Understanding Discourse Acts: Political Campaign Messages Classification on Facebook and Twitter

Feifei Zhang, Jennifer Stromer-Galley✉, Sikana Tanupabrungsun, Yatish Hegde, Nancy McCracken, and Jeff Hemsley

School of Information Studies, Syracuse University, Syracuse, NY, USA

{fzhang09,jstromer,stanupab,yhegde,njmccrac,jjhemsle}@syr.edu

Abstract. To understand political campaign messages in depth, we developed automated classification models for classifying categories of political campaign Twitter and Facebook messages, such as calls-to-action and persuasive messages. We used 2014 U.S. governor's campaign social media messages to develop models, then tested these models on random selected 2016 U.S. presidential campaign social media data. Our classifiers reached .75 micro-averaged F value on training set and .76 micro-averaged F value on test set, suggesting that the models can be applied to classify English-language political campaign social media messages. Our study also suggests that features afforded by social media help improve classification performance in social media documents.

Keywords: Automated classification · Political campaign · Social media ·Supervised learning · Text mining

1 Introduction

Since U.S. political campaigns have incorporated social media like Twitter and Facebook into their strategic messaging, it has become more challenging to have a full appreciation for the style of campaigning or its substance. To understand campaign messages in depth, we built models to classify each campaign-generated message into a category based on what the message is trying to do: urging people to act, changing their opinions through persuasion, informing them about some activity or event, featuring an endorsement, honoring or mourning people or holidays, or on Twitter having a conversation with members of the public. In this way, we can provide a more expansive and comprehensive lense for understanding political discourse.

To develop classifiers that automatically categorize political campaign message type, we used 2014 governor's campaign data to develop a codebook to categorize campaign message categories, generate training data, and build initial models. Then, we applied these to the 2016 presidential campaign messages to test their generalizability. Finally, we used combined governor's and presidential social media campaign datasets to rebuild more generalizable models, aimed at predicting other political

campaign messages. Our classifiers reached .75 micro-averaged F value or better on combined training set and .76 micro-averaged F value or better on randomly selected presidential test set. These results suggest that our models can be applied to categorize political campaign social media messages written in English. Our study also suggests that considering characteristics of social media messages and adding features afforded by social media help improve classification performance in social media documents.

2 Relevant Literature

Previous studies suggest adding features afforded by social media to the common bag-of-words/ngrams models helps improve classification performance. Conover et al. [1] classified the political alignment of tweets as either politically left, right or ambiguous. They developed two Support Vector Machines (SVM) classifiers with different sets of features. The first classifier focused on representing messages with bag-of-words, but removing common stop words, hashtags, mentions and URLs. The second classifier was trained with features consisting of a bag-of-hashtags. This classifier performed better, with 83.5% accuracy over the first classifier, which was 72.6% accurate. The results suggest that the hashtags of tweets contain more significant semantics indicating the political alignment of the users than the message words. Sriram et al. [8] classified tweets by author's intentions - daily chatter, conversations, sharing information and reporting news. They compared three models: bag-of-words, bag-of-words plus author, and bag-of-words plus author and tweet specific features (e.g. the presence of shortened words and slang, time-event phrases, emphasis on words, and mentions at the beginning and within the tweets). Results showed that the third model performed better than the first two models by 10% to 23%.

3 Data and Content Analysis

We used two open source toolkits [3], [4] to collect social media data covering all 36 states during the 2014 U.S. gubernatorial elections (78 candidates) and all major party candidates (26 campaigns) of the 2016 U.S. presidential election. We collected 34,275 gubernatorial campaign tweets and 9,133 Facebook posts between September 14th and November 11th, 2014. We collected 79,102 presidential campaign tweets and 29,503 Facebook posts from when they declared their presidential bids until election day.

We developed a codebook for categorizing 2014 gubernatorial social media campaign messages through deductive analysis following prior political campaign studies [5]. We developed additional categories through inductive analysis of our corpus. Our final codebook contains 5 message categories for Facebook and 6 categories for Twitter: calls-to-action (CTA), persuasive (PER), informative (INF), endorsements (END), ceremonial (CER), and conversational (CON), in Twitter only.

Four annotators were trained to apply the codebook to sub-datasets of gubernatorial election data. The final inter-coder agreement on a random sample of 648 messages reached .79 agreement or better on message categories. When Krippendorff's alpha is .75 or higher, the coding of a variable is considered reliable [6].

We randomly selected 4,147 tweets and 2,494 Facebook posts from gubernatorial data for annotators to code and adjudicate. The adjudicated data is called gold standard data and used for model training. The distribution of the gold standard data is skewed on both Twitter and Facebook, with more messages of types CTA, PER, INF than of types END, CER and CON, as shown in Table 1 below.

