Abstract

In pursuit of computer-run emotion detection, Albu et al. successfully created BERT and SVM models to predict emotion from tweets [1]. These models were designed to pick up on emotional cues from tweets, ignoring irrelevant information like usernames and focusing primarily on emoticons. Both models expressed individual success; however, their ensemble showed the highest accuracy. For this project, we aimed to replicate Albu et al's findings by reproducing their code and hoping to obtain similar results. Despite some computational obstacles and slight numerical differences, our reproduction yields very similar results, expressing high accuracy for both the BERT and SVM models individually, but even higher accuracy at their intersection.

Introduction

This report acts as a partial reproduction of Albu et al.'s paper "Emotion Detection From Tweets Using a BERT and SVM Ensemble Model" [1]. The paper's main focus is emotion detection, which is a branch of semantic analysis, and the authors trained and tested their models on the WASSA dataset [2], a benchmarking emotion detection dataset. 1500 tweets from each category of fear, sadness, joy, and anger were extracted, and the authors also added 1500 neutral tweets from CrowdFlower to account for tweets that cannot be confidently categorized into one of the aforementioned groups. The study utilized SVM and BERT models to investigate their emotion detection performances separately, as well as in an ensemble framework. To build the final ensemble model, SVM and BERTweet [3] were combined via vector addition of the output log probabilities of each model. The results revealed a superior performance of the BERTweet-SVM model on the test set compared with each individual model. Upon reproduction of the paper's results, we found that our code revealed issues in a few sections. However, we managed to successfully run the paper's code and observed similar results, namely the superior performance of the BERTweet-SVM model.

Dataset

The training and testing of each tweet's focal emotion was conducted using data obtained from the WASSA dataset, which was provided to the participants in the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-2017). 1,500 tweets were selected to represent each emotion of fear, sadness, joy, and anger, without considering the magnitude of the emotion. Additionally, a class of 1,500 neutral tweets from CrowdFlower4 were included, as having a neutral class is important for the designation of tweets with unknown emotions during the classification process. Finally, dataset divisions were tuned to ensure a well-balanced dataset with an equal number of samples per class, which helps prevent the model from favouring larger classes during the classification process.

Models

SVM

The preprocessing phase of the SVM model involved traditional machine learning preprocessing operations: emoticon to word conversion, Unicode to ASCII conversion, stop words filtering, sentence tokenization and vectorization, and label encoding. The tweets were first preprocessed by eliminating unnecessary words and artefacts from tweets (usernames, links, hashtag symbols), and then translating emoticons into their meaning using the Demoji Python library for Unicode emoticons, and Wikipedia for western-style emoticons. Emoticon-to-word conversion was used due to the fact that emoticons are usually strong indicators of emotion in tweets, as they are frequently used in text messages to convey feelings. Also, to convert tweets into numerical values, the authors used tokenization as well as term-frequency times inverse document-frequency (tf-idf).

BERT

The paper examined three variations of BERT: the original BERT (referred to as vanilla BERT), an improved version of BERT called RoBERTa (Robustly Optimised BERT approach) [4], and a RoBERTa model trained on tweets called BERTweet [3]. With consideration to the SVM model, the BERT models had different preprocessing methods.. In particular, only links and usernames were removed, as they were deemed unimportant and could potentially bias the model. Additionally, each BERT version underwent specific preprocessing techniques such as word tokenization, padding, building the attention mask, and adding BERT tokens.

BERT-SVM ensemble model

Throughout recent history, ensemble models have been used in various fields with the aim of combining the benefits of several ML models concurrently, and this process usually achieves better results when the models have different architectures. This study combined the BERTweet model and SVM model with the RBF (Radial Basis Function) kernel to generate the study's ensemble model. To achieve this, the authors performed vectorized addition on the log probability outputs of both models.

Results

The paper concluded that the ensemble model provided better performance compared with individual models. In our reproduction, we also observed superior performance of the ensemble model in comparison with the BERTweet model and SVM models (Figures 1, 2).

