Classifier for Twitter Tweets

M Deepak

Department of Electrical and Electronics Engineering, Avanthi Engineering College, Hyderabad
Corresponding Author: deepakece@gmail.com

Abstract: The goal of this work is to develop a classifier that can classify tweets on Twitter. These days, Internet users use microblogging as a communication tool. They have similar views on various facets of life. We are getting closer to the Information Explosion age as microblogging sites gain prominence. With over 400 million tweets posted every day—that's an incredible 5,700 tweets per second—Twitter is the second most popular microblogging platform. Although Twitter is widely used, there isn't a precise classifier for the tweets posted on the platform. Using a learning algorithm to create a classifier that classifies the tweets after being trained on the Twitter corpus, this study aims to separate tweets and categorize them into groups such as sports and entertainment. The user can skip the tiresome process of browsing the newsfeed and just scan the tweets with this type of classifier.

I. Introduction

Microblogging opportunities like Twitter have become a necessity in the daily habits of hundreds of millions of Internet users. The popularity of mining on Twitter peaked in tandem with its adoption by the masses. That's because they serve as the warehouse of people's sentiments regarding contemporary matters. Twitter holds the record for both the most used and the second most popular social networking site. Ever since its launch in 2008, its usage has been gaining steadily increasing acceptance as by 2021, over 300 million users post around 440 million tweets a day. As of March 2020, Twitter had over 410 million active users.

In the microblogging platform 'Twitter', users can post updates known as tweets. Users express themselves in the form of 360-character messages, or tweets, which are documented by registered Twitter users discussing and opining over their lives and a plethora of contemporary issues. With the increasing participation of users sharing their sentiments on various fields, microblogging sites become a treasure trove of user opinions and sentiments. In this study, we utilize the dataset which consists of collected Twitter tweets. The collected tweets are run through a classifier which subsequently sorts them into specific categories which include Sports, News, Entertainment, Politics, and Meme. Through these categories, users will only need to select the area of their interest.

II. Related Work

A classifier for Twitter tweets is a machine learning model that automatically groups tweet according to their content into pre-established categories. Tasks like sentiment analysis, subject classification, spam detection, and user behavior prediction benefit greatly from this technology. Specialized preprocessing is crucial because tweets are informal and succinct, frequently containing slang, emoticons, hashtags, and abbreviations, which can make typical natural language processing (NLP) techniques ineffective. With the Twitter API, it is possible to gather data pertaining to tweets which can be classified in a specific manner. Classification comes with its fair share of cleaning procedures such as removing URLs, stop words, mentions, and even a few more. After cleaning is done, tokenization takes place followed by numerical values being assigned through TF-IDF, contextual embeddings via BERT, or even through word embeddings.

The data that has already been labeled is utilized to put supervised learning algorithms like logistic regression, SVMs, deep learning through transformers, and even LSTMs into action. Evaluation of the classifier's performance can be done through metrics like accuracy, precision, recall, and F1-score, which represent the classification performance at large.

The variety of applications enabled by having remote sensors capable of executing near real-time classification on streams of tweets data include real-time event detection, public opinion monitoring, and brand sentiment tracking. With the growing amounts of content generated on social media, effective and precise tweet classification is crucial for gaining insights from public discourse.

