Next Sentence Prediction for Emoji Sequences

Sammie Kim, Max Needle, Antonio Robayo
New York University
{sk7327, mn1931, amr1059}@nyu.edu

Abstract

NLP research on emojis has mostly focused on emotional information that emojis capture or has attempted to associate large chunks of text with a single emoji. We hypothesize that a single emoji is unable to distinctly capture and represent meaning that is context-specific when used in text-based messages. We experiment with an NSP-style binary classification task, leveraging different embedding techniques, and in support of our hypothesis, we find that our models learn better representations when trained on sequences of emojis rather than single emojis.

1 Introduction

In the last two decades, emojis have become integral to how we communicate electronically and on social media. While commonly associated with sentiment or emotion, emojis encompass a wide variety of ideas that include food, geography, action, occupation, and miscellaneous objects. As such, emojis can be rich with textual expressiveness, often supplementing an already articulated idea or adding new information in the sentences they appear in.

Recent years have seen an increase in NLP research and constructed tasks geared toward gaining a better understanding of emoji-use and user-generated text, typically from Twitter. These studies, while important for advancing our understanding of emoji-usage, limit emojis to the scope of sentiment or attempt to align textual meaning with a single emoji. Barbieri et al. (2020), for example, introduce an emoji-prediction task that relates a body of text to a single emoji. Furthermore, they limit the label set to only 20 emojis. Emojis, however, are not limited to single-use per sentence or body of text. Emojis can be used in sequences of varying length, with each distinct emoji conveying a different meaning.

In an effort to capture emoji meaning, we look to sentence-level tasks. The ability to reason about the relationship between two sentences is a particular area of interest in the NLP community as indicated by the breadth of natural language understanding tasks, such as question answering, NLI, and determining semantic similarity (Wang et al., 2018, 2019). We draw inspiration from these studies and propose a sentence-level task aimed at relating a text sentence to a sequence of emojis, with the goal of training a classifier to identify whether the sequence of emojis follows the associated text. This task can be thought of as NLI-adjacent, though it is more in line with next-sentence prediction (NSP) (Devlin et al., 2019) as the choice of emojis are not limited to an entailment relation.

By leveraging this NSP-style classification task on tweets containing multiple emojis and single emojis separately, we find that multiple emojis allow language models to capture sentence meaning better than single emojis, which supports our hypothesis that single emojis are unable to fully capture sentence meaning. Our implementation is publicly available on GitHub[1].

2 Related Work

There is a broad swath of emoji-related NLP work done to analyze and leverage emojis as text. As for creating vectors to represent emojis, Eisner et al. (2016) approach learning emoji representations by summing the word embeddings from `word2vec` for each word in an emoji's description. Their study finds that augmenting `word2vec` with emoji embeddings substantially improves tweet classification performance for tweets that contain emojis. Singh et al. (2015), on the other hand, replace emojis with their textual descriptions as opposed to

[1] https://github.com/amauriciorr/emojinsp

Textual Sentence	Emoji Sentence	IsNext
Good morning	☕	Y
That what I was thinking too. But I wasn't sure...	🤜🏾🤛🏾😉	Y
Congratulations Can't wait to watch you shine in this next adventure	🤦‍♂️	N
HAHAHA it's Bc I drank vodka	🏎️🏎️🚀	N

Table 1: Example sentence pairs; including published tweets with their original corresponding emojis and others that have been shuffled.

creating emoji embeddings. Their research shows that incorporating emoji descriptions also improves tweet classification.

In an attempt to better understand information provided by emojis, Donato and Paggio (2017) construct a carefully annotated dataset. Specifically, they aim to determine whether emojis are used in a redundant way (e.g., using "France" and 🇫🇷 in the same sentence) to emphasize or add information. Felbo et al. (2017) similarly focus on utilizing emojis for learning representations of emotional contents in tweets. They limit their corpus to tweets with single emojis with the task of predicting the emoji originally associated with the tweet.

Barbieri et al. (2020) explore tweet-specific language models for classification tasks such as sentiment analysis, emoji prediction, and irony detection. In their study, RoBERTa is used as it does not employ NSP as a pretraining task, thus making it more suitable for brief, single sentence inputs such as tweets. Corazza et al. (2020) learn representations by implementing a hybrid emoji-based masked language model (MLM) that targets emojis as candidates for masking when they are present, otherwise the standard MLM task is implemented. Their results yield that the hybrid approach shows advantages over the standard MLM when using social media data.

