PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter

Kornraphop Kawintiranon and Lisa Singh

Georgetown University

LREC 2022 Marseille

Introduction

Motivation

Language models are widely used in various natural language processing (NLP) tasks, including word prediction and stance detection.

While language models are commonly pre-trained on general-domain texts, pre-trained models for specific domains that are of public interest are needed to improve text analysis.

Contributions

1. Propose a language model pre-trained on a large amount of Twitter data related to US politics, specifically the US election 2020.
2. Compare variations of language models on different NLP tasks: perplexity, mask token prediction and stance detection.
3. Publicly release models and stance data sets.

Data Sets

*All data sets are non-overlapping

Unlabeled Data

- Consists of over 83 million unique English tweets related to the 2020 US Presidential election, not including quotes and retweets.
- Collected using Twitter APIs (Streaming and Enterprise Decahose) between January 2020 and February 2021.

Evaluation Data

Poli-Test: 10,000 politics-related tweets collected using keywords and hashtags.

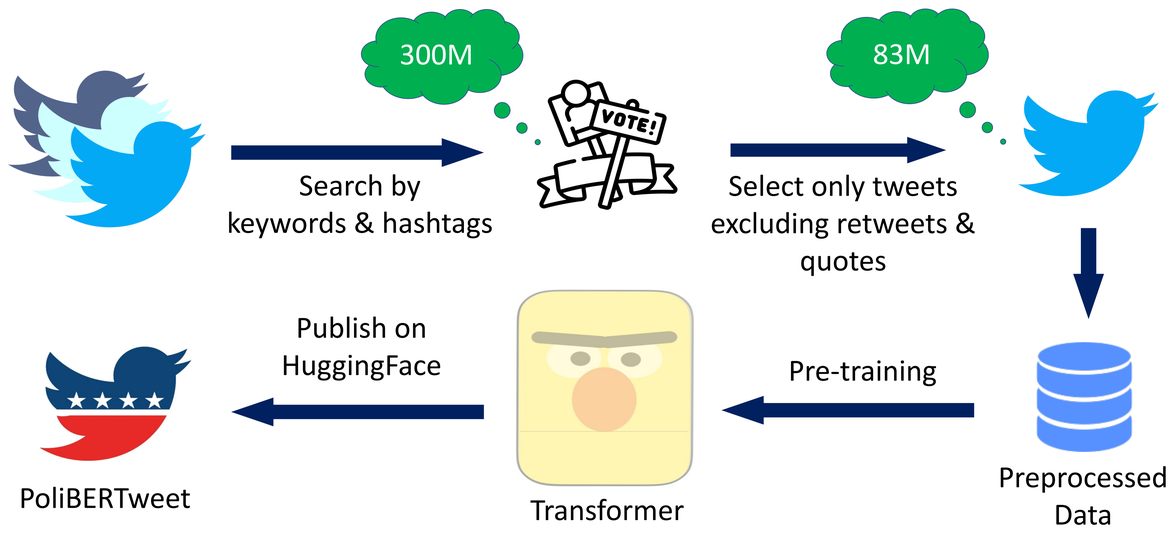
NonPoli-Test: 10,000 non-political tweets sampled using Decahose API.

Stance Data

We use a stance-labeled data set that focuses on analyzing tweets about the US 2020 election proposed by [Kawintiranon & Singh, 2021].

	Split	Total	Support	Oppose	Neutral
Biden	Train	875	266	279	330
	Test	375	112	106	157
Trump	Train	875	243	347	285
	Test	375	98	152	125

Methodology



	Keyword/Hashtag	Count
Candidates	trump	56,542,935
	biden	31,377,347
	#joebiden	482,903
	#biden2020	328,037
	#donaldtrump	323,607
	#dumptrump	158,569
	#fucktrump	97,800
	#nevertrump	32,920
	#lovetrump	1,093
Parties	#gop	378,131
	#democrat	104,129
	#republican	100,881
	#conservative	42,647
	#liberal	29,795
Election	#voteblue	238,502
	#rnc2020	210,862
	#voteblue2020	143,672
	#dnc2020	104,419
	#dnc	70,534
	#notmypresident	68,037
	#qanon	64,099
	#bluewave	62,429
	#votebluenomatterwho	44,145
	#makeamericagreatagain	37,913
	#democraticnationalconvention	33,109

Table 1: Keywords and hashtags used to collect tweets. They are organized by three categories: candidates, political parties/leaning, and election.

