Ryan Dimaranan (rtd180003)
Hannah Valena (hcv180000)

ACL Paper Summary

Title: Assessing Emoji Use in Modern Text Processing Tools

Authors: Abu Awal Md Shoeb, Dept. of Computer Science, Rutgers University

Gerard de Melo Hasso, Plattner Institute / University of Potsdam

In this paper, the authors seek to determine how effective popular NLP tools are on datasets that include emojis. Emojis introduce a variety of complications when doing key NLP tasks due to their effect on the text and grammar rules in general. While able to be encoded as Unicode, there are some subtleties that need to be considered such as skin tone modifiers and composite emojis. Also, the use of emojis to replace words, provide missing context and even changing the emotion of the text adds difficulty in performing NLP tasks using tools that were created before the age of emojis and social media.

There have been many studies conducted on how emojis affect the human language and how we communicate since their incorporation into Unicode. Work includes how the increasing availability of emojis have changed grammar rules [3], how skin tone modifiers are used in social media [4], and various studies that compare the features of popular NLP tools such as NLTK and Adobe Stanza to help determine which is the best tool to use when processing text that includes emojis. Given the growing popularity of emojis in everyday communication, there have also been innovations in NLP models that target text with emojis such as prediction models, part of speech taggers and dependency parsers. [4] This paper, however, focuses only on industry standard and widely used open source tool kits.

While the use of emojis and their effect on semantics and human communication have been widely studied since their inception, no study has been done on the performance of popular open source tools on datasets that include emojis out of the box. Also, since some properties of modern emojis are recent additions, use of skin tone modifiers, zero-width joiners, and BMP vs non-BMP emojis have not been adequately investigated on current tools. The results gathered from this paper and their recommendations will surely help future-proof current NLP tasks and tools to better support any current and future innovations in communication using emojis.

The authors evaluated the emoji processing capabilities of open source toolkits based on four basic NLP tasks: tokenization, part of speech tagging, dependency parsing, and sentiment analysis. For relevant tasks, they considered the positions of emojis in a sentence, the number of emojis used, the use of skin tone modifiers, and the use of zero width joiners (which combine multiple emojis into one), to evaluate the accuracy of NLP tools in processing emojis. They created EmoTag, a collection of 22.3 million tweets containing at least one frequently used emoji, that can be found at http://emoji.nlproc.org/ [2]. Representative subsets of this collection were chosen for each NLP task tested, and only clear-cut cases of emoji use were studied in order to evaluate basic support.

The results of this study showed that many out-of-the-box NLP tools have shortcomings with processing emojis, and that a "mix-and-match approach" to using libraries could work best in practice [1]. SpaCy, SpaCyMoji, and NLTK-TT performed the best with tokenization; NLTK-TT and TextBlob were the best with POS tagging; Stanford CoreNLP and Stanza were the best with dependency parsing; and VADER was the best with sentiment prediction. More research needs to be performed for more ambiguous and diverse cases of emoji use, as this paper focused only on unambiguous usage.

Assessing Emoji Use in Modern Text Processing Tools has received five citations on Google Scholar. The authors, Abu Awal Md Shoeb and Gerard de Melo, have received a combined 6,246 citations, with Shoeb receiving 103 citations and de Melo receiving 6,143 citations. Abu Awal Md Shoeb was a PhD student at Rutgers University, specifically studying natural language processing, machine learning, and social media analytics. Gerard de Melo is a professor at the Hasso Plattner Institute and the University of Potsdam, specializing in artificial intelligence, natural language processing, and web mining. The research they have conducted on the emoji-processing capabilities of popular NLP tools will help push these tools to increase support for emojis. Increased proficiency in processing emojis can provide rich data insights, especially with regards to digital communication and social media.

Works Cited

[1] Abu Awal Md Shoeb and Gerard de Melo. 2021. Assessing Emoji Use in Modern Text Processing Tools. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1379–1388, Online. Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.110/

[2] Abu Awal Md Shoeb, Shahab Raji, and Gerard de Melo. 2019. EmoTag – Towards an Emotion-Based Analysis of Emojis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1094–1103, Varna, Bulgaria. INCOMA Ltd. https://aclanthology.org/R19-1126/

[3] Kaye, Linda K., et al. "Emojis: Insights, Affordances, and Possibilities for Psychological Science." Trends in Cognitive Sciences, vol. 21, no. 2, 2017, pp. 66–68., https://doi.org/10.1016/j.tics.2016.10.007.

[4] Robertson, Alexander, et al. “Emoji Skin Tone Modifiers.” *ACM Transactions on Social Computing*, vol. 3, no. 2, 2020, pp. 1–25., https://doi.org/10.1145/3377479.

[4] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.