ML Practical - Project Proposal

Vlad Udrea (s4645014), David Vanghelescu (s4683889), Hristo Karagyozov (s4796683)

November 24, 2023

1 Preliminary domain knowledge

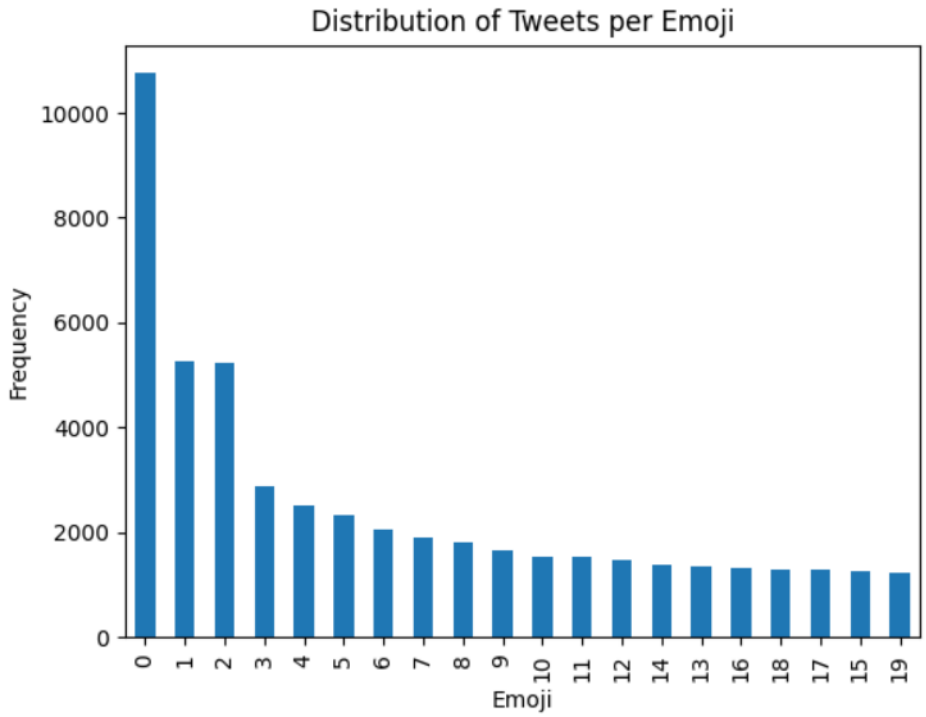
Our machine learning project is centered around building a model which classifies tweets into specific emoji categories. The task we are trying to accomplish is predicting the most appropriate emoji given a certain inputted tweet. Our corresponding research question is therefore as follows: "Can a machine learning classifier accurately predict the most appropriate emoji given a certain tweet?"

Previous work into emoji recognition tasks [3] suggests that there will be a significant difficulty in the ambiguous nature of emojis. Elements like irony are very prevalent as some studies suggest [2]; this, alongside with some emojis just being more popular than others, may create uneven spread in the data. This may turn out to be an issue because it could skew the results and make the model less accurate in predicting the correct emoji for a given tweet [1].

For example, if the training data is dominated by the few most frequently used emojis, the model may become very adept at predicting those while struggling to make sense of the less common uses of emojis. This could severely compromise our model's ability to handle the real-life diversity of emojis. Moreover, the internet culture has a very fast-paced dynamic which means that trends may change rapidly. Different emojis are being repurposed all the time which means that staying up to date could be a great difficulty for our model in the long run.

2 Preliminary data exploration

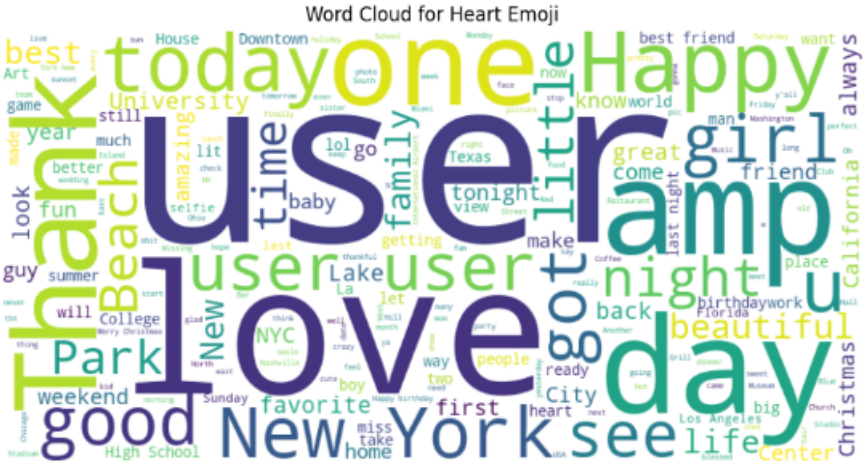
To test the hypothesis we made before looking into the data, we made the following histogram which shows the number of tweets per emoji:



This confirms our hypothesis of the heavily uneven spread of the tweets across the different emojis. The red heart, smiling face with hearteyes and face with tears of joy emojis are present in almost 40% of the tweets. This makes sense, as the face with tears of joy emoji is used in many ironic contexts which heavily increases its usage compared to more specific emojis.

