Design and Development of an Emoji Sentiment Lexicon

Fabian Haak

Institute for Information Management, Technische Hochschule Köln
Claudiusstraße 1, 50678 Cologne, Germany
ORCID https://orcid.org/0000-0002-5982-7104, fabian.haak@th-koeln.de

Abstract

Emojis represent an essential means of expressing sentiments such as opinions and attitudes in computer-mediated communication, especially in chats and social media. To effectively capture these sentiments, the sentiments associated with the emojis used must be known. Previous approaches to determining the sentiments expressed with emojis require a large amount of manual annotation. For many emojis, especially less frequently used platform-specific emojis, studies on expressed sentiments do not yet exist. Therefore, these emojis cannot be considered in sentiment analyses so far. In this work, a method for effective and efficient determination of emojis' sentiments and their compilation in a sentiment lexicon was developed. The determined sentiments are compiled as a sentiment lexicon. For this purpose, software was created in Python to process collections of texts into a corpus. The software derives the emojis' sentiments as valence values based on the sentiments of the texts in which the emojis appear. The lexicons produced by the method can be used in lexicon-based sentiment analysis approaches. The method also derives other information on the emojis and their usage that can be used to assess the sentiment lexicon produced and the usage of the emojis. Using the developed method, two analyses were conducted with corpora of different text sources. The results and subsequent comparisons with existing sentiment lexicons have shown that the developed method is able to efficiently produce similar results as sentiment lexicons produced with manual annotation.

Keywords: sentiment analysis; emojis; computer-mediated communication; natural language processing

1 Introduction

Sentiment analysis is an essential tool for natural language processing. *Sentiments* are opinions, feelings, or tonality expressed in texts and can be described in various ways. Sentiment analysis describes the process of extracting these sentiments from texts. Usually, sentiment is derived by sentiment words recorded in lists, so-called *sentiment lexicons*, or by machine learning processes. The capture of sentiments as a measure of polarity and intensity or subjectivity, so-called *valence-based* sentiment analysis, is a fundamental method for analyzing texts. Especially in social media and chats, many opinions are expressed. These are valuable sources of information for social and information science and other academic disciplines. However, the short length of texts from these sources is a challenge for performing practical sentiment analysis. Conventional approaches to capture valence-based sentiments, both lexicon-based and machine learning, have had problems effectively determining the valence of short texts.

An important feature of text communication in social media and chats is the heavy use of emojis. In these informal text forms, they take over many functions of nonverbal communication. Thus, they are also an essential means of expressing sentiments in computer-mediated communication. Opinions and tonality are often conveyed exclusively through emojis. In order to capture these sentiments effectively, the sentiments associated with the emojis used must be known. Previous approaches to determine the sentiments expressed with emojis require a great deal of manual annotation. For many emojis, especially platform-specific emojis, no studies on expressed sentiments exist. Besides, domain-specific knowledge is required to determine the emojis' expressed sentiments. Therefore, these emojis cannot yet be included in sentiment analysis, and determining their valences would be very labor-intensive, as demonstrated by the development of the emoji sentiment lexicon by Kobs et al. (2020). Furthermore, new Unicode emojis are introduced every year. Emoji sentiment lexicons, such as the Emoji Sentiment Ranking (ESR) by Kralj Novak et al. (2015), lose their relevance for sentiment analysis over time and would ideally be updated regularly. Due to these challenging factors, emojis are rarely considered in lexicon-based sentiment analysis. To overcome them, a solution is presented that aims to

1. minimize the need for manual annotation for emoji sentiment determination and

2. improve the results of lexicon-based sentiment analysis, especially of emoji-rich, domain-specific texts such as chats.

