Twitter Sentimental Analysis Model Comparison

Eliana Mugar
emugar@bu.edu

Samantha Rigor
srigor@bu.edu

Abstract

This paper describes the usage of sentiment analysis models on the TweetEval dataset from HuggingFace (Barbieri et al., 2020). The *Sentiment Analysis in Twitter task* is a popular task run yearly at SemEval since 2013 ((Nakov et al., 2013); (Rosenthal et al., 2014); (Rosenthal et al., 2015); (Nakov et al., 2016); (Rosenthal et al., 2017a)). Using the TweetEval dataset, we ran two different sentiment analysis models. Model 1 consisted of using Cardiff NLP's "twitter-roberta-base-sentiment" model (Rosenthal et al., 2017b), and Model 2 consisted of using Cardiff NLP's "bertweet-base-sentiment" model. While Model 2 had slightly lower numbers than Model 1, we found that the two models had similar results for each of the 4 metrics we recorded: accuracy, F1, precision, and recall. We explore the quality of the dataset, noticing the difficulties of annotating and labeling the sentiments to their respective tweets based on the annotators and the sentiment of certain topics of contemporary society.

1 Introduction

Like other social media platforms, Twitter gives its users an opportunity to express themselves by posting tweets and retweeting others' tweets. However, in recent years, Twitter has transformed from a place to catch up with friends and family to a platform with many categories of posts: news, misinformation, pop culture, sports, and so much more. With this in mind, it can be easy to fall victim to the opinions and sentiments with which other users post online. For this reason, we have chosen to study how sentiment analysis models evaluate tweets for our midterm assignment.

TweetEval is a multi-class tweet classification dataset that consists of seven heterogenous tasks in Twitter which include irony, hate, offensive, stance, emoji, emotion, and sentiment. All the tasks have been unified into the same benchmark, with each dataset presented in the same format with fixed training, validation, and test splits. The main supported task with this dataset involves text classification, with specific subtasks of intent classification, multi-class classification, and sentiment classification. Because we were interested in tasks involving sentiment classification, TweetEval was a good dataset to select to run sentiment analysis models. Since we chose two models from the same source, Cardiff NLP, we hypothesize that the models will report similar metrics.

2 Model Evaluation

The two models we chose are two different sentiment analysis models that have the same number and type of labels. We want to compare the metrics of sentiment analysis models that have similar label systems, since other sentiment analysis models seemed to have a five-label Likert scale.

In choosing the two models, we found two sentiment analysis models from the same developers, Cardiff NLP. By comparing these two models, we can also determine whether there is a significant difference in metrics between Model 1 and Model 2, which would indicate which model is better at sentiment analysis as well as show whether these developers have improved their models over time.

2.1 Model 1

Model 1 is Cardiff NLP's 'twitter-roberta-base-sentiment' model. It is a roBERTa-base model trained on about 58 million tweets and finetuned for sentiment analysis with the TweetEval benchmark that is suitable for English tweets. There are three labels: negative, neutral, and positive. To test this model, we used the datasets, evaluate, and transformers libraries in Python to compare the model's predictions to the dataset's assigned labels. Additionally, rather than testing the entire dataset, we also made a smaller, random set of 1,000 tweets out of the 'test' data split to determine the accuracy, F1, precision, and recall of the 'twitter-roberta-base-

sentiment' model. After running the model, the code reported that the accuracy is 0.735, the F1 is 0.735, the precision is 0.737, and the recall is 0.735.

2.2 Model 2

Model 2 is Cardiff NLP's 'cardiffnlp/bertweet-base-sentiment' model. There is not much information about the model on HuggingFace, so it is not clear what data, if any, was used to train and fine-tune this model. As with the first model, there are three labels: negative, neutral, and positive. To test this model, we used the same code we used to run the first model, which includes making the smaller, randomized 1,000 tweet dataset from the 'test' split in the larger TweetEval dataset. After running the model, the code reported that the accuracy is 0.725, the F1 is 0.724, the precision is 0.727, and the recall is 0.725.

