PoliBERT: Classifying political social media messages with BERT

Shloak Gupta [0000-0002-5947-5452], Sarah E. Bolden[0000-0002-8992-1580], Jay Kachhadia [0000-0003-0961-751X], Ania Korsunska [0000-0003-3158-8436] and Jennifer Stromer-Galley [0000-0001-6079-8788]

Syracuse University, Syracuse NY 13244, USA
sebolden@syr.edu

Abstract. Political candidates use social media extensively in campaigning. In an effort to better understand the content of political messages, we have explored various machine learning models and methods to classify US election candidates' social media content into message types. Despite its burgeoning popularity, the pre-trained language model BERT has not yet been used to classify social media political campaign messages. Using BERT, we gained significant improvements in the results of our models for classifying types of political messages in short social media texts. Overall, our findings suggest that leveraging pre-trained classifiers like BERT to classify types of short-text political messages from Facebook and Twitter may substantially improve the performance of automated machine classification. Such improvements can help social scientists better understand human communication and persuasion on social media.

Keywords: Text Mining, Automated Classification, Deep Learning, Social Media, Political Campaign

1 Introduction

Whether they are produced by political campaigns or the general public, discussions about politics and public affairs are increasingly mediated through digital communication channels [1]. At the same time, computational techniques are growing increasingly sophisticated and are being applied to the study of political messaging. Some of the current research in this area uses human-supervised machine learning (ML) techniques to classify political content [2–4]. As deep learning techniques become more widely available, they can be leveraged to further improve the accuracy of these current methods. One specific advance, the pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) [5], has proven beneficial in improving the accuracy of identifying and classifying human communication in several contexts, such as health communication online [6]. To the best of our knowledge, BERT has not yet been applied to the domain of politics.

This study leverages the pre-trained language model BERT [5] in the classification of political candidates' social media content. Our findings demonstrate that the use of BERT can improve machine learning classification of political content produced on social media. Specifically, our application of BERT to United States political candidates' Facebook and Twitter content produced an improved overall F1 score for each of the seven classes that we tested. In addition to discussing overall improvement, we highlight noteworthy variations in the degree to which BERT improved individual categories' results. In doing so, we consider the current strengths and limitations of the use of BERT in the classification of political social media content, and we draw attention to several avenues for further research.

The application of BERT to political content on social media provides an opportunity to assess BERT's ability to classify content that differs in form and substance from the Wikipedia posts and English literature that it was originally trained on. Although social media content is more informal, more emotive, and much shorter in length than the content that BERT was trained on, the results of our study suggest that BERT is nevertheless capable of strengthening the human-supervised machine learning techniques that are currently used to classify this content. Improvement in the classification of social media content in general, and political content specifically, stands to benefit both academic and industry research. Strengthened classification performance can help scholars not only to better understand that content of political communication on social media, but to more reliably convey these findings to scholars and the public.

This paper is structured as follows: In the next section, we review literature on both traditional supervised machine learning algorithms as well as the deep learning algorithm BERT. We then review the methods that were used to annotate, train, and classify our data; in this section, we also review the findings of our BERT-trained model. We conclude by discussing why, while there were overall improvements in our model's performance, individual results for each category varied; we also describe possible avenues for future research.

2 Relevant literature

2.1 Traditional classification algorithms

Researchers have leveraged various automatic classification methods to label political campaigning content on social media. Many of these approaches rely on supervised ML algorithms, which train on human-annotated samples in order to independently label larger volumes of data. To classify social media content, researchers have had to adapt pre-existing Natural Language Processing (NLP) models to capture the unique textual, syntactic, morphosyntactic, and lexical attributes found in social media texts [7]. Compared to news, congressional records, and other long-form content that traditional supervised ML algorithms were developed to classify, social media content is shorter in form and tends to be more informal and multimodal [7, 8]. Researchers have also observed variations in communication patterns across different social media platforms [9]. Developing and applying supervised ML algorithms to social media data therefore requires configuring models that are sensitive to the unique features of the platform(s) being studied.

Although there are many supervised ML algorithms that can be trained to classify social media data, such as Naïve Bayes and Logistic Regression, we wish to focus specifically on the Support-Vector Machine (SVM) algorithm, given that it is the model that produced the best results when applied to our political candidates' social media corpus. In the context of political campaigning, researchers have utilized SVM models to classify content produced by political candidates themselves [8] as well as the general public [10]. These models have classified a variety of categories found in political social messages, including, among others, topic [11], sentiment [12] and toxicity [7]. Current SVM algorithms used for research on political campaigning have achieved F1 scores ranging from 65% to 80%, which speaks toward the efficacy of adapting traditional supervised ML algorithms to social media content [8].

