Non-lexical Features Encode Political Affiliation on Twitter

Rachael Tatman
Amandalynne Paullada
University of Washington
Linguistics Department
rctatman@uw.edu
paullada@uw.edu

Leo G. Stewart
University of Washington
Human Centered Design
Engineering
lgs17@uw.edu

Emma S. Spiro
University of Washington
Information School
Department of Sociology
espiro@uw.edu

Abstract

Previous work on classifying Twitter users' political alignment has mainly focused on lexical and social network features. This study provides evidence that political affiliation is also reflected in features which have been previously overlooked: users' discourse patterns (proportion of Tweets that are retweets or replies) and their rate of use of capitalization and punctuation. We find robust differences between politically left- and right-leaning communities with respect to these discourse and sub-lexical features, although they are not enough to train a high-accuracy classifier.

1 Introduction

Characterizing social media users based on their political affiliation is an ongoing challenge in Natural Language Processing and Computational Social Science (Conover et al., 2011; Cohen and Ruths, 2013; Sylwester and Purver, 2015; Wong et al., 2016). In addition, linguistic reflections of political identity are of interest to sociolinguists (Hall-Lew et al., 2010; Labov, 2011). However, the approaches of these two communities of researchers with respect to identifying political affiliation are somewhat different. Large-scale computational work has generally focused on the classification of Twitter users based on social network and lexical features. Conover et al. used unigrams (excluding punctuation) and social networks (Conover et al., 2011), while Cohen and Ruths used a large feature set including words, stems, bi- and trigrams, and hashtags (Cohen and Ruths, 2013). Sylwester and Purver, who were interested in characterizing psychological differences between Democrats and Republicans, focused on word frequency, friend-follower ratio and Linguistic Inquiry and Word Count (Pennebaker et al., 2001)–although they also excluded punctuation from their data. Another study by Wong et al. used no linguistic features at all, relying instead on social network relations with users whose political affiliation was known (Wong et al., 2016).

Much of the sociolinguistic work, on the other hand, has focused on sub-lexical features that encode political identity. Hall-Lew et al., for instance, found that American political party affiliation was strongly associated with whether a speaker produced the final syllable in "Iraq" to rhyme with "rock" or "rack" (Hall-Lew et al., 2010). Kirkham and Moore found that British politician Ed Miliband modulated his use of t-glottalling depending on his audience (Kirkham and Moore, 2016).

While the bulk of the sociolinguistic work has focused on speech, there is a growing body of evidence that, unsurprisingly, sociolinguistic variation is also reflected in text (Eisenstein, 2015; Grieve, 2016; Nguyen, 2017). Punctuation in particular has been used as a feature in a variety of tasks, including authorship identification (Chaski, 2005; Abbasi and Chen, 2005) and predicting users' gender (Bamman et al., 2012) and personality (Pennebaker et al., 2015; Golbeck et al., 2011). In addition to punctuation, there is some evidence that variation in capitalization is an important stylistic feature in informal computer-mediated communication (Ling, 2005).

What has not been investigated is whether these sub-lexical text features, like capitalization and punctuation, vary with users' political affiliation. Our central question is this: while earlier work shows that it possible to identify a user's political affiliation with high accuracy using lexical and social-network features, can we also do so using sub-lexical features and without relying on social

network relationships?

This approach has several advantages. The main one is the promise of a classifier that will remain accurate over time. One reason for word-based models' high accuracy is that they are capturing underlying differences in the topics each community is discussing. However, given that the topics of political discussion change frequently, these models may only be useful for a limited time frame. There is little reason, however, to suppose that non-lexical features (like patterns of use of capitalization or punctuation) would change at the same rate. In addition, if the features proposed here can successfully be applied to classifying political alignment, they may prove useful in identifying troll accounts. If a user from one political affiliation creates a fake account for the purpose of trolling users of an opposing political affiliation, they may consciously adopt vocabulary and hashtags from the community they intend to impersonate. However, it is possible that these users will not be adopt stylistic norms of capitalization and punctuation, which may aid in identifying them.