2 Data and method for emoji sentiment lexicon generation

In this master thesis on Market and Media Research at TH Köln, a novel method for the effective and efficient automatic determination of emojis' sentiments is introduced. It substantially facilitates the determination of emoji sentiment for the application in lexicon-based sentiment analysis. A software implementation realized in Python derives the emojis' sentiments as valence values based on the sentiments of texts in which the emojis occur. This methodology utilizes emojis' property that they are usually utilized to reinforce sentiment expressed in messages. As Hu et al. (2017) stated, "[…] expressing sentiment, strengthening expression, and adjusting tone are the top three most popular intentions of using emojis". Therefore, we assume that their occurrence in sentimental messages can identify emojis used for expressing sentiment. Furthermore, we assume that these emojis feature the same sentiment as the texts containing them. Previous research, such as by Hu et al., has shown that emojis are used to express irony and sarcasm by adding emojis with valences opposite to the message's. Therefore, a second assumption is that a bimodal distribution of the valences of texts, in which an emoji appears, with local maxima in both polarities, is an indicator for emojis that express irony and sarcasm.

For the implementation of the developed method, two text collections were compiled. The first corpus consists of 2.86 million messages taken from chat logs of the video streaming platform Twitch.tv (2009). Characteristic for chat communication on Twitch is the frequent use of platform-specific emojis, called *Twitch emotes*. Hence, the Twitch corpus ought to demonstrate the developed method's ability to determine platform-specific emoji valences. Due to the large number and their ambiguous nature, the determination of sentiments expressed by Twitch emotes would be very complex using conventional methods. To better evaluate the effectiveness of the developed method and the quality of the assigned valences, a second corpus consisting of 1.48 million publicly accessible Twitter messages (2018) was generated.

This way, valences of Unicode emojis are derived, which can be compared with existing emoji sentiment lexicons valences.

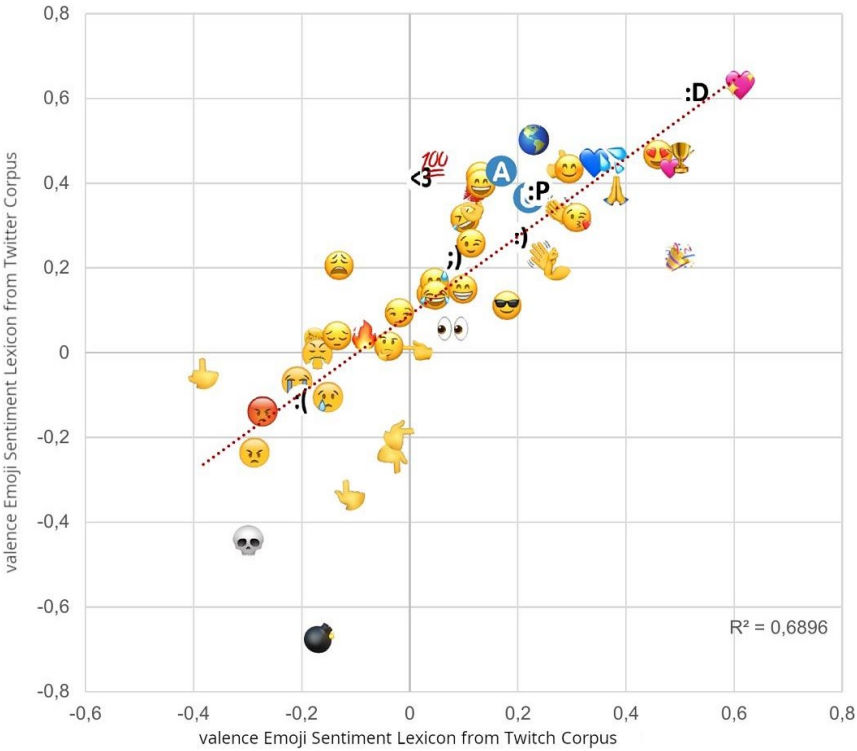

Fig. 1 Comparison of determined emoji valences of emojis, for which both produced lexicons contain valence scores

3 Results

The valences of the emojis are derived by the valences from the texts containing the respective emojis. These are analyzed using the VADER method developed by Hutto and Gilbert (2014), which currently represents the best lexicon-based approach for determining the valences of short informal texts. The VADER method has been modified so that all emojis whose valence must be determined have first been removed from the VADER sentiment

