Understanding Emojis for Sentiment Analysis

Byungkyu Yoo, Julia Rayz
Purdue University
West Lafayette, IN
yoob@purdue.edu, jtaylor1@purdue.edu

Abstract

Many people use emojis to express themselves in a more clear and efficient way and the usage of such expressions has grown significantly. In recent years, there have been many research papers that analyze the meanings of individual emojis or show the accuracy of sentiment analysis when emojis are included. However, there is limited research done on understanding how emojis are used to show sentiment and how it affects sentiment analysis. In this paper, we analyze the usage of emojis in Tweets and their effects on the overall sentiment of the Tweet. We also introduce a pre-processing method for emojis that increases the effect of emojis and as a result, improves the sentiment analysis accuracy.

Introduction

Billions of social media posts are created every day. With the increase of social media there has been an increase of use of emojis and emoticons. Emojis and emoticons are widely used in social media to express emotions, moods, and ideas, especially when people struggle to express their emotions through pure text, or they show sarcasm. These emojis and emoticons have great value as they have a big impact on expressing the sentiment of a sentence. Thus, the emojis and emoticons are valuable resources for sentiment analysis.

Sentiment analysis is a natural language processing technique used to analyze a text and identify the attitude behind it. In most cases, it will be classified into three categories, positive, neutral, and negative. It is widely applied to understanding customer satisfaction, monitoring social media, and predicting stock price using social media data.

Emojis are not only being used to improve sentiment analysis but also for other composite tasks like sarcasm detection. The idea that one single emoji could have such an impact on the sentiment of a statement was the motivation for this paper. This paper aims to find how emojis are used in social media, how emojis affect the sentiment detection of a sentence, and how to improve the sentiment analysis accuracy by using emojis.

Literature Review

Although many people use emoticons and emojis interchangeably and emoticons and emojis often have identical meanings, the two are different. Emoticons are a pictorial representation of a facial expression using characters written only with characters that are on the keyboard. Examples include ":)" and ":(". Emojis are graphic symbols that represent a concept like "😀". Each emoji is registered as a Unicode character and cannot be typed using keyboards. The literature review section will include both emoticons and emojis as they are both used interchangeably, and emojis are generally assumed to be the new generation of emoticons.

Many devices display the emoji using different graphics and therefore may slightly change the meaning(Gupta, Singh, and Ranjan 2020). Twiiter data is slightly different as Twitter has its own emoji graphics called Twemoji. Twemoji is displayed identically in all devices except for the native Twitter applications in iPhones. The meaning of emojis may also differ by culture(Financieras 2020). For example, emoji of a smiling face generally has a positive sentiment in China but has a negative sentiment in Argentina.

Wang and Castanon (2015) performed a study on analyzing the role of emoticons in both building sentiment lexicons and in training learning classifiers. The model described in the paper had approximately 15% improvement in terms of sentiment accuracy. The paper also concluded that large groups of emoticons convey complicated sentiment and should be treated with extreme caution. Emoticons are not perfectly consistent and often depend on the context and the person that uses them. Similarly, there was research done to quantify the effects of emoji in sentiment analysis (Ayvaz and Shiha 2017). The paper discovered that utilization of emojis in sentiment analysis also resulted in a higher sentiment score. The paper hypothesized that emoji characters seemed to have a higher impact on overall sentiments of positive opinions in comparison to negative opinions. This knowledge can be used to help track products, improve services and also predict upcoming events.

Guibon, Ochs, and Bellot (2016) put forward the idea that the effects of emojis are not limited to just sentiment expression but they can be expanded to various other avenues such as sentiment enhancement and sentiment modification. Usage of emojis make them very ambiguous and unreliable when taken out of context, so different usages of emojis can be determined by comparing their sentiment to the sentiment of the sentence they are a part of.

There were several papers that worked with Twitter data

for training their models using emojis. Chen et al. (2018) created an RNN model for Twitter sentiment analysis with bi-sense emoji embedding. With this model, the authors show that emojis can be useful to express more subtle sentiments like anger, sadness, happiness, etc.

To summarize the findings, emoticons and emojis are useful to express a wide range of emotions and not just positive or negative polarity. Using these emoticons and emojis, the sentiment analysis score can be increased. However, there is a gap in research when it comes to understanding the impact of emojis combined with text for sentiment analysis. This paper will try to bridge the gap and provide an analysis of emojis in sentiment analysis that results in a higher accuracy of sentiment analysis.

Methodology

As Twitter is a widely used platform for expressing thoughts and emotions, including the use of emojis and emoticons, Twitter datasets were used for the experiments.

Dataset

Several Twitter datasets were used for the project: one was obtained from Github, two from Kaggle, and the last was obtained by directly from Twitter.

