Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations

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

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

Abstract

We investigate whether wording, stylistic choices, and online behavior can be used to predict the age category of blog authors. Our hypothesis is that significant changes in writing style distinguish pre-social media bloggers from post-social media bloggers. Through experimentation with a range of years, we found that the birth dates of students in college at the time when social media such as AIM, SMS text messaging, MySpace and Facebook first became popular, enable accurate age prediction. We also show that internet writing characteristics are important features for age prediction, but that lexical content is also needed to produce significantly more accurate results. Our best results allow for 81.57% accuracy.

1 Introduction

The evolution of the internet has changed the way that people communicate. The introduction of instant messaging, forums, social networking and blogs has made it possible for people of every age to become authors. The users of these social media platforms have created their own form of unstructured writing that is best characterized as informal. Even *how* people communicate has dramatically changed, with multitasking increasing and responses generated immediately. We should be able to exploit those differences to automatically determine from blog posts whether an author is part of a *pre-* or *post-social media* generation. This problem is called *age prediction* and raises two main questions:

- Is there a point in time that proves to be a significantly better dividing line between pre and post-social media generations?
- What features of communication most directly reveal the generation in which a blogger was born?

We hypothesize that the dividing line(s) occur when people in *generation Y*[1], or the *millennial generation*, (born anywhere from the mid-1970s to the early 2000s) were typical college-aged students (18-22). We focus on this generation due to the rise of popular social media technologies such as messaging and online social networks sites that occurred during that time. Therefore, we experimented with binary classification into age groups using all birth dates from 1975 through 1988, thus including students from generation Y who were in college during the emergence of social media technologies. We find five years where binary classification is significantly more accurate than other years: 1977, 1979, and 1982-1984. The appearance of social media technologies such as AOL Instant Messenger (AIM), weblogs, SMS text messaging, Facebook and MySpace occurred when people with these birth dates were in college.

We explore two of these years in more detail, 1979 and 1984, and examine a wide variety of

[1] http://en.wikipedia.org/wiki/Generation_Y

features that differ between the pre-social media and post-social media bloggers. We examine lexical-content features such as collocations and part-of-speech collocations, lexical-stylistic features such as internet slang and capitalization, and features representing online behavior such as time of post and number of friends. We find that both stylistic and content features have a significant impact on age prediction and show that, for unseen blogs, we are able to classify authors as born before or after 1979 with 80% accuracy and born before or after 1984 with 82% accuracy.

In the remainder of this paper, we first discuss work to date on age prediction for blogs and then present the features that we extracted, which is a larger set than previously explored. We then turn separately to three experiments. In the first, we implement a prior approach to show that we can produce a similar outcome. In the second, we show how the accuracy of age prediction changes over time and pinpoint when major changes occur. In the last experiment, we describe our age prediction experiments in more detail for the most significant years.

2 Related Work

In previous work, Mackinnon (2006) , used LiveJournal data to identify a blogger's age by examining the mean age of his peer group using his social network and not just his immediate friends. They were able to predict the correct age within +/-5 years at 98% accuracy. This approach, however, is very different from ours as it requires access to the age of each of the blogger's friends. Our approach uses only a body of text written by a person along with his blogging behavior to determine which age group he is more closely identified with.

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. In our study on determining the ideal age split we did not find 18 (bloggers born in 1986 in their dataset) to be significant.

Prior work by Schler et al. (2006) has examined metadata such as gender and age in blogger.com bloggers. 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"), all defined by the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007). They did not use characteristics of online behavior (e.g., friends). 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%. We re-implement Schler et al.'s work in section 5.1 with similar findings. 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). Neither Schler et al., nor we, attempt to determine how a person's writing changes over his lifespan (Pennebaker and Stone, 2003; Robins et al., 2002). 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. This work shows that n-grams are useful features for detecting age and it is difficult to detect differences between consecutive groups such as teens and 20s, and this