There are two features of supervised ML algorithms, including SVM, that can negatively impact model performance. First, the performance of a given model is contingent on the size of its training corpus. Model refinement (via hyper parameter tuning or weight balancing, for example) can produce small to modest improvements in classification results, but increasing the training corpus, to-date, is one of the most effective strategies to improve model results [7, 11]. Human annotation, however, is a time-consuming and costly endeavor, which means that increasing the size of the training corpus is not always a feasible solution. Second, supervised ML models treat words individually as standalone, decontextualized features. Creating features from words is not problematic in and of itself; it is the basis for models such as Term Frequency - Inverse Document Frequency (TF-IDF), which can be extremely effective depending on the corpus and ML algorithm. However, such models, by removing words from the context in which they were originally situated, can be problematic in several ways. For example, supervised algorithms are unable to differentiate between homonyms and other forms of textual polysemy. Similarly, supervised algorithms do not have any pre-existing recognition of a word's definition and/or synonyms, which can pose challenges to classification models when frequently used terms are exchanged for synonyms that are unrecognized by the algorithm. Here, increasing the size of the algorithm's training corpus does not provide any additional information about the meaning of a given synonym and, consequently, is unlikely to improve precision and recall results.

In the context of political campaigning, audiences such as journalists stand to benefit from real-time updates about the content that politicians and the public are producing. That said, the speed with which updates are delivered matters to the extent that the content that is shared is reliably labeled. There is, in short, exigence to develop automatic classification tools that can more effectively balance the need for efficiency and reliability. In the next section, we introduce the BERT model as a tool that can be leveraged to overcome some of the limitations of traditional classification algorithms such as SVM.

2.2 BERT Model

To further improve the classification performance of our SVM model, we employed a new semi-supervised approach using the BERT model. BERT is a pre-trained algorithm that leverages an extensive corpus of English-language Wikipedia text passages (2,500M words) and BooksCorpus (800M words) books to improve the accuracy of various NLP tasks [5]. Unlike other language representation models, which primarily keep the left context (and, in fewer cases, the right context), BERT is the first truly bidirectional representation model in which both the right and left contexts are jointly considered while calculating attention in all layers [5]. BERT utilizes a Masked Language Model (MLM) pre-training objective, in which some percentage of the input tokens are masked at random and then used for prediction. BERT also uses Next Sentence Prediction (NSP) to train for context and sentence relationships.

BERT has achieved state-of-the-art results on eleven NLP tasks and can be easily fine-tuned to create models for a wide range of NLP tasks [5]. For example, DocBERT, the BERT model fine-tuned for document classification tasks, has shown great results on datasets such as Reuters-21578, arXiv Academic PaperS, IMDB reviews, and Yelp 2014 reviews, which include both multi-label (Reuters and arXiv) and single-label documents (IMDB, Yelp) [13]. Researchers have also recently used BERT to improve sentiment analysis with an aspect-based sentiment analysis (ABSA) model, which considers text sentiment alongside other important aspects, such as the entity or topic to which the sentiment is directed. These experiments showed that applying the ABSA model at the top of the pre-trained BERT encoder outperforms previous state-of-the-art approaches to sentiment classification [14]. BioBERT, a BERT model designed to classify biomedical texts [15], has been successfully used to identify personal health mentions in Tweets [6]. As Gondane observes, the Twitter data classified by BioBERT differs substantially from the BERT data that it was trained on [6]. Nevertheless, the model performed well on Tweets, which is noteworthy because, unlike BERT's training data, Tweets are very short and are more likely to contain misspellings, sarcasm, and slang terms [6].

To the best of our knowledge, BERT has not yet been used in the classification of political campaign social media messages. We set out to see if BERT can improve our performance, despite the peculiarities and mismatch of the model training corpus and our social media corpus.

3 Methods

3.1 Data Collection

We used open-source tools to collect real-time data from Twitter [16] and Facebook [17] on our multi-server infrastructure. Separate collectors were set up on multiple servers to build a resilient collection system that does not miss any data in the collection process. All of the data that we collected went through our preprocessing pipelines. Because Facebook accounts have different identifiers than Twitter (e.g., @realdonaldtrump on Twitter vs. Donald Trump on Facebook), sizeable effort was required at this stage to integrate both Facebook and Twitter candidate information. Once preprocessing was completed, our process shifted to final collection, at which point it was ready to be used for analysis.

