Gender, Topic, and Audience Response: An Analysis of User-Generated Content on Facebook

Yi-Chia Wang
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA
yichiaw@cmu.edu

Moira Burke
Facebook
1601 Willow Road,
Menlo Park, CA 94025
mburke@fb.com

Robert Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA
robert.kraut@cmu.edu

ABSTRACT
Although users generate a large volume of text on Facebook every day, we know little about the topics they choose to talk about, and how their network responds. Using Latent Dirichlet Allocation (LDA), we identify topics from more than half a million Facebook status updates and determine which topics are more likely to receive audience feedback, such as likes and comments. Furthermore, as previous research suggests that men and women use language for different purposes, we examine gender differences in topics, finding that women tend to share more personal issues (e.g., family matters) and men discuss more general public events (e.g., politics and sports). Post topic predicts how many people will respond to it, and gender moderates the relationship between topic and audience responsiveness.

Author Keywords
Social networking sites; Facebook; computer-mediated communication; gender; topics; natural language analysis

ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces - Web-based interaction.

INTRODUCTION

The majority of Internet users participate in social networking sites (SNS) such as Facebook and Twitter, sharing personal stories, political views, and what they had for lunch [9]. Although users generate a large volume of text on SNS every day, we know little about this content and how user characteristics influence what they talk about. In this paper, we examine whether male and female SNS users talk about different topics, and how their audience of friends and followers respond.

There are longstanding differences in how men and women communicate [2, 14]. Research on face-to-face communication in the early twentieth century documented women's conversations centering on people and relationships, and men's focusing on work and money [2]. Though social roles and technology have changed, persistent differences in how men and women speak have been observed. Computer-mediated communication (CMC) may remove many visual and temporal cues between writer and reader, but writers "give off" gender signals [7], both in discourse style—women are more supportive, use more emoticons indicating smiles and hugs, write shorter posts, and use less profanity—and in topic, with women writing about personal issues and men more likely to write about matters external to their lives [10, 11].

Most research on gender differences in CMC has been performed on discussion groups, games, and blogs, which may attract a self-selected audience of relatively savvy web users, or those focused on a specific topic, but what about SNS like Twitter and Facebook, where nearly everyone is a producer? One's audience may include close friends who want to hear personal news and weaker ties who may only care about shared interests. Furthermore, broadcast content on SNS is typically short (e.g., 140 characters on Twitter), forcing users to choose their text carefully. Given these distinctions, does traditional gendered discourse appear in social network site posts? In the present work, we apply an automated, large-scale approach to examine topic differences between the sexes, including separate analyses of teens and adults, and aim to answer the following research question:

RQ1. Are there gender differences in topics on SNS?

Next, we examine how audiences respond to these different topics. When users broadcast content on SNS, their friends and followers may leave positive feedback, such as @replies or "favorites" on Twitter, or comments and "likes" on Facebook. Previous research has shown that this kind of feedback has positive consequences on the users that receive it, and for the site as a whole. For example, users who receive feedback from their friends feel greater social capital [5] and share more content in the future [6]. How does the poster's language elicit responses? In Usenet newsgroups, a message's rhetorical strategy, language complexity, and word choice are all related to whether it receives a reply [1]. Polite messages receive more responses in technical groups, while rude posts spur longer discussions in political forums [4]. However, we know less about the actual topics people choose to talk about and how those topics are related to feedback. The popular press describes "annoying Facebookers" like the "The Let-Me-Tell-You-Every-Detail-of-My-Day Bore" and "The Sympathy-Baiter" [8], yet do the data support these stereotypes? And do audiences reward certain topics with positive feedback?

RQ2.Which topics are associated with greater audience responsiveness?

Submitted to CHI 2013.

This paper consists of two parts. In order to understand what people talk about on SNS, we first applied machine learning to discover hidden topics of user-generated content on Facebook. We then examined topic differences between men and women, and the impact of content topics on audience responsiveness.

