Social and Emotional Correlates of Capitalization on Twitter

Sophia Chan
University of Victoria
schan1@uvic.ca

Alona Fyshe
University of Victoria
afyshe@uvic.ca

Abstract

Social media text is replete with unusual capitalization patterns. We posit that capitalizing a token like THIS performs two expressive functions: it marks a person socially, and marks certain parts of an utterance as more salient than others. Focusing on gender and sentiment, we illustrate using a corpus of tweets that capitalization appears in more negative than positive contexts, and is used more by females compared to males. Yet we find that both genders use capitalization in a similar way when expressing sentiment.

1 Introduction

Gender lines divide language use in speech (Eckert and McConnell-Ginet, 2003); in writing (Koppel et al., 2002); and on social media (Koppel et al., 2006; Bamman et al., 2014). Unsurprisingly, genders differ in their use of emotive language as well (Volkova et al., 2013; Hovy, 2015). Volkova et al. (2013) give the example of *weakness*. Whereas females are more likely to use the word in a positive context, as in *chocolate is my weakness*, males are more inclined to use it when speaking negatively.

Orthographic choices in particular, such as lengthening (*coool*) and coda deletion (*walkin*), have been shown to be socially meaningful (Androutsopoulos, 2000; Eisenstein, 2015) and tied to sentiment (Brody and Diakopoulos, 2011). To the best of our knowledge, however, the use of capitalization has not yet been examined in this context.

Social media text is replete with non-standard capitalization. While many agree that capitalization has some communicative function (Vandergriff, 2013; Nebhi et al., 2015), in practice this information is frequently interpreted as noise and removed by text normalization procedures early on in natural language processing (NLP) pipelines (Eisenstein, 2013).

We posit that capitalization (operationalized here as the number of fully capitalized words in a tweet) has two functions. Capitalizing a token like THIS marks a person socially, and marks certain parts of the utterance as more salient than others. Capitalization thus encodes information about the user and their attitude that can be useful for NLP tasks, such as sentiment analysis.

With these suggested functions in mind, we focus on examining how capitalization patterns vary with respect to two variables: the gender of the user and sentiment of the tweet. We are also interested in possible interaction effects.

Our analysis extends existing literature on orthographic variation in social media, filling the research gap in capitalization. We define a meaningfulness criteria to differentiate between when capitalization is used for convention (e.g. in acronyms) and when it is used creatively to add expressive value, since we are only interested in the latter.

The results indicate that capitalization on Twitter does indeed vary with respect to gender and sentiment, and that effects are strengthened when you consider only *meaningfully* capitalized tokens. We find no interaction effects, suggesting that both genders use capitalization in a similar way when it comes to expressing sentiment.

2 Data

For the purpose of training a gender classifier Burger et al. (2011) built a corpus of approximately 213 million tweets from 18.5 million users and annotated them for gender by following links to users' Facebook or MySpace profiles, where self reporting of gender was required. Volkova et al. (2013) later refined the corpus by excluding re-tweets and non-English tweets, and selecting a random, gender-balanced sample of 1 million tweets. We were able to retrieve 85.50% of

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 10–15
New Orleans, Louisiana, June 6, 2018.

the tweets from this sample using the Twitter API.

Apart from this sample, we collected 1% of all tweets in North America using Twitter's streaming API from January 2017 to July 2017 and randomly sampled a set of 15 million tweets to be used to approximate true frequency distributions.

2.1 Emoticons as sentiment labels

The first step in examining possible interactions between gender and sentiment was to obtain sentiment labels for each tweet. We refrained from relying on text-based features (e.g. "happy" words versus "sad" words) to annotate our gender-labeled dataset for sentiment, as we are interested in examining the distribution of capitalization, a text-based feature itself. Rather, we assumed that the polarity of emoticons found in a tweet is a valid proxy for the sentiment of the tweet.

