Analyzing Linguistic Differences between Owner and Staff Attributed Tweets

Daniel Preoţiuc-Pietro
Bloomberg LP
dpreotiucpie@bloomberg.net

Rita Devlin Marier
Bloomberg LP
rdevlin5@bloomberg.net

Abstract

Research on social media has to date assumed that all posts from an account are authored by the same person. In this study, we challenge this assumption and study the linguistic differences between posts signed by the account owner or attributed to their staff. We introduce a novel data set of tweets posted by U.S. politicians who self-reported their tweets using a signature. We analyze the linguistic topics and style features that distinguish the two types of tweets. Predictive results show that we are able to distinguish between owner and staff attributed tweets with good accuracy, even when not using any training data from that account.

1 Introduction

Social media has become one of the main venues for breaking news that come directly from primary sources. Platforms such as Twitter have started to play a key role in elections (Politico, 2017) and have become widely used by public figures to disseminate their activities and opinions. However, posts are rarely authored by the public figure who owns the account; rather, they are posted by staff who update followers on the thoughts, stances and activities of the owner.

This study introduces a new application of Natural Language Processing: predicting which posts from a Twitter account are authored by the owner of an account. Direct applications include predicting owner authored tweets for unseen users and can be useful to political or PR advisers to gain a better understanding on how to craft more personal or engaging messages.

Past research has experimented with predicting user types or traits from tweets (Pennacchiotti and Popescu, 2011; McCorriston et al., 2015). However, all these studies have relied on the assumption that tweets posted from an account were all written by the same person. No previous study has looked at predicting which tweets from the *same* Twitter account were authored by different persons, here staffers or the owner of the Twitter account. Figure 1 shows an example of a U.S. politician who signs their tweets by adding '-PM' at the end of the tweet.

Figure 1: Example of a politician account where signed tweets are attributed to the account owner.

Staff posts are likely to be different in terms of topics, style, timing or impact to posts attributed to the owner of the account. The goal of the present study is thus to:

- analyze linguistic differences between the two types of tweets in terms of words, topics, style, type and impact;
- build a model that predicts if a tweet is attributed to the account owner or their staff.

To this end, we introduce a novel data set consisting of over 200,000 tweets from accounts of 147 U.S. politicians that are attributed to the owner or their staff.[1] Evaluation on unseen accounts leads to an accuracy of up to .741 AUC. Similar account sharing behaviors exists in several other domains such as Twitter accounts of entertainers (artists, TV hosts), public figures or CEOs who employ staff to author their tweets or with organi-

[1]The data is available at: https://github.com/danielpreotiuc/signed-tweets

zational accounts, which alternate between posting messages about important company updates and tweets about promotions, PR activity or customer service. Direct applications of our analysis include automatically predicting staff tweets for unseen users and gaining a better understanding on how to craft more personal messages which can be useful to political or PR advisers.

2 Related Work

Several studies have looked at predicting the type of a Twitter account, most frequently between individual or organizational, using linguistic features (De Choudhury et al., 2012; McCorriston et al., 2015; Mac Kim et al., 2017). A broad literature has been devoted to predicting personal traits from language use on Twitter, such as gender (Burger et al., 2011), age (Nguyen et al., 2011), geolocation (Cheng et al., 2010), political preference (Volkova et al., 2014; Preoţiuc-Pietro et al., 2017), income (Preoţiuc-Pietro et al., 2015), impact (Lampos et al., 2014), socioeconomic status (Aletras and Chamberlain, 2018), race (Preoţiuc-Pietro and Ungar, 2018) or personality (Schwartz et al., 2013a; Preoţiuc-Pietro et al., 2016).

Related to our task is authorship attribution, where the goal is to predict the author of a given text. With few exceptions (Schwartz et al., 2013b), this was attempted on larger documents or books (Popescu and Dinu, 2007; Stamatatos, 2009; Juola et al., 2008; Koppel et al., 2009). In our case, the experiments are set up as the same binary classification task regardless of the account (owner vs. staffer) which, unlike authorship attribution, allows for experiments across multiple user accounts. Additionally, in most authorship attribution studies, differences between authors consist mainly of the topics they write about. Our experimental setup limits the extent to which topic presence impacts the prediction, as all tweets are posted by US politicians and within the topics of the tweets from an account should be similar to each other. Pastiche detection is another related area of research (Dinu et al., 2012), where models are trained to distinguish between an original text and a text written by one who aims to imitate the style of the original author, resulting in the documents having similar topics.