4 Automated Text Classification

4.1 Model building

Because Facebook posts and tweets are different in terms of length and other characteristics, we trained multi-class classifier models separately for each application. We built models using the Scikit-learn toolkit [7]. Before training, we pre-processed tweets and Facebook posts by parsing each message to tokens using the ARK Twitter Tokenizer [2], and converted all tokens to lowercase to avoid superfluous features.

We performed many experiments using different multi-class classification algorithms, e.g. SVM with a linear kernel, Naïve Bayes (NB), and Multinomial Logistic Regression (MaxEnt). Among the three, SVM performs best, followed by MaxEnt and NB. As such, we chose SVM for our study.

We tested different combinations of widely used document and feature representation techniques to identify the optimal combination. For document representation, a combination of unigrams and bigrams gave us the highest performance. For four feature representations we compared, results showed that normalized frequency performed worst and the performances of Boolean, term frequency and term frequency–inverse document frequency make little difference. We ended up representing our data with a combination of unigrams and bigrams using Boolean features.

To avoid bloating feature space and over-fitting models, we set a threshold of N-grams representations and filtered out the low-frequency features. Our experiments suggested we keep the 3,000 most frequent unigrams when their frequency is higher than 2, and keep the 1,000 most frequent bigrams.

Given the skewed data distribution in our study, we used micro-averaged F value (Micro-F1), which is a weighted-average over the binary models of the multi-class classifier, to measure overall performance of models. It weights raw scores based on the number of instances in each class [9], which makes it possible to compare averages across result sets. We evaluated classification tasks with 5-fold cross validation.

Using the settings noted above, the F1 scores of our first models are over .70 for all categories except CER and CON (see Model 1 in Table 1). We then developed the second version of models by including the characteristics of social media data based on Model 1. We canonicalized some word tokens by replacing numbers, emoticons, and URLs (e.g. http://abc) with the general tokens. This improved the F1 score of CON category from .48 to .66. We also tried to canonicalize hashtags, but it degraded performance of all categories. We noted that many CON messages started with @mention, and thus added this as a feature. This improved the F1 score of CON from .66 to .76 (see Model 2 in Table 1). Model 2 is our best Twitter model, with .72 Micro-F1, and Model 1 is our best Facebook model, with .74 Micro-F1.

Table 1. Machine learning experiments and performance on governor's data (M: model; P: precision; R: recall; N: Number of training data)

M	Features	Category	Twitter				Facebook			
			P	R	F1	N	P	R	F1	N
1	Lowercase TF=3 Unigram=3000 Bigram =1000	CTA	.78	.78	.78	991	.81	.79	.80	999
		PER	.74	.73	.73	1393	.74	.73	.73	755
		INF	.73	.71	.72	1342	.67	.73	.70	526
		END	.76	.75	.76	166	.79	.73	.76	106
		CER	.41	.40	.40	135	.46	.38	.42	108
		CON	.42	.57	.48	120	n/a	n/a	n/a	n/a
		Micro-F1	.72	.72	.72	4147	.74	.74	.74	2494
2	Lowercase TF=3 Unigram=3000 Bigram =1000 Canonical_form @mention	CTA	.78	.79	.78	991	.81	.80	.80	999
		PER	.74	.74	.74	1393	.74	.74	.74	755
		INF	.74	.73	.73	1342	.67	.70	.68	526
		END	.71	.73	.72	166	.78	.78	.78	106
		CER	.50	.48	.49	135	.52	.45	.49	108
		CON	.76	.77	.76	120	n/a	n/a	n/a	n/a
		Micro-F1	.72	.72	.72	4147	.74	.74	.74	2494

4.2 Codebook and Model Testing

We tested whether the codebook developed and the models trained and validated on governor's data were generalizable to presidential campaign messages. We applied the codebook and models against a randomly selected subset from the early stage of 2016 presidential campaign, 2,989 tweets and 2,638 Facebook posts. We also did another round of model testing using representative campaign data when all the candidates finished their campaigns (see details below).

We used our best classification models to predict message categories of the selected subset. Two annotators who had achieved good intercoder agreement (.79) separately corrected machine predictions. Annotators did not find differences on message categories between governors and presidential data, suggesting our codebook approach was applicable for presidential data.

We then compared differences between machine-predicted and human-corrected categories to evaluate the generalizability of models. Our models achieved at .70 Micro-F1 value for both tweets and Facebook posts. This suggested that the gubernatorial and presidential datasets share many features in common. As such, we constructed a new gold standard dataset by combining the gubernatorial gold standard data and the human-corrected presidential data for each social media platform. These new gold datasets were used for re-building the more generalizable models for prediction. The new set comprises of 7,136 tweets and 5,132 Facebook posts.