2 Data

Our data, including collection and clustering methods, are borrowed from (Stewart et al., under review). Using the Twitter Streaming API, we collected Tweets containing the terms 'shooting,' 'shooter,' 'gun shot,' or 'gun man,' as well as plural and contracted forms of each term, in order to control for topic. This collection lasted roughly nine months, from December 31, 2015 to October 5, 2016 and yielded 58,812,322 Tweets. From this larger set of Tweets, we selected all Tweets containing "#blacklivesmatter", "#bluelivesmatter" or "#alllivesmatter" (the first strongly indicative of Left-leaning politics, the latter ones more characteristic of the Right), which left us with a smaller dataset of 248,719 Tweets. Each of these Tweets contains both a shooting-related term and one of the three hashtags.gun

We next collected user data to construct a social graph. We collected only the user data for the 8,524 users who contributed at least four Tweets to the sampled dataset. For each user, we collected their followers list, capped at 100,000 followers. Followers were collected between one and three months after the end of Tweet collection: November 15, 2016 to January 17, 2017.

Using the follower data from the 8,524 users, we constructed the shared audience graph in Figure 1. In comparison to friend/follower networks, the shared audience network elicits communities of shared attention (i.e. audience), or potential influence. In this graph, each node is an account, and each edge represents the shared audience between two accounts. The shared audience metric is defined as the Jaccard similarity of followers lists (audiences) for any two accounts (see Equation 1). To prioritize the strongest connections while preserving the nuances of smaller edge weights, we select the top 20th percentile of edges by edge weight, or roughly 5 million of the 25 million original edges. Of the 5 million edges, the minimum edge weight represented an audience overlap of 1.78%.

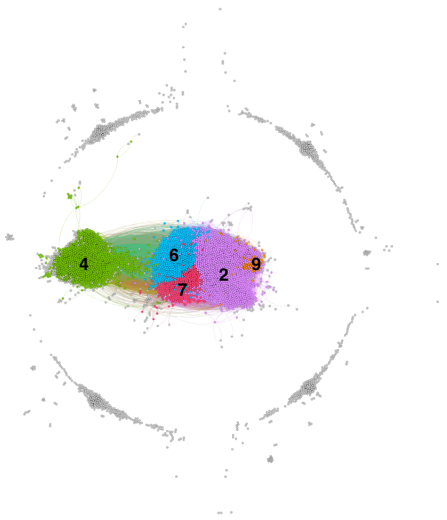

Figure 1: Shared Audience clusters.

$$jaccard(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \quad (1)$$

Our final step in constructing our graph was using Louvain clustering to elicit closely connected communities (clusters) (Blondel et al., 2011). We used Gephi (Bastian et al., 2009) to run the clustering algorithm and visualize the resulting graph. As shown in Figure 1, the clustering algorithm produced five large clusters, along with a multitude of smaller clusters and disconnected nodes.

Our analysis focuses on the five most prominent clusters. For each of these clusters, we identify them by the most commonly used hashtags in user account descriptions, shown in Table 2. Based on the frequent use of such hashtags as "#feelthebern" and "#imwithher," which refer

Cluster ID	Size	Common hashtags
4	2153	#imwithher, #feelthebern
2, 6, 7, 9	4689	#maga, #trump2016

Table 1: Clusters and commonly-used hashtags in user account descriptions from each cluster.

to support for 2016 Democratic Party presidential primary candidates Bernie Sanders and Hillary Clinton, respectively, we define cluster 4 as largely left-leaning, while clusters 2, 6, 7, and 9 are largely right-leaning, as evident by frequent use of "#trump2016" and "#maga," an acronym for Donald Trump's campaign slogan "Make America Great Again." For the binary classification task, we define cluster 4 as the Left and collapse clusters 2, 6, 7, and 9 into a composite Right category.

3 Features

Four features were calculated on a per-user basis: the proportion of Tweets that were replies, the proportion that were retweets, the average number of punctuation marks per Tweet and the average number of capital letters per Tweet.

3.1 Discourse Features

The first two are discourse features that may represent group interaction norms. A higher proportion of replies suggests that a user is engaging in a more conversations (compared to broadcasting), while a higher proportion of retweets suggests that a user is instead amplifying other users.

While there was not a significant difference between the Left and Right Twitter accounts in terms of retweets ($t(3942)=-3.06$, $p > 0.0001$), there was a very robust difference in proportion of replies ($t(3656)=6.45$, $p < 0.0001$). This can be seen in Figure 2. In particular, users from the Right were more likely to have no replies in the dataset than users from the Left.

