiment of opinions." in *In Coling*, 2004.

[18] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[19] N. Yu and S. Kübler, "Filling the gap: semi-supervised learning for opinion detection across domains," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, ser. CoNLL '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 200–209. [Online]. Available: http://dl.acm.org/citation.cfm?id=2018936.2018959

[20] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proceedings of WWW*. New York, NY, USA: ACM Press, 2007, pp. 171–180.

[21] A. S. Rao and M. P. Georgeff, "Modeling rational agents within a bdi-architecture," in *KR*, 1991, pp. 473–484.

[22] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguist.*, vol. 26, no. 3, pp. 339–373, Sep. 2000. [Online]. Available: http://dx.doi.org/10.1162/089120100561737

[23] M. T. Diab, L. S. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo, "Committed belief annotation and tagging," in *Linguistic Annotation Workshop*, 2009, pp. 68–73.

[24] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.