LDK 2023

Sentiment analysis with emojis: a model for Brazilian Portuguese

Vinícius Moitinho da Silva Santos
Raquel Meister Ko Freitag
Hector Julian Tejada Herrera
Ayla Santana Florêncio
Pedro Paulo Oliveira Barros Souza
Túlio Sousa de Gois
Federal University of Sergipe

Abstract

Sentiment analysis in multimodal texts that include emojis is a complex task because of a lack of tools and cultural specificities. There are some options available (VADER or the emoticons library), but most fail to perform this analysis accurately and efficiently in languages other than English, such as Brazilian Portuguese. This study presents a model based on the sum of polarities to contribute to the improvement of sentiment analysis in different languages. A set of the 100 most used emojis in 2021 by Unicode (Daniel, 2021) was judged by a group (n = 13) into three categories: positive, negative, and neutral. Based on the agreement results, a sentiment analysis model using Python was run, which consisted of summing the polarities of the emojis. Two training databases from Twitter: one in Brazilian Portuguese about the Brazilian elections of 2019 (n = 61,590) and another in English about the 2022 World Cup (n = 22,525). After the filter to exclude tweets without emojis from the agreement test was applied, a dataset with 511 tweets in Brazilian Portuguese and 2,531 tweets in English was used. A sentiment analysis model was run with the datasets to classify the sentiments based on the sum of polarities developed in the previous stage. The results were compared with those of VADER (Hutto, C.J., & Gilbert, 2014), a natural language processing tool that has been validated by linguists and data scientists for performing similar tasks. The results show that the new model for Brazilian Portuguese has slightly lower performance. However, for English, the new model had an accuracy of 51% compared to VADER's accuracy of 41%. This suggests that the new model may be a useful tool for sentiment analysis in English texts containing emojis. Improvements in Brazilian Portuguese are required to broaden the accuracy of sentiment analysis in texts that include emojis. In addition, it is necessary to expand the range of emojis covered by the model and perform classification using machine-learning techniques, which may further improve the accuracy of the model. This study was developed following Open Science standards, with data and code available to the scientific community for enhanced transparency and reproducibility, while also promoting the digital inclusion of nonhegemonic languages such as Brazilian Portuguese.

1 Introduction

The area of natural language processing has considered only the textual linguistic clue, without any support for embodied resources that make up the expressivity of human language (Bühler, 2011), like facial gestures, body gestures and prosodic signals, which make up, together with the linguistic clue, the human interaction. The linguistic system has textual resources of expressiveness, such as punctuation marks (exclamation marks, quotation marks, italics, and bold, for example), but they are not always enough to express the demand for meaning related to the emotional state of users, especially in written interaction situations. It is in this context that multimodal resources emulate the expressive dimension of natural human language, with its embodied resources. This makes that in situations of written interaction, such as in social networks and microblogs like Twitter, the use of visual resources such as emoticons, emojis and memes have been recurrent in the construction of meaning, especially feelings.

One of the most widely used features, not only on Twitter but also in other social networks and instant messaging applications, are emojis (Bai et al., 2019). An emoji is classified as a pictogram or ideogram, an image that conveys a message. When it comes to communicative expressiveness, emojis are often associated with representing the emotional state of their users (Alexandrino, 2016). The search for cues of emotional states in written texts has been the focus of sentiment analysis, which determines the polarity of texts based on values associated with each word. Going beyond the lin-

guistic clue, the use of emojis can also assist in the classification of texts according to their polarity (Cavalcante, 2017), especially in situations of irony where the linguistic clue does not fully express the user's sentiment. Another important aspect to consider is that emotions and embodied resources are sensitive to cultural context (Tejada et al., 2022) and are not universal; therefore, the interpretations and meanings attributed to emojis can vary, requiring specific libraries for each language.

This paper presents the procedures for a sentiment analysis that, considering the expressive nature of emojis in social networks, includes emojis in the polarity classification, specifically, the construction of a lexicon dictionary composed of emojis and their respective polarities, validated on a dataset of brazilian portuguese, a language still underrepresented in terms of technologies for natural language processing.