Table 1: Distribution of gender and sentiment in our dataset of tweets.

	positive		**negative**	
	count	**%**	**count**	**%**
male	4798	25.06	4569	24.01
female	4746	24.94	4945	26.00

For each tweet that contained at least one emoticon, we determined its sentiment by matching emoticons to human-annotated sentiment labels (positive, negative, or neutral) (Hogenboom et al., 2015). From this set, we retained only positive and negative tweets for which there were no conflicts in emoticon sentiment. In other words, we excluded tweets if they contained both positive and negative emoticons.

This process yielded 75,670 tweets labeled for both gender and sentiment. From these tweets, we obtained a random sample of 19,028 tweets balanced across gender (male or female) and sentiment (positive and negative) groups. The distribution of our dataset is summarized in Table 1.

3 Methodology

3.1 Preprocessing

All tweets were tokenized using Natural Language Toolkit (NLTK)'s TweetTokenizer[1] (Bird et al., 2009). We removed non-alphabetic tokens and tokens that consisting of fewer than three characters.

[1] `nltk.tokenize.TweetTokenizer( preserve_case=False, reduce_len=True)`

3.2 Identifying meaningful capitalization

While we claim that capitalization has expressive function, this does not apply across the board to all capitalized tokens. Acronyms, for example, are frequently capitalized by convention to signal to the reader that the token is a stand-in for some longer string, as opposed to being a creative language resource that users can draw on to express themselves.

Nonetheless, it is clear that in certain cases capitalizing a word causes a change in interpretation—as in *that's so cool* versus *that's SO cool*—that may serve the purpose of mimicking real-life conversational cues such as intonation or volume (Vandergriff, 2013).

To operationalize this intuition, we set a threshold designed to filter out acronyms from our data. We obtained counts for how often a token appeared in uppercase and non-uppercase (lowercase or title case) forms in the corpus of 15 million tweets, and called a token meaningfully capitalized if it appeared in its uppercase form less than 10% of the time. The definition for *meaningful* capitalization is shown below.

$$\frac{Count(upper)}{Count(upper) + Count(nonupper)} < 0.1$$

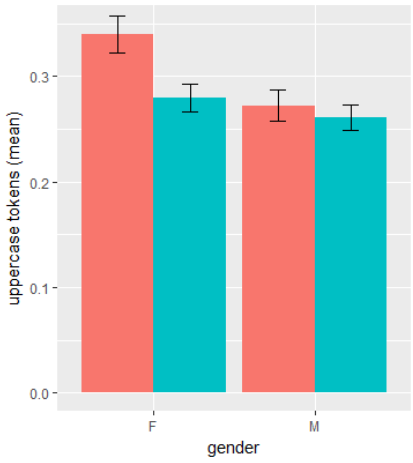
3.3 Analysis

We ran two ANOVAs ($gender \times sentiment$) on our data, using as response variables (1) the number of uppercase tokens and (2) the number of meaningfully capitalized tokens in each tweet, as identified by the metric described in Section 3.1. Data analysis was performed in R 3.4.2 (R Core Team, 2013).

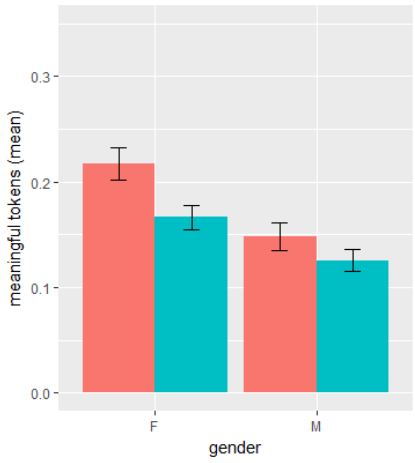
Within the categories of male, female, positive, and negative, we identified tokens that are most likely to be capitalized by calculating each specific token's probability of being capitalized. For example, if *rip* was capitalized 9 times out of 10 in our corpus, it was assigned a probability of 0.9. To reduce noise in our findings, we only considered tokens that appeared at least 10 times within the category under analysis. We also identified tokens most likely to be *meaningfully* capitalized.

