Emotion Recognition by Textual Tweets Classification Using Voting Classifier

Shaik Salma Sultana

Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Recently, social media platforms such as Twitter have generated enormous amounts of structured, unstructured and semi-structured data. Twitter provides an opportunity to its users to analyze its data on a large and broader point of view. Sentiment analysis is a technique used to analyze the attitude, emotions and opinions of different people towards anything, and it can be carried out on tweets to analyze public opinion on news, policies, social movements, and personalities. To further validate stability of the proposed approach on two more datasets, one binary and other multi-class dataset and achieved robust results. As Twitter is very fast and an efficient micro-blogging examination that facilitates the end users to transmit small posts are said to be tweets. Twitter is a highly demanding app in the world and is a successful platform in social media. Sentiment analysis inspires corporations to define clients' preferences about products, services, and brands. Further, it plays an important role in interpreting information about industries and corporations to reserve them in making entity review.*

I. INTRODUCTION

Automatic emotion recognition, pattern recognition and computer vision have become significantly important in Artificial Intelligence lately with applications is a wide range of areas. Recently, social media platforms such as Twitter have generated enormous amounts of structured, unstructured and semi-structured data. One of the most recent examples is COVID-19 info emic that shows misinformation in social media can be far more important and devastating than a disaster such as a pandemic. There is a need to analyze to accurately assign sentiment classes on a large scale. To perform such tasks, accurate NLP techniques and machine learning (ML) models for text classification are required. Twitter provides an opportunity to its users to analyze its data on a large and broader point of view. Efficient methods are important to automatically label text data due to its noisy nature. In the past many studies have been performed on Twitter sentiment classification [1]. As Twitter is very fast and an efficient micro-blogging examination that facilitates the end users to transmit small posts are said to be tweets. Twitter is a highly demanding app in the world and is a successful platform in social media. Free account can be created by using Twitter that can provide an enormous audience potential. With the purpose of business and marketing, Twitter can be proved as the best platform, through which one can get in touch with very rich and famous personalities like stars and celebrities, so their purchasing can be very charming for them as well as for advertisers. Using Twitter, every celebrity is linked with fans as well as to grant a communication to followers. Such a platform is one of the superlative approaches for lovers as well. But it has a short note range; only 140 letters for each post and it can type a post or link on the website since it has no cost and also open as the advertisements as well. There is no problem with clusters of personal ads which are similar to other social networking sites. It is quick because as a tweet is posted on Twitter, the public who is subsequent to respective business will get it without delay. Companies and advertisers can compose utilization of this source to check the diverse operational point of views which are very considerable. With help of this, they will obtain an immediate response from their followers. Remarkably, a lot of businesses with the intention of purchase, Twitter followers increase their deals. Twitter facilitates the followers by making them identify regarding fresh business, products, services, websites, blogs, eBooks etc. Consequently, Twitter clients might tick lying on link and also optimistically endow in a manufactured goods or examine the products presented and to get share in profit. It is extremely effortless to utilize as people can follow to get the news and updates, as organizations can tweet or re-tweet, they can mark favourite or selected people to send the tweets, also know how to propel the posts plus to be able to endow their money and instance through it. Academy, Industry, super bowls and Grammy Awards of such major Sports and Entertainment events generate a lot of buzz in the global world by using it. Competition is rising among different products on Twitter. People love to express their feelings about a particular product on social networks like twitter. Product owners are ready to spend more money on social media platforms to better advertise their products and to generate more revenue. When a person shares experience about a product, it helps the owner to change their market strategy, selling schemes, and improving the quality. Customer reviews serve as feedback to the owners or manufacturers too. The data generated in such a way is of large amount and requires an analysis expert team to classify the customer sentiment from the reviews. Experts can make a human error in sentiment analysis therefore it requires machine learning and ensemble learning classifiers to accurately classify the sentiment of the customers. This study compares various machine learning models for

emotion recognition by tweet classification using TF and TF-IDF. This research presents a voting classifier (LR-SGD) and aims to estimate the performance of famous ML classifiers on twitter datasets.