This work aims to develop and validate a lexical dictionary to identify the polarity of emojis, aiming for the implementation of a sentiment analysis model, the EmojiMapper. Through the presence of emojis in texts, the tool will be able to assign a polarity to the sentence based on the balance of the polarities of the present emojis. Moreover, the model will be validated through tests using previously analyzed datasets and comparing the results with another sentiment analysis model that has support for emojis, using VADER.

The paper is divided as follows: section 2 presents the theoretic foundation on sentiment analysis, its approaches and the VADER analyzer; section 3 deals with the methodology applied for the development of the proposed model, elucidating tools, techniques and used databases; section 4 presents the obtained results; and finally, section 5 deals with the authors' conclusions, based on the results, and presents possible future work. .

2 Sentiment Analysis

Sentiment analysis (SA) is a process that seeks to identify and categorize, using computational methods, the emotions, opinions, and attitudes people express through text (Medhat et al., 2014). Considered a type of text classification, SA is an important part of Natural Language Processing, a field of study resulting from the intersection between linguistics and computation that mainly deals with the linguistic interaction between human and machine (Devika et al., 2016).

SA involves the lexicon-based approach and the machine learning-based approach (Bonta and Janardhan, 2019). For the inclusion of emojis in SA, the lexicon-based approach, with the VADER tool, was the starting point.

The lexicon-based approach uses the classification of each lexicon item for sentiment to describe the polarity of a textual content, which can be positive, negative, or neutral. The classification of items can be dictionary-based or corpus-based (Sadia et al., 2018).

The construction of the lexicon starts by compiling the words of interest and assigning their respective polarities. In the case of lexicon dictionaries, the construction of the list is initially performed manually, with the collection and classification of the objects of interest, creating a dictionary-like structure containing the object (word or symbol) and their polarity (Bonta and Janardhan, 2019). Unlike the corpus-based approach, the dictionary does not consider the context of the selected objects. However, the dictionary-based list allows the selection of specific terms from a field, while the more comprehensive corpus-based approach considers a large volume of data in different contexts, and may lose precision.

Starting from the lexicon-based approach, it is possible to find some well described and tested tools in the literature, as discussed in Bonta et al. (2019). However, considering the model proposed in this paper, the tool that has parameters able to be compared and tested is VADER (Hutto and Gilbert, 2014).

VADER (Valence Aware Dictionary for Sentimental Reasoning) is a model built for sentiment analysis that uses quantitative and qualitative methods, combining a list of lexicon attributes, and syntactic and grammatical rules (Hutto and Gilbert, 2014). Unlike other tools, VADER can also assign polarities to emojis and demonstrates good performance in texts originating from social networks (Bonta and Janardhan, 2019).

3 Methodology

3.1 Lexicon dictionary

The construction of the lexicon dictionary used in the proposed model started with the selection of the 100 most used emojis in 2021 according to Unicode. For the classification of the selected data, a concordance test was performed in which expert judges (n = 13) rated their perception of

polarity for each individual symbol, which could assume three different values presented in Table 1. After the individual categorization, a table was constructed containing the emoji identifications and their respective polarities established based on the majority choice.

After the construction of the lexical dictionary, a coding step was performed in Python language to develop the application. To build the tool, the following functions were implemented. In the tool, functions were implemented to do the cleaning of datasets and individual texts, classify an input set based on the polarity of the emojis present in the text, and validate the result. Validation occurs through a routine that calculates the accuracy of the model against the test data.

3.2 Database

Once the lexical dictionary was completed and integrated into the python script, two datasets consisting of tweets previously classified based on the polarity presented in each text were selected. The data was obtained from the Kaggle platform and used to test and validate the EmojiMapper tool.

The first dataset selected was 'Twitter in Portuguese - Elections 2019[1],' which initially contained texts from 61.590 tweets in Brazilian Portuguese, focusing on the 2019 elections as the central theme. The second selected database was 'FIFA World Cup 2022 Tweets[2],' which contained 22.525 tweets in English, with the central theme being the 2022 World Cup. A different language was chosen to observe how the model performed on datasets with diverse natures.

Both databases were filtered using an internal function of EmojiMapper called 'cleanData.' This function takes the database to be filtered as a parameter and returns a new dataset suitable for the model application. The filtered datasets only contain tweets that have one or more of the 100 emojis mapped in the lexical dictionary step. After filtering, two sub-databases were obtained: one derived from the first dataset, containing 511 tweets (Dataset 1), and another derived from the second dataset, containing 2,531 tweets (Dataset 2)."