3.2 Punctuation and Capitalization

Punctuation, as discussed above, is an established feature in text analysis. While most analyses look at the use of individual punctuation characters, in order to maintain parallelism with capitalization we instead used the average number of punctuation marks per Tweet for each user. This calculation was done on Tweets which had URLs and mentions (which contain the @ symbol) removed.

Capitalization was included as a feature based on empirical observations of differences between

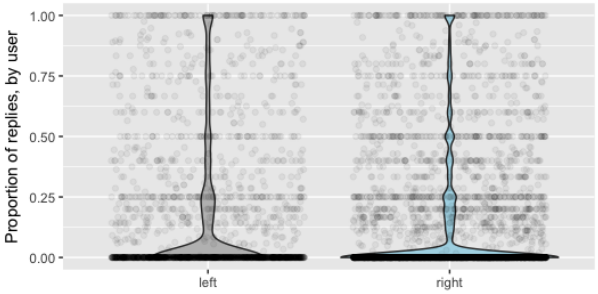

Figure 2: Proportion of Tweets that are replies, per user and per affiliation. A greater number of right-affiliated users have a smaller proportion of replies, i.e., replies make up a relatively small proportion of their Tweets.

these communities. Less capitalization is associated with an informal, casual or nonchalant writing style, but also seems to be a marker of Left-leaning identity. This is explicitly discussed in a viral Tweet (currently >120 thousand favorites) by Twitter user @PatrickCharlto5. The Tweet reads "when you accidentally type a capital letter at the beginning of a sentence" with an attached stock photo of a man with his head in his hands, with the caption "oh no my aloof and uninterested yet woke and humorous aesthetic" (Charlton, 2017). The term "woke" refers to an awareness of social justice issues that are especially prevalent in Left-leaning communities, and the Tweet directly indexes the evocation of the "woke aesthetic" via casual writing style.

Users in these two group used significantly different amounts of both punctuation ($t(5006)=-6.22$, $p < 0.0001$) and capitalization ($t(4465) = -16.051$, $p < 0.0001$). The distribution of users by group can be seen in Figure 3. In keeping with earlier observations, users from the Right tended to use more capitalization and more punctuation marks. In addition there was a strong positive correlation between the amount of punctuation and the amount of capitalization used per Tweet over all users ($r(6831)= .33$, $p < 0.0001$). This covariance suggest that these may both reflect the same underlying stylistic differences.

Our findings have interesting implications in

that they suggest that Left-aligned Twitter users, whether consciously or not, adopt a casual writing style more than Right-aligned users do. We do not have information on age or education level, which may be confounding factors in stylistic choices online.

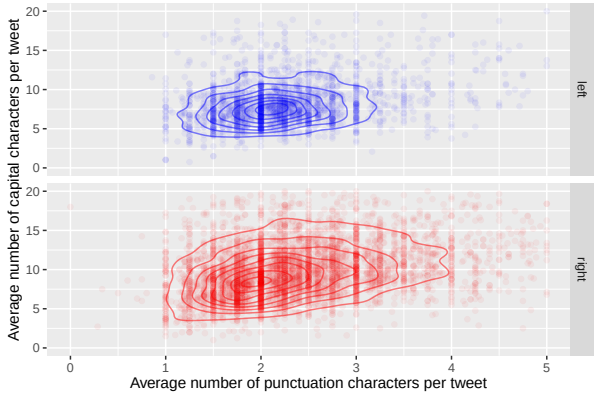

Figure 3: Use of punctuation and capitalization by affiliation. Each dot represents the average number of punctuation marks and capital letters per Tweet by an individual user. Right-affiliated users tend to use more punctuation and capitalization overall.

4 Classification

With the exception of the proportion of a user's Tweets which are retweets, all of the features discussed above are robustly different between these communities. However, it is possible that these differences are not great enough to aid in classification. In order to asses this, we constructed two classifiers were trained using the significant features discussed above.

Because the number of users from the Left and the Right are imbalanced in full data set, we trained and tested on a balanced subset of the data. We randomly sampled users from the Right to create a subset that had as many users as the Left. 90% of each subset was assigned to training set, and the remaining 10% was used as the test data for cross-validation.

