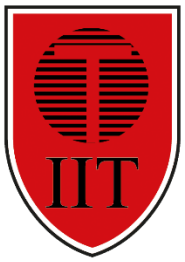

INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Tweet Reach Prediction Through Combination of Natural Language Processing and Tweet Statistics

A dissertation by

Mr. Chehan Lakvindu Sivaruban

Supervised by

Mrs. Sapna Kumarapathirage

Submitted in partial fulfilment of the requirements for the BEng in Software Engineering degree at the University of Westminster.

May 2023

Abstract

An engagement and reach prediction system for text-based Tweets is the main goal of this project. The need for a social media engagement and reach forecast system using the post content and user account features is the issue this project attempts to solve. However, it can be difficult to predict engagement and reach with accuracy, so a system that can offer trustworthy and precise predictions is required to help users improve their social media performance. The project makes predictions regarding tweet engagement using machine learning methods. The model was trained using a dataset that included details on user accounts on social media and their activities. The dataset contains information on users' followers, following, total likes, user verification status, and tweet content. As for the prototype, sentiment models, Neural Network models, LDA topic modeling, KeyBert, and Decision Tree Regression models were used to approach predicting numerical values.

The system's performance can be enhanced by using advanced NLP techniques and deep learning be enhanced by the inclusion of more features and a larger variety of data. In terms of estimating the audience and reach of social media activities, the project's early implementation has generally produced good results. However, there is still potential for progress, and the accuracy and usability of the prediction system need to be improved by the inclusion of more in-depth systems and methodologies in the upcoming advancements.

Keywords: Social media, Reach prediction, Audience estimation, Web application, Tweet text content, Account features, Sentiment analysis, Keyword extraction, Topic detection

Subject descriptors:

- Information Systems: social media, web-based applications
- Human-centered computing: User interfaces, User experience
- Software and its engineering: Software development process, software architecture, software algorithms
- Computing methodologies: Machine learning, Predictive modeling, data analysis
- Computer applications: Digital marketing, audience reach prediction.