3 Dataset Auditing

In observing the TweetEval dataset truncated to 1,000 random tweets, each datapoint seems to be correctly annotated. Sentiment is quite subjective, so some tweets may seem more negative or positive rather than neutral label it has been assigned to, but overall, the overall negative and positive sentiment labels seem correctly annotated.

Each datapoint is relevant to the task it intends to evaluate and fitting for sentiment analysis models. In checking other sentiment analysis models, most of them tended to use five labels for their label mapping to provide a wider scale of opinions and sentiments, but the simplicity of three labels (negative, neutral, positive) restricts the dataset with reasonable limitations.

If multiple qualified annotators provided annotations, there would most likely be disagreement with the positive and negative labels. The concern regarding the positive and negative labels would be whether the tweet itself is overall positive or negative for the agent or the patient of the tweet. Regarding the neutral label, it makes sense if there are is a high disagreement because it may not be neutral for the agent of the tweet, the patient of the tweet, or the annotators evaluating the tweet. The more annotators there are, the more susceptibility to disagreement there may be.

The datapoints provide sufficient coverage of the phenomena described by the task. The tweets range from pop culture references to news outlet stories. There is a wide variety of subjects and topics discussed in the tweets that it gives a good overview of the Twitterverse.

The tweets selected in this dataset are English only (not multilingual), so there is a geographic bias towards English, specifically in the United States. Therefore, majority of negative, positive, and neutral sentiments will be regarding the state of affairs in the United States and how the general American media may feel towards certain topics. The dataset does a decent job and labeling certain news headlines as neutral, but politically aggressive or supportive tweets tend to get their respective positive and negative labels, but in other parts of the world, they may not receive the same positive or negative sentiments.

4 Conclusion

After running the sentiment analysis models on the TweetEval dataset, the metrics overall have similar numbers, with Model 1 being slightly higher than Model 2. In analyzing the specific metrics, the accuracy indicates the number of correct cases out of the total number of cases. The accuracy of both models score similarly with Model 1 being 73.5% accurate and Model 2 being 72.5% accurate. As mentioned in the model evaluation section of this paper, Model 1 was specifically stated to be trained on the TweetEval dataset while Model 2 did not contain any information on training sets, so this higher accuracy may be attributed to the specific training and fine-tuning processes used in developing Model 1.

Though both sets of metrics reported high accuracy, running these models on other Twitter datasets would provide better insight on the models' real capabilities of analyzing sentiment. As we ran these models, it was not clear as to whether or not the 'test' data split contained information from the 'train' and 'validation' data splits, which were, at the very least, used in the training and fine-tuning of Model 1. Finding another Twitter dataset or developing our own custom dataset would help us to better understand these models and determine whether they serve as good models for Twitter sentiment analysis. While these models may be high-performing on datasets they are familiar with, their true value will shine through in their ability to be generalized to larger, more varied datasets.

Limitations

In analyzing these models, we acknowledge that there are limitations in how we tested the two sentiment analysis models we chose. For example, as stated in the Dataset Auditing section, multiple annotators may disagree on the assigned "correct" labels to each of the tweets in the TweetEval dataset. On a similar note, we ourselves have not annotated or contributed to the dataset, so our perception of sentiments may vary greatly from what has been assigned by the annotators or by the models.

Additionally, we did not use the same 1,000-tweet dataset from the 'test' data split while running these models and recording their performance metrics. Although these models are being run on the same overall dataset, it is possible that one model could have received a test set that contains more tweets with more ambiguous sentiments. Future comparisons of these models would benefit from running the same data with the two models to determine if the two sets of metrics would remain similar or appear significantly different from one another.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. pages 1–18.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. *Proceedings of the 7th International Workshop on Semantic Evaluations (SemVal 2013)*, pages 312–320.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017a. Semeval-2017 task 4: Sentiment analysis. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 502–518.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017b. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80.