- The first dataset is a Twitter sentiment labeled dataset with emojis. It has 6,600 positive and negative Tweets each, a large enough dataset for accurate sentiment analysis (Prusa, Khoshgoftaar, and Seliya 2015).
- The second dataset is a Twitter sentiment labeled dataset without emojis. 26,400 positive and negative Tweets each were used from the dataset. The second dataset was combined with the first dataset for sentiment analysis.
- The third dataset is a list of Tweets with emojis. 1.8 million Tweets with emojis were used to train the Word2Vec model (Mikolov et al. 2013).
- The fourth dataset is tens of thousands of Tweets with emoticons and or emoji retrieved using the Twitter API.

Preprocessing

Data was preprocessed for sentiment analysis or Word2Vec. To clean the data, hashtags (#Target), URLs, user mentions (@username), and numbers were deleted as they do not have any sentiment value. Punctuation was not deleted to keep any emoticons and other meaningful tokens such as "..." and ":)". Spaces were added between emojis for better tokenization. Tokenization was done using the tokenization method in the NLTK library version 3.5.

Word2Vec Model

Word2Vec (Mikolov et al. 2013), a word embedding model, was used to find out what emojis represent. Both CBOW and Skip-Gram was trained using the the third dataset and fourth dataset. CBOW predicts the likelihood of a word given a context. Skip-Gram predicts a context given a word. The Gensim library version 3.8.3 was used to train the Word2Vec models. The top 100 emojis were searched to find the most similar tokens that represent the emoji.

Sentiment Analysis

The first dataset and a mixed dataset was used for sentiment analysis. The mixed dataset is a combination of the first and second dataset to make the dataset similar to a real Twitter dataset, a mix of Tweets that has emojis with Tweets that do not. The Tweets with emoji and the Tweets without the emoji were combined with a ratio of 1 to 4 as about 20 percent of the Tweets in 2020 had at least one emoji in it (Broni 2020). In total, the mixed dataset had 66,000 sentiment labeled Tweets, half labeled positive and the other half labeled negative.

Both datasets were divided into 6 versions. The intuition behind the versions was to find out what emojis most closely represent. What are emojis used most frequently used with? Would replacing the emojis with its name, meaning, or representative word have any effect on sentiment analysis? These were some of the question this paper attempted to answer. The list of top 100 emojis and its meanings were retrieved from the web (Keely 2020).

- The first version removed all emojis.
- The second version kept all the emojis as it is.
- The third version replaced the top 100 most frequently used emoji with the multiple meanings of the emoji.
- The fourth version replaced the 100 most frequently used emojis with the name of the emoji or representative meaning of the emoji. The name was shortened to one or two words by removing the words like "sign" or "face".
- The fifth version replaced the 100 most frequently used emoji with the closest representative word that was found using the Word2Vec model.
- The last version added a newly introduced rule after the tokenization. If an identical emojis were next to each other, the token on the left was changed to have two same emojis in a single token. For example if there are two smile face emojis, the token on the left would have two smile face emojis and the token on the right would have one smile emoji.

Tokens were vectorized using two method: counting the number of tokens (referred as simple count), and TF-IDF algorithm (Term Frequency–Inverse Document Frequency). TF-IDF gives more weights to tokens that have a higher impact in the corpus(Krouska, Troussas, and Virvou 2016). Both methods were done using the Scikit-learn library version 0.23.2.

The data was randomly split into training set of 80% and testing set of 20%. Multiple machine learning algorithms were used for sentiment analysis to see how the different processing methods of emojis would effect the sentiment accuracy of different machine learning algorithms. All algorithms was done using the Scikit-learn library.

Observations of Emoticon and Emoji

Emojis are indeed the next generation of emoticons and can be treated almost equally. Emojis are more commonly used than emoticons and is replacing emoticons. Manual observations of the fourth dataset was done to confirm several patterns of emoji and emoticon usage.

Both emoticons and emojis are generally used at the end of a clause or the sentence. Observations of three hundred Tweets mostly showed the same pattern. For example, in the sentence “my day started nice, but it ended bad”, a positive emoji or emoticons, like smile face, would be placed at the of the nice. A negative emoji or emoticons, like sad face, would be placed after bad.

Emojis mostly replaced emoticons, not used together. From four thousand Tweets with a happy face emoji “😀”, only 1 Tweet had both emoji and emoticon. However, out of several thousand Tweets with a happy face emoticon ”:)”, there were several hundred Tweets with emojis. Emojis that are used with emoticons mostly express things and concepts that emoticons cannot express easily. Emojis like monkey, cake, and dogs (🐵🎂🐶) are some examples. Emojis that could be easily shown using emoticons were rarely used by emoticon users.