Our data collection included Facebook posts and Tweets produced by presidential candidates in the 2016 United States General Election. We restricted the collection of this data to the active election campaign period of 08/02/2015 to 11/10/2016. The resulting corpus included 27,591 Facebook posts and 88,980 tweets from 33 candidates. The distribution of higher-level Twitter and Facebook messaging activity, separated by candidate, can be seen in Figure 1.

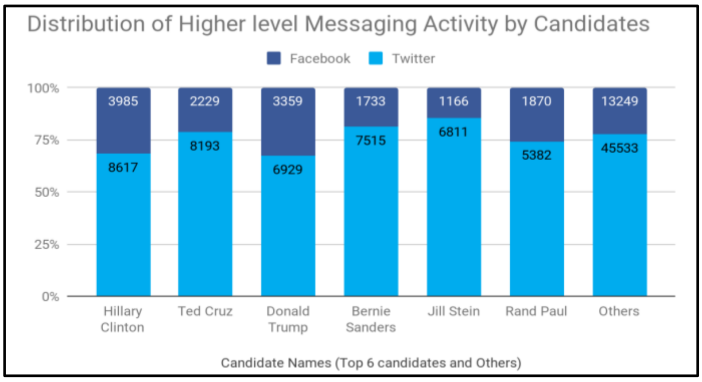

Fig.1 Distribution of Higher-Level Messaging Activity by Candidates

Candidates tend to be more active on Twitter than on Facebook by a ~3:1 ratio. Although the volume of content produced by candidates varies by platform, the distribution of messaging activity for each platform is relatively similar across all of the candidates. It should be noted that whereas the production of content is relatively similar between candidates, the *substance* of candidates' content (e.g., the distribution of classification labels) varies both within and between platforms.

3.2 Classification Development

To perform content analysis on our data, we relied on a preexisting codebook that was developed in 2014 to classify political speech acts found in political candidates' social media posts [8]. Using this codebook, coders could apply up to seven mutually exclusive, binary labels to a given message. Four of these labels (*Advocacy*, *Attack*, *Image*, and *Issue*) describe forms of persuasive messaging that attempt to persuade the reader to support the candidate and/or reject their opponent. The *Call to Action* classifier identifies messages that include a directive for readers to take (e.g., watch, retweet, share, etc.). *Campaigning Information* describes messages about the campaign's organization, ground game, and strategy, and *Ceremonial* captures messages that include social elements (e.g., honoring, praising, joking, etc.).

Prior to creating any gold-standard datasets, human coders trained in pairs on pre-existing sets of gold-standard messages until achieving consistent intercoder agreement (Krippendorf's alpha ≥ 0.70) for every category. Once coders reached intercoder agreement, pairs annotated and adjudicated two random samples of messages: 5,231 Tweets and 4,434 Facebook posts. It is worth reiterating that although candidates produce, roughly, the same volume of content on Facebook and Twitter, the substance of the content that they produce differs between the two platforms. Given these differences, rather than combine Twitter and Facebook content, the two datasets were kept separate throughout both the human annotation and machine classification process.

3.3 Automated Classification

Our initial classification models used supervised learning algorithms such as Naïve Bayes and Logistic Regression, but improved results were obtained using SVM. Using Python's NLTK package, we followed basic preprocessing steps for NLP, such as tokenization and stopword removal[1]. We tested different techniques[2] to convert text to features and found that the best model results were obtained using TF-IDF.

We restricted our input features to only include unigrams and bigrams—we did this both to avoid overfitting and to limit feature space. We set a maximum Document Frequency (DF) of 0.9 to remove additional stopwords and a minimum DF of 0.2 to capture relevant features. In order to optimize the number of features and overall performance, we used five-fold cross-validation to tune the SVM model's hyperparameters. Given the differential distribution of codes[3], we weight-balanced the classes with weights that were inversely proportional to the label proportions of our data. Finally, as both precision and recall were equally important for our research, we used the F1 score to optimize our SVM models. The results for this model are presented in Table 1.