Both an SVM and KNN were trained and evaluated in R (using the e1071 (Meyer et al., 2015) and Class (Venables and Ripley, 2002) packages, respectively). To select K for the KNN classifier, models were trained with K's of 1 through 200 (inclusive) and the most accurate selected, in this case 77.

Table 2: Though neither of our classifiers beat the state-of-the art, they did classify users well above chance using only three non-lexical features.

Study	**Accuracy**
Conover	87%
Cohen (politically active accounts)	84%
Wong (no linguistic features)	94%
KNN classifier (this study)	64%
SVM classifier (this study)	65%

As can be seen in Table 2, neither model reached the same accuracy as those used in earlier work. However, both models classified the political affiliation of accounts in the test set at well above chance. Results would likely be improved by incorporating other features known to aid in predicting political affiliation.

5 Conclusion

This study provided evidence that certain discourse and character-level features are sociolinguistically active markers that vary with users' political affiliation. This suggests several interesting areas for future work, especially in looking at the sociolinguistic role of sub-lexical text features.

We have also shown that it is possible to classify Twitter users' political affiliation well above chance without using lexical or social network features. Further work is necessary to determine whether the features discussed here are stable over time. It is possible that they may be more stable than lexical features, especially if the latter are capturing differences in what topics each community discusses. These results strongly suggest that researchers looking at political affiliation should reconsider stripping punctuation from Tweets, as they contain useful information on community norms.

Finally, it should be noted that the analysis in this paper was done on Tweets which contained hashtags. This is an important consideration, as previous work has found that Tweets which contain hashtags are less likely to include sociolinguistically-marked forms, even if the user uses them in other Tweets (Shoemark et al., 2017; Goldman, 2017). Rather than invalidating these results, however, this strengthens them: if this sociolinguistic variation survives in an environment which discourages the use of social markers, this suggests that it is very robust.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20(5):67–75.

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

Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. 2009. Gephi: an open source software for exploring and manipulating networks. *ICWSM* 8:361–362.

VD Blondel, JL Guillaume, R Lambiotte, and E Lefebvre. 2011. The louvain method for community detection in large networks. *J of Statistical Mechanics: Theory and Experiment* 10:P10008.

Patrick W. Charlton. 2017. "when you accidentally type a capital letter at the beginning of a sentence" Tweet ID: 846179708765130763.

Carole E Chaski. 2005. Whos at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence* 4(1):1–13.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on twitter: It's not easy! In *ICWSM*.

Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, pages 192–199.

Jacob Eisenstein. 2015. Written dialect variation in online social media. *Charles Boberg, John Nerbonne, and Dom Watt, editors, Handbook of Dialectology. Wiley* .

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, pages 149–156.

Nora Goldman. 2017. #yesallwomens language: construction feminist identity on twitter. Annual meeting of the Linguistics Society of America. https://www.youtube.com/watch?v=wIpUNP52qVw.

Jack Grieve. 2016. *Regional variation in written American English*. Cambridge University Press.

Lauren Hall-Lew, Elizabeth Coppock, and Rebecca L Starr. 2010. Indexing political persuasion: variation in the iraq vowels. *American Speech* 85(1):91–102.

Sam Kirkham and Emma Moore. 2016. Constructing social meaning in political discourse: Phonetic variation and verb processes in ed miliband's speeches. *Language in Society* 45(01):87–111.

William Labov. 2011. *Principles of linguistic change, cognitive and cultural factors*, volume 3. John Wiley & Sons.

Rich Ling. 2005. The sociolinguistics of sms: An analysis of sms use by a random sample of norwegians. In *Mobile communications*, Springer, pages 335–349.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2015. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. https://CRAN.R-project.org/package=e1071.

Dong-Phuong Nguyen. 2017. *Text as social and cultural data: a computational perspective on variation in text*. Ph.D. thesis, University of Twente.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 1239–1248. http://www.aclweb.org/anthology/E17-1116.

Leo G. Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. under review. From aggregation to virtual armbands: The shifting roles of hashtags in organizing political activism on twitter .

Karolina Sylwester and Matthew Purver. 2015. Twitter language use reflects psychological differences between democrats and republicans. *PloS one* 10(9):e0137422.

W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2016. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering* 28(8):2158–2172.