

Senior Design Project Report

CSE/EEE/ETE 499

Tweet Popularity Prediction Using Deep Learning

Submitted By

Leyon Ibn Kamal – 1811050042

Nujhat Tabassum – 1731175042

Raihan Hasan Joy – 1712340642

Mahafuza Anzum Lia – 1711261642

Supervisor

Dr. Shahnewaz Siddique – SNS1

Associate Professor

ELECTRICAL AND COMPUTER ENGINEERING

NORTH SOUTH UNIVERSITY

Fall 2022

Agreement Form

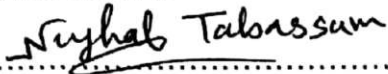
We take great pleasure in submitting our senior design project report on *Tweet Popularity Prediction Using Deep Learning*. This report is prepared as a requirement of the Capstone Design Project CSE/EEE/ETE 499 A & B which is a two semester long senior design course. This course involves multidisciplinary teams of students who build and test custom designed systems, components, or engineering processes. We would like to request you to accept this report as a partial fulfillment of Bachelor of Science degree under Electrical and Computer Engineering Department of North South University.

Declared By:



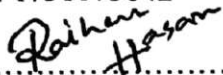
.....
Name: Leyon Ibn Kamal

ID: 1811050042



.....
Name: Nujhat Tabassum

ID: 1731175042



.....
Name: Raihan Hasan Joy

ID: 1712340642



.....
Name: Mahafuza Anzum Lia

ID: 1711261642

Approved By:

.....

Supervisor

Dr. Shahnewaz Siddique

Associate Professor, Department of Electrical and Computer Engineering

North South University, Dhaka, Bangladesh

.....

Dr. Rajesh Palit

Chair, Department of Electrical and Computer Engineering

North South University, Dhaka, Bangladesh

Tweet Popularity Prediction Using Deep Learning

Shahnewaz Siddique
ECE Department
North South University
Dhaka, Bangladesh

Email: shahnewaz.siddique@northsouth.edu

Leyon Ibn Kamal
ECE Department
North South University
Dhaka, Bangladesh

Email: leyon.kamal@northsouth.edu

Nujhat Tabassum
ECE Department
North South University
Dhaka, Bangladesh

Email: nujhat.tabassum@northsouth.edu

Raihan Hasan Joy
ECE Department
North South University
Dhaka, Bangladesh

Email: raihan.joy@northsouth.edu

Mahafuza Anzum Lia
ECE Department
North South University
Dhaka, Bangladesh

Email: mahafuza.anzum@northsouth.edu

Abstract—Twitter is one of the most popular social media platforms. It is used for a wide range of uses, from just sharing information to promoting brands, and even by many governments and politicians to promote their policies and opinions. Regardless of the use, the most common expectation of the users is for their tweets to get higher likes and retweets, which eventually leads to a greater reach in the audience. From general observation, we can notice that on Twitter some tweets seem to get better reach than others. So our goal is to create a deep learning model that would allow us to predict if a certain text a user wants to tweet out will get popular so that they can optimize their tweets for maximum benefit. We will use text classification to create a rating system to allow us to get a general idea of how popular a tweet could get.

Index Terms—Text Classification, NLP, Deep Learning, CNN, RNN, LSTM, GRU, Multi-class classification TensorFlow, Twitter

I. INTRODUCTION

Social media can play an important role in revenue generation for companies. There is a connection between the amount of marketing done on social media and the increase in revenue for the companies [1]. Twitter is often used for marketing purposes by companies and individuals alike to build a brand and promote new content, and the revenue generated by Twitter was about 5 billion dollars in 2021 and is only expected to grow even more [2]. This shows that the platform has a diverse audience and how important Twitter is for marketing, so it is important to have the tools to best use the platform. From general usage of the platform, we can see that there are some tweets that are promoted as advertisements, and there are lots of tweets that are not promoted and naturally have a good amount of activity through likes, comment replies, etc., while others don't perform very well. Marketing is expensive, so, if it was possible to predict in advance if one's tweet will grab enough activity or not, then it would help build a brand faster and cheaper. So, we shall try to solve this problem using deep learning. To build a model to predict if a tweet will get popular, we will first have to build a large dataset that should contain tweets from varied

topics with enough data for the model to try to find pattern in it. Then to create a deep learning model we will test different deep learning models and select the best one which will be the one that gets the best accuracy is less time, and then compare our model to those created

II. BACKGROUND

In the paper "Improved Twitter virality prediction using text and RNN-LSTM" [3], they worked on Mexican news. For the dataset, they collected about 133,000 tweets from 46 Mexican news accounts, ranked either most influential or least influential for binary classification. From those, they only used 5000 tweets, where they selected 2500 tweets labeled as most influential and 2500 tweets as least influential, to create a balanced dataset. To compare the influence of the tweets, they compared the ratio of retweets and favorites, with their logic being that if the retweets are more than likes, most likely the tweet is being criticized, and more likes mean it has a positive impact, and an equal number of retweets and favorites meaning, it is neutral. They also looked at the ratio of followers and following and used these four values to create a metric for measuring influence which they used as a label with the tweet as a feature. Then they use RNN-LSTM for classification which gave about 84% accuracy.

In the paper "Prediction of likes and retweets using text information retrieval" [4], they worked on data science-related tweets which they collected 2 million of, using tweepy python library. They used the text, likes, retweets, and time created as features. For pre-processing the dataset, they removed links, @RT annotations, emoticons, and stop words to reduce unnecessary information in the tweets. They also removed duplicate tweets and tweets with zero likes or retweets from the dataset. Then they categorized the likes and retweets into categories ranked from 1-4 for both likes and retweets. After that, they created different models using Logistic Regression, Random Forest, SVM, Neural Network, and Multinomial Naive Bayes. They used both TF-IDF and Doc2Vec for feature processing. After training the models, under TF-IDF, the Multinomial