Exploring Text Classification and Emotion Detection Using Keras Models on Reddit Data

Yue Shao
Troy High School, California, U.S.A.

Keywords: Text Classification, Machine Learning, Keras.

Abstract: Human communication has evolved alongside technology, with the internet playing a pivotal role in contemporary digital communications. This research focuses on the emotional aspect of text communication by developing a machine learning model capable of classifying emotional polarity within Reddit posts. Using a TensorFlow Keras model, the study explores how varying the number of epochs and batch sizes influences model accuracy. TextBlob was used to generate polarity labels for the large Reddit dataset, providing a supervised learning framework for the study. Despite initial issues with a Keras layer incompatibility and processing limitation, the final model achieved an accuracy of 0.8619 on a test sample of 24,053 Reddit posts. The research highlights the challenges encountered during model development, particularly related to time constraints and the computational limitations of Google Colab. The findings suggest that further optimization and larger datasets could improve performance in future iterations. This study demonstrates the potential of AI to analyze emotional content in large-scale communication data, contributing to the growing field of sentiment analysis and emotion classification in social media contexts.

1 INTRODUCTION

Communication has always been important to mankind as a whole, setting human apart from other animals. While almost all animals can communicate, humans excel at it. Communication in the society plays a key role on a macro level, facilitating the advancement of technology for example, but also on a micro level, allowing individuals to share their emotions, opinions, and feelings. Throughout the years, communications have facilitated advancements in technology, and technology has also facilitated advancements in technology. The telegram allows messages to be transferred in seconds, the telephone allows for audio to be transferred, the radio allows for audio to be transferred wirelessly, the television allows video to be displayed across the globe, and now people have the internet which is far more complex than any of these previous systems, and arguably have a far greater impact.

The internet is responsible for almost all digital communications in contemporary times. Long distance communication is almost done entirely using the internet and it facilitates operations that can only be sustained with itself. This was highlighted during the 2020 pandemic lockdowns which saw almost all communications moved online. The role that it played professionally was impressive, increasing efficiency to various degrees despite the global conditions. However, its role in personal relations should not be understated. The internet has been used to connect those that would have never met without it and facilitate the exchange of ideas on a scale never before seen. This high volume of communication data has been analyzed by major companies and governments already, but many elements of it remain unanalyzed.

With the resurgence of artificial intelligence in recent times (Holzinger, 2019; Holmes, 2004; Kaul, 2020), it can be noted that there weren't a lot of models that were focused on the human aspect of communication despite all the information and data that exists. Companies and developers are much more interested in the data such as user interest and other information that can benefit them. Recognizing this issue, this project decided to create a model that could focus on the human aspect seen in communications, emotions and feelings.

There exist many platforms that contain the information necessary to train the model. In this study, reddit was ultimately picked to be the source for the training data for the model. This is due to the

Shao, Y.
Exploring Text Classification and Emotion Detection Using Keras Models on Reddit Data.
DOI: 10.5220/0013337100004558

In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 413-416
ISBN: 978-989-758-738-2

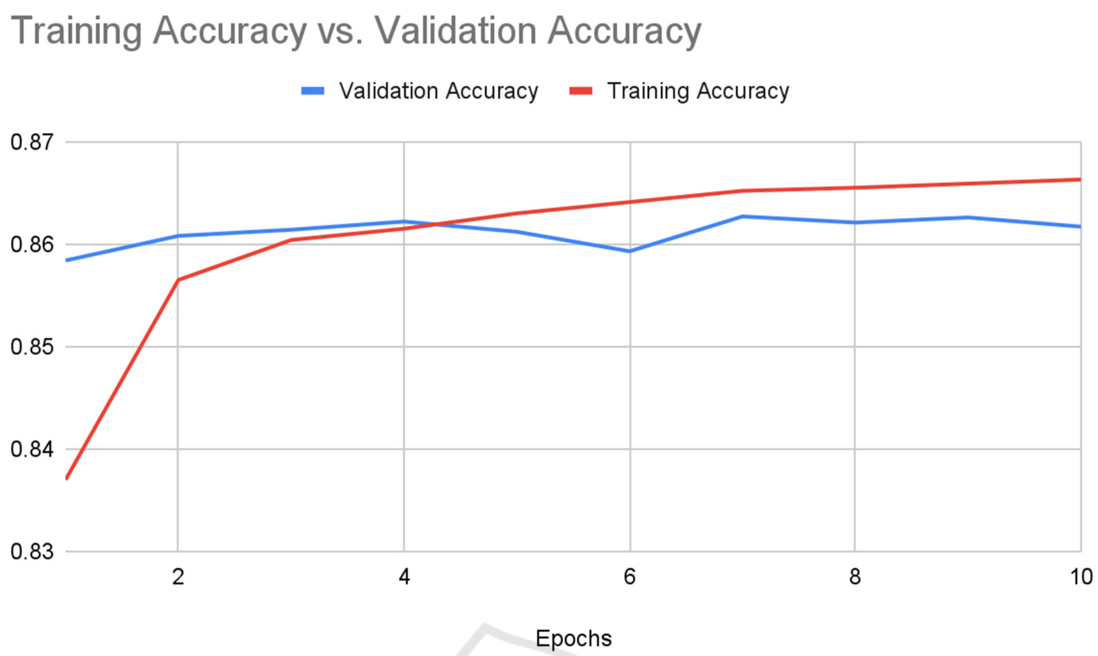

Figure 1: Training accuracy and validation accuracy of the model (Photo/Picture credit: Original).