1.1 What is Data Engineering?

The key to understanding what data engineering lies in the "engineering" part. Engineers design and build things. "Data" engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state. These pipelines must take data from many disparate sources and collect them into a single warehouse that represents the data uniformly as a single source of truth.

1.2 How Did Data Engineering Come About?

Many would say that data engineering as a profession has been around for well over a decade, maybe a couple, ever since databases, Microsoft SQL Servers and ETL came to be. Some would say ever since IBM popularized database management systems in the 1970s. With that, here's a very brief history recap.

In the 1980s the term "information engineering" was coined to largely describe database design and to include software engineering in data analysis. Somewhere after the rise of the internet in the 1990s and 2000s, 'big data" came to be. Yet DBAs, SQL Developers and IT professionals working in the field were not labelled "Data Engineers" at that time.

1.3 Why the Critical Need for Data Engineering Now?

By now you've heard/read about Gartner's determination back in 2017 that 85% of big data projects fail. This was largely due to a lack of reliable data infrastructures. Data could not be trusted enough to base key business decisions on it. Fast forward to 2019 and things had not improved. The CTO of IBM said that 87% of data science projects never make it into production. Gartner reiterated its prediction that now just 80% of projects would fail. A New Vantage Report produced similar stats.

Over the last decade, most companies have completed a digital transformation. This has produced unimaginable volumes of new types of data and much more complicated data at a higher frequency. While it was previously apparent that Data Scientists were needed to make sense of it all, it was less apparent that someone needs to organize and ensure this data's quality, security, and availability for the Data Scientists to do their jobs.

So, in the early days of big data analytics, Data Scientists were very often expected to build the necessary infrastructure and data pipelines to do their work. This was not necessarily in their skill sets or expectations for the job. The result was that data modelling would not be done correctly. There would be redundant work and inconsistency in the use of data among Data Scientists. These kinds of issues prevented companies from being able to extract optimal value from their data projects, so they failed. It also led to a high rate of Data Scientist turnover that still exists today.

Today with the onslaught of completed corporate digital transformations, the Internet of Things and the race to become AI-driven, it is crystal clear that companies need Data Engineers in abundance to provide the foundation for successful data science initiatives.

This is why will we continue to see the role of Data Engineers grow in importance and breadth. Companies need teams of people whose sole focus is to process data in a way that allows them to extract value from it.

1.4 What Is the Relationship and Difference between Data Scientists and Data Engineers?

Much has been written about the relationships between these two roles, so we'll be brief. In the past, companies thought that they could get away with having Data Scientists do the role of Data Engineers. This is what has caused much of the "unicorn effect" and shortage in Data Scientist recruitment.

Some Data Scientists also sold themselves as being able to do a Data Engineer's job. Many fell short – see the image to the right courtesy of O'Reilly.com.

Today, the volume and speed of data have driven Data Scientist and Data Engineer to become two separate and distinct roles albeit but with some overlap.

It's now widely recognized that companies need both Data Scientists and Data Engineers in an advanced analytics team. It's difficult to do any meaningful data science without Data Engineers to support this function. There's frequent collaboration between Data Engineers and Data Scientists however the priority skills and knowledge of tools are different.

Data Engineer Ability:

Data Scientists are focused on advanced analytics of data that is generated and stored in a company's databases. Data Engineers design, manage and optimize the flow of data with those databases throughout the organization. So, Data Scientists will be highly skilled in math and statistics, R, algorithms and machine learning techniques. Data Engineers will be more versed in SQL, MySQL, and NoSQL, architecture and cloud technologies and frameworks such as agile and scrum.

Both will likely know Python, visualization techniques and have other coding languages in common.

Foundation software engineering – Agile, devOps, architecture design, service-oriented architecture.

Distributed systems – This would include software engineer skills and software architect skills.