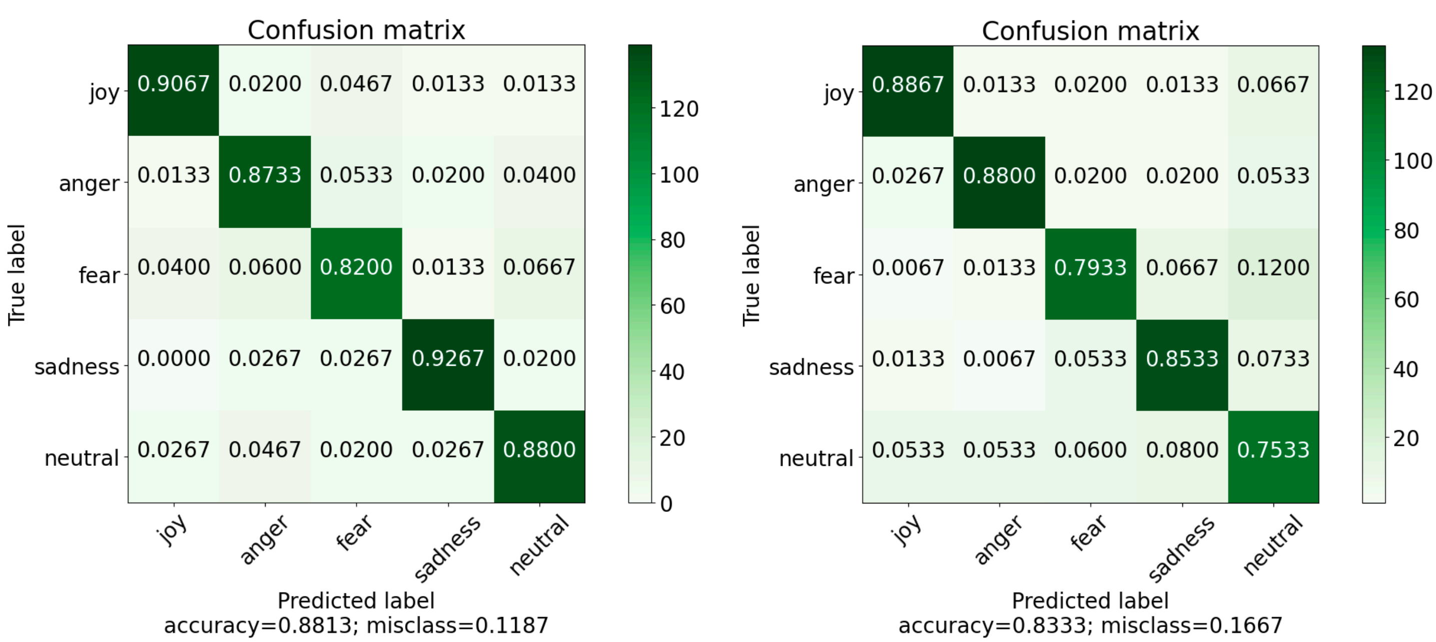

Figure 1. Confusion matrix and performance of the BERTweet (**A**) and SVM (**B**) models on the test set.

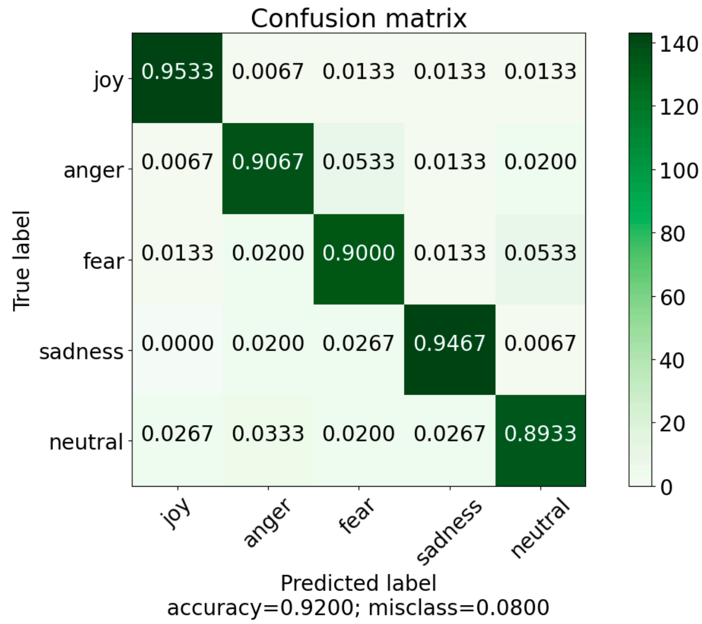

Figure 2. Confusion matrix and performance of the ensemble model on the test set.

Although the numerical values for accuracy and the confusion matrix were slightly different than those presented in the manuscript, they provide very similar results and match the same accuracy ranking of the ensemble being the highest, followed by BERTweet, then SVM with the lowest accuracy (Table 1).

Model	Accuracy	Accuracy (Albu et al.)
SVM	0.83	0.84
BERTweet	0.88	0.89
Proposed ensemble model	0.92	0.91

Table 1. Accuracy comparison between selected emotion detection models in our reproduction and Albu et al.'s study.

Reproduction challenges

The reproduction of Albu et al.'s results was relatively straightforward. However, we encountered one major issue that took some time to resolve, which was that one of the methods from the SVM model seemed to have been changed in later versions of scikit-learn. The authors did not specify which version of the scikit library they used, but after some investigation, we managed to downgrade Colab's scikit version to a suitable one. Additionally, between the .ipynb files provided by the previous study, there were inconsistencies in the shape of prediction arrays. Specifically, one of the BERT models had its prediction NumPy array of the dimensions (600,4), while the SVM prediction array had dimensions (600,5), creating errors upon the vector addition of these two arrays for the ensemble. This was because one of the provided BERT models disregarded the "neutral" class in obtaining the results, while the SVM did not, creating a necessity to edit the code.

Conclusion

"Emotion Detection From Tweets Using a BERT and SVM Ensemble Model" from Albu et al. displays the power of ensemble models where even a classical model, such as SVM, could enhance the project's performance when combined with more sophisticated models, such as BERTweet. As classical models tend to be less expensive in terms of computation time and resources, combining these models with complex neural networks could enhance task performance while having a mild impact on computation time and resources. It's also evident that these two models individually perform worse than their ensemble, further emphasizing the common world idea that the whole is greater than the sum of its parts. With regard to the original study, it is undeniable that Albu et al.'s BERT, SVM, and ensemble models can be easily reproduced and yield very similar results when trained on the same dataset. Therefore, their study was successfully completed to scientific research standards. as our project emphasizes the reliability and reproducibility of their work.

Contributions

Mahdi Mahdavi: Running the code. Writing and editing of the report.
Taylor Fergusson: Editing the code. Writing and editing of the report.
Yu Cheng: Writing and editing of the report.

References

1. Albu I.-A., and Spînu, S., *Emotion Detection From Tweets Using a BERT and SVM Ensemble Model*. ArXiv:2208.04547, 2022.

2. Mohammad, S.M. and Bravo-Marquez, F., *WASSA-2017 Shared Task on Emotion Intensity*. arXiv preprint arXiv:1708.03700, 2017.

3. Nguyen, D.Q., Vu, T., and Nguyen, A.T., *BERTweet: A pre-trained language model for English Tweets*. arXiv preprint arXiv:2005.10200, 2020.

4. Liu, Y., et al., Roberta: *A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692, 2019.