Category	Twitter 2016				Facebook 2016			
	P	R	F1	*n*	P	R	F1	*n*
Advocacy	0.7	0.71	0.71	2259	0.81	0.8	0.81	2163
Attack	0.67	0.66	0.67	1020	0.73	0.66	0.7	670
Image	0.64	0.65	0.65	1582	0.72	0.69	0.7	1537
Issue	0.77	0.73	0.75	1424	0.85	0.79	0.82	1629
Call to Action	0.87	0.81	0.84	1269	0.91	0.86	0.88	1681
Campaign Info	0.78	0.78	0.78	1830	0.63	0.56	0.59	770
Ceremonial	0.79	0.67	0.73	632	0.85	0.72	0.78	744

[1] Although lamenting and stemming are common NLP preprocessing steps, our results indicated that these steps did not contribute to improved model performance.

[2] E.g., Bag-of-Words and Count-Vectorizer.

[3] For example, whereas 43% of gold-standard messages were labeled as advocacy, only 12% were labeled as attack

Table 1. Results of SVM Classifier

3.4 Improvement with BERT Model

We used a pre-trained BERT Base model (12-layer, 768-hidden, 12-heads, 110M parameters) for our classification of 2016 political candidates' social media messages. We took a stratified split of our gold-standard data: 80% for training and 20% for testing. The maximum length of input for BERT after tokenization was 512, though we chose to keep it at 128 because this value covered a majority of content and reduced the duration of training time. The BERT model was fine-tuned for 2 epochs on the training data; any further epochs resulted in overfitting. We initially found that our F1 results for the *Attack* and *Ceremonial* classifiers were lagging. Because BERT requires a significant amount of data for training and event fine-tuning, the lower performance for these two classes could be attributed to their smaller sets of gold-standard labels. To counter this, we sampled an additional 300 Tweets and 300 Facebook posts; we had human coders annotate and adjudicate this content in order to increase the training data for these two categories. The addition of these messages produced significant improvement: our performance and F1 scores for Facebook improved for *Attack* by 7% and for *Ceremonial* by 4%; for Twitter content, *Attack* improved by 1% and *Ceremonial* by 9%. The results for our BERT model classification on the test data are available in Table 2.

Category	Twitter 2016				Facebook 2016			
	P	R	F1	*n*	P	R	F1	*n*
Advocacy	0.76	0.79	0.77	2259	0.83	0.81	0.82	2163
Attack	0.80	0.80	0.80	1020	0.77	0.82	0.79	670
Image	0.75	0.72	0.73	1582	0.70	0.79	0.74	1537
Issue	0.85	0.88	0.86	1424	0.93	0.84	0.89	1629
Call to Action	0.92	0.92	0.92	1269	0.94	0.96	0.95	1681
Campaign Info	0.86	0.86	0.86	1830	0.71	0.73	0.72	770
Ceremonial	0.90	0.84	0.87	850	0.82	0.88	0.85	744

Table 2. BERT Model Classification Results

After getting a substantial boost in results using the BERT model for our classification task, we wanted to see if our model for the 2016 Presidential Election could be applied to content from other election cycles. We took a sample of 1,000 Tweets from the 2020

US Presidential candidates[4] and applied our 2016 model. It is important to note that between 2018 and 2020, significant changes were made to the *Campaigning Information* category. Given that the 2016 model has not been retrained to encompass the changes to this category, we have excluded it from the results in Table 3.

Category	Twitter 2020			
	P	R	F1	*n*
Advocacy	0.72	0.85	0.78	507
Attack	0.81	0.61	0.69	236
Image	0.57	0.54	0.56	237
Issue	0.92	0.90	0.91	543
Call To Action	0.86	0.95	0.90	93
Ceremonial	0.74	0.65	0.69	199

Table 3. BERT Classification of Twitter 2020 data using Twitter 2016 Models

In terms of overall F1 scores, the shift from supervised ML to a semi-supervised approach improved classification performance by 9.0% for Twitter and 5.7% for Facebook. A full breakdown of the semi-supervised model's performance improvement, separated by category, is shown in Table 4.

Category	Twitter			Facebook		
	F1- SVM	F1- BERT	*n*	F1-SVM	F1-BERT	*n*
Advocacy	0.71	0.77	2259	0.81	0.82	2163
Attack	0.67	0.80	1020	0.70	0.79	670
Image	0.65	0.73	1582	0.70	0.74	1537
Issue	0.75	0.86	1424	0.82	0.89	1629
Call to Action	0.84	0.92	1269	0.88	0.95	1681
Campaign Info	0.78	0.86	1830	0.59	0.72	770
Ceremonial	0.73	0.87	850	0.78	0.85	744
Weighted F1	0.73	0.82		0.78	0.83	

Table 4. Result Comparison between SVM and BERT across Twitter and Facebook

[4] This sample was collected early in the 2020 election cycle, prior to the Iowa caucus. Consequently, this corpus is not intended to be representative of the entire course of the 2020 campaigning period.