4 Results

The mean number of capitalized tokens and *meaningfully* capitalized tokens for each group are shown in Figures 1a and 1b, respectively. Across



(a) The mean number of capitalized tokens in each tweet (without our meaningfulness criteria applied) across sentiment and gender. Within female tweets, the mean is **0.28** in positive contexts and **0.34** in negative contexts. For male tweets, the mean is **0.26** in positive contexts and **0.27** in negative contexts.



(b) The mean number of *meaningfully* capitalized tokens in each tweet, across sentiment and gender. Within female tweets, the mean is **0.17** in positive contexts and **0.22** in negative contexts. For male tweets, the mean is **0.13** in positive contexts and **0.15** in negative contexts.

both genders, capitalization is employed more in negative contexts.

As shown in Table 2, we find a main effect of both gender (p <0.01) and sentiment (p <0.05) for capitalized tokens, but no interaction. Similarly, Table 3 displays main effects of gender (p <0.001) and sentiment (p <0.01) for meaningfully capitalized tokens and no interaction.

Table 4 shows the 10 tokens most likely to be capitalized, and to be *meaningfully* capitalized within each gender and sentiment category.

5 Discussion and conclusion

Our results in Table 2 show that capitalization varies systematically with respect to gender and sentiment, but that these two factors do not interact. On average, capitalization is used more by females, and used to express negativity as opposed to positivity.

Crucially, the use of capitalization functions as both a marker of identity and a marker of sentiment, following a similar pattern to other types of non-standard orthography, such as lengthening or phonologically-motivated variation (Brody and Diakopoulos, 2011; Eisenstein, 2015).

We also provide an operational definition of meaningful capitalization. A token was considered *meaningfully* capitalized if, in a corpus of 15 million tweets, it was capitalized less than 10% of the time.

The value of our meaningfulness criteria can be seen by comparing capitalized to *meaningfully* capitalized tokens in Table 4. Acronyms such as *rip*, *nyc*, *dvd* are stripped out. Because these tokens are capitalized out of convention, orthography does not reflect user attributes or attitudes.

Several abbreviations appear in the meaningful columns in Table 4, such as *lol*, *lmao*, and *smh*. Our intuition is that people have stopped uppercasing these for the most part, probably due in part to their high frequency. In fact, it has been suggested that the status of *lol* is shifting from abbreviation to discourse marker (Tagliamonte and Denis, 2008; Markman, 2017). Our threshold of 10% appears to filter out most acronyms in our data, but it would be valuable to systematically test different thresholds to quantitatively validate our method. We leave this for future work.

The use of capitalization may serve another function in addition to signaling acronyms and encoding user attitudes. If a token can refer to multiple entities, capitalization can help differentiate one meaning from another, allowing users to refer, say, to the band *TOOL* as opposed to the category of *tools*. While we were not interested in detecting such cases, the insight that capitalization has functions beyond what is discussed here provides future avenues for research.

As shown in Table 3, the effects of gender and sentiment are stronger when we apply our mean-

Table 2: ANOVA table for testing the significance of all capitalized tokens, without our meaningfulness criteria applied. We find a main effect of sentiment and gender, but no interaction. * = p <0.05 and ** = p <0.01.

	sum of squares	mean square	F	p
gender	9.0	9.044	8.085	**.003** **
sentiment	6.2	6.616	5.999	**.014** *
gender:sentiment	2.8	2.827	2.755	.097

Table 3: ANOVA table for testing the significance of meaningfully capitalized tokens. We find a main effect of sentiment and gender, but no interaction. Using our meaningfully capitalized token filter increases the margin of significance for gender and sentiment. ** = p <0.01 and *** = p <0.001.

	sum of squares	mean square	F	p
gender	14.7	14.686	19.319	**.000** ***
sentiment	6.6	6.587	8.665	**.003** **
gender:sentiment	0.9	0.989	1.222	.269

ingfulness criteria, corroborating our intuition that we need to consider each token separately, taking its capitalization distribution into account in order to differentiate between capitalization as convention, and capitalization as a creative resource.