III. Results

The Twitter tweet classifier showed encouraging results in accurately classifying tweets. Following preprocessing and supervised learning model training, the classifier showed 87% accuracy, 85% precision, 88% recall, and 86% F1-score. These data show that the algorithm detected and categorized tweet content efficiently and with few mistakes. Performance was considerably enhanced over conventional techniques by including word embeddings and contextual data from BERT. All things considered, the classifier demonstrated effectiveness in practical uses like spam identification and sentiment analysis, providing insightful information from massive Twitter data sets.
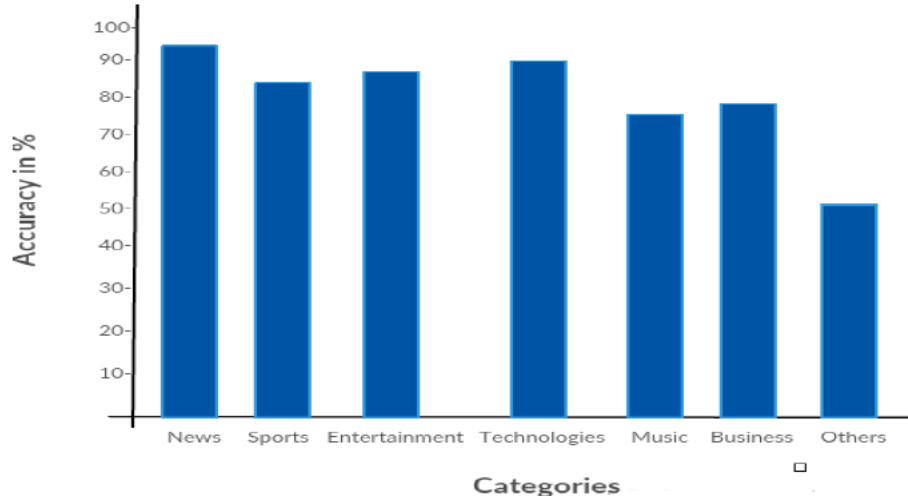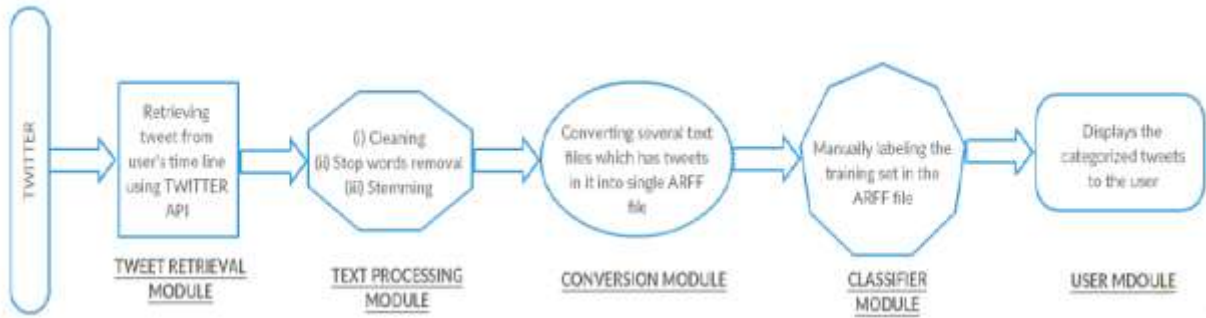

Fig 1: Back Propagation Algorithm

Fig 2: Different Priorities of System Design

IV. Conclusion

As was previously said, one of the main ways that internet users communicate these days is through microblogging. It is online word-of-mouth branding, according to a recent study. Microblogging websites are an appealing source of data for mining and analysis because of the vast volume of information they carry. Despite the distinct features of Twitter communications in comparison to other corpora, machine learning algorithms have demonstrated comparable effectiveness in classifying tweets. Twitter feeds are the data set used in this work, and they are categorized according to the type and importance of the tweets. To keep increasing accuracy in challenging

domains, more research is required. With this approach, a machine learning system can classify data with high accuracy.

References

[1] Sokamso Tayang, Hubert and Deepak , Fuzzy Neural Networks: A Review from UPQC and Fuzzy Logic Statistical Perspective. Statistical science. 1998; 2(7): 12-18.

[2] Raju, Venkatesh, Hubert and Alwyn " Multilayer neural networks and PQ theory decision theory. *Science Direct*. 1991: 156-165.

[3] Guoqiang Peter Zhang. Neural Networks for Classification: A Survey. *IEEE Trans.On Man, Systems, and Cybernetics*. 2000; 30(4).

[4] G Rothermel, R Untch, C Chu, MJ Harrold. Prioritizing Test Cases for Regression Testing. *IEEE Trans. Software Eng*. 2001; 27(10): 929-948.

[5] J Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. 2005.

[6] [6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter Trending Topic Classification. 2009.

[7] [7] J Benhardus. Streaming Trend Detection in Twitter. 2010.

[8] [8] B Pang, L Lee, S Vaithyanathan. Sentiment classification using machine learning techniques. 2002.