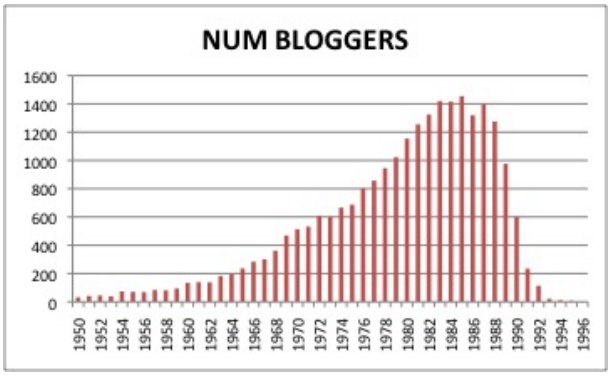

Figure 1: Number of bloggers in 2010 by year of birth from 1950-1996. A minimal amount of data occurred in years not shown.

provides evidence for the need to find a good classification split.

Other researchers have investigated weblogs for differences in writing style depending on gender identification (Herring and Paolillo, 2006; Yan and Yan, 2006; Nowson and Oberlander, 2006). 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. We also use a supervised machine learning approach, but classification by gender is naturally a binary classification task, while our work requires determining a natural dividing point.

3 Data Collection

Our corpus consists of blogs downloaded from the virtual community LiveJournal. We chose to use LiveJournal blogs for our corpus because the website provides an easy-to-use format in XML for downloading and crawling their site. In addition, LiveJournal gives bloggers the opportunity to post their age on their profile. We take advantage of this feature by downloading blogs where the user chooses to publicly provide this metadata.

We downloaded approximately 24,500 LiveJournal blogs containing age. We represent age as the year a person was born and not his age at the time of the experiment. Since technology has different effects in different countries, we only analyze the blogs of people who have listed US as their country. It is possible that text written in a language other than English is included in our corpus. However, in a manual check of a small portion of text from 500 blogs, we only found English words. Each blog was written by a unique individual and includes a user profile and up to 25 recent posts written between 2000-2010 with the most recent post being written in 2009-2010. The birth dates of the bloggers range in years from 1940 to 2000 and thus, their age ranges from 10 to 70 in 2010. Figure 1 shows the number of bloggers per age in our group with birth dates from 1950 to 1996. The majority of bloggers on LiveJournal were born between 1978-1989.

4 Methods

We pre-processed the data to add Part-of-Speech tags (POS) and dependencies (de Marneffe et al., 2006) between words using the Stanford Parser (Klein and Manning, 2003a; Klein and Manning, 2003b). The POS and syntactic dependencies were only found for approximately the first 90 words in each sentence. Our classification method investigates 17 different features that fall into three categories: online behavior, lexical-stylistic and lexical-content. All of the features we used are explained in Table 1 along with their trend as age decreases where applicable. Any feature that increased, decreased, or fluctuated should have some positive impact on the accuracy of predicting age.

4.1 Online Behavior and Interests

Online behavior features are blog specific, such as number of *comments* and *friends* as described in Table 1.1. The first feature, *interests*, is our only feature that is specific to LiveJournal. Interests appear in the LiveJournal user profile, but are not found on all blog sites. All other online behavior features are typically available in any blog.

	Feature	Explanation	Example	Trend as Age Decreases
1	Interests	Top[3] interests provided on the profile page[2]	disney	N/A
2	# of Friends	Number of friends the blogger has	45	fluctuates
	# of Posts	Number of downloadable posts (0-25)	23	decrease
	# of Lifetime Posts	Number of posts written in total	821	decrease
	Time	Mode hour (00-23) and day the blogger posts	11/Monday	no change
	Comments	Average number of comments per post	2.64	increase
3	Emoticons	number of emoticons[1]	:)	increase
	Acronyms	number of internet acronyms[1]	lol	increase
	Slang	number of words that are not found in the dictionary[1]	wazzup	increase
	Punctuation	number of stand-alone punctuation[1]	...	increase
	Capitalization	number of words (with length > 1) that are all CAPS[1]	YOU	increase
	Sentence Length	average sentence length	40	decrease
	Links/Images	number of url and image links[1]	www.site.com	fluctuates
4	Collocations	Top[3] Collocations in the age group.	to [] the	N/A
	Syntax Collocations	Top[3] Syntax Collocations in the age group.	best friends	N/A
	POS Collocations	Top[3] Part-of-Speech Collocations in the age group.	this [] [] VB	N/A
	Words	Top[3] words in the age group	his	N/A