fact that Reddit is a social media platform with a large community, resulting in being able to obtain more data as a result. Reddit also has a structure called subreddits, which are narrower and more focused. These subreddits contain more personalized content and therefore those that regularly utilize it are likely to be more passionate about the topics. This brings in data of those that are more emotionally invested in the topic meaning a higher variety of emotions that the model can be trained on. This data is also publicly available unlike conversations that take place within private messages like text messages and direct messages. These factors stated above are only a few of many reasons why Reddit was selected as the platform of choice to use the training data from. The actual dataset used is the Reddit dataset from Tensorflow. This study aims to complete the text classification and it is an important branches of machine learning (Kowsari, 2019; Mirończuk, 2018; Gasparetto, 2022; Minaee, 2021).

2 METHOD

The Reddit dataset that was used had information for the author, body, content, id, normalizedBody, subreddit, subreddit_id, and summary. These are all presented as string types. The later model used textblob to augment the dataset with a polarity label.The description given by the documentation says that “TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and more.” This method was necessary due to the amount of time it would take to assign a label by hand to such a large dataset. This was also only done for the second model

Google Colab was the platform that the research group decided to conduct research through (Bisong, 2019; Kuroki, 2021; Gunawan, 2020; Kanani, 2019). The intended output is an indicator of the degree of emotion contained inside a text message, this study designated this as the polarity label. The project originally aimed to get a polarity result ranging from 0.0-1.0. However, the project ultimately had to settle on 1 or 0 for the polarity after the model took too long to train a small sample from the dataset and it was realized that due to time and processing power limitations it would be unfeasible to train a model that can output more specific data.

The first model used Neural-Net Language Models by Google to create text embeds for the model. The model also extensively used Keras. This model ran into many issues handling Keras layer and was abandoned due to this fact. Specifically, the problem with the first model was the Kerastensor being incompatible with tensorflow functions and mismatches between the Keraslayer and Tensorflow hub’s Keraslayer.

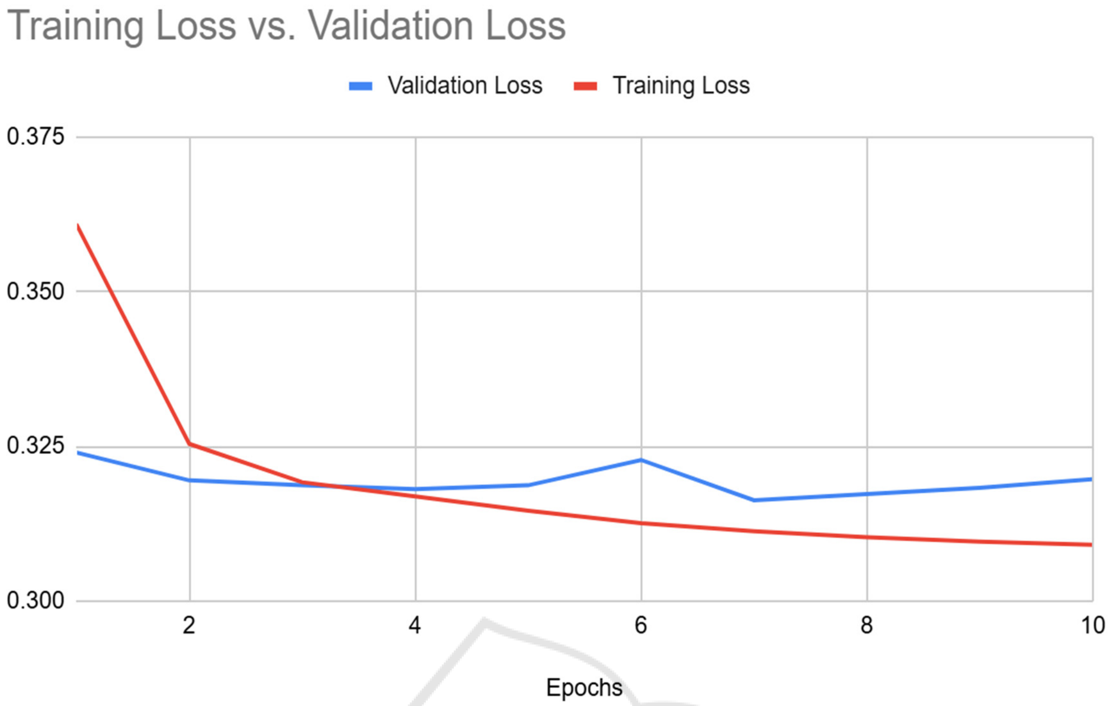

Figure 2: Training loss and validation loss of the model (Photo/Picture credit: Original).