3 Data

We build a data set of Twitter accounts used by both the owner (the person who the account represents) and their staff. Several Twitter users attribute the authorship of a subset of their tweets to themselves by signing these with their initials or a hashtag, following the example of Barack Obama (Time, 2011). The rest of the tweets are implicitly attributed to their staff.

Thus, we use the Twitter user description to identify potential accounts where owners sign their tweets. We collect in total 1,365 potential user descriptions from Twitter that match a set of keyphrases indicative of personal tweet signatures (i.e., *tweets by me signed*, *tweets signed*, *tweets are signed*, *staff unless noted*, *tweets from staff unless signed*, *tweets signed by*, *my tweets are signed*). We then manually check all descriptions and filter out those not mentioning a signature, leaving us with 628 accounts. We aim to perform our analysis on a set of users from the same domain to limit variations caused by topic and we observe that the most numerous category of users who sign their messages are U.S. politicians, which leaves us with 147 accounts. We download all the tweets posted by these accounts that are accessible through the Twitter API (a maximum of 3,200). We remove the retweets made by an account, as these are not attributed to either the account owner or their staff. This results in a data set with a total of 202,024 tweets.

We manually identified each user's signature from their profile description. To assign labels to tweets, we automatically matched the signature to each tweet using a regular expression. We remove the signature from all predictive experiments and feature analyses as this would make the classification task trivial. In total, 9,715 tweets (4.8% of the total) are signed by the account owners. While our task is to predict if a tweet is attributed to the owner or its staff, we assume this as a proxy to authorship if account owners are truthful when using the signature in their tweets. There is little incentive for owners to be untruthful, with potentially serious negative ramifications associated with public deception.

We use DLATK, which handles social media content and markup such as emoticons or hashtags (Schwartz et al., 2017). Further, we anonymize all usernames present in the tweet and URLs and replace them with placeholder tokens.

4 Features

We use a broad set of linguistic features motivated by past research on user trait prediction (Preoţiuc-Pietro et al., 2015, 2017) in our attempt to predict and interpret the difference between owner and staff attributed tweets. These include:

LIWC. Traditional psychology studies use a dictionary-based approach to representing text. The most popular method is based on Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) consisting of 73 manually constructed lists of words (Pennebaker et al., 2015) including some specific parts-of-speech, topical or stylistic categories. Each message is thereby represented as a frequency distribution over these categories.

Word2Vec Clusters. An alternative to LIWC is to use automatically generated word clusters. These clusters of words can be thought of as *topics*, i.e., groups of words that are semantically and/or syntactically similar. The clusters help reduce the feature space and provide good interpretability. We use the method by Preoţiuc-Pietro et al. (2015) to compute topics using Word2Vec similarity (Mikolov et al., 2013) and spectral clustering (Shi and Malik, 2000; von Luxburg, 2007) of different sizes. We present results using 200 topics as this gave the best predictive results. Each message is thus represented as an unweighted distribution over clusters.

Sentiment & Emotions. We also investigate the extent to which tweets posted by the account owner express more or fewer emotions. The most popular model of discrete emotions is the Ekman model (Ekman, 1992; Strapparava and Mihalcea, 2008; Strapparava et al., 2004) which posits the existence of six basic emotions: anger, disgust, fear, joy, sadness and surprise. We automatically quantify these emotions from our Twitter data set using a publicly available crowd-sourcing derived lexicon of words associated with any of the six emotions, as well as general positive and negative sentiment (Mohammad and Turney, 2010, 2013). Using these models, we assign sentiment and emotion probabilities to each message.

Unigrams. We use the bag-of-words representation to reduce each message to a normalised frequency distribution over the vocabulary consisting of all words used by at least 20% of the users (2,099 words in total). We chose this smaller vocabulary that is more representative of words used by a larger set of users such that models would be able to transfer better to unseen users.

Tweet Features. We compute additional tweet-level features such as: the length in characters and tokens (**Length**), the type of tweet encoding if this is an @-reply or contains a URL (**Tweet Type**), the time of the tweet represented as a one-hot vector over the hour of day and day of week (**Post Time**) and the number of retweets and likes the tweet received (**Impact**). Although the latter features are not available in a real-time predictive scenario, they are useful for analysis.

