Twitter Sentiment Analysis through the Fine-Tuning of TwHIN-BERT

Ivan Nikitovic
Department of Computer Science, Boston University
(in@bu.edu)

Abstract

During the past several years, Twitter has become a critical communication medium of celebrities, influencers, and major industry and political leaders such as Donald Trump, Elon Musk, and many others. The platform provoked various debates at the end of October 2022, when it was acquired by one of the richest known persons in the world, Elon Musk. Because tweets usually contain mostly textual content, analyzing this social media should be attempted using techniques from natural language processing. We propose a system that performs sentiment analysis on Twitter data (tweets) and infers public opinion (of Twitter users) on certain topics of interest, such as international developments and controversies. Our model implements the pre-trained TwHIN-BERT language model based on BERT (Zhang et al., 2022), fine-tuned on the Sentiment140 dataset for the purposes of predicting sentiment as either *positive* or *negative*. We collect new tweets using snscrape[1], an open-source python library, and perform sentiment analysis on the scraped data to estimate Twitter users' aggregated sentiment of any topic at any point in time. After just two epochs performed on 500 thousand tweets, the model achieves an F1-score of 0.845 on balanced, human-labeled test data.

1 Introduction

Before using a supervised learning approach, our first attempt at Twitter sentiment analysis was with unlabeled data. This approach was largely motivated by the lack of large human-labeled datasets and therefore the cost effectiveness of using an unsupervised learning algorithm. After several trial runs using K-means clustering, we were dissatisfied with the results and opted for supervised learning. Given our budget of zero, fine-tuning was an effective solution given its ability to reuse weights trained on large datasets (TwHIN-BERT was trained on 7 billion tweets), and achieve satisfactory results using a relatively small labeled dataset.

2 Data Collection

The Sentiment140 dataset was created at Stanford University in 2009 by automatic (machine) labeling of tweets based on emoticons. This approach allowed them to label 1.6 million tweets, much more than other publicly available hand-labeled datasets.

To collect new tweets, we used snscrape instead of TwitterAPI to avoid query limits and the use of API keys. This makes the system more accessible, and anyone with the code can easily try it out.

[1] https://github.com/JustAnotherArchivist/snscrape

3 Model Training

Our model was trained using two epochs over 450 thousand training and 50 thousand evaluation samples. Using cross-validation, we eventually used the hyperparameters shown in Table 1.

Parameter	Value
Number of Epochs	2
Batch Size	32
Learning Rate	3e-5
Weight Decay	0.01

Table 1: Hyperparameters chosen from Cross-Validation

We stopped at the 2nd epoch as the model began overfitting after just several epochs. Therefore, the training loss was still relatively high, but the metrics were satisfactory as shown in Table 2.

Epoch	Train. Loss	Valid. Loss	Accur.	F1 Score
0	0.3158	0.3165	0.8642	0.8697
1	0.2565	0.3142	0.8703	0.8695

Table 2: Training arguments

4 Evaluation

We evaluated the performance of the model using a manually labeled subset of Sentiment140 consisting of 359 tweets from 2009 with a target label of "0" meaning negative and "4" meaning positive.

Considering that the training data was from 2009, we expected to see degraded performance on recent data due to change in language, slang, and talked about topics. Therefore, we also evaluated the model on a dataset containing 22 thousand hand-labeled tweets collected between 2013 and 2016. The dataset is a subset of the SemEval-2017 Task 4 dataset (Rosenthal et al., 2017). The evaluation results are described in Table 3.

Dataset	Accur.	Precis.	Recall	F1
s140	0.8356	0.8059	0.8901	0.8459
sEval	0.8132	0.8892	0.8459	0.8670

Table 3: Evaluation metrics (s140 - Sentiment140 subset, sEval - SemEval2017 subset)

Examples of execution on new data from snscrape are described in Table 4.

5 Conclusion and Future Work

Twitter will continue to be an important medium of communication in the future, and therefore be used as a political tool by many. This is why it is critical to detect and analyze the public sentiment on Twitter, as many people use it as a source of information which could influence various aspects ranging from lifestyle habits to even voting preferences. In the future, we aim to advance our model by using more data and trying different hyperparameters to achieve better results. We also wish to develop a public platform that displays Twitter sentiment on topics of interest, such as economic factors and geopolitical events.

Keyword	Sentiment
“Joe Biden”	0.3908
“Donald Trump”	0.5128
“Vladimir Putin”	0.2336
“Russia”	0.2886
“Ukraine”	0.3340
“Novak Djokovic”	0.8814

Table 4: Estimated sentiment of Twitter users’ (0 - negative, 1 - positive) of sample topics with data from December 1st, 2022 scraped using snscrape. **NOTE**: the score for a specific topic does not necessarily or plausibly mean that the topic is directly referred to negatively. For the complete details on how this data was generated, refer to the codebase.[2]

[2] https://colab.research.google.com/drive/1HZ1nmOF76vHX7aKgTxr4gF2XilJWFg-K?usp=sharing

References

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., & El-Kishky, A. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. *arXiv preprint arXiv:2209.07562.*

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.