Tweets sentiment analysis

Huailin Tang
ht35@rice.edu

1. Introduction

Sentiment analysis is a powerful tool that enables social media to understand the emotions of their users. In recent years, users on many platforms have used a set of graphic symbols in online conversations to express their emotions. These graphic symbols are called emojis. In general, these emojis provide the same amount of information compared to words. However, most of the sentiment analysis removes emojis as noisy labels. This project aims to explore two forms of incorporating emojis in the sentiment analysis, keeping original emojis or replacing emojis with descriptions, and how these incorporations influence the accuracy of sentiment analysis compared to removing emojis.

2. Literature Survey

”Emojis and Emoticons in Sentiment Analysis: A Review” by H. D. Le, T. Q. Dang, and D. T. Nguyen (2020) [1] provides a comprehensive review of the use of emojis and emoticons in sentiment analysis, including their advantages and limitations, and discusses potential future research directions. Some important advantages are emojis can add additional context to the text, providing a more nuanced understanding of the sentiment being expressed; emojis can improve the accuracy of sentiment analysis models, especially when used in combination with text; Captures emotional intensity: Emojis can help capture the intensity of emotions being expressed, such as the degree of happiness or sadness. Some important limitations are: emojis can vary in their interpretation across different platforms and contexts, making it difficult to standardize their use in sentiment analysis; emojis have a limited vocabulary and cannot express complex sentiments or ideas, making them less useful in analyzing more complex texts.

Liu et al. (2021) [2] studied incorporating emojis in sentiment analysis of Chinese text. They found that emojis are effective as expanding features for improving the accuracy of sentiment analysis algorithms, and the algorithm performance can be further increased by taking different emoji usages into consideration. They also developed the LSTM model to achieve the best performance (95% accuracy) when analyzing Chinese texts.

3. Hypothesis

Based on the papers about emojis, it is reasonable to hypothesize that incorporating emojis could improve the accuracy of the sentiment analysis and keep the original emojis producing the best result.

4. Experiment

The dataset this experiment use is the Twitter dataset which has a large number of tweets without sentiment labels. You can download the dataset from this website.

The first step is preprocessing. Firstly, combining data from the individual JSON files to one Pandas dataframe. Secondly, identifying the language of tweets using fasttext library and only keeping English tweets. Third, remove URLs, mentions, hashtags, retweets, extra spaces, numbers, punctuations, and empty strings or NA. Fourthly, identifying the emotions of the tweets using VADER library. Fifthly, lowercase the words, remove the stopwords using NLTK library, and lemmatize the words in the tweets using NLTK library. Now, I have the first dataset that has processed texts with emojis, which is the first form of incorporation. I generate the second dataset by replacing the emojis with the text description of the emojis using the emoji library, which is the second form of incorporation. Lastly, removing emojis from the text from the first dataset and forming the third dataset as the benchmark data to compare.

The second step is applying the same machine learning classifiers to three datasets and comparing their performance for bag-of-words representation. The classifiers are Naïve Bayes, Logistic Regression, Linear SVC, Random Forest, XGBoost, and LSTM. Due to the limitation of time and computation resources, this project does not aim to find the best hyperparameters for each classifier, instead it uses the same hyperparameters for each classifier on each dataset and tests their performance. For Naïve Bayes, Linear SVC, and XGBoost, I use default hyperparameters; For Logistic Regression, I specify solver = ‘saga’, because ‘saga’ solver is useful for large-scale learning, particularly when the number of samples is significantly larger than the number of features; For Random Forest, bootstrap is set to be false, the number of trees is set to be 100, and use entropy for the splits; For LSTM, I set the dimensionality of the output space of the embedding layer to 128, the number of output units in the LSTM layer to 196, amount of dropout regularization to apply to the LSTM layer are both 0.2, 64 samples per gradient update, and 7 times the model is trained on the entire dataset.