For the new gold standard data, the optimal sets of features are the same as the gubernatorial models. Specifically, the basic features still give the highest performance for Facebook (Micro-F1 of .76), and adding canonical form and @mention features to the base set is best for Twitter (Micro-F1 of .75), as shown in Table 2.

Table 2. Machine learning experiments and performance on combined governor's and presidential data (M: model; P: precision; R: recall; N: Numbers of training data)

M	Features	Category	Twitter				Facebook			
			P	R	F1	N	P	R	F1	N
1	Lowercase TF=3 Unigram=3000 Bigram =1000	CTA	.78	.81	.80	1575	.85	.82	.83	2058
		PER	.76	.76	.76	2780	.75	.75	.75	1660
		INF	.72	.68	.70	2187	.68	.72	.70	1065
		END	.73	.73	.73	181	.73	.76	.75	127
		CER	.41	.44	.43	219	.68	.72	.70	222
		CON	.42	.51	.46	194	n/a	n/a	n/a	n/a
		Micro-F1	.73	.73	.73	7136	.76	.76	.76	5132
2	Lowercase TF=3 Unigram=3000 Bigram =1000 Canonical_form @mention	CTA	.83	.80	.82	1575	.84	.83	.84	2058
		PER	.78	.76	.77	2780	.76	.75	.75	1660
		INF	.72	.73	.72	2187	.66	.71	.69	1065
		END	.70	.70	.70	181	.72	.70	.71	127
		CER	.40	.48	.44	219	.54	.49	.51	222
		CON	.68	.75	.71	194	n/a	n/a	n/a	n/a
		Micro-F1	.75	.75	.75	7136	.76	.76	.76	5132

We tested the reliability of these more generalizable models in Dec 2016, when all the candidates finished their campaign. We randomly selected 1,000 tweets (994 written in English) and 1,000 Facebook posts (987 written in English) over the course of the campaign as test sets. After comparing differences between machine-predicted and human-corrected categories, results suggested that our models work well on the test set, especially for CTA, PER and INF with Micro-F1 scores of .81, .81, .73 for Twitter and .84, .77, .74 for Facebook, as shown in Table 3.

It is not surprising that the CER category still was not predicted accurately, given the limited number of messages in training data and the lack of obvious patterns in text. The END category was also not predicted well. The good performance from cross-validation on training data but poor with the test data indicates that the model might over-fit the training data for this category. We expect that more training data would be helpful to improve the performance of this category.

Table 3. Model testing on presidential data over the campaign period (P: precision; R: recall; N: Numbers of training data)

Category	Twitter				Facebook			
	P	R	F1	N	P	R	F1	N
CTA	.78	.84	.81	173	.86	.82	.84	331
PER	.85	.77	.81	574	.76	.77	.77	342
INF	.71	.74	.73	189	.67	.82	.74	215
END	.71	.50	.59	10	.67	.38	.48	21
CER	.62	.41	.49	32	.69	.44	.54	78
CON	.85	.69	.76	16	n/a	n/a	n/a	n/a
Micro-F1	.80	.78	.78	994	.77	.77	.76	987

5 Conclusion and Future Work

To better understand political discourse on social media, this paper built classification models to categorize campaign messages on Twitter and Facebook. The good model performance on training data (Micro-F1 of .75) and randomly selected test data (Micro-F1 of .76) suggests that the models are applicable to categorize other political campaign social media messages. Our study supports prior research [8] that considering social media messages characteristics and including features afforded by social media platforms helps improve model performance. In our experimentation, using canonical forms and Twitter's @mention is helpful for classify conversational tweets. We also found that the classifier trained with a feature space of hashtags is better, since they may include important information. For future research, external aspects of the political campaign might be beneficial, such as including the mentions of the candidate's opponent or self might improve the performance of persuasive category.

Acknowledgements. We thank Bei Yu's helpful feedback on this paper. The project was partly supported by the Tow Center for Digital Journalism at Columbia University and the Center for Computational and Data Sciences at the School of Information Studies at Syracuse University.

References

1. Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F. (2011). Predicting the political alignment of twitter users. In Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 192-199. IEEE.
2. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pp. 42-47. Association for Computational Linguistics.
3. Hegde, Y. (2016). fb-page-scraper: Version 1.33 Released. https://doi.org/10.5281/zenodo.55940
4. Hemsley, J., Ceskavich, B. and Tanupabrungsun, S. (2014). Syracuse Social Media Collection Toolkit. https://github.com/bitslabsyr/stack
5. Jamieson, K. H., Waldman, P., & Sherr, S. (2000). Eliminate the negative? Categories of analysis for political advertisements. Crowded airwaves: Campaign advertising in elections, pp. 44-64.
6. Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. Human communication research, 28(4), pp. 587-604.
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, pp. 2825-2830.
8. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841-842. ACM.
9. Van Asch, V. (2013). Macro-and micro-averaged evaluation measures [basic draft].