METHOD

Data Collection

Facebook users share many kinds of content, including photos, links, location check-ins, private messages, songs played, and short responses to the prompt, "What's on your mind?" The latter are known as status updates, and are the focus of the present work because they are text-based, directed at more than one friend, and have the potential to receive comments and likes, potential metrics of quality.

We randomly sampled one million English status updates posted by U.S. Facebook users in June 2012 from Facebook's server logs. For each status update, we obtained metadata including post time, number of viewers, and number of comments and likes within three days. Demographic information about the poster was also included: gender, age, friend count, and days since registration. The data processing procedures described below were validated by analyzing the authors' own status updates. All posts in the dataset were analyzed in aggregate for privacy; researchers built models from counts of topic terms, such as those in Table 1.

Text Processing and Topic Modeling

To identify the topics common in Facebook users' status updates, we applied Latent Dirichlet Allocation (LDA). LDA is a statistical generative method that can be used to discover hidden topics in documents as well as the words associated with each topic [3]. It analyzes large amounts of unlabeled documents by clustering words that frequently co-occur and have similar meaning into "topics".

We went through several steps to pre-process and clean the data before constructing topic models. Our experience suggests that this pre-processing and pruning result in far superior topic models than those from unpruned data. Status updates were tokenized with the OpenNLP toolkit [12], stemmed with a Porter stemmer [13], and lowercased. We removed punctuations and replaced URLs, email addresses, and numbers with tags. Updates were then represented as an unordered set of unigrams (single words) and bigrams (word pairs).

Across all terms in the one million status updates, 71% of unigrams only appeared once, and 500 unigrams accounted for 55% of all text. For example, 10% of updates contain "love," the most frequent unigram. Though "love" is a meaningful word, its sheer popularity makes it unhelpful in topic modeling, because so many different terms co-occur with it. Similarly, very low frequency terms are not helpful, as they do not co-occur often enough with other terms to distinguish clear topics. This skew of words is a well-known phenomenon in natural language known as Zipf's law [15]. Therefore, we pruned high and low frequency unigrams and bigrams (those that occurred in more than 0.5% or less than 0.01% of the updates) to reduce noise and vocabulary size. In addition, we excluded all unigrams from a 500-word stopword list (e.g. "the" and "in"); bigrams were filtered if both words were stopwords. After pruning, approximately 50% of the status updates had fewer than eight n-grams; these documents were too short for successful model training. Therefore, we built topic models from the remaining status updates (N=521,636).

Topic	Sample Vocabulary
Sleep	last night, wake up, bed, nap, asleep
Food	lunch, coffee, chicken, ice cream
Clothing	shop, dress, bag, shoe, size, shirt
House	door, window, floor, my house, yard
Work	at work, get back, come home, work on
Weather/travel	road, weather, cold, city, fly, storm
Family fun	great day, kid, swim, cousin, have fun, enjoy
Girlfriend/boyfriend	best friend, boyfriend, my girlfriend, love her
Birthday	happy birthday, love it, anniversary, today I
Father's Day	happy father, father day, daddy, love you
Sports	beat, fan, ball, king, miami, game, player
Politics	country, president, vote, law, tax, obama
Love	my heart, in love, my love, touch, open your
Thankfulness	thank you, god bless, a bless, thank everyone
Anticipation	celebrate, can't wait, so excited, look forward
Asking for support/prayers	worry about, help me, pray for, support, I hope, please pray, faith, advice, favor, cancer
Medical	doctor, hospital, shot, blood, surgery, patient
Memorial	I miss, memory, peace, grandma, rip, wish you
Negativity about people	some people, piss, idiot, annoy, bother, rude
Complaining	I hate, tried of, hate when, don't want, sick of
Deep thoughts	idea, human, goal, universe, achieve, value
Christianity	the lord, church, christ, god is, spirit, amen
Religious imagery	die, star, born, angel, earth, the sun, fear, dark
Slang	yo, em, yu, bro, tryna, rite, cuz, yur, gunna
Swearing	fucking, dumb, dick, a bitch, bullshit, shit

Table 1. Samples of vocabulary in LDA topic dictionaries

To identify topics in status updates, we built an LDA model treating each status update as a document. The model was set to derive 50 latent topics; this parameterization produced models with greater interpretability to human judges than models deriving 10, 30, 60, or 100 topics. Topic dictionaries were generated from the 500 terms most strongly associated with each topic, and two experts familiar with SNS content manually labeled each dictionary (e.g. *Food*). Because the status updates were from a single month, several topics were clearly associated with popular memes in that month. These topics were excluded from analysis because of their limited generalizability, as were topics that were uninterpretable to the judges. In the end there were 25 topics, with representative terms shown in Table 1.

