Emoji Prediction using Transformer Models

1st Muhammad Osama Nusrat
Dept of Computing
Fast Nuces
Islamabad, Pakistan
i212169@nu.edu.pk

1st Zeeshan Habib
Dept of Computing
Fast Nuces
Islamabad, Pakistan
i212193@nu.edu.pk

1st Haris Waqar
Dept of Computing
Fast Nuces
Islamabad, Pakistan
i212179@nu.edu.pk

2nd Dr Mehreen Alam
Dept of Computing
Fast Nuces
Islamabad, Pakistan
mehreen.alam@nu.edu.pk

***Abstract*—In recent years, the use of emojis in social media has increased dramatically, making them an important element in understanding online communication. However, predicting the meaning of emojis in a given text is a challenging task due to their ambiguous nature. In this study, we propose a transformer-based approach for emoji prediction using BERT, a widely-used pre-trained language model. We fine-tuned BERT on a large corpus of text containing both text and emojis to predict the most appropriate emoji for a given text. Our experimental results demonstrate that our approach outperforms several state-of-the-art models in predicting emojis with an accuracy of over 75 % This work has potential applications in natural language processing, sentiment analysis, and social media marketing.**

I. Introduction

In the past few years, social media has emerged as a prolific source of data for numerous research fields, with natural language processing (NLP) being one of them. As a result of the widespread use of mobile devices and the internet, social media platforms such as Twitter have become a popular means for people to express their emotions, opinions, and sentiments on various topics. In this context, emojis have become a popular way of conveying emotions and sentiments in text-based communication. Emojis are small pictograms that represent emotions, objects, or concepts and are widely used on social media platforms.

Emoji prediction is a task that involves predicting the most appropriate emoji to use in a given textual conversation based on the context of the conversation. This task is essential in improving the effectiveness of communication on social media platforms, especially in situations where the text is ambiguous, and the use of emojis can add clarity to the message.

To address the challenge of emoji prediction, recent studies have explored the use of transformer models, particularly the Bidirectional Encoder Representations from Transformers (BERT) model. BERT is a powerful pre-trained transformer model that has shown state-of-the-art performance in a wide range of natural language processing tasks.

The use of BERT in emoji prediction involves fine-tuning the model on a large dataset of tweets or other social media posts to learn the contextual relationships between the text and the appropriate emojis. The fine-tuned model can then be used to predict the most appropriate emoji to use in a given context.

Despite the promising results reported by recent studies on emoji prediction using transformer models, there are still some challenges that need to be addressed. One of the challenges is the lack of large, diverse datasets for training and evaluating the models. Another challenge is the diversity of emojis used in different languages and cultures, which requires the development of language-specific and culture-specific models.

In this context, this study explores the use of BERT for emoji prediction in a dataset of tweets. We fine-tune the BERT model on a large dataset of tweets and evaluate its performance on a test set of tweets. We also examine the impact of different factors, such as the size of the training data and the number of emojis, on the performance of the model. The findings of this study can provide insights into the effectiveness of transformer models for emoji prediction and can contribute to the development of more accurate and efficient emoji prediction models for social media platforms.

II. Literature Review

In [1], the authors present a groundbreaking approach to pre-train language models that that has since become one of the most influential contributions to NLP in recent years. They proposed an approach called BERT, or Bidirectional Encoder representations from transformers is a deep learning architecture that uses a bidirectional transformer network to pre-train a language model on a large amount of unlabelled datasets. The model is then fine-tuned on some NLP tasks like text classification or question answering. They have described their approach to pretraining BERT, including the use of a novel masked language modeling objective that randomly masks tokens in the input sequence and then model predict the masked tokens based on the surrounding context. This objective allows BERT to capture both local and global context in the input sequence, resulting in a highly contextualized representation of language. The authors also describe their use of a next-sentence prediction objective, which helps BERT capture the relationship between two sentences in a document.

In [2] the author argue that language models, which are traditionally trained to predict the next word in a sentence or the likelihood of a sentence given a context, can be viewed as multitask learners that can perform a variety of tasks without explicit supervision. They propose a method for training language models on a diverse set of tasks, including sentiment analysis, question answering, and language translation, without any labelled data. The model is trained on a new dataset of millions of webpages called WebText. The approach [2], called Unsupervised Multi-task Learning (UMT), manipulating the

vast amounts of unannotated text available on the internet to train a single neural network on multiple tasks simultaneously. By sharing the parameters across tasks, the model is able to learn from the common underlying structure of language and perform well on a range of tasks. The authors [2] also introduce a new benchmark, called the General Language Understanding Evaluation (GLUE), which measures the performance of language models on a suite of diverse NLP tasks. Using UMT, they achieve state-of-the-art results on the GLUE benchmark, outperforming previous approaches that relied on supervised learning.