Notably, although the weighted F1 scores show overall improvements in our model's performance, individual results for each category varied. Our BERT model produced substantial F1 improvements for *Attack*, *Campaigning Information*, and *Ceremonial*, but there are two categories that did not show substantive improvement: *Advocacy* and *Image*. To reiterate, our data was significantly different from the Wikipedia and English literature content that was used for BERT's pre-training. One possible explanation for the lower improvement for the *Advocacy* and *Image* categories is that these particular categories differ more substantially than others from the content that BERT was pre-trained on. However, it could also be the case that our training sets for these categories were not large enough to capture the full range of modalities in which content belonging to these categories is communicated. Investigating the reasons behind the variation in our model's performance for each category is an exciting avenue for future research.

4 Conclusion

Overall, our findings suggest that leveraging BERT to classify types of political messages over short texts from Facebook and Twitter may substantially improve the accuracy of machine classification. Deep learning techniques have the promise of further helping social scientists to better understand human communication and persuasion on social media. Although there are some limitations in its application on a few categories, the overall improvements are promising. Future applications include continuing to explore various avenues to improve model performance across all classification categories, as well as applying this technique to the classification of text in Facebook and Instagram paid political ads.

References

1. Stromer-Galley, J.: Presidential campaigning in the Internet age. Oxford University Press (2019).
2. Guess, A., Munger, K., Nagler, J., Tucker, J.: How Accurate Are Survey Responses on Social Media and Politics? Political Communication. 36, 241–258 (2019). https://doi.org/10.1080/10584609.2018.1504840.
3. Suiter, J., Culloty, E., Greene, D., Siapera, E.: Hybrid media and populist currents in Ireland's 2016 General Election. European Journal of Communication. 33, 396–412 (2018). https://doi.org/10.1177/0267323118775297.
4. Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S.A., Parnet, O.: A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates: Incivility in Interactions With Candidates on Twitter. Journal of Communication. 66, 1007–1031 (2016). https://doi.org/10.1111/jcom.12259.
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. (2019).

6. Gondane, S.: Neural Network to Identify Personal Health Experience Mention in Tweets Using BioBERT Embeddings. In: Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. pp. 110–113. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-3218.
7. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on Facebook. 10.
8. Zhang, F., Lee, D., Lin, Y.-R., Osgood, N., Thomson, R., Stromer-Galley, J., Tanupabrungsun, S., Hegde, Y., McCracken, N., Hemsley, J.: Understanding Discourse Acts: Political Campaign Messages Classification on Facebook and Twitter. In: Social, Cultural, and Behavioral Modeling. pp. 242–247. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-60240-0_29.
9. Lin, H., Qiu, L.: Two Sites, Two Voices: Linguistic Differences between Facebook Status Updates and Tweets. In: Rau, P.L.P. (ed.) Cross-Cultural Design. Cultural Differences in Everyday Life. pp. 432–440. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39137-8_48.
10. Yilmaz, K.E., Abul, O.: Inferring Political Alignments of Twitter Users: A case study on 2017 Turkish constitutional referendum. arXiv:1809.05699 [cs]. (2018).
11. Khandelwal, A., Swami, S., Akhtar, S.S., Shrivastava, M.: Classification Of Spanish Election Tweets (COSET) 2017: Classifying Tweets Using Character and Word Level Features. In: IberEval@ SEPLN. pp. 49–54 (2017).
12. Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., Shah, J.: Sentiment analysis of extremism in social media from textual information. Telematics and Informatics. 48, 101345 (2020). https://doi.org/10.1016/j.tele.2020.101345.
13. Adhikari, A., Ram, A., Tang, R., Lin, J.: DocBERT: BERT for Document Classification. arXiv:1904.08398 [cs]. (2019).
14. Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-Based Sentiment Analysis Using BERT. In: NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa). pp. 187–196. Linköping University Electronic Press, Turku, Finland (2019).
15. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. (2019). https://doi.org/10.1093/bioinformatics/btz682.
16. Hemsley, J., Ceskavich, B., Tanupabrungsun, S.: Syracuse Social Media Collection Toolkit. (2014). https://github.com/jhemsley/Syr-SM-Collection-Toolkit
17. Hedge, Y.: fb-page-scraper. (2016). https://doi.org/10.5281/zenodo.55940