Emoji Prediction based on Twitter Data

Andrew Joseph Magri   
B. Sc. I.T. in Artificial Intelligence  
University of Malta   
andrew.magri.18@um.edu.mt

Daniel Attard   
B. Sc. I.T. in Artificial Intelligence  
University of Malta   
daniel.attard.18@um.edu.mt

# Abstract

# Introduction

Social Media is an ever-growing platform, which offers a wide range of utility such as posting digital media, sharing personal content and in particular communicating with other people on a day-to-day basis. This is a useful tool which helps people to integrate with each other, be it in an informal way or a more formal way. Prior to 1999, textual communication on social media was totally reliant on text-based communication. The notion of an emoticon did exist using various symbols such as ‘:-)’ and ‘:-(’, but unfortunately, it is hard to express certain emotions using only these symbols. Luckily, this type of communication was revolutionised in 1999 by Shigetaka Kurita, a Japanese artist who designed the first emoji set for the Japanese mobile phone operator company NTT DoCoMo [1]. After this, emojis were adopted by all communication companies, from online platforms like Facebook and Instagram, to mobile application such as SMS on smart phones, as a means of expressing emotions within the text which one writes.

It was found that an estimated 19.6% of all tweets on Twitter contain the usage of an emoji. Furthermore, 37.6% of all Twitter users make use of emojis in their tweets [2]. It can be said without a doubt, that the usage of emojis in social media is becoming more and more popular, especially with the younger and growing generations. This increasing usage of emojis, produces the need for new studies about emojis in the field of Natural Language Processing to be performed. Despite this, research in this field is relatively limited and only covers a very small portion of the NLP area.

In this paper, we attempt to implement classification techniques, which given a Tweet, determine which is the most appropriate emoji to used based on the content of the Tweet. The classification task being solved in this paper is a multi-class classification task, where a single Tweet can only be assigned a single label, unlike a multi-label classification task where each Tweet can be assigned an array of labels [3]. In order to see which classification methodology is best, we implement multiple classifiers, all using different techniques, to predict the best emoji for a specific Tweet. We then compare which of the classifiers performed best when compared to each other in order to determine which is the best framework for the task at hand. As a baseline for the results, we consider the results of the implemented Naïve Bayes classifier and the Random Forest classifier.

In the following sections, we will go through the procedure of implementing such a system. In Section 2, we will have a look at previous work done in the area of natural language processing paired up with the usage of emojis. In Sections 3 and 4, we describe the dataset which will be used, the pre-processing which is performed on the data, and the different models which were implemented in this paper. Finally, in Section 5 we will discuss and evaluate the results achieved by each individual model implemented.

# Literature Review

## Naïve bayes

## Random Forest

## N-Grams

## Bag of Words

## Support Vector Machines

## Long short-term memory

## Gated recurrent units

# Dataset

# Methodology

## Pre-processing

Removal/no removal of location

## Baselines

## Support Vector Machine

## Neural Networks

# Experiments and Evaluation

1. Macro and micro

# Conclusion

# References

1. Oxford's 2015 Word of the Year Is This Emoji
2. A Global Analysis of Emoji Usage
3. An Active Learning Based Emoji Prediction Method in Turkish