Initial Report on improving emoji predictions in tweets via sentiment analysis

By Fani Noncheva

Introduction & Brief Overview

The project focuses on the problem of emoji prediction in tweets. The idea is: given a tweet, the machine learning model will learn to find patterns that would yield different predictions as to what emoji would follow. With that base, I intend to add a sentiment analysis component to the emoji prediction system, that would help the model discern what emoji is most suited for the given text. To keep the sample size manageable, I intend to only focus on the twenty most used emojis. The findings of this project are likely to find use in assisted messaging, text analysis tools, autocomplete bots etc.

Objective

The objective is to compare how does a machine learning model compare to a similar model with the added functionality of sentiment analysis, when it comes to predicting an emoji given a tweet. Essentially, the comparison consists of the results the machine learning model predicts prior to having information on the sentiment analysis, and after being given access to it, and whether this improves the predictions.

Ethics

To accomplish the aforementioned, I'll use a dataset of tweets that are suitable for analysis on emoji prediction, that was previously used in a published paper [1]. I've obtained said dataset through one of the task organizers, which would mean that the ethics of the data used are pre-approved for this kind of processing. Therefore, it is safe to assume that no ethical approval will be required for the usage of this data, as my supervisor seconded when I brought up the question. I completely understand the constraints that human data can have, as I've previously done a project that required ethics approval, and I am thus acquainted with the do-s and don't-s when it comes to processing it and using it.

Milestones

There are four major milestones that I would like to achieve through the course of the upcoming twelve academic weeks. I want to ensure that the stages I am completing deliver an optimal result that I am happy with, which would also allow me to work and progress towards the next step.

The first milestone is to research the methods and tools that I intend to use and build a more concrete plan than the one listed below, with smaller and much more particular tasks. After the initial state of the research is done, I'll formulate three separate options for how to proceed with my project depending on the time I have left, and the different degrees of completion achieved at given stages. That will allow me more flexibility and security to properly assess my options and current situation.

The second milestone is the creation of the base neural network, that would involve looking for patterns in the pre-collected data. Firstly, I need to develop a model that incorporates only the given data and uses it to find correlations between the use of the emojis and the context provided by the text. This'll be the bare minimum I'm striving to achieve, as I'll record the success of this model and compare it to the improved ones that are associated with the later milestones.

The third milestone is the implementation of a sentiment analysis tool within the neural network and using it to determine the associations of specific emojis with a certain sentiment. By implementing a sentiment analysis algorithm to help the neural network identify emojis, I think the accuracy will be drastically improved. Due to the nature of tweets they don't provide much context and don't offer a large volume of data, hence the algorithm used must be very sensitive to provide an adequate measure of the tweet content's sentiment. By creating this dependency, I hope to improve the performance and ensure a better result at the same time, while not compromising with the quality of the method.

The fourth milestone is compiling my findings into the final report. That includes carefully analysing the data, findings and overall experience and making clear distinctions between what is relevant to the final product and what hurdles I've had to overcome in the development and research stages of my project. I'll collect the data for analysis in each separate stage of development as I make improvements to the system, in order to achieve a clear comparison. Overall, I intend to pursue this milestone throughout developing and researching, as I will be able to reflect on my experience more clearly if I document it from the start to finish.

Work Plan

The current plan is based on the above section of milestones. It's currently but a rough estimate, as nothing apart from the research has been started to date. I've arranged weekly meetings with my supervisor to keep track of progress and possible improvements for the remainder of the semester and mostly have a vision of how I want to develop this project.

Weeks one and two that are in yellow are allocated for research. Weeks three to five offer time for me to develop the first neural network prototype. The Easter recess and the period between weeks six and nine is devoted to refining the sentiment tool, which is essential to the project. Weeks ten to twelve are left as buffers, that giving the other tasks a little room for overflow in case a problem arises in the earlier allocations.

The overall idea for the final report is for me to start writing it concurrently with coding, as it will give a lot more pointers and overview of how my work progressed by noting issues that I've encountered early on, that would otherwise be forgotten if I were to start later on.

Week 1	Write up initial report, discuss techniques to use for future development of the project, set up weekly meetings with supervisor.
Week 2	Do further research in detail on how to accomplish desired result in practice. Learn essential frameworks and decide on implementation model to work towards.
Week 3	Start working on the first prototype of the neural network, noting progress in the final report. Discuss any possible changes from initial plan and receive feedback on report.
Week 4	Allocated to ensure that the neural network is properly implemented and scalable for the remainder of the project.
Week 5	Achieve stable version of the first prototype and discuss following steps with supervisor.
Week 6	Start working towards adding the sentiment analysis component of the neural network.
Week 7	Continuation of week stated above, as refined sentiment analysis will require several different algorithms tested to determine which one gives best results.
Week 8	Ensure sentiment analysis is stable and gives reasonable output for the data given.
Week 9	Start gradual work towards improving the accuracy and adequacy of the sentiment analysis once an algorithm has been decided on and stable version of initial plan is satisfied.

Easter	Continue working on refining the sentiment analysis and refining the final report.
Week 10	Finalize report, presentation and work towards completing demo.
Week 11	
Week 12	

Challenges

The initial issue will be the implementation of the machine learning model. I've got a variety of options, however, currently I've set my eyes on a neural network model. I intend to build up on it with other tools to improve its accuracy regarding text processing [1],[5]. I'll base my decision on previous research done on the topic and I'll discuss my choice with my supervisor if in doubt about the optimal choice.

As noted in the timetable above, I believe the sentiment analysis to be key to this project, hence I've allocated most of the time to its refinement. I've read several papers that deal with sentiment analysis tweets, all with varying success and outcomes, but so far I think that the sentiment analysis will be challenging to implement as it would take up most of the time to accurately synchronize with the neural network, that I'll use as foundation.

To ensure success of that specific aspect, I've investigated several different ways to implement the sentiment analysis [1],[2],[4],[6],[7]. I'm going to carefully compare the implementations on performance and implementation difficulty and use that as guidelines on which ones I can afford to implement given the time I have left when I reach this stage. I will first implement a simpler model to ensure the overall goal is complete and try to build a more complex algorithm further down the line.

Bibliography

[1] Barbieri, Francesco, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. "Semeval 2018 task 2: Multilingual emoji prediction." In Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 24-33. 2018.

[2] Dos Santos, Cicero, and Maira Gatti. "Deep convolutional neural networks for sentiment analysis of short texts." Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014.

[3] Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139084789

[4] Chikersal, Prerna, Soujanya Poria, and Erik Cambria. "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning." In Proceedings of

the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 647-651. 2015.

[5] Coletta, Luiz Fernando Sommaggio, Nádia Félix Felipe da Silva, Eduardo Raul Hruschka, and Estevam Rafael Hruschka. "Combining classification and clustering for tweet sentiment analysis." In 2014 Brazilian Conference on Intelligent Systems, pp. 210-215. IEEE, 2014.

[6] Na'aman, Noa, Hannah Provenza, and Orion Montoya. "Varying linguistic purposes of emoji in (twitter) context." In Proceedings of ACL 2017, Student Research Workshop, pp. 136-141. 2017.

[7] Jefferson, Chris, Han Liu, and Mihaela Cocea. "Fuzzy approach for sentiment analysis." In 2017 IEEE international conference on fuzzy systems (FUZZ-IEEE), pp. 1-6. IEEE, 2017.