5 Prediction

Our hypothesis is that tweets attributed to the owner of the account are different than those attributed to staff, and that these patterns generalize to held-out accounts not included in the training data. Hence, we build predictive models and test them in two setups. First, we split the users into ten folds. Tweets used in training are all posted by 80% of the users, tweets from 10% of the users are used for hyperparameter tuning and tweets from the final 10% of the users are used in testing (**Users**). In the second experimental setup, we split all tweets into ten folds using the same split sizes (**Tweets**). We report the average performance across the ten folds. Due to class imbalance – only 4.8% of tweets are posted by the account owners – results are measured in ROC AUC, which is a more suitable metric in this setup.

In our predictive experiments, we used logistic regression with Elastic Net regularization. As features, we use all feature types described in the previous section separately as well as together using a logistic regression model combining all feature sets (**Combined**). The results using both experimental setups – holding-out tweets or users – are presented in Table 1.

Results show that we can predict owner tweets with good performance and consistently better than chance, even when we have no training data for the users in the test set. The held-out user experimental setup is more challenging as reflected by lower predictive numbers for most language features, except for the LIWC features. One potential explanation for the high performance of the LIWC features in this setup is that these are low dimensional and are better at identifying general patterns which transfer better to unseen users rather than overfit the users from the training data.

	ROC AUC	
Feature Set	**Users**	**Tweets**
Majority Class	.500	.500
Tweet Features		
Length	.619	.664
Tweet Type	.654	.660
Post Time	.554	.585
Impact	.573	.718
LIWC	.720	.724
W2V Clusters	.676	.744
Sentiment & Emotions	.568	.567
Unigrams	.649	.857
Combined	.741	.872

Table 1: Predictive results with each feature type for classifying tweets attributed to account owners or staffers, measured using ROC AUC. Evaluation is performed using 10-fold cross-validation by holding out in each fold either: 10% of the tweets (**Tweets**) or all tweets posted by 10% of the users (**Users**).

6 Analysis

In this section we investigate the linguistic and tweet features distinctive of tweets attributed to the account owner and to staff. A few accounts are outliers in the frequency of their signed tweets, with up to 80% owner attributed tweets compared to only 4.8% on average. We perform our analysis on a subset of the data, in order for our linguistic analysis not to be driven by a few prolific users or by any imbalance in the ratio of owner/staff tweets across users. The data set is obtained as follows. Each account can contribute a minimum of 10, maximum of 100 owner attributed tweets. We then sample staff attributed tweets from each account such that these are nine times the number of tweets signed by the owner. Newer messages are preferred when sampling. This leads to a data set of 28,150 tweets with exactly a tenth of them attributed to the account owners (2,815).

We perform analysis of all previously described feature sets using Pearson correlations following Schwartz et al. (2013a). We compute Pearson correlations independently for each feature between its distribution across messages (features are first normalized to sum up to unit for each message) and a variable encoding if the tweet was attributed to the account owner or not. We correct for multiple comparisons using Simes correction.

Top unigrams correlated with owner attributed tweets are presented in Table 3, with the other group textual features (LIWC categories, Word2Vec topics and emotion features) in Table 2. Tweet feature results are presented in Table 4.

LIWC Features		
r	**Name**	**Top Words**
.111	FUNCTION	to, the, for, in, of, and, a, is, on, out
.102	PRONOUN	our, we, you, i, your, my, us, his
.101	AFFECT	great, thank, support, thanks, proud, care
.098	SOCIAL	our, we, you, your, who, us, his, help, they
.107	PREP	to, for, in, of, on, at, with, from, about
.095	VERB	is, are, be, have, will, has, thank, support
Word2Vec Clusters Features		
r	**Top Words**	
.079	great, thank, support, thanks, proud, good, everyone	
.049	led, speaker, charge, memory, universal, speakers	
.047	happy, wishing, birthday, wish, miss, wishes, lucky	
.042	their, families, protesc, children, communities, veterans	
.042	an, honor, win, congratulations, congrats, supporting	
.042	family, friends, old, mom, daughter, wife, father	
Sentiment & Emotion Features		
r	**Name**	**Top Words**
.090	Positive	join, proud, working, good, happy
.038	Negative	tax, fight, fighting, small, violence, gun

Table 2: Pearson correlations of group features (maximum six per type) with owner attributed tweets. No features are significantly correlated with staff attributed tweets. All correlations are significant at $p < .01$, two-tailed t-test, Simes corrected.