Emojis and emoticons are similarly used in a very similar context. However, when doing sentiment analysis, emojis and emoticons should be processed slightly differently. One thousand Tweets with smile emoticons ”:)” and smile emojis “😀” each showed similar usage patterns. About 7% of the Tweets with smile emojis had another smile emojis next to each other and 8% of Tweets with smile emoticons had another smile emoticons next to each other. About 1% of the Tweets showed that both Tweets with emojis and Tweets with emoticons were used more than twice in different parts of the sentence. The same pattern was showed for the sad emojis and emoticons, “😕” and ”:(”. About 27% of the sad emoji or sad emoticon were used next to each other and showed similar usage patterns. However they should be processed slightly differently. Sad emojis are shown next to each other like “😕😕😕😕” while sad emoticon only added parentheses like ”:((((“. Another difference when processing the data was that emoticons could easily be mistaken. For example, in a Tweet “My favorite food:omelet”, the surprised emoticon ”:o” would have been mistakenly recognized by a computer.

Experimental Results

Emoji Sentiment

Many emojis have a higher sentiment ratio than words. Sentiment ratio was calculated by comparing the number of emojis or words in negative Tweets compared with positive Tweets, or the other way around. A sentiment ratio of 10 to 1 of negative to positive means that the emojis was used in negative Tweets 10 times more frequently than the emoji being used in positive Tweets. The top ten highest sentiment ratio were all emojis. only 3 tokens were words from the top 30, and only 8 were words from the top 30.From the tokens that had a sentiment ratio of 20 to 1, only 5 tokens were words, and the remaining 19 tokens were emojis.

Emojis that are positioned next to each other tend to show stronger sentiment. Sentiment analysis of the first dataset, labeled tweets with emojis, using the version 6 of the dataset, adding identical emojis that are next to each other as a single token, was done. The top 10 highest sentiment ratio is shown on table 1 below.

Token	Sentiment	Ratio
😭😭😭	neg : pos	154.4 : 1.0
😭😭	neg : pos	137.8 : 1.0
😱	neg : pos	135.3 : 1.0
🔥🔥	neg : pos	88.4 : 1.0
💔	neg : pos	86.7 : 1.0
😂😂😂	pos : neg	73.4 : 1.0
😘	pos : neg	61.4 : 1.0
😂😂	pos : neg	52.8 : 1.0
😭	neg : pos	51.7 : 1.0
😰	neg : pos	47.8 : 1.0

Table 1: Highest Sentiment Ratio

The result of version 6 in Table 1 shows that two or three emojis combined as a single token tend to show higher sentiment ratio than a single emoji in a token. As shown with yellow highlights, a single crying emoji compared to several crying emojis next to each other showed a significant sentiment ratio difference. However, comparing two and three emojis next to each other, there was no significant sentiment ratio difference. Another interesting finding was that emoji only show stronger sentiment when they are next to each other, not in different parts of the sentences. If the number of emojis were simply counted and the analysis was done, the sentiment ratio significantly went down.

Sentiment analysis was done using the tokenization method of version 6 and version 2. The accuracy using version 6 of the dataset, two emojis or three emojis in single token, was higher than version 2, the one emoji per token. The table shows the result of the sentiment analysis using the first dataset. The column named accuracy single is the accuracy using version 2. The next column is the accuracy using version 6. “LogReg” in the table below stands for logistic Regression. ”SVM” stands for support vector machine. ”MultiNB” stands for multinomial Naïve Bayes. ”BinaryNB” stands for Bernoulli Naïve Bayes algorithms.

Algorithm	Token Weight	Accuracy Single	Accuracy Double
LogReg	Count	0.9844	0.9848
LogReg	TF-IDF	0.9829	0.9852
SVM	Count	0.9825	0.9829
SVM	TF-IDF	0.9859	0.9863
MultiNB	Count	0.9810	0.9814
MultiNB	TF-IDF	0.9731	0.9765
BinaryNB	Count	0.9712	0.9780
BinaryNB	TF_IDF	0.9685	0.9765

Table 2: Sentiment Analysis Comparison Double Token

Result of the analysis in Table 2 show that the accuracy of the version 6, newly introduced tokenization method, was slightly higher than the version 2. The accuracy raised by 0.04% to 0.8 %. The increase of accuracy was the biggest when using Bernoulli Naïve Bayes algorithms. The increase was smaller when the sentiment analysis was done using the

mixed dataset.