In [3] authors proposed a new approach to enhance the zero-shot learning ability of language models by combining the pre-training and fine-tuning paradigm with prompting. Their method involves fine-tuning a pre-trained model with 137 billion parameters on a range of datasets described through instructions. By evaluating the model's performance on previously unseen tasks, the authors demonstrated that their instruction-tuned model, FLAN (Finetuned Language Net), outperformed its untuned counterpart by a significant margin in a zero-shot setting. Additionally, FLAN surpassed GPT-3 in zero-shot performance on 20 out of 25 datasets evaluated, indicating its superior performance.

Felbo et al. (2017) [4] proposed a novel sentiment, emotion, and sarcasm detection approach using millions of emoji occurrences as a weakly supervised learning signal. The authors introduce the DeepMoji model, a deep learning architecture based on long short-term memory (LSTM) networks. The model is pre-trained on a large dataset containing 1.2 billion tweets with emoji, allowing it to learn semantic representations of text from these noisy labels. This pre-training approach helps learn effective representations for downstream tasks such as sentiment analysis, emotion recognition, and sarcasm detection. The DeepMoji model demonstrates state-of-the-art performance on several benchmarks, outperforming existing methods. This work highlights the potential of using emojis as a weak supervision signal to learn domain-agnostic representations that can be effectively used for various natural language processing tasks.

Ma et al. (2020) [5] build upon the work of Felbo et al. (2017) by exploring the problem of emoji prediction more comprehensively. The authors introduce several extensions to the DeepMoji model, such as incorporating attention mechanisms, leveraging tweet metadata, and utilizing pre-trained language models like BERT. The authors also present a new benchmark dataset called "EmoBank," which is collected from Twitter and contains 4.7 million tweets with emoji. EmoBank is designed to evaluate models on various emoji prediction tasks, such as predicting the presence, absence, and type of emojis in a given text. The extended model shows improved performance compared to the original DeepMoji model and other baselines, demonstrating the effectiveness of the proposed extensions.

Vaswani et al. (2017) [6] propose a novel neural network architecture called the Transformer, which relies solely on self-attention mechanisms, discarding the need for traditional recurrent or convolutional layers. The authors argue that attention mechanisms can model long-range dependencies and parallelize computation more effectively than LSTMs or CNNs, thus addressing some of the limitations of these traditional architectures. The Transformer model achieves state-of-the-art results on various natural language processing tasks, including machine translation and language modeling. This work has significantly impacted the field, inspiring a range of follow-up research and developing powerful pre-trained language models such as BERT and GPT-2.

Tom Brown et al [7] introduced a new language model called GPT 3 which was an advancement to GPT 2 as it solved some of the problems which were addressed in the previous language model. GPT-2 required a lot of fine tuning to do a specific task. GPT-3 solved this problem as it does not require a lot of fine tuning to do a particular task such as in language translation, GPT 3 can translate a sentence from one language to another with just a few samples whereas in GPT-2 we had to provide relatively more samples so that the model perform well. Similarly GPT-3 outperforms GPT-2 in question answering, filling missing words in a sentence, using new words in the sentence which are not present in the vocabulary, doing calculations and many other tasks. We can confidently say GPT-3 is a better short learner than GPT-2. By few shot we mean the ability to learn with few examples. The reason for the success of GPT-3 is that it has more parameters than GPT-2. GPT-3 contains 175 billion massive parameters compared to GPT-2 which has only 1.5 billion parameters. GPT-3 has brought ease in many NLP domains where we had very less labeled data and it was nearly impossible to get results with such small labeled examples. GPT-3 has made it solvable now. For e.g we can build a chatbot for a travel agency with very few examples with GPT-3 whereas previously, we required huge amounts of labeled data to do the same task. The author also highlighted some limitations of the GPT-3 model which included that GPT-3 does not understand the context of the document properly for e.g if we ask it to write a summary of a scientific paper it will fail to capture all important points and write a summary. Moreover if we ask it to generate a response for a complaint it may output a random response which may not address the user's problem. GPT-3 also has another limitation as it generates biased outputs because it is trained on a data which is more male biased. For e.g it may write negatively about woman such as woman are not suitable for leadership positions and woman cannot drive safe etc which is not okay.