[1] https://www.kaggle.com/datasets/adilmar/twitter-pt

[2] https://www.kaggle.com/datasets/tirendazacademy/fifa-world-cup-2022-tweets?resource=download

Classification	Start of range	End of range
Negativo	-1	-0.05
Neutro	-0.05	0.05
Positivo	0.05	1

Table 1: Rating ranges

3.3 Experiment

After filtering the data, it was possible to apply it to the model and evaluate its accuracy on the test sets. To establish a comparison parameter with another similar tool, we utilized VADER (Hutto and Gilbert, 2014). The classification of each database was performed using EmojiMapper's internal function called 'classify.' This function takes the filtered dataset as a parameter and returns a list containing the predicted responses, classified as positive, negative, or neutral. The same procedure was carried out using VADER, and the data classification thresholds for this tool are shown in Table 1. After the classification of the data by both models, a performance metric analysis was conducted.

4 Results

After applying the model to the test sets, metrics were obtained to compare the effectiveness of the tools. Table 2 presents the results obtained from the experiment. It can be observed that EmojiMapper had lower accuracy than VADER for Dataset 1, while the opposite was true for Dataset 2, favoring EmojiMapper. A possible cause for the discrepancy between the metrics may be related to the implementation of each tool. Unlike EmojiMapper, VADER does not directly assign polarity values to emojis. Instead, it employs a methodology to describe the emoji and assigns polarity to the descriptive terms.

Considering this hypothesis, it is possible to explain the phenomenon of VADER's better performance in Dataset 1, as it consists of texts related to politics, and the nature of this dataset is ironic, this can be verified by analyzing the provaIronia.csv dataset present in the tool repository. By examining the emoji description, VADER obtains a well-defined context of the message, while EmojiMapper tends to interpret the symbol literally. However, it is worth noting that in datasets without a predominance of irony, EmojiMapper performs better. This can be attributed to EmojiMapper considering the literal meaning of the emoji and its strong correlation with the text, making it an indicator of the

Dataset	EmojiMapper	VADER
Dataset 1	44%	46%
Dataset 2	51%	41%

Table 2: Model Accuracy

message's polarity.

The model used in this study is available at https://github.com/vmoitinhoss/Emojimapper and can be accessed freely, adhering to the principles of open science.

5 Conclusion

Based on the results obtained in the experiment, it can be concluded that EmojiMapper demonstrates itself as a viable and effective alternative for performing sentiment analysis on datasets without an ironic nature. It may even outperform a validated and prestigious tool. However, it is important to note that this work is still a proof of concept, as its scope of application is limited to texts that contain one or more of the 100 emojis present in the model. Moreover, an improvement is urgently needed to better deal with the irony phenomenon, considering the relationship between joint emojis.

For future research, it would be interesting to explore the feasibility of machine learning methods for weighting the importance of emojis based on the nature of the data. This approach could help overcome the performance issues encountered in datasets with an ironic nature.

References

Ana Débora Oliveira Alexandrino. 2016. Análise de redes sociais aplicada a tweets sobre séries de tv.

Qiao Bai, Qibin Dan, Zhonghai Mu, and Meina Yang. 2019. A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10:2221.

Vandana Bonta and N. K. N. Janardhan. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6.

Karl Bühler. 2011. Theory of language. the representational function of language, translated by donald fraser goodwin in collaboration with achim eschbach. *Amsterdam/Philadelphia, Benjamins*.

Pedro Emanuel Cunha Cavalcante. 2017. Um dataset para análise de sentimentos na língua portuguesa.

M. D. Devika, C. Sunitha, and A. Ganesh. 2016. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, 87:44–49.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Ayesha Sadia, Faiza Khan, and Fahad Bashir. 2018. An overview of lexicon-based approach for sentiment analysis. In *2018 3rd International Electrical Engineering Conference (IEEC 2018)*, pages 1–6.

Julian Tejada, Raquel Meister Ko Freitag, Bruno Felipe Marques Pinheiro, Paloma Batista Cardoso, Victor Rene Andrade Souza, and Lucas Santos Silva. 2022. Building and validation of a set of facial expression images to detect emotions: a transcultural study. *Psychological Research*, 86(6):1996–2006.