After constructing topic models, we applied the resulting dictionaries to all status updates and considered an update to be "about" that topic if it contained at least three n-grams from the corresponding LDA topic dictionary. For example, the post: *"Weekend plans include camping, rafting, and total domination of the mud obstacle course. West Virginia, consider yourself warned."* would map to the *Weather/travel* topic because it contains the terms raft, west, virginia, and warn. By this standard, 50% of status updates had two topics. This dictionary-based approach was used rather topic distributions of updates resulting from the LDA model because it can be applied quickly and at scale.

RESULTS AND DISCUSSION

Topics of Status Updates

Figure 1(a) shows the distribution of topics, indicating the percentage of status updates about each topic. The overall pattern suggests Facebook users frequently disclose personal information (e.g. *Thankfulness*, *Asking for support*), talk about holidays and family events (*Father's day, Birthday, Family fun*), and wax philosophical (*Deep thoughts, Christianity*). Some details of their daily lives are common (*Work, Sleep*), but most banal topics are less common (*House, Food*).

To understand whether men and women talk about different topics on Facebook, we calculated the percent of topics by each sex about each topic. Fig 1(b) presents these differences ranked by topic popularity for adults aged 25 and older. Women's posts are disproportionately about relationships and personal details (*Father's Day, Family fun, Birthday, Anticipation*), while men are more likely to write about sports and abstract concepts, like *Politics*, *Deep thoughts*, and *Christianity* ($p < 0.001$ for each). Despite the differences in format and audience on SNS, this finding is consistent with previous work on face-to-face communication and blogs.

Teens, on the other hand, were more homogeneous across genders. Figure 1(c) presents the topic distributions between teen girls and boys aged 13 to 17. In contrast to the gender differences among the adults shown in Fig 1(b), the teenage sample had fewer gender differences. For example, *Complaining*, *Girlfriend/Boyfriend*, and *Slang* were the most popular topics for both teen girls and boys. One possible explanation for the topic similarity among teens is the cohort effect: Teens spend most of their time in school, and primarily communicate with other teens online, so they are more susceptible to peer influence and become more similar to each other. However, some patterns seen in adults are evident in teens, as well; teen girls are less likely to talk about sports, and boys are less likely to talk about family events.