Table 1: List of all features used during classification divided into three categories (1,2) online behavior and interests, (3) lexical - content, and (4) lexical - stylistic [1] normalized per sentence per entry, [2] available in LiveJournal only, [3] pruned from top 200 features to include those that do not occur within +/- 10 position in any other age group

We extracted the top 200 interests based on occurrence in the profile page from 1500 random blogs in three age groups. These age groups are used solely to illustrate the differences that occur at different ages and are not used in our classification experiments. We then pruned the list of interests by excluding any interest that occurred within a +/-10 window (based on its position in the list) in multiple age groups. We show the top interests in each age group in Table 2. For example, "disney" is the most popular unique interest in the 18-22 age group with only 39 other non-unique interests in that age group occurring more frequently. "Fanfiction" is a popular interest in all age groups, but it is significantly more popular in the 18-22 age group than in other age groups.

18-22		28-32		38-42	
disney	39	tori amos	49	polyamory	40
yaoi	40	hiking	55	sca	67
johnny depp	42	women	61	babylon 5	84
rent	44	gaming	62	leather	94
house	45	comic books	67	farscape	103
fanfiction	11	fanfiction	58	fanfiction	138
drawing	10	drawing	25	drawing	65
sci-fi	199	sci-fi	37	sci-fi	21

Table 2: Top interests for three different age groups. The top half refers to the top 5 interests that are unique to each age group. The value refers to the *position of the interest in its list*

Amongst the other online behavior features, the number of friends tends to fluctuate but seems to be higher for older bloggers. The number of *lifetime posts* (Figure 2(d)), and *posts* decreases as bloggers get younger which is as one would expect unless younger people were orders of magnitude more prolific than older people. The *mode time* (Figure 2(b)), 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. We found time to not be a useful feature in this manner and it is difficult to come to any conclusions from its change as year of birth decreases.

4.2 Lexical - Stylistic

The Lexical-Stylistic features in Table 1.2, such as *slang* and *sentence length*, are computed us-