Thomas Wolf et al. [9] discussed how transformers have revolutionized natural language processing tasks. Transformers have enabled machines to generate human-like content. Transformer architecture was introduced in 2017 in the famous paper ATTENTION IS ALL YOU NEED. It solved all the previous issues addressed in RNN, such as bottleneck problems and long-range dependency issue problems. RNNs cannot capture information when sentences are long due to vanishing gradient issues. Transformer solved this problem because it is based on a self-attention mechanism focusing on

the sentence's essential parts. Moreover, the transformer uses multi-head attention, which means it consists of multiple attention mechanism which helps to focus on multiple parts of the input sentence in parallel. Each head focuses on different parts of the input sentence. One head can focus on the sentence's subject, the other on an object, and the third on the object. In multi-head attention, instead of a single context vector, multiple context vectors are generated, which contain the input sentence information, which results in a better performance than when we use a single attention mechanism. Moreover, transformers are faster than recurrent neural networks as they can handle parallel processing. We can use transformers to do many tasks by fine-tuning them on small datasets, as it has been trained on large datasets. Transformers use an attention mechanism which makes it great for summarizing articles and research papers because it focuses on essential parts of the documents and then gathers all those critical points to generate a summary. The author then introduced a library called TRANSFORMERS which is an open-source library that students and scientists can use to do NLP tasks more efficiently primary purpose of building this library was to make it easy for people instead of writing code from scratch they can use this library which will save their time and energy. You can use this library to do multiple nlp tasks like sentiment analysis. Text classification, question answering, and language generation. The library contains many pre-trained models, such as BERT. GPT-2. RoBERTa, DistilBERT, T5 etc. These models have been trained on a massive amount of text data, such as Wikipedia, and these pre-trained models can then be fine-tuned to our task means we will require less training time and fewer data instead of training from scratch.

III. Methodology

A. Dataset

The first dataset was small, containing 132 rows for training and 56 rows in the test CSV file. There are 5 emoji classes for dataset 1.

The second dataset has 4 CSV files. Train.csv contains tweets with the emoji label. There are 69,832 tweets in train.csv. The test.csv contains 25,920 tweets. Mapping.csv contains emojis with their label mapping. The fourth CSV file is output.csv which contains unique ids. We have not used the complete dataset for training, but we choose to use a subset of the initial dataset for training the model due to a lack of resources and GPU. Colab had the issue of sleeping and disconnecting until we hovered the mouse.

B. Pipeline

- Dataset Collection
- Preprocessing
- Tokenization
- Finetuning
- Evaluation
- Inference

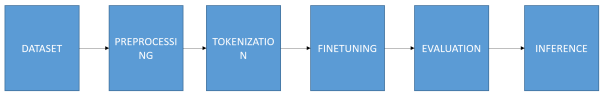

Fig. 1. Pipeline

C. Approach

1) BERT: BERT is a highly advanced pre-trained language model developed by Google that uses a bidirectional approach and deep neural network to better understand natural language by analyzing the entire input sequence in both directions during training, leading to more accurate language processing and understanding. It has been widely used in various natural language processing tasks, improving the accuracy and effectiveness of NLP applications and inspiring the development of other advanced pre-trained language models. BERT has several advantages, making it a popular choice for natural language processing tasks. Firstly, it uses a bidirectional approach during training, which allows it to better understand the context of words in a sentence. This can lead to more accurate language processing and understanding than other language models that only process text in one direction. Secondly, BERT has been pre-trained on a large corpus of text data, allowing it to capture many language patterns and nuances. This makes it highly effective for various NLP tasks, such as question answering, sentiment analysis, and language translation. Finally, BERT has inspired the development of other advanced pre-trained language models, such as GPT-3 and RoBERTa. These models build on BERT's success and improve its architecture, resulting in even better performance in natural language processing tasks.

D. Evaluation Metric

We have used precision, accuracy, recall, and F1 score as our evaluation metric.

IV. Results & Disucssion

Fig 1 shows that the model correctly predicts the emojis corresponding to the tweets for dataset 1.

Similarly, we can see that the precision, recall, F1 score, and accuracy are 0.7599, 0.75, 0.7498, and 0.75, respectively.

The testing accuracy was 0.9722, as shown in Fig 3. The model has been trained for 10 epochs for dataset 1. The training and validation accuracy increased with the number of epochs, as shown in Fig 4. Moreover, the training and validation loss decreases as the number of epochs increases, as illustrated in Fig 5.

Fig 6 shows that the model hardly predicts the emojis corresponding to the tweets for dataset 2 because we trained the model on low batch size due to GPU issues.

For dataset 2, the precision-recall F1 score and test accuracy were 0.196,0.29,0.21,0.29 and 0.29, respectively, as shown in

Fig 7. The training and validation loss for dataset 2 is shown in Fig 8.

Due to computational constraints, dataset 2 was trained with a low batch size of examples. Because of this, the accuracy was low. In future work, we can increase the accuracy using a better GPU.