3 Data Collection

As mentioned, most of the research has focused on aligning text with a single emoji. We hypothesize that a single emoji is unable to fully capture sentence meaning in every case and therefore models will learn better representations when trained on sequences of emojis. We collect 24,000 English tweets that use only a single emoji over a two-day period to validate our assumption. We then collect another set of 24,000 English tweets containing at least two emojis as a contrast. Lastly, another 24,000 tweets containing a mix of both aforementioned emoji use-cases are collected over a different two-day period to further examine our assumptions. We create one additional dataset using a subset from the last data collection by removing repeated instances of emojis, such that "😂😂😂" becomes "😂". Our assumption here is that emojis can be repeated for emphasis but the choice and frequency of repeated emojis introduce noise that can be too random to correctly predict. As a result, we produce four distinct datasets, (1) **Single**: tweets with a single emoji only; (2) **Multi**: tweets with at least two emojis ; (3) **Full (single and multi)**: tweets containing at least one emoji; and (4) **Full no repeat**: "Full (single and multi)" with emoji repetitions removed.

Each dataset is constructed by treating each tweet as a pair of sentences; pairs consist of a textual sentence and what we refer to as an "emoji sentence", which is one or a series of emojis that have been extracted from the textual sentence. We develop a NSP-like task using these pairs with the goal of training a classifier to determine whether the "emoji sentence" in the pair follows the textual one. To establish our labels we split our data in half, leaving one half as is (correct pairs) and shuffling the other half (creating a mismatch between textual sentence and emojis). We add an additional constraint that the shuffled sentence pairs cannot include emojis from its original emoji sentence. We then split this into training, validation and testing sets under the ratio 70:10:20, maintaining an equal number of correct and incorrect pairs within each set. We show example sentence pairs in Table 1.

In order to reduce noise, we only include tweets with at least three tokens. Emulating (Felbo et al., 2017), we remove tweets that include URLs, anticipating that these links provide further content explaining associated emojis. Lastly, user handles are replaced with the token `[USER]` in order to treat each username uniformly.

Model	Full (single and multi)	No repeats	Single	Multi
Baseline results				
emoji2vec + word2vec + LR (avg)	0.517	0.516	0.488	0.522
emoji2vec + word2vec + LR (concat)	0.517	0.516	0.500	0.522
RoBERTa-base	0.516	0.498	0.492	0.502
Finetune results				
emoji2vec + word2vec + LR (avg)	0.524	0.519	0.488	0.519
emoji2vec + word2vec + LR (concat)	0.520	0.515	0.497	0.522
RoBERTa-base	0.701	0.701	0.663	**0.818**

Table 2: Baseline and finetuned accuracy on different emoji datasets

4 Models

We first create a model with pre-trained embeddings from emoji2vec (Eisner et al., 2016) and word2vec, combined with a simple Logistic Regression ("LR") for the binary classification task. However, since this model lacks an understanding of the context in which emojis appear, we doubt that it can be the correct tool for our task. Inspired by Barbieri et al. (2020) and with an aim to incorporate context for emojis, we train and fine-tune a RoBERTa model for each dataset separately. As a result, we create 8 different models at the end of our experiments (4 LR models and 4 RoBERTa models).

4.1 emoji2vec + word2vec + LR

Tweets consist of short sentences, sometimes just a few words, due to Twitter's character limit. In our datasets, 95 percent of the text sentences are no longer than 37-43 words. As such, we start with a simple model with static embeddings, without considering long-term dependencies.

First, we conduct additional preprocessing: (1) lowercase all words as we do not have a large enough dataset; (2) remove punctuation as it is not typically associated with an emoji; and (3) remove white space. We then tokenize the preprocessed tweets using NLTK[2]. Next, we obtain emoji vectors from emoji2vec for the "emoji sentence". emoji2vec embeddings are created by summing word2vec embeddings for the emoji's description in the Unicode emoji standard[3]. Of the 3,521 emojis in the Unicode Standard emoji[4], over a third appeared in our study, and a list of our most common emojis is shown in Figure 2 in the Appendix. As for the "textual sentence", we use the standard 300-dimensional word2vec embeddings trained on Google News that have a vocabulary of 3M words. We then average the emoji embeddings and word embeddings to represent a tweet. In addition, as explored in Meng et al. (2020), we also concatenate the vectors so as to keep averaged textual and averaged emoji representations explicitly, not aggregated. Lastly, we pass these embeddings through a simple LR model to establish a baseline accuracy. We then use GridSearchCV to find the best regularization parameters and obtain the finetuned results.