This study was limited by the availability of Twitter data that are labeled for both gender and sentiment. Alongside, our dataset is composed entirely of tweets that contain emoticons, which may be biasing the sample towards users who are predisposed to use language (including capitalization) in a specific way. By selecting tweets on the basis of whether they contain emoticons, we may be introducing age, gender, and/or sentiment biases. In a study involving blogging data, for example, Rosenthal and McKeown (2011) found that younger users were more likely to use both emoticons and capitalization. In the future, these biases could be mitigated by incorporating human-annotated sentiment labels.

We suspect that capitalization is a type of conversational cue which serves to clarify the meaning of an utterance over text-based communication and help the reader select one of the possible interpretations. According to Vandergriff (2013), these cues are difficult to study because they are often "subtle, highly variable, and relatively infrequent".

Notwithstanding these limitations, our analysis suggests that capitalization encodes information about speaker attributes and attitudes, calling into question the pervasive practice of complete lowercasing in NLP.

Our work displays a computational approach for analyzing the special orthographic characteristics that permeate social media, and positions capitalization as a type of orthographic variation that warrants further, and more detailed analyses in terms of function and distribution. The use of capitalization may be related to other demographic factors, such as age, and may serve different functions depending on the context it appears in.

Acknowledgments

During the course of this work Sophia Chan received support from an NSERC Undergraduate Research Award (USRA).

We would like to thank Alexandra D'Arcy and David Medler for their insightful discussion and feedback, as well as Sam Liu for collecting and sharing with us a large corpus of Twitter data.

References

Jannis K Androutsopoulos. 2000. Non-standard spellings in media texts: The case of german fanzines. *Journal of Sociolinguistics*, 4(4):514–533.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Table 4: For each category, we show the 10 tokens most likely to be capitalized (**upper**) compared to the 10 tokens that are most likely to be meaningfully capitalized (**mng-ful**). A token is considered meaningful if it occurs in uppercase less than 10% of the time in a corpus of 15 million tweets.

male		**female**		**positive**		**negative**	
upper	**mng-ful**	**upper**	**mng-ful**	**upper**	**mng-ful**	**upper**	**mng-ful**
rip	tool	nyc	lmao	tool	tool	rip	lmao
tool	nom	dvd	smh	nyc	lol	nyc	smh
omg	lol	omg	lol	dvd	lmao	omg	lol
nom	lmao	huge	goodness	omg	halloween	asap	entire
wtf	thx	hahah	kill	huge	exactly	lmfao	burn
yay	note	wtf	none	lol	heck	gah	concert
lol	fire	lmao	fuck	lmao	damn	wtf	tour
btw	goin	smh	bless	hahahaha	hugs	lmao	none
lmao	joke	lol	nuts	halloween	ice	smh	yikes
thx	idk	hahahaha	jonas	exactly	note	aim	fail

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570. Association for Computational Linguistics.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Penelope Eckert and Sally McConnell-Ginet. 2003. *Gender and language*. Cambridge: Cambridge University Press.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 359–369.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1&2):22–40.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 752–762.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 1–7.

Kris M Markman. 2017. Exploring the pragmatic functions of the acronym lol in instant messenger conversations.

Kamel Nebhi, Kalina Bontcheva, and Genevieve Gorrell. 2015. Restoring capitalization in# tweets. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1111–1115. ACM.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Sali A Tagliamonte and Derek Denis. 2008. Linguistic ruin? lol! instant messaging and teen language. *American speech*, 83(1):3–34.

Ilona Vandergriff. 2013. Emotive communication online: A contextual analysis of computer-mediated

communication (cmc) cues. *Journal of Pragmatics*, 51:1–12.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.