Open Frameworks – Apache Spark, Hadoop, perhaps Hive, MapReduce, Kafka and others…

SQL – This is a database staple and remains that way.

Programming – Python has become the favoured language for working with data. Java on the other hand, while still widely sought has fallen out of favour with most data scientists and engineers. Scala is another language that Apache Spark and Kafka are based on.

Pandas – a Python library for cleaning and manipulating data.

Visualization/dashboards

Cloud platforms – AWS is probably the most prevalent cloud skill set for Data Engineers to know. Google Cloud Data Engineering and Microsoft Azure are right behind.

Analytics – While mainly the realm of data scientists, statistical analysis skills or understanding of some of the different mathematical principles or probabilistic principles are necessary for being able to properly manipulate the data so that it is in a shape that is accessible for the people who are doing the end analysis on it.

Data modeling – Data modeling knowledge is quite important now in the sense that a Data Engineer needs to know how they are going to structure tables, partitions, where to normalize and deformalize data in the warehouse, etc. and how to think about retrieving certain attributes.

II. LITERATURE SURVEY

Kwon, S., Cha, K., Jung, W.C., and Wang, Y et. al. [1] had published Aspects of Rumour Spreading on a Micro blog Network Rumours have been studied for several decades in social and psychological fields, where most studies were theory-driven and relied on surveys due to difficulties in gathering data. Rumour research is now gaining new perspectives, because online social media enable researchers to examine closely various kinds of information dissemination on the Internet. In this paper, we review social psychology literature on rumours and try to identify the key differences in the dissemination of rumours and non-rumours. The insights from this study can shed light on improving automatic classification of rumours and better comprehending Rumour theories in online social media.

Kwon, S., Cha, K., Jung, W.C., and Wang, Y et. al. [2] had published Prominent Features of Rumour Propagation in Online Social Media The problem of identifying rumours is of practical importance especially in online social networks, since information can diffuse more rapidly and widely than the offline counterpart. In this paper, we identify characteristics of rumours by examining the following three aspects of diffusion: temporal, structural, and linguistic. For the temporal characteristics, we propose a new periodic time series model that considers daily and external shock cycles, where the model demonstrates that rumour likely have fluctuations over time. We also identify key structural and linguistic differences in the spread of rumours and non-rumours. Our selected features classify rumours with high precision and recall in the range of 87% to 92%, that is higher than other states of the arts on rumour classification.

Oazvinian,VB.Rosengren,E., and Radev, R et. al. [3] had published Rumour has it: Identifying Misinformation in Micro blogsrumour is commonly defined as a statement whose true value is unverifiable. Rumours may spread misinformation (false information) or disinformation (deliberately false information) on a network of people. Identifying rumours is crucial in online social media where large amounts of information are easily spread across a large network by sources with unverified authority. In this paper, we address the problem of rumour detection in micro blogs and explore the effectiveness of 3 categories of features: content-based, network-based, and micro blog-specific memes for correctly identifying rumours. Moreover, we show how these features are also effective in identifying dis informers, users who endorse a rumour and further help it to spread. We

perform our experiments on more than 10,000 manually annotated tweets collected from Twitter and show how our retrieval model achieves more than 0.95 in Mean Average Precision (MAP). Finally, we believe that our dataset is the first large-scale dataset on rumour detection. It can open new dimensions in analysing online misinformation and other aspects of micro blog conversations.

Castillo, C., Mendoza , M., and Poblete, P et. al. [4] had published a "Information Credibility on Twitter" We analyse the information credibility of news propagated through Twitter, a popular micro blogging service. Previous research has shown that most of the messages posted on Twitter are truthful, but the service is also used to spread misinformation and false rumours, often unintentionally. On this paper we focus on automatic methods for assessing the credibility of a given set of tweets. Specifically, we analyse micro blog postings related to "trending" topics, and classify them as credible or not credible, based on features extracted from them. We use features from users' posting and re-posting ("re-tweeting") behaviour, from the text of the posts, and from citