University of Peradeniya

Towards Cyberbullying Identification on Social Media Using Text And Emojis

D. C. D. Menike[1]
Supervised by: H. R. O. E. Dayaratna[1]
[1]*Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya.*

Abstract- Cyberbullying is an extremely disturbing social problem. Current studies mostly focus on identifying text-based harassment only. This study identifies cyberbullying using text and emoji-based embedded technique and three classification models: an RNN, a LSTM and a SVM.

Introduction

Today, a large number of people from different socioeconomic groups are inclined to use social media for different purposes. In the current context, due to the frequency and the severity, Cyberbullying in Social Media has become a major concern despite most demographic factors.

Cyberbullying in social media is not limited to text. Emojis, punctuation marks, sarcasm and various other methods are viable in cyberbullying. Emojis can change the entire meaning of a sentence. Fig 1 shows how a single emoji changes the whole meaning of a sentence and Fig. 2 shows how emojis can also be a language itself.

Objectives

Detect abuse, hateful, aggressive and harmful comments, which are inclined to cyberbullying on social media platforms based on emojis and text.

Data Acquisition and Cleansing

Approximately 80 000 replies are scraped from Twitter.

The following steps are followed in cleaning process.

- Remove non-English words, URLs, hashtags (#) and At symbols(@).
- Remove all punctuations.
- Convert all words into simple letters.
- Corrected Misspelled words using the TextBlob Python library.
- Tokenized done sentence-wise.
- Null values and duplicates are removed.

Methodology

The methodology is shown in Fig.3

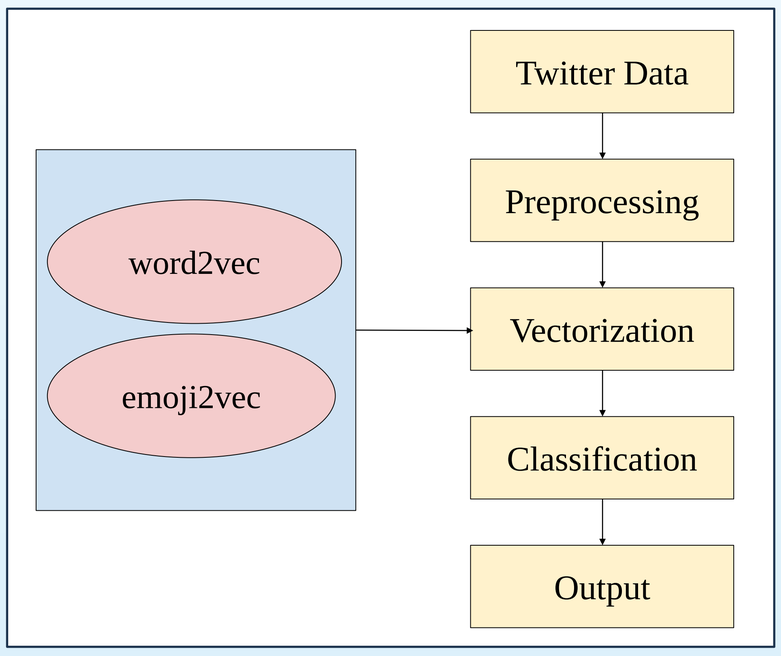

Fig. 3: Overview of the methodology

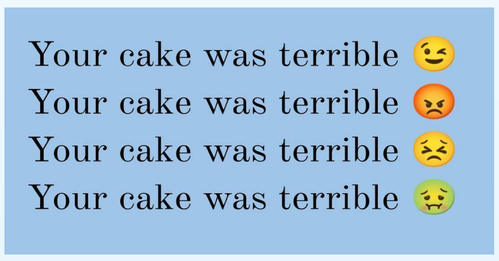

Fig.1: emojis affecting the meaning of a sentence.

Fig. 2: emojis as a language itself

Result and Discussion

After training each classification model with 350 epochs with an early stop function, the f1 scores and accuracies are as in Table 1. By considering the F1 score LSTM shows the best result which is an 80.01% .