The metric I use to measure their performance is accuracy. I also present the f-1 score and confusion matrix in the code for reference.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F - score = \frac{2 * Recall * Precision}{Recall + Precision}$$

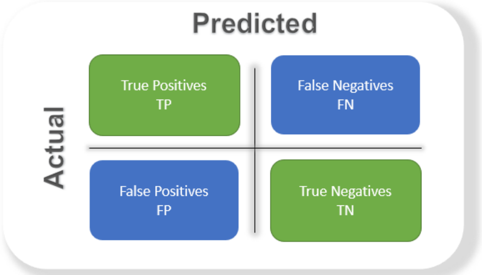
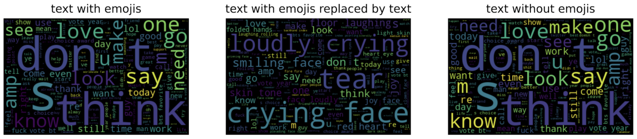


Figure 1: Confusion matrix. [3]

5. Result

The visualization of the Wordcloud of three data. The Wordcloud of text with emojis and text without emojis are similar, while Wordcloud of text with emojis replaced by descriptions is different.

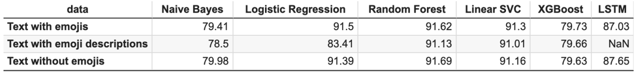

text with emojis

text with emojis replaced by text

text without emojis

Figure 2: Word Cloud

The comparison of the performance of classifiers on the three data. From the accuracy of the classifiers with the same hyperparameters, we can see including emojis with text improves some of the classifiers' accuracy, such as Logistic Regression, Linear SVC, and XGBoost, while decreasing other classifiers' accuracy, such as Naive Bayes and LSTM. Replacing emojis with descriptions generally decreases the classifiers' accuracy, such as Naive Bayes, Logistic Regression, Random Forest, and Linear SVC, while only increasing the accuracy of XGBoost slightly.



data	Naive Bayes	Logistic Regression	Random Forest	Linear SVC	XGBoost	LSTM
Text with emojis	79.41	91.5	91.62	91.3	79.73	87.03
Text with emoji descriptions	78.5	83.41	91.13	91.01	79.66	NaN
Text without emojis	79.98	91.39	91.69	91.16	79.63	87.65

Figure 3: accuracy

note: LSTM for text with emojis descriptions is not trained due to computation resources

6. Conclusion

Overall, the experiment result does not justify the idea that incorporating emojis in the text could be helpful and improving the accuracy of the sentiment analysis as we see the mixture of improvement and decline of the accuracy by including emojis in text and mostly decline of accuracy by including descriptions of emojis.

The explanation could be Emojis can be ambiguous and their meaning can depend on the context in which they are used. For example, the ”smiling face with sunglasses” emoji can indicate happiness or a feeling of coolness, depending on the context. Another explanation could be emojis are not always used consistently: Emojis can have different meanings in different cultures and contexts. Moreover, people often use different emojis to express the same emotion. This makes it difficult to create a comprehensive dictionary of emoji sentiments, especially for the relatively small dataset that is used for this project.

7. future study

This experiment explores how incorporating emojis could affect the accuracy of the sentiment analysis in the bag-of-words representation. It is worth discovering the impact on other representations, such as TF-IDF and Word2Vec.

It is also interesting to see the difference in the accuracy of the classifiers that have optimized hyperparameters between datasets that include and do not include emojis.

The future study should use a larger dataset, if possible, due to the diverse use and meaning of the emojis.

References

[1] H. D. Le, T. Q. Dang, and D. T. Nguyen. Emojis and emoticons in sentiment analysis: A review. *Journal of Information Science Theory and Practice*, 8(2):14–30, 2020.

[2] Y. Liu, X. Li, Y. Li, and W. Liu. Incorporating emojis in sentiment analysis of chinese text. *IEEE Access*, 9:78111–78119, 2021.

[3] S. Narkhede. Understanding confusion matrix, Jun 2021.