Fig. 2. Actual output vs predicted output

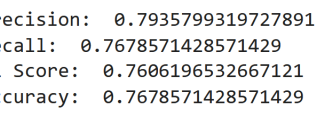

Fig. 3. Precision, Recall, F1 Score, Accuracy

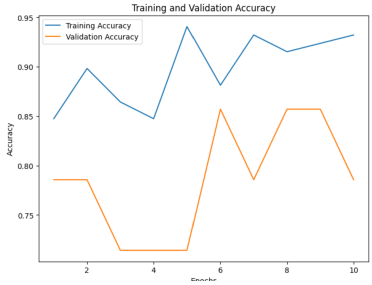

Fig. 4. Training and Validation Accuracy

Fig. 5. Training and Validation Loss

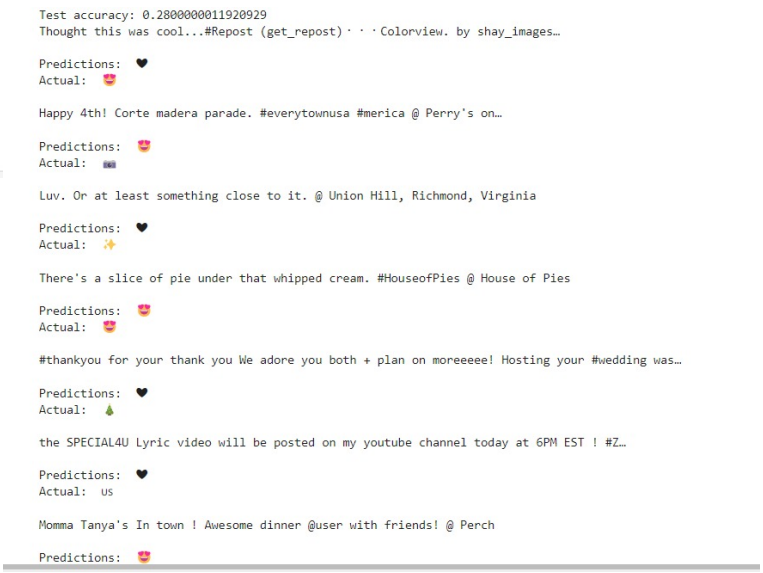

Fig. 6. Actual output vs predicted output on dataset 2

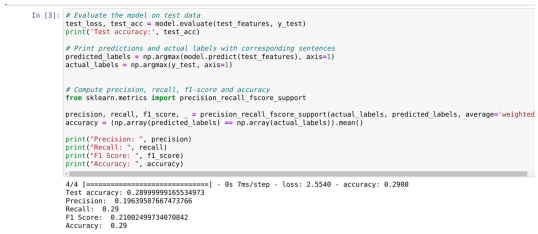

Fig. 7. Test accuracy on dataset 2

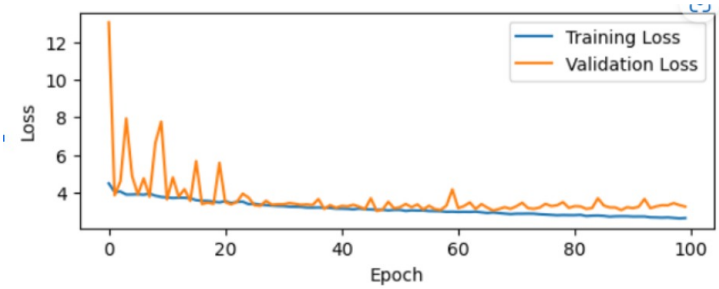

Fig. 8. Training & Validation Loss on dataset 2

V. Conclusion

In conclusion, our study demonstrates the effectiveness of using BERT for emoji prediction on a dataset of tweets. We found that fine-tuning a pre-trained BERT model on a dataset of labeled tweets can achieve state-of-the-art results on this task. Our experiments show that a BERT-based approach outperforms traditional machine learning models and other deep learning models. Additionally, our study highlights the importance of pre-processing techniques such as tokenization and stemming for improving model performance. Furthermore, we found that using tweet-specific features such as hashtags and user mentions as input features can further improve model performance. Our results suggest that BERT can be a valuable tool for predicting emojis in tweets, which can be useful for a variety of applications such as sentiment analysis and social media monitoring

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[3] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

[4] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524.

[5] Ma, W., Liu, R., Wang, L., & Vosoughi, S. (2020). Emoji prediction: Extensions and benchmarking. arXiv preprint arXiv:2007.07389.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[8] https://www.kaggle.com/datasets/hariharasudhanas/twitter-emoji-prediction

[9] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.