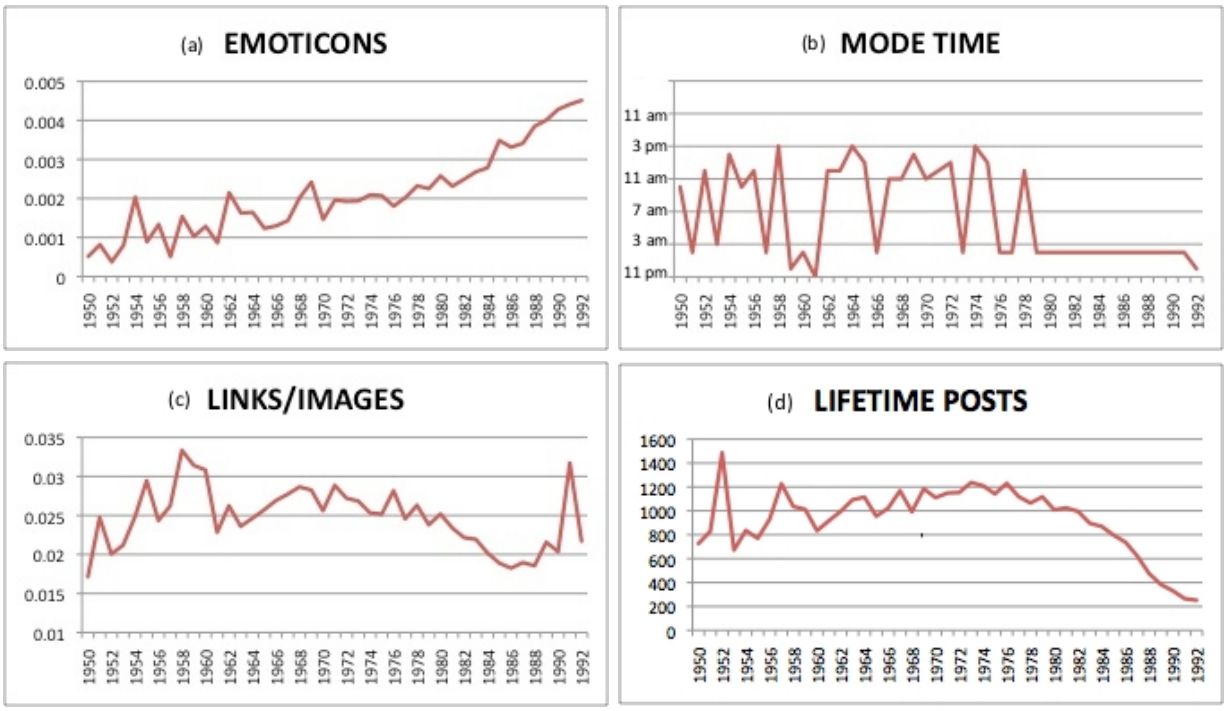

Figure 2: Examples of change to features over time (a) Average number of emoticons in a sentence increases as age decreases (b) The most common time fluctuates until 1982, where it is consistent (c) The number of links/images in a sentence fluctuates (d) The average number of lifetime posts per year decreases as age decreases

ing the text from all of the posts written by the blogger. Other than *sentence length*, they were normalized by sentence and post to keep the numbers consistent between bloggers regardless of whether the user wrote one or many posts in his/her blog. The number of *emoticons* (Figure 2(a)), *acronyms*, and *capital words* increased as bloggers got younger. *Slang* and *punctuation*, which excludes the emoticons and acronyms counted in the other features, increased as well, but not as significantly. The *length of sentences* decreased as bloggers got younger and the number of *links/images* varied across all years as shown in Figure 2(c).

4.3 Lexical - Content

The last category of features described in Table 1.3 consists of collocations and words, which are content based lexical terms. The top words are produced using a typical "bag-of-words" approach. The top collocations are computed using a system called *Xtract* (Smadja, 1993).

We use Xtract to obtain important lexical collocations, syntactic collocations, and POS collocations as features from our text. *Syntactic collocations* refer to significant word pairs that have specific syntactic dependencies such as subject/verb and verb/object. Due to the length of time it takes to run this program, we ran Xtract on 1500 random blogs from each age group and examined the first 1000 words per blog. We looked at 1.5 million words in total and found approximately 2500-2700 words that were repeated more than 50 times.

We extracted the top 200 words and collocations sorted by *post frequency* (pf), which is the number of posts the term occurred in. Then, similarly to interests, we pruned each list to include the features that did not occur within +/-10 window (based on its position in the list) within each age group. Prior to settling on these metrics, we also experimented with other metrics such as the number of times the collocation

18-22		28-32		38-42	
ldquot (')	101	great	166	may	164
t	152	find	167	old	183
school	172	many	177	house	191
x	173	years	179	world	192
anything	175	week	181	please	198
maybe	179	post	190	-	-
because	68	because	80	because	93
him	59	him	85	him	73

Table 3: Top words for three age groups. The top half refers to the top 5 words that are unique to each age group. The value refers to the *position of the interest in its list*

	blogger.com			livejournal.com		
download year	2004			2010		
# of Blogs	19320			11521		
# of Posts[1]	1.4 million			256,000		
# of words[1]	295 million			50 million		
age	13·17	23·27	33·37	18·22	28·32	38·42
size	8240	8086	2994	3518	5549	2454
majority baseline	43.8% (13-17)			48.2% (22-32)		

Table 4: Statistics for Schler et al.'s data (blogger.com) vs our data (livejournal.com) [1] is approximate amount.

occurred in total, defined as *collocation* or *term frequency* (tf), the number of blogs the collocation occurred in, defined as *blog frequency* (bf), and variations of TF*IDF (Salton and Buckley, 1988) where we tried using inverse blog frequency and inverse post frequency as the value for IDF. In addition, we also experimented with looking at a different number of important words and collocations ranging from the top 100-300 terms and experimented without pruning. None of these variations improved accuracy in our experiments, however, and thus, were dropped from further experimentation.