Token	r	**Token**	r
.	.102	&	.049
to	.081	I	.045
offer	.071	”	.045
my	.070	prayers	.043
and	.060	a	.042
for	.065	you	.042
leadership	.061	in	.040
the	.057	your	.040
of	.054	our	.039
,	.0511	thank	.039
all	.050	have	.038

Table 3: Unigrams with the highest Pearson correlations to owner tweets. No unigrams are significantly correlated with staff attributed tweets. All correlations are significant at $p < .01$, two-tailed t-test, Simes corrected.

Feature	μ **Owner**	μ **Staff**
# Chars	105.4	102.4
# Tokens	23.2	21.4
Contains URL	45.7%	73.9%
@-reply	4.2%	9.5%
Sent on Weekends	23.5%	20.7%
# Retweets	29.4	38.0
# Likes	82.3	79.1

Table 4: Mean values of tweet features in owner and staff attributed tweets. All differences between means shown in this table are significant at $p < .001$, Mann-Whitney U test, Simes corrected.

Our analysis shows that owner tweets are associated to a greater extent with language destined to convey emotion or a state of being and to signal a personal relationship with another political figure. Tweets of congratulations, condolences and support are also specific of signed tweets. These tweets tend to be retweeted less by others, but get more likes than staff attributed tweets.

Tweets attributed to account owners are more likely to be posted on weekends, are less likely to be replies to others and contain less links to websites or images. Remarkably, there are no textual features significantly correlated with staff attributed tweets. An analysis showed that these are more diverse and thus no significant patterns are consistent in association with text features such as unigrams, topic or LIWC categories.

7 Conclusions

This study introduced a novel application of NLP: predicting if tweets from an account are attributed to their owner or to staffers. Past research on predicting and studying Twitter account characteristics such as type or personal traits (e.g., gender, age) assumed that the same person is authoring all posts from that account. Using a novel data set, we showed that owner attributed tweets exhibit distinct linguistic patterns to those attributed to staffers. Even when tested on held-out user accounts, our predictive model of owner tweets reaches an average performance of .741 AUC. Future work could study other types of accounts with similar posting behaviors such as organizational accounts, explore other sources for ground truth tweet identity information (Robinson, 2016) or study the effects of user traits such as gender or political affiliation in tweeting signed content.

References

Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In *Proceedings of the 29th on Hypertext and Social Media*, HT, pages 20–24.

D. John 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*, EMNLP, pages 1301–1309.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, CIKM, pages 759–768.

Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW, pages 241–244.

Liviu P Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012. Pastiche detection based on stopword rankings: exposing impersonators of a romanian writer. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 72–77.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4):169–200.

Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology*, 60(1):9–26.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 405–413.

Ulrike von Luxburg. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416.

Sunghwan Mac Kim, Qiongkai Xu, Lizhen Qu, Stephen Wan, and Cécile Paris. 2017. Demographic Inference on Twitter using Recursive Neural Networks. volume 2 of *ACL*, pages 471–477.

James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. ICWSM, pages 650–653.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 746–751.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, NAACL, pages 26–34.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ACL, pages 115–123.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 281–288.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Mahway: Lawrence Erlbaum Associates.

Politico. 2017. Trump credits social media for his election. https://www.politico.com/story/2017/10/20/trump-social-media-election-244009.

Marius Popescu and Liviu P Dinu. 2007. Kernel Methods and String Kernels for Authorship Identification: The federalist Papers Case. In *Proceedings of the 2007 International Conference Recent Advances in Natural Language Processing*, RANLP.

Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the Dark Triad of Personality using Twitter Behavior. In *Proceedings of the 25th ACM Conference on Information and Knowledge Management*, CIKM, pages 761–770.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Conference of the Association for Computational Linguistics*, ACL, pages 729–740.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING, pages 1534–1545.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*, 10(9).

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 729–740.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

David Robinson. 2016. Text analysis of Trump's tweets confirms he writes only the (angrier) Android half. http://varianceexplained.org/r/trump-tweets/.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. 2013a. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE*, 8(9).

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes Eichstaedt, and Lyle Ungar. 2017. DLATK: Differential Language Analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP, pages 55–60.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013b. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891.

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1556–1560.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume 4 of *LREC*, pages 1083–1086.

Time. 2011. Obama Is Actually Writing His Own Tweets Now. http://techland.time.com/2011/06/20/obama-is-actually-writing-his-own-tweets-now/.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring User Political Preferences from Streaming Communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 186–196.