Reproducing

“Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm”

Members:

Ajit Kumar Chaudhary

Avinash Pandey

Moiz Ahmed Khan

Saroj Bogati

Table of Contents

[Abstract 3](#_Toc55482190)

[Brief description of the source paper and justification 3](#_Toc55482191)

[Evaluation Framework 3](#_Toc55482192)

[Justification 4](#_Toc55482193)

[Description of original dataset 4](#_Toc55482194)

[Replication of original work 4](#_Toc55482195)

[Issues of original work and resolving those 5](#_Toc55482196)

[Construction of new data 5](#_Toc55482197)

[Twitter API 6](#_Toc55482198)

[Results on new data 6](#_Toc55482199)

[Any other reflections 6](#_Toc55482200)

[References 7](#_Toc55482201)

# Abstract

Reproducibility, a similar act as replicability, performs as a repetitive method of proving correctness, or reaching to similar suppositions as the expected results. When a published result is replicated with the usage of data, codes as well as papers delivered by the authors leading to expected results or similar, a research is said to be reproducible. Not just as a justification, but it also facilitates us as a revelation of research roadmaps as well as the transmission of knowledge. Here in this report, we will explain about our reproducible research based on a paper that used millions of emoji occurrences to study any domain illustrations in order to detect sentiment, emotion and sarcasm underlying.

# Brief description of the source paper and justification

It is published at EMNLP and globally ranked 1st in Sarcasm detection. It uses 1246 million tweets containing about 64 common emojis on 8 benchmark datasets to predict sentiment, emotion, and sarcasm.

<<<<<<<<<<<<<<<<<<<<<<<<<<<PICTURE>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

<<<<<<AROUND 150 WORDS>>>>>>

## Evaluation Framework

<<<<<<<<<<<<AROUND 200 WORDS>>>>>>>>

## Justification

<<<<<<<<<<<AROUND 150 WORDS>>>>>>>>>

# Description of original dataset

<<<<<<<<<<<<<<<<<<<<<<<<<<<PICTURE>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

In the paper, there are 8 benchmark datasets.

Olympic: Negative and high control, positive and high control,

negative and low control, positive and high control.

PsychExp: joy, fear, anger, sadness, disgust, shame, guilt

SCv1: not sarcastic, sarcastic

SCv2-GEN: not sarcastic, sarcastic

SE0714: fear, joy, sadness

SS-Twitter: negative, positive

SS-Youtube: negative, positive

kaggle-insults: neutral, insult

# Replication of original work

The concept of empirical generalization greatly implies on any research being able to replicate. For our approach of replication, duplication or even extending the original work, there are sufficient tools available that facilitate us on the process. However, prior to the initiation, a good plan is needed that ease us with rectifying and evaluating necessary changes on the original work. One of our major goals in this project is to properly replicate TorchMoji model and ensure it runs properly as done in the original work. This model will help us to understand how language has a usage of expressing emotions. Likewise, with the usage of transfer learning, the model will help us to obtain state of art performance on many text modeling tasks related to emotions. With reference to original work, we were able to successfully install and run source code using torchmoji model. Following to that, TorchMoji model is trained and tested among various tweet texts that we fetched as a new dataset.

<<<<<<<<<<<<<<<<<<<<<<<<<<<PICTURE>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

As per paper, there are following accuracy on the benchmark datasets.

We were able to replicate following accuracy on the benchmark datasets

The original work consisted of emoji as well as score prediction. Emoji prediction consist of 5 most relevant emoji based on the sentence which you can see here. Similarly, Score prediction consists of accuracy score of the emoji in relevant to the emojis. Likewise, going through evaluation part, we had 8 benchmark datasets, and we could only replicate 3 datasets. Because it took approximately 12 hours for us to replicate those. Likewise, we made usage of 1000 epochs along with each epoch having sample size of 1000. The metrics used here is accuracy score along with labels. Similarly, we used fine tune model of torchmoji to facilitate us with the execution part which all comes under last method. So, talking about accuracy on these benchmark datasets that we executed, we got accuracy score of 31.95% in YouTube dataset, 41.20% in twitter dataset and 49.92% in Sarcasm dataset. Even though we tried evaluating for these 3 datasets, our main aim was to check and implement this model for twitter dataset, since we managed to create a new dataset from tweets fetched from twitter which will be explained further later part of the presentation.

## Issues of original work and resolving those

The original work is implemented in Python programing language along with PyTorch, NumPy, and Scikit-learn as dependencies. To execute the python files, there was need of running command line arguments along with the adjustment of dataset size, location and other attributes. Overall work was performed on Google Colaboratory with the creation of Jupyter Notebook, since the datasets are relatively large and also won’t descend performance of the local machine while program execution. Another great advantage of Google Colaboratory is that, there is no need of installing any modules for running the code since modules are already installed within it. Using Google Colaboratory made it convenient for us to share the notebook file within team members. The below mentioned issues are addressed while execution:

Note: - The model took too long to train (approximately 12 hours to train only 40% of the dataset).

There were lots of bugs on the original code, which require manually fixes.

We referred common issues and the comments on the GitHub code repository to fix the issue. We were only able to train 3 datasets (40% of the total datasets) as the model took more than 10 hours to train these datasets. The accuracy score is not matching with the paper as we were not able to run full dataset. The original data file was in .pickle, which consist of text, label, validation, testing data. We made some changes in the script to load text and label for performing evaluation.

# Construction of new data

Precisely, there is need of a reproducibility so as to aid the crisis of reproducibility. Replicability, can also facilitate to fulfill this gap, where there is a requirement of replication of original research paper along with the code, followed by the achievement of similar results, however on a changed set up. Likewise, the data can either be downloaded from any public source or be scrapped using certain tools and techniques. Here in this project work, we fetched our dataset through specific techniques and procedures of raw data extraction. Extracting raw tweets data from Twitter’s database has been our major source of forming a dataset.

## Twitter API

With known fact of Twitter’s massive database, it has always been one of the major hubs for data extraction. The process initiates with the usage of Twitter API and tweepy library, thanks to Python libraries that made us easier to access and analyze data from the twitter. To be more specific, we have extracted 3200 tweets of Bill Gates from his twitter account using these libraries. One of the prominent reasons to extract this dataset was due to its characteristics of consisting various tweet texts written by the same user. Likewise, the dataset will also ease us with variance and affirmation, since we can evaluate the predictions of 5 most relevant emojis based on each sentence extracted from various tweets.

<<<<<<<<<<<<<<<<<<<<<<<<<<<PICTURE>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Following to that, there was a need of data cleaning process in order to get clean texts from raw tweets, excluding hashtags, mentions and other elements from the dataset. Similarly, we also divided tweet texts into each separate sentence with the purpose of getting more specific results on relevant emojis. Correspondingly, after cleaning the texts, we generated labels with positive and negative sentiments using TextBlob library. Besides, we also converted paragraphs of a tweet to several sentences, resulting to 5515 rows of dataset.

<<<<<<<<<<<<<<DESCRIPTION ON SUBJECTIVITY, POLARITY AND LABEL>>>>>>>>

# Results on new data

<<<<<<<<<<<<<<<<<<<<<<<<<<<PICTURE>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Evaluation of new data:

We have performed evaluation using finetune of torchmoji model.

We have received 92.20% accuracy on the new data.

# Any other reflections

# References