Table 3 shows the top words for each age group; older people tend to use words such as "house" and "old" frequently and younger people talk about "school".

In our analysis of the top collocations, we found that younger people tend to use first person singular (*I,me*) in subject position while older people tend to use first person plural (*we*) in subject position, both with a variety of verbs.

5 Experiments and Results

We ran three separate experiments to determine how well we can predict age: 1. classifying into three distinct age groups (Schler et al. (2006) experiment), 2. binary classification with the split at each birth year from 1975-1988 and 3. Detailed classification on two significant splits from the second experiment.

We ran all of our experiments in Weka (Hall et al., 2009) using logistic regression over 10 runs of 10-fold cross-validation. All values shown are the averages of the accuracies from the 10 cross-validation runs and all results were compared for statistical significance using the t-test where applicable.

We use logistic regression as our classifier because it has been shown that logistic regression typically has lower asymptotic error than naive Bayes for multiple classification tasks as well as for text classification (Ng and Jordan, 2002). We experimented with an SVM classifier and found logistic regression to do slightly better.

5.1 Age Groups

The first experiment implements a variation of the experiment done by Schler et al. (2006). The differences between the two datasets are shown in Tables 4. The experiment looks at three age groups containing a 5-year gap between each group. Intermediate years were not included to provide clear differentiation between the groups because many of the blogs have been active for several years and this will make it less common for a blogger to have posts that fall into two age groups (Schler et al., 2006).

We did not use the same age groups as Schler et al. because very few blogs on LiveJournal, in 2010, are in the 13-17 age group. Many early demographic studies (Perseus Development, 2004; Herring et al., 2004) show teens as the dominant age group in all blogs. However, more recent studies (Nowson and Oberlander, 2006; Lenhart et al., 2010) show that less teens blog. Furthermore, an early study on the LiveJournal

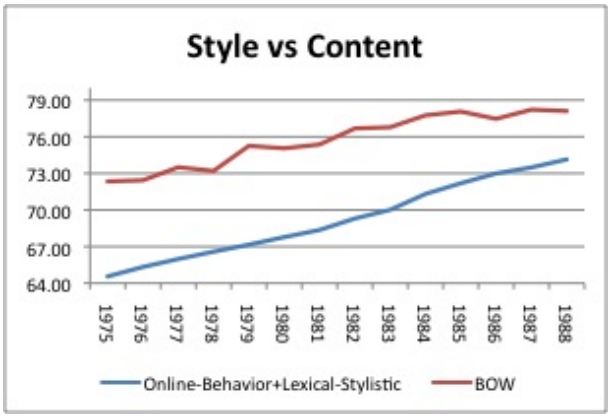

Figure 3: Style vs Content: Accuracy from 1975-1988 for Style (Online-Behavior+Lexical-Stylistic) vs Content (BOW)

demographic (Kumar et al., 2004) reported that 28.6% of blogs are written by bloggers between the ages 13-18 whereas based on the current demographic statistics, in 2010[2], only 6.96% of blogs are written by that age group and the number of bloggers in the 31-36 age group increased from 3.9% to 12.08%. We chose the later age groups because this study is based on blogs updated in 2009-10 which is 5-6 years later and thus, the 13-17 age group is now 18-22 and so on.

We use style-based (lexical-stylistic) and content-based features (BOW, interests) to mimic Schler et al.'s experiment as closely as possible and also experimented with adding online-behavior features. Our experiment with style-based and content-based features had an accuracy of 57%. However, when we added online-behavior, we increased our accuracy to 67%. A more detailed look at the better results show that our accuracies are consistently 7% lower than the original work but we have similar findings; 18-22s are distinguishable from 38-42s with accuracy of 94.5%, and 18-22s are distinguishable from 28-32s with accuracy of 80.5%. However, many 38-42s are misclassified as 28-32s with an accuracy of 72.1%, yielding overall accuracy of 67%. Due to our findings, we believe that adding online-behavior features to Schler et al.'s dataset would improve their results as well.