Emoji Meaning

Emojis are much more closer to other emojis or urban phrases than dictionary words. Examples of urban phrases include “lmao” (laughing), “lol” (laughing), “smh” (shaking head), “ily”(I love you), “yikes” (surprised sound), and “haha” (laughing sound). The 100 most used emojis were checked to see the most similar tokens. Word2Vec model was trained using the third dataset, non-labeled Tweets with emojis. Almost all the words that had a higher similarity score of above 0.5 were either emojis or urban phrases. The exceptions were emojis that represent specific objects or concepts, not emotions. For example, rainbow emoji “🌈” showed similarity of 0.61 with token “rainbow” and 0.576 with token “rainbows” followed by LGBT related words. The emoji sweat had “💦” had many sexual words with above 0.5 similarity scores. Other emojis like “💥”, “☀️”, and “❤️” are some examples of these. Even emojis that show strong sentiment like “😭” and “😱” did not have any dictionary words like sad and crying of over 0.5 similarity. Emojis like “😂” and “🤣” were much more similar to tokens like ”lmao” and ”lol” than other dictionary words.

Meaning of emojis generally do not change in a short amount of time as long as the graphics of the emojis do not change. However, there may be exceptions when a nationwide event occurs. Twitter data from 2018, 2020, and before 2018 was gathered to see the difference in meaning. The Twitter data from 2020 was gathered while a pandemic was at its peak. Each emoji had at least 10,000 tweets per emoji to train the Word2Vec model. Emoji such as “😂” and “😱” did not show a big changes in similar words. However the emoji with mask “😷” had much more words similar to the pandemic in 2020 compared to 2018 and before. The different word with high similarity include ”careful”, ”gloves”, ”covid”, and ”pandemic”.

Many researchers replace emojis with its name or several meanings before doing sentiment analysis. Although emojis have many meanings, it should not be replaced with multiple words. Instead, emojis should be replaced with one or few representative meanings. Sentiment analysis was done using the mixed dataset with first five versions mentioned above with the 4 machine algorithms in Table 2. Version 1, Tweets without emoji, showed an accuracy of 75.3% to 76.6%. Version2, Tweets with emoji, showed an accuracy of 76.7% to 78.8%. Compared to version 2, version 3 had a minor change and version 5 had a slight decrease in accuracy. Only Version 5, changing the emojis to one or few words, showed an increase in accuracy compared to version 2 of 0.5% to 0.8% and gave the highest accuracy. Simple count and TF-IDF token method both showed similar results. The main reason is hypothesized to be because the word that replaced the emoji helped trained other words in the dataset. However, multiple meanings are not accurate representations of a single emoji when doing sentiment analysis. Although the increase in accuracy may not be significant, the benefit is that the preprocssing method can be applied to any sentiment analysis or other analysis.

Conclusion

This paper used various tokenization methods, machine learning algorithms, and word embedding models in an attempt to gain a better understanding of emojis and increase the sentiment analysis score.

There were several findings that we saw from this research paper. Emojis generally replace emoticons; they are not used together. Emojis show stronger sentiment compared to words. Emojis next to each other show stronger sentiment ratio and, using this knowledge, sentiment analysis accuracy can be slightly increased. Emojis are generally close to other emojis or modern phrases like “lol” or “haha,” but these words should not be used to replace emojis while doing sentiment analysis. Replacing emojis with one or few words leads to the highest accuracy as sentiments of emoji help train other English words.

References

Ayvaz, S., and Shiha, M. 2017. The effects of emoji in sentiment analysis. *International Journal of Computer and Electrical Engineering* 9:360–369.

Broni, K. 2020. Emoji use in the new normal.

Chen, Y.; Yuan, J.; You, Q.; and Luo, J. 2018. Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM ’18, 117–125. New York, NY, USA: Association for Computing Machinery.

Financieras, C. N. 2020. World emoji day: which are the most commonly used and what meanings have according to cultures. Copyright - CE Noticias Financieras English, Latin America - Distributed by ContentEngine LLC; Last updated - 2020-12-30.

Guibon, G.; Ochs, M.; and Bellot, P. 2016. From Emojis to Sentiment Analysis. In *WACAI 2016*. Brest, France: LabSTICC and ENIB and LITIS.

Gupta, S.; Singh, A.; and Ranjan, J. 2020. Sentiment analysis: Usage of text and emoji for expressing sentiments. In Kolhe, M. L.; Tiwari, S.; Trivedi, M. C.; and Mishra, K. K., eds., *Advances in Data and Information Sciences*, 477–486. Singapore: Springer Singapore.

Keely, J. 2020. The 100 most popular emojis explained.

Krouska, A.; Troussas, C.; and Virvou, M. 2016. The effect of preprocessing techniques on twitter sentiment analysis. In *2016 7th International Conference on Information, Intelligence, Systems Applications (IISA)*, 1–5.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

Prusa, J.; Khoshgoftaar, T. M.; and Seliya, N. 2015. The effect of dataset size on training tweet sentiment classifiers. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 96–102.

Wang, H., and Castanon, J. A. 2015. Sentiment expression via emoticons on social media. In *2015 IEEE International Conference on Big Data (Big Data)*, 2404–2408.