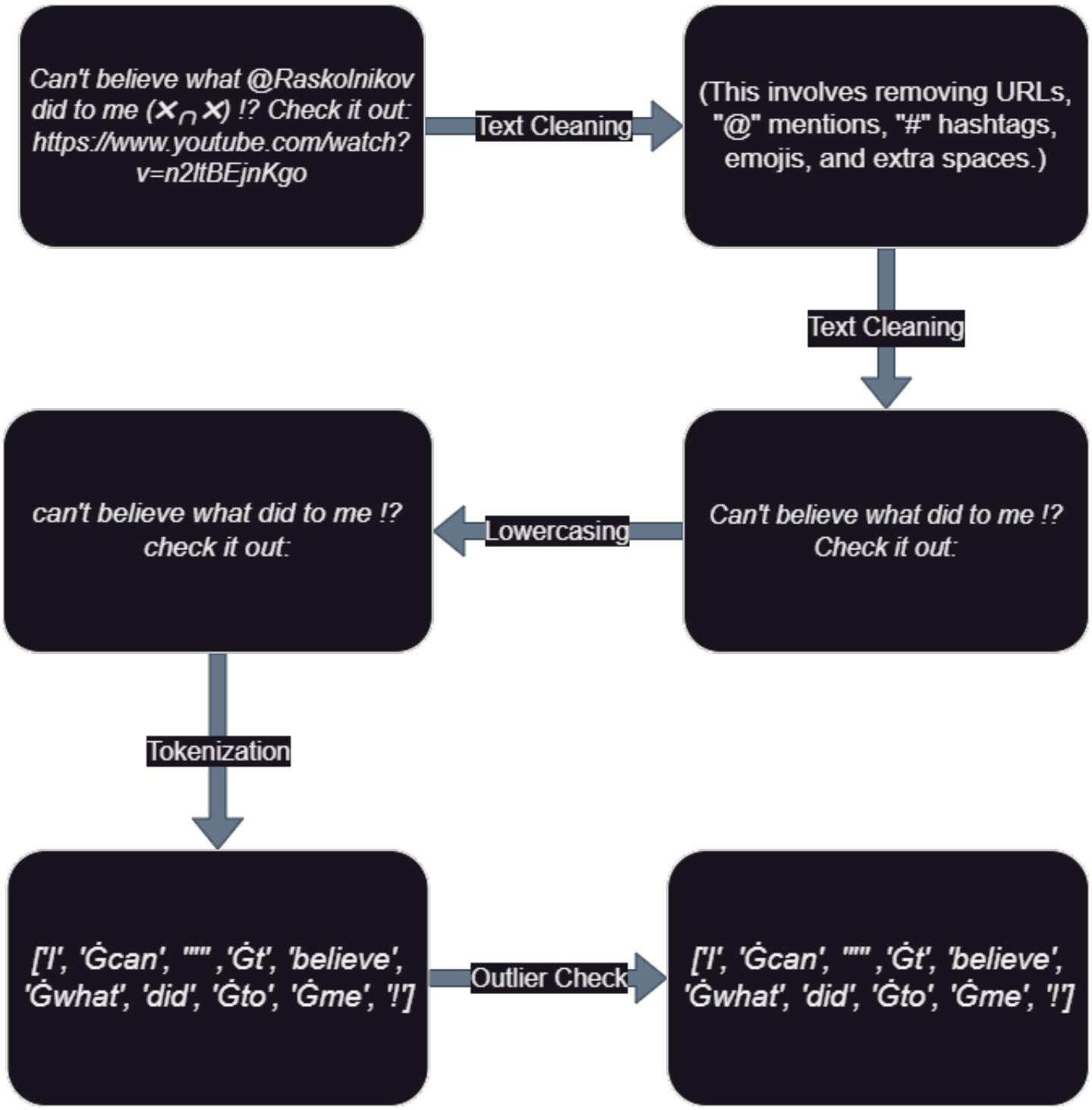
This heavy skew could negatively impact our model and could be an issue should we decide to delve into sentiment analysis which notoriously struggles with irony which is very hard to properly capture [4].

To represent the frequency of the different words per emoji, we also created a word cloud:



User being the most frequent word is of course due to the lack of preprocessing, which will remove mentions and special characters so that the truly defining and semantically-important words will remain.