4.2 RoBERTa

We use a RoBERTa-base model (Liu et al., 2019) that has been pre-trained on roughly 58M tweets as part of TweetEval (Barbieri et al., 2020) so as to leverage embedding representations that already have exposure to emojis instead of having to learn embeddings from scratch. This model is a perfect fit for our use-case as we also use Twitter data, thus making the data distribution on which the RoBERTa base was pre-trained the same as that which it is fine-tuned on. We extend the embeddings to include [USER] as a learnable token, but otherwise leave the lookup table as is before fine-tuning. A linear layer is added to be able to use RoBERTa for our NSP-style binary classification task. After observing the quantiles for textual and emoji sentence length, we decide to use a maximum token length of 225–this ensures that we are capturing the majority of tweet content in its entirety while also accounting for byte pair encoding. Our model is fine-tuned up to 15 epochs with early-stopping using a learning rate of 5e-6 (which proved to be the most successful across our experiments). We experimented using 1e-4, 5e-4, 1e-5,

[2] https://www.nltk.org/

[3] http://www.unicode.org/emoji/charts/full-emoji-list.html

[4] https://blog.emojipedia.org/emoji-use-in-the-new-normal/

5e-5, 1e-6 as learning rates. We also incorporate a weight decay of 1e-6 to handle any overfitting.

5 Evaluation

5.1 Results

Table 2 shows baseline and fine-tuned performance for our `emoji2vec + word2vec` + LR and RoBERTa-base models. In spite of having some learned representation for emojis, both models do no better than random guessing prior to training. After training, we observe a stark difference in performance; the LR model does not show any improvement, whereas RoBERTa sees an increase in accuracy of at least 0.17 across all datasets.

5.2 Analysis

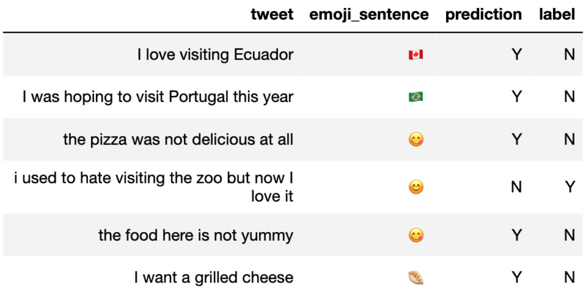
As we hypothesized, emojis fail to correctly represent meaning in varying contexts when using embeddings from `emoji2vec`. By contrast, RoBERTa models are able to account for the sequential relationship between words and emojis in a sentence. Generally, the RoBERTa-base model gains an understanding of sentiment associated with emojis as well as the meanings of emojis displaying objects and activities.

While our results from the RoBERTa model look promising and indeed indicate some degree of efficacy with our constructed task, there are limitations in the model's ability to fully capture sentence meaning as seen in Figure 1. For example, when encountering flags (e.g., 🇪🇨), the model seems to understand that these can be used to reference locations but doesn't always correctly associate the flag with its origin country.

Similarly, when evaluated against negation (i.e., "the food was good" vs "the food was *not* good") the model can be easily fooled. It is clear that in some cases the model has learned to associate specific words with emojis, regardless of the context, such as "😋" with "yummy" or "delicious".

5.3 Further Experimentation

Based on the observed success of the RoBERTa model trained on the multi-emoji dataset, we decide to evaluate it further using more stringent criteria. To accomplish this we employ five workers[5] to construct positive and negative examples for our emoji-NSP task. We expect this approach to provide a more accurate measure of performance, as the negative examples now correctly capture irrelevant relationships between a text sentence and an emoji sentence as we intended. We end up with a total of 238 handcrafted examples (130 positive, 108 negative). On this dataset, the model achieves an accuracy of 0.706.



tweet	emoji_sentence	prediction	label
I love visiting Ecuador	🇨🇦	Y	N
I was hoping to visit Portugal this year	🇧🇷	Y	N
the pizza was not delicious at all	😋	Y	N
i used to hate visiting the zoo but now I love it	😊	N	Y
the food here is not yummy	😋	Y	N
I want a grilled cheese	🥟	Y	N

Figure 1: Examples illustrating limitations of the RoBERTa model

6 Conclusions

We introduced an emoji-focused NSP-style task constructed by extracting the emoji portion of tweets and training a model to determine whether the "emoji sentence" follows the "textural sentence". With a modest sized dataset of 72,000 total tweets, we provide support for our hypothesis that a single emoji is unable to fully capture sentence meaning in every case by showing that tweets with multiple emojis perform better on this task than tweets with a single emoji.