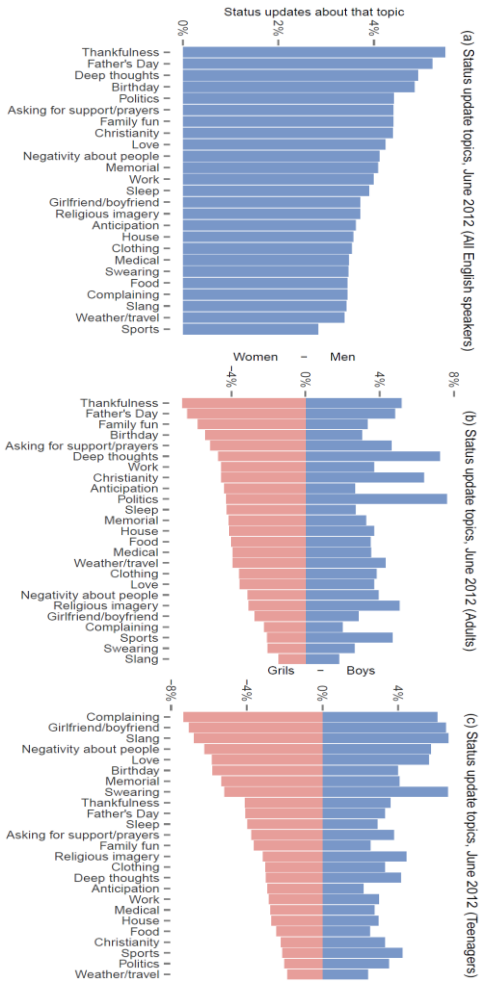

Figure 1. Distribution of status update topics

Gender, Topics and Audience Responsiveness

To determine whether certain topics were associated with greater responsiveness (e.g., likes and comments) from a poster's audience, and whether gender moderates that relationship, we built a linear model on responsiveness.

Audience responsiveness is the dependent variable, operationalized as the number of comments a status update received within three days. Because the distribution of responses was highly skewed, we used the logged number of comments, base 10, after adding a value of 1. Results are qualitatively similar using likes rather than comments, with some exceptions discussed below.

Each **topic** is an independent variable. A status update had a binary value for each of the topics, indicating whether the update was about the topic (1) or not (0).

As previously demonstrated, topic choice is not independent of the author's gender and age, and so we control for these features in the model. That way, we can make claims about the topic, rather than about the poster. We also control for other demographic information, including days since joining Facebook, friend count, and the average number of comments per post the author received the prior week, a rough measure of poster popularity or interestingness. We also controlled for the day of the week the update was posted (1 for weekday; 0 for weekend), the number of times the update had been seen, and the word length of the update.

Table 2 presents a linear regression predicting audience responsiveness. The intercept indicates a message with all variables at their means and all binary variables set to zero, so in this case, an average-length status update written by a woman on the weekend would receive 1.04 comments. Betas represent the effect on comments (plus one and logged base 10) from a one-unit increase in continuous independent variables, or a binary variable having a value of 1. To make the results more interpretable, we include the estimated number of comments an update will receive. For example, Table 2 shows that updates posted by males received .02 fewer comments than by females ($10^{-.006}$). Older posters received more comments, as do posts on weekdays, and longer posts. The poster's previous number of comments received is highly predictive, suggesting that users who produce posts that elicit many responses continue to write "interesting" or evocative content, or that audiences that tend to be responsive in one week continue to be responsive the next. Posts that receive more views receive more comments, a cyclical relationship related to the site's algorithmic ranking of content, such that posts that receive comments are likely to be high-quality, and

are thus more likely to be shown to more people, resulting in even more comments. After controlling for the number of views a post received, the friend count of the poster is negatively correlated with comments.

Topics are listed in order of comments received, with *Medical* posts receiving the most feedback and *Christianity* receiving the least. All topics shown in Figure 1 were included in the model; topics with beta absolute values smaller than 0.025 and those not statistically significant at the $p < 0.001$ level were omitted from Table 2; these topics are generally neutral in terms of audience responsiveness. An otherwise average post *Asking for support or prayers* receives 1.17 comments, while an average post about *Sleep* receives 70% as many comments (0.83 comments). While the results are qualitatively similar using likes rather than comments, there are some exceptions: negatively-valenced topics, such as *Medical*, are associated with fewer likes; more abstract topics, like *Love* and *Christianity*, get more likes but fewer comments, which might be because audience don't know how to respond to them but still want to show their support.

To identify whether gender moderates the relationship between topic and responsiveness, for each topic, we calculated a *male score*, the ratio of the percent of status updates written by men about that topic to the percent of status updates written by women about that topic. By using within-gender percentages, this score takes into account the relative post frequencies of men and women; women posted two times as many updates as men. After that, each update was assigned a post-level male score, the average male score across all of the post's topics. The interaction between gender and the post's male score in Table 2 shows that although men generally receive fewer comments than women, "male" topics generally receive more comments, and the effect is slightly greater for female posters.