The improved model didn't utilize Google's nnlm. The second model used a modified version of the training dataset containing a polarity label by TextBlob. The version of Tensorflow and Tensorflow Hub was also downgraded in order to resolve the issues that the project ran into for the first model. Initially this led to errors that made the model unable to be trained due to KerasLayer being added to a sequential model.

3 RESULTS AND DISCUSSION

The final model had a final accuracy of 0.8619 and loss of 0.3188 when used on a test sample of 24053 samples. The result of the 10 training epochs can be seen in Figure 1 and Figure 2 showing the accuracy and loss respectively. Each epoch used 96209 samples. The training accuracy for the first epoch was 0.8370 with the validation accuracy being 0.8584 as shown in Figure 1. Figure 2 also shows the training loss for the first epoch was 0.3609 and the validation loss was 0.3240. The training accuracy for the 10th epoch was 0.8663 and the validation accuracy was 0.8617. The training and validation loss for the 10th epoch was 0.3091 and 0.3197 respectively.

This model performed slightly worse than expected. This is because the output that expected was already simplified. Therefore, it was expected that the accuracy would be higher as a result. However, the project also hypothesized it might have been the cause for the lower-than-expected accuracy due to less detail in the label.

One of the primary limitations that the project ran into was the time and processing power. This was due to the project being done using Google Collab Notebook and the natural limitations with the platform. Time limitations were due to the large size of the dataset and therefore some data was left out of the training set which may have led to better accuracy. Without these limitations the model can be modified to produce a float output determining the emotion rather than the integer in this study.

Another potential cause for lower-than-expected performance could be the data in the 6th epoch. The accuracy and loss for the 6th epoch did not follow the general trend for the training, as can be seen in both Figure 1 and Figure 2. Improvements that can be applied in future variations of the code can be to use a dataset with a more accurate label. The project also seeks to apply the model on other text datasets to get a wider variety of text.

Despite the below than expected performance, this project achieved the intended purpose of creating a model that is capable of rating emotional polarity and producing a simple output. If given an opportunity in the future without the time and processing limitations, it would be more than plausible to improve the model to produce a more accurate output.

4 CONCLUSIONS

This project successfully developed a machine learning model capable of analyzing emotional polarity in Reddit posts, achieving an accuracy of 0.8619 after ten epochs of training. While the model performed reasonably well, the results were somewhat lower than expected due to simplified polarity labels and limitations imposed by the Google Colab platform. One of the key challenges in this project was the handling of the large dataset, which required adjustments in training parameters, such as batch size and epochs. Time and processing power limitations prevented the research from experimenting with more granular output (float values for polarity), which could have provided a more detailed emotional analysis. Instead, the project settled on binary output (0 or 1) for polarity, which likely contributed to the reduced accuracy.

Despite these limitations, the project achieved its primary goal of creating a functional model for emotional polarity classification. The errors encountered, particularly the anomaly in the sixth epoch, highlight the complexity of training deep learning models and the importance of sufficient computational resources. Future research should aim to overcome these barriers by utilizing more robust computing platforms, allowing for finer-tuned models and the inclusion of additional data.

One promising direction for future work is to experiment with more accurate and varied datasets, which could improve the model's ability to recognize complex emotional patterns in text. Furthermore, expanding the project to include more nuanced outputs for emotional intensity could increase its usefulness in real-world applications, such as customer sentiment analysis or mental health monitoring. Overall, the research demonstrates the potential of artificial intelligence to analyze human communication, contributing valuable insights into emotion detection and text classification in the digital age.

REFERENCES

Bisong, E., & Bisong, E. 2019. Google colaboratory. Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners, 59-64.

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. 2022. A survey on text classification algorithms: From text to predictions. Information, 13(2), 83.

Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. 2020. Development of video-based emotion recognition using deep learning with Google Colab. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(5), 2463-2471.

Holmes, J., Sacchi, L., & Bellazzi, R. 2004. Artificial intelligence in medicine. Ann R Coll Surg Engl, 86, 334-8.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

Kanani, P., & Padole, M. 2019. Deep learning to detect skin cancer using google colab. International Journal of Engineering and Advanced Technology Regular Issue, 8(6), 2176-2183.

Kaul, V., Enslin, S., & Gross, S. A. 2020. History of artificial intelligence in medicine. Gastrointestinal endoscopy, 92(4), 807-812.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. 2019. Text classification algorithms: A survey. Information, 10(4), 150.

Kuroki, M. 2021. Using Python and Google Colab to teach undergraduate microeconomic theory. International Review of Economics Education, 38, 100225.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. 2021. Deep learning-based text classification: A comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40.

Mirończuk, M. M., & Protasiewicz, J. 2018. A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36-54.