As part of future work, we hope to observe how training is affected with a much larger dataset, one that consists of our automatically shuffled sentence pairs as well as handcrafted examples. Our original instructions for constructing text and emoji sentence pairs leave much to be desired in terms of the different meanings they can target—for example, capturing an understanding of emojis depicting activities, additionally flags and foods have geographic associations that have yet to be fully learned, and lastly ensuring that emojis of different skin tones and genders are appropriately represented is another concern (as well as probing for social biases from their usage).

For this initial study, we focused solely on evaluating our model against the NSP-style task; however in the future, we seek to evaluate the embeddings trained on this task against the TweetEval benchmark tasks to get a supplementary measure for the validity and efficacy of this task.

[5] workers are college-educated friends of one of the authors with no background in NLP or machine learning

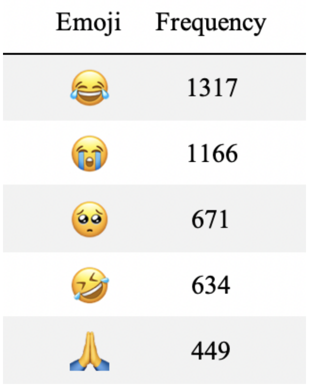
7 Appendix

Ethical Considerations

Research on emojis has become an increasingly important topic in the academic field. The present research provides insight into how emojis contribute to sentence meaning, especially when a sequence of emojis is used. These techniques could be applied to a range of applications, including natural language processing and recommender systems to be able to leverage meaning associated with emojis. However, with the expansion of emoji options to include multiple skin tones and genders, there is concern over whether learned associations between emojis and text may allow bad actors to identify and surveil people of marginalized identities based on their use of the textual language associated with these emojis depicting certain races and genders or simply based on their use of these emojis, as skin tone emojis are more widely used by people of those skin tones (Coats, 2018). By learning emoji embeddings that ignore skin tones and genders prior to experimentation, resulting models would not capture associations between text and emojis relating to these identities. By evaluating language models on both the complete emoji embeddings and on the emoji embeddings stripped of race and gender information, future work could assess the race and gender bias in datasets containing text and emojis before learning dangerous associations between them.

Contribution Statements

- Sammie Kim (sk7327) - data preprocessing, implemented Logistic Regression model, conducted preliminary error analysis
- Max Needle (mn1931) - deepmoji/torchmoji experimentation, not included in paper because model treated emojis as classes to predict rather than representations to learn
- Antonio Robayo (amr1059) - emoji and Twitter scraping, data preprocessing, implemented RoBERTa emoji-NSP model



Emoji	Frequency
😂	1317
😭	1166
🥺	671
🤣	634
🙏	449

Figure 2: Most common emojis seen across our four datasets

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweet-Eval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, pages 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148.

Steven Coats. 2018. Skin tone emoji and sentiment on twitter. http://arxiv.org/abs/1805.00444.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. Hybrid emoji-based masked language models for zero-shot abusive language detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, pages 943–949. https://doi.org/10.18653/v1/2020.findings-emnlp.84.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Giulia Donato and Patrizia Paggio. 2017. Investigating redundancy in emoji use: Study on a Twitter based corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Copenhagen, Denmark, pages 118–126. https://doi.org/10.18653/v1/W17-5216.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Austin, TX, USA, pages 48–54. https://doi.org/10.18653/v1/W16-6208.

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*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1615–1625. https://doi.org/10.18653/v1/D17-1169.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. https://arxiv.org/abs/1907.11692.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, and Jiawei Han. 2020. Unsupervised word embedding learning by incorporating local and global contexts. In *Unsupervised Word Embedding Learning by Incorporating Local and Global Contexts*. Frontiers in Big Data, Urbana, Illinois. https://doi.org/10.3389/fdata.2020.00009.

Abhishek Singh, Eduardo Blanco, and Wei Jin. 2015. Incorporating emoji descriptions improves tweet classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, page 2096–2101. https://doi.org/10.18653/v1/N19-1214.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR* abs/1905.00537. http://arxiv.org/abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, pages 353–355. https://doi.org/10.18653/v1/W18-5446.