[2]http://www.livejournal.com/stats.bml

5.2 Social Media and Generation Y

In the first experiment we used the current age of a blogger based on when he wrote his last post. However, the age of a person changes; someone who was in one age group now will be in a different age group in 5 years. Furthermore, a blogger's posts can fall into two categories depending on his age at the time. Therefore, our second experiment looks at year of birth instead of age, as that never changes. In contrast to Schler et al.'s experiment, our division does not introduce a gap between age groups, we do binary classification, and we use significantly less data.

We approach 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:

- get 1500 blogs (∼33,000 posts) balanced across years BEFORE X
- get 1500 blogs (∼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 amongst 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

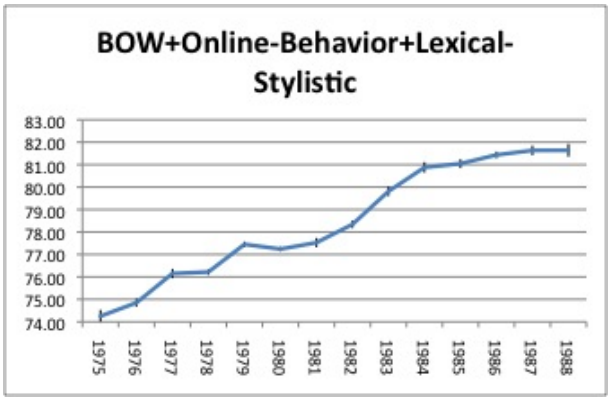

Figure 4: Style and Content: Accuracy from 1975-1988 using BOW, Online Behavior, and Lexical-Stylistic features

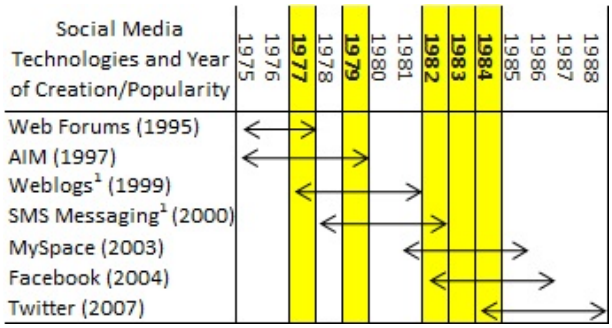

Figure 5: The impact of social media technologies: The arrows correspond to the years that generation Yers were college aged students. The highlighted years represent the significant years. [1]Year it became popular (Urmann, 2009)

online-behavior, lexical-stylistic, and BOW features.

We found accuracy to increase as year of birth increases in various feature experiments which is consistent with the trends we found while examining the distribution of features such as emoticons and lifetime posts in Figure 2. We experimented with style and content features and found that both help improve accuracy. Figure 3 shows that content helps more than style, but style helps more as age decreases. However, as shown in Figure 4, 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.

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 5 compares the significant years 1977,1979, and 1982-1984 against when each technology was created or became popular amongst 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.

5.3 A Closer Look: 1979 and 1984

Our final experiment provides a more detailed explanation of the results using various feature combinations when splitting pre- and post- social media bloggers by year of birth at two of the significant years found in the previous section; 1979 and 1984. The results for all of the experiments described are shown in Table 5.

We experimented against two baselines, online behavior and interests. We chose these two features as baselines because they are both easy to generate and not lexical in nature. We found that we were able to exceed the baselines significantly using a simple bag-of-words (BOW) approach. This means the BOW does a better job of picking topics than interests. We found that including all 17 features did not do well, but we were able to get good results using a subset of the lexical features. We found the best results to have an accuracy of 79.96% and 81.57% for 1979 and 1984 respectively using BOW, interests, online behavior, and all lexical-stylistic features.

In addition, we show accuracy without interests since they are not always available.