lexicon. This is to prevent that the valences of these emojis are derived from the valences recorded for them in the VADER dictionary. For each emoji, the valence values of all texts containing the emoji are analyzed, and the standard deviation and arithmetic mean of these values calculated. All emojis whose standard deviations are below a certain threshold (0.625) and for which there are enough occurrences (more than ten) are considered valid. These emojis and their corresponding valence values (mean valence of texts they appear in) are exported as a sentiment lexicon. The produced lexicon can easily be inserted into lexicon-based sentiment analysis models. By doing so, the model is extended by information on platform-specific emoji valences, thus increasing the effectiveness of sentiment analysis of texts containing these emojis. The developed method also determines additional information such as TF-IDF values and n-grams about the emojis. Analyses of the two corpora (Twitch, Twitter) were performed, producing two emoji sentiment lexicons. In both analyses, close to 200 valid emojis valences were assigned (Twitter corpus: 198, Twitch corpus: 185, 107 of which are Twitch emotes). Both produced lexicons are publicly available at Zenodo (2020), along with a script that incorporates them into the VADER sentiment lexicon to use them in a standard sentiment analysis toolchain.

3 Evaluation

To evaluate the quality of the produced sentiment lexicons and to be able to judge the effectiveness of the developed method for determining emoji valences, the two lexicons produced were first compared with each other. The comparison of the valence values of the mutual emojis contained in both of the two dictionaries (see Fig. 1) and their Spearman's rank and Pearson correlation show that there is a statistically significant correlation between the two dictionaries. This is a strong indication that the determined valences represent the actual emoji valences and that the method is reliable. For a better evaluation of the quality of the generated lexicons and the developed method, the lexicons were compared with the ESR, and a sentiment lexicon for Twitch emotes developed by Kobs et al. (2020), both produced with manual annotation. The emoji sentiment lexicon generated using the Twitter corpus showed a strong positive Spearman's rank correlation ($r = 0.74$, $p < 0.01$) and Pearson correlation ($r = 0.723$, $p < 0.01$) with the ESR. Comparing the

emoji sentiment lexicon generated by analyzing the Twitch corpus with the *emote lexicon* produced by Kobs et al. also shows a strong and highly significant Pearson correlation coefficient ($r = 0.717$, $p < 0.01$) as well as rank correlation according to Spearman ($r = 0.744$, $p < 0.01$).

4 Conclusion and outlook

The results and subsequent comparison with existing sentiment dictionaries have shown that the developed method can efficiently produce results similar to those of sentiment dictionaries produced with manual annotation. Thus, the approach of determining emojis' valences via the texts in which they occur is a promising method to improve the precision of sentiment analysis of short informal texts. A test with the emoji sentiment lexicon produced by the Twitch corpus analysis, integrated into the VADER model, identified twice as many sentiment expressing texts as the stand-alone VADER model.

In his current work in the ESUPOL project as a Ph.D. student at TH Köln, Fabian Haak's research is related to bias detection in search query suggestions and the impact of query suggestions for search queries of political topics on opinion formation. Sentiment analysis of short texts is highly relevant for bias detection, showing the developed method's potential and future use-cases.

References

Haak, F. (2020). Emoji Sentiment Lexicons [Data set]. http://doi.org/10.5281/zenodo.4159573

Hu, T., Guo, H., Sun, H., Nguyen, T., & Luo, J. (2017). Spice up Your Chat: The Intentions and Sentiment Effects of Using Emoji. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (pp. 102–111). Palo Alto, CA: AAAI Press.

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. http://github.com/cjhutto/vaderSentiment/ blob/master/vaderSentiment/vaderSentiment.py

Kim, J. (2019). Twitch.tv Chat Log Data. Harvard Dataverse [Data set]. https://doi.org/10.7910/DVN/VE0IVQ

Kobs, K., Zehe, A., Bernstetter, A., Chibane, J., Pfister, J., Tritscher, J., & Hotho, A. (2020). Emote-Controlled: Obtaining Implicit Viewer Feedback through Emote based Sentiment Analysis on Comments of Popular Twitch.tv Channels. *ACM Transactions on Social Computing, 3*(2), 1–34.

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLOS ONE 10*(12), e0144296. http://dx.doi.org/10.1371/journal.pone.0144296

Kramer, S. (2018). Social Media Bot Detection by Paragon Science [Data Set]. https://data.world/drstevekramer/social-media-bot-detection-by-paragon-science

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 432–438. DOI: doi.org/10.5283/epub.44960.