		Comments received [1]		
		Beta	Std. Err.	Est. # of Comments
	(Intercept)	0.311***	0.001	1.04
Controls	Male	-0.006***	0.001	1.02
	Age (years) [3]	0.030***	0.001	1.19
	Days since registration [3]	0.010***	0.001	1.09
	Friend count [2]	-0.048***	0.001	0.83
	Comments per prev. post [2]	0.066***	0.001	1.38
	IsWeekday	0.045***	0.001	1.27
	Post views [2]	0.134***	0.001	1.78
	Post length (words) [2]	0.025***	0.001	1.17
Topics	Medical	0.084***	0.002	1.48
	Swearing	0.028***	0.002	1.18
	House	0.028***	0.002	1.18
	Asking for support/prayers	0.026***	0.002	1.17
	Deep thoughts	-0.031***	0.002	0.90
	Family fun	-0.032***	0.002	0.90
	Religious imagery	-0.042***	0.002	0.86
	Love	-0.047***	0.002	0.83
	Sleep	-0.047***	0.002	0.83
	Christianity	-0.069***	0.002	0.75
	Male score of post [3]	0.006***	0.001	1.07
	Male * Male score of post	-0.003**	0.001	1.03

1: Logged (base 10) 2: Logged (base 10), standardized, centered
3: standardized and centered. *p<0.05, **p<0.01, ***p<0.001

Table 2. Linear regression model predicting comments on Facebook status updates based on topic.

CONCLUSION

This study demonstrates that there are demographic differences in topics of user-generated content on SNS. Using topic modeling, we find that women are more likely to broadcast personal issues, while men are more likely to post philosophical topics. Although men get fewer comments than women, "masculine" topics receive more comments. One interpretation is that women are more likely to catch their audience's attention when subtly defying gender expectations.

One limitation of this work is that we do not have the information about the potential audience and people who made the responses. We can only make the claim that men and women have different topic choices when broadcasting content. Future research will examine whether men and women respond differently to topics. The limited data collection period, one month, may generate less generalizable topics than could be generated from a longer time period.

Our findings have several implications. First, our analysis indicates that certain topics evoke more feedback from audiences (e.g. *Asking for support*), while others may be poorer (e.g. *Sleep*). Designers of SNS may want to take these into consideration when designing feed ranking algorithms, promote content with topics that viewers are more likely to respond to. Second, that topics matter for responsiveness suggests that assisting individuals with content construction might improve their experiences on SNS.

REFERENCES

1. Arguello, J., Butler, B., Joyce, E., Kraut, R., Ling, K. S., & Wang, X. (2006). Talk to me: foundations for successful individual-group interactions in online communities. In *CHI 2006*, 959-968.
2. Bischoping, K. (1993). Gender differences in conversation topics, 1922-1990. *Sex Roles,* 28, 1-18.
3. Blei, David M., Ng, Andrew Y., & Jordan, Michael I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.
4. Burke, M. and Kraut, R. (2008). Mind your Ps and Qs: The impact of politeness and rudeness in online communities. In *CSCW 2008*, 281-284.
5. Burke, M., Kraut, R., & Marlow, C. (2011). Social capital on Facebook: Differentiating uses and users. In *CHI 2011*, 571-580.
6. Burke, M., Marlow, C., and Lento, T. (2009). Feed me: Motivating newcomer contribution in social network sites. In *CHI 2009*, 945-954.
7. Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY.
8. Griggs, B. (2008). The 12 most annoying types of Facebooker. *CNN*. Retrieved September 18, 2012, from http://edition.cnn.com/2009/TECH/08/20/annoying.facebook.updaters/
9. Hampton, K. N., Sessions, L., & Her, E. J. (2011). Core networks, social isolation, and new media: How Internet and mobile phone use is related to network size and diversity. *Information, Communication & Society, 14*(1).
10. Herring, S.C. (2000). Gender Differences in CMC: Findings and Implications. *Computer Professionals for Social Responsibility Journal*, Winter.
11. Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459.
12. OpenSource. (2010). OpenNLP: http://opennlp.apache.org.

13. PorterStemmer: http://tartarus.org/martin/PorterStemmer.
14. Tannen, D. (1990). *You just don't understand: Women and men in conversation*. Ballantine, NY: Morrow.
15. Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.