Experiment	1979	1984
Online-Behavior	59.66	61.61
Interests	70.22	74.61
Lexical-Stylistic	65.38[2]	67.28[2]
Slang+Emoticons+Acronyms	60.57[2]	62.10[2]
Online-Behavior + Lexical-Stylistic	67.16[2]	71.31[2]
Collocations + Syntax Collocations	53.47[1]	73.45[2]
POS-Collocations + POS-Syntax Collocations	55.54[1]	74.00[2]
BOW	75.26	77.76
BOW+Online-Behavior	76.39	79.22
BOW + Online-Behavior + Lexical-Stylistic	**77.45**	**80.88**
BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations	74.8	80.36
BOW + Online-Behavior + Lexical-Stylistic + POS-Collocations + POS Syntax Collocations	74.73	80.54
Online-Behavior + Interests + Lexical-Stylistic	74.39	77.20
BOW + Online-Behavior + Interests + Lexical-Stylistic	**79.96**	**81.57**
All Features	71.26	74.07[2]

Table 5: Feature Accuracy. The top portion refers to the baselines. The best accuracies are shown in bold. Unless otherwise marked, all accuracies are statistically significant at p<=.0005 for both baselines. [1] not statistically significant over Online-Behavior and Interests. [2] not statistically significant over Interests.

BOW, online-behavior, and lexical-stylistic features combined did best achieving accuracy of 77.45% and 80.88% in 1979 and 1984 respectively. This indicates that our classification method could work well on blogs from any website. It is interesting to note that collocations and POS-collocations were useful, but only when we use 1984 as the split which implies that bloggers born in 1984 and later are more homogeneous.

6 Conclusion and Future Work

We have shown that it is possible to predict the age group of a person based on style, content, and online behavior features with good accuracy; these are all features that are available in any blog. While features representing writing practices that emerged with social media (e.g., capitalized words, abbreviations, slang) do not significantly impact age prediction on their own, these features have a clear change of value across time, with post-social media bloggers using them more often. We found that the birth years that had a significant change in writing style corresponded to the birth dates of college-aged students at the time of the creation/popularity of social media technologies, AIM, SMS text messaging, weblogs, Facebook and MySpace.

In the future we plan on using age and other metadata to improve results in larger tasks such as identifying opinion, persuasion and power by targeting our approach in those tasks to the identified age of the person. Another approach that we will experiment with is the use of ranking, regression, and/or clustering to create meaningful age groups.

7 Acknowledgements

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

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346.

John D. Burger and John C. Henderson. 2006. An exploration of observable features related to blogger age. In *AAAI Spring Symposia*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *In LREC 2006*.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers'

age and gender. In *International AAAI Conference on Weblogs and Social Media*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update.

Susan C. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Susan C. Herring, L.A. Scheidt, S. Bonus, and E. Wright. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences*.

Dan Klein and Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

Dan Klein and Christopher D. Manning. 2003b. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. 2004. Structure and evolution of blogspace. *Commun. ACM*, 47:35–39, December.

Amanda Lenhart, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. 2010. Social media and young adults.

Ian Mackinnon. 2006. Age and geographic inferences of the livejournal social network. In *In Statistical Network Analysis Workshop*.

Andrew Y Ng and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Neural Information Processing Systems*, 2:841–848.

Scott Nowson and Jon Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: language use over the life span. *J Pers Soc Psychol*, 85(2):291–301.

J.W. Pennebaker, R.E. Booth, and M.E. Francis. 2007. Linguistic inquiry and word count: Liwc2007 Đ operatorÕs manual. Technical report, LIWC, Austin, TX.

Perseus Development. 2004. The blogging iceberg: Of 4.12 million hosted weblogs, most little seen and quickly abandoned. Technical report, Perseus Development.

R.W. Robins, K. H. Trzesniewski, J.L. Tracy, S.D Gosling, and J Potter. 2002. Global self-esteem across the lifespan. *Psychology and Aging*, 17:423–434.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523.

J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Jenny Tam and Craig H. Martell. 2009. Age detection in chat. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, ICSC '09, pages 33–39, Washington, DC, USA. IEEE Computer Society.

David H. Urmann. 2009. The history of text messaging.

Xiang Yan and Ling Yan. 2006. Gender classification of weblog authors. In *AAAI Spring Symposium Series on Computation Approaches to Analyzing Weblogs*, pages 228–230.

Kathryn Zickuhr. 2010. Generations 2010.