3 Proposed preprocessing



The majority of the preprocessing steps we will perform on our model is described in the graph above. The main steps are as follows:

Data cleaning - this method takes care of URLs, mentions, special symbols, stop words (e.g. "and, so, the"), unnecessary symbols or spaces, emojis (in various forms, like text for example) and so on.

Lowercasing - this method converts all the text to lowercase. This is because the model is case-sensitive, and lowercasing ensures that the model treats the same word in different cases as the same word.

Tokenization is an essential preprocessing technique to almost every NLP task. In addition to that, we will perform lemmatization on the tweets to bring all of the different derivatives of the words to a standard form. We believe this will help our model.

One thing that we noticed is that certain emojis can be used in a multitude of contexts. As such, we propose to perform sentiment analysis on the tweets to establish a potential relationship between the emojis in a tweet and the sentiment of that tweet.

Sequence Padding is a crucial preprocessing step when working with models like RoBERTa, where input sequences must have uniform lengths. We omitted this step from the diagram as it aims to represent preprocessing in a more tangible way. The process involves adding extra tokens to sequences to ensure they all match a specified length. In our case, which requires consistent sequence lengths, this becomes necessary to facilitate efficient batch processing.

4 Proposed model(s) + baseline

Our baseline model is going to be word2vec with logistic regression. It is a classic text classification approach because it has simplistic preprocessing and it is relatively straightforward but effective.

Our comparison model is going to be built upon the improved BERT model RoBERTa which is a transformer-based model. It is a state-of-the art model which is excellent at capturing context which is the part which we think will demonstrate a great improvement over the baseline model. RoBERTa can potentially outperform simple models especially when we are talking about the intricacies of natural language processing.

For the baseline model we are going to train Word2Vec embeddings on the training data, then use these embeddings as features for logistic regression. Training should be relatively fast and straightforward, as the preprocessing.

For our advanced model we are going to fine-tune a pre-trained RoBERTa model on the task at hand. This means we are going to adjust the last few layers to adapt to the model.

In terms of hyperparamters we are going to potentially tune regularization strength and class imbalances for the baseline and learning rate, batch size and epoch number for the advanced model.

5 Proposed evaluation

Our model will be assessed by looking at the accuracy of each emoji prediction. Furthermore, we will also look at the confusion matrix to see if certain emojis are used in similar cases. We believe that certain emojis may be used in different cases such as ironically which may affect the accuracy of our model. We also want to compare a version of our model with sentiment analysis with another version without sentiment analysis in order to see how important the overall sentiment of a given tweet affects the selection of a certain emoji.

6 Model usage

Our model does not have a real-life usage outside of testing different preprocessing methods/NLP algorithms. It learns to map a certain input(tweet) to a certain emoji. A similar but much more advanced model can be used in natural language processing to add the correct emojis to the outputed text. As such, it can communicate with the user not only by using natural sounding text but also by using adequate emojis for that situation.

7 Risk assessment

Our model only uses a small part of the entire emoji collection that is available to the users (only the 20 most used ones) and will be limited in terms of performance in the real world. In order to fix this problem, a much larger data set will be needed as well as a lot more computational resources. One societal problem that can arise as a result of deploying such a model is that it can become really hard to discern AI-generated text on social media, which in turn can lead to dangerous parasocial relationships between human users who think they are conversing with another human being, when they are actually conversing with an AI.

8 Individual learning outcomes

David: I want to know more about working in NLP, such as how to build a good model and how to create an efficient preprocessing pipeline.
Hristo: I am mainly interested in learning the decision-making process of an AI specialist.
Vlad: Given the proclivity of chat-bots in the last year alone (e.g. GPTs) and the importance of NLP, not just to natural language interactions between AI models and humans, but also to my own Bachelor's Thesis, I wish to expand and perfect my knowledge of this topic, delving deeper than I ever have before, from both a practical and a theoretical point of view.

9 Member contributions

Vlad contributed for the plot and preprocessing. David contributed for the model + baseline, risk assessment and model usage. Hristo contributed for the preliminary domain knowledge, proposed evaluation and the diagram.

References

[1] Qiyu Bai, Qi Dan, and Zhe Mu. Emoji and its use in chinese social media. *Psychological Reports*, 124(2):354–362, 2020.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020.

[3] Muhammad Osama Nusrat, Zeeshan Habib, Mehreen Alam, and Saad Jamal. Emoji prediction in tweets using bert. 2023.

[4] Leila Weitzel, Raul A. Freire, Paulo Quaresma, Teresa Gonçalves, and Ronaldo Cristiano Prati. How does irony affect sentiment analysis tools? In *Portuguese Conference on Artificial Intelligence*, 2015.