INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Emotional Sentiment recognition through emoji-based sentiment analysis for Sinhala Language using supervised learning techniques.

A Thesis by

Mr. Sudesh Dilshan Peiris

Supervised by

Mr. Iresh Bandara

May 2023

ABSTRACT

According to the results of several studies conducted focusing on, exploring the relationship between cultural values and emotional expensiveness, Sri Lankans tend to value emotional expensiveness, which is reflected in their communication styles and social interactions. Since social media platforms have become a central hub for people to express their emotions, opinions, and reviews on variety of topics related to culture, entertainment, politics, healthcare and many more. The increasing usage of emojis in addition to textual expressions is proving that interactions through digital media has shifted towards more visual an image-based communication, which makes it easier and quicker in conveying emotions and opinions through emojis. This paper proposes a novel approach to emoji based emotional and sentiment analysis to recognize the actual emotion of text (review) containing emojis for Sinhala language. The main research aim of this study is an attempt to recognize the patterns of expressions with respect to the polarity scores of the textual data and the emotional expression of the emojis used in social media reviews(comments) and to build an LSTM based RNN to infer expression patterns through considering the relationship between the textual data and the emojis. Since a supervised learning technique is followed data is collected through publicly available comment sections of YouTube content via a developed python function wrapping the YouTube data API, are filtered for Sinhala texts containing emojis and appended to csv in the intension of creating a satisfiable dataset, which is preprocessed, manually annotated, feature engineered and then fed into the created deep neural network. 48 unique expression patterns were identified as an outcome of engineering for new features out of the annotated data. Due to the limited amount of sample data the expression patterns had to be reduced and generalized for a much lesser number of categories. A major trend discovered while annotation of emotion over extracted unique emojis in a text was that no face emoji rather than a simple happy face was expressing a positive emotion, for example almost all the instances where a grinning or face with tears of joy were expressing an emotion mixed with sarcasm. Based on this analysis of emojis used in social media reviews, this research has been able to identify a massive number of expression patterns that contain potential link among specific collections of emojis and the capability of conveying sarcasm. While findings of this research declare the initial insights to the usage of emojis as a tool for identifying the combinations of sentiment and emotion of social media reviews. Collection of more sarcasm specific data and further study of sarcastic expression patterns respective to its context, collection of emojis

incorporated and generalization of those considering the similarity would significantly enhance the detection of sarcasm while increasing the potential of developing more accurate and reliability insured deep learning models for automated sentiment emotion analysis.

Keywords: Recurrent Neural Network, Long-Short-Term-Memory, Emoji Emotion Analysis, Sentiment Analysis