Fig.4 – a and b are the lost curves for RNN and LSTM models respectively. The confusion matrixes for RNN, LSTM and SVM models are shown in Fig.5 – a, b and c respectively.

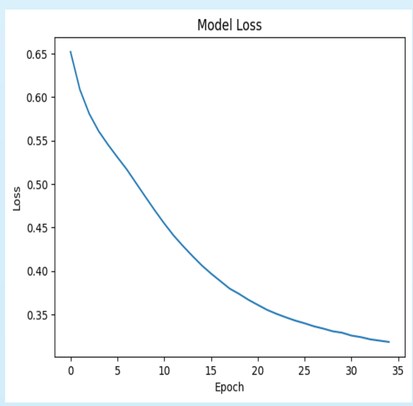

Fig. 4-a : The lost curve for the RNN model

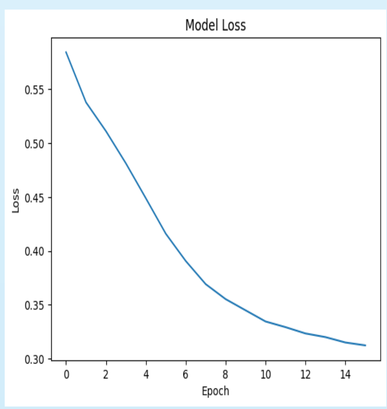

Fig. 4-b : The lost curve for the LSTM model

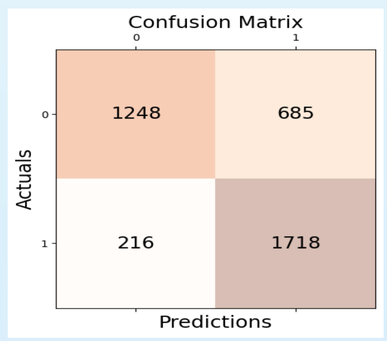

Fig. 5-a : the Confusion Matrix for the RNN model

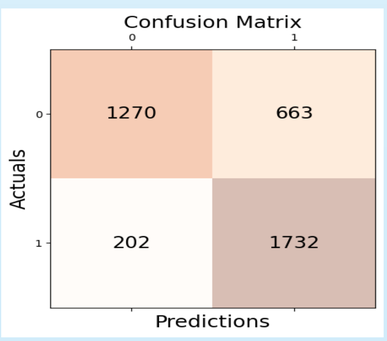

Fig. 5-b : the Confusion Matrix for the LSTM model

Table 1: The F1 score and the accuracy corresponding to each model

Model	F1 Score (%)	Accuracy (%)
RNN	79.22%	76.70%
LSTM	80.01%	77.67%
SVM	79.35%	79.75%

Conclusion

The study's forward steps include data collection, preprocessing, vectorization of sentences using Word2vector and emoji2vector and training three classification models. The best model was LSTM and the f1 score for the model was 80.01%. The accuracy and F1 score are expected to be enhanced in the future by improving the emoji database and looking out for improved vectorization techniques.

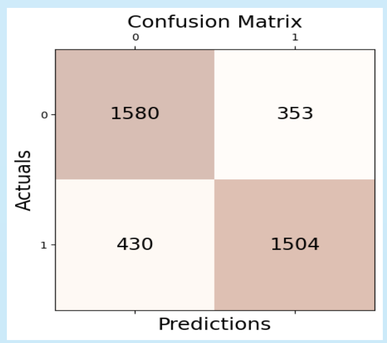

Fig. 5-c : the Confusion Matrix for the SVM model

Contact details
Name : Dr. H. R. O. E. Dayaratna
Tel. No.: 0766986500
Email : erunika.dayaratna@sci.pdn.ac.lk

Multidisciplinary AI Research Centre (MARC)
University Research Council
University of Peradeniya
Peradeniya, 20400, Sri Lanka