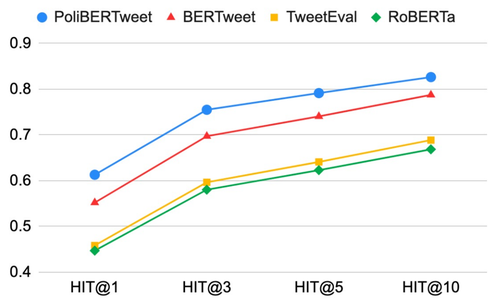
Evaluation Tasks and Results

Perplexity Scores

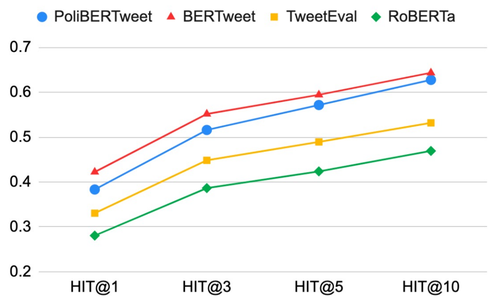
We find that our model outperforms baselines on the political data and performs comparably on the non-political data. Lower scores are better.

Model	Political	Non-political
RoBERTa	25.1158	37.8536
TweetEval	18.4893	24.4660
BERTweet	8.0463	13.5791
PoliBERTweet	**4.4846**	**13.1037**

Masked Token Prediction



(a) Political Tweets



(b) Non-political Tweets

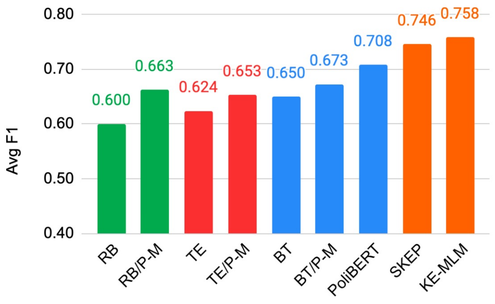
We design our evaluation as a ranking task, where for each tweet, the model ranks the list of potential tokens in decreasing order of relevancy. We evaluate the models using the *Hits@k* metric.

- Our model outperforms the other models on the political data by 4 to 17%.
- On the non-political data, BERTweet performs best because it is pre-trained on a more general set of tweets. Our our model performs comparably to BERTweet.

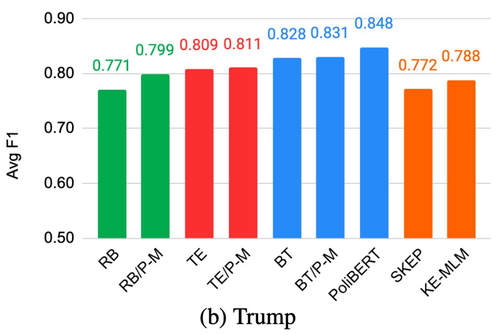
Stance Detection

We use macro-avg F1 to evaluate the models: RoBERTa (RB), TweetEval (TE), BERTweet (BT), our PoliBERTweet (PoliBERT), SKEP by Tian et al. and KE- MLM by Kawintiranon et al. P-M indicates Poli-Medium.

- Our model outperforms the state-of-the-art models on Trump data set by up to 8%.
- On the Biden data set, KE-MLM performs better than our model by 5%. We hypothesize that It could perform even better if it was trained using our model instead of the vanilla BERT.



(a) Biden



(b) Trump

Conclusions

- We present **PoliBERTweet**, a pre-trained language model trained on a large corpus of **political tweets for over 1000 hours** using the starting weights from BERTweet.
- We evaluate PoliBERTweet and other state-of-the-art models on different NLP metrics and tasks.
- Our model is valuable for both politics-related domains but also general semantic understanding on Twitter.
- Our evaluation data sets and models are available at https://github.com/GU-DataLab/PoliBERTweet.

Available on

GitHub HUGGING FACE SCAN ME

Acknowledgements

This research was funded by National Science Foundation awards #1934925 and #1934494, and the Massive Data Institute (MDI) at Georgetown University. We would like to thank our funders, the MDI staff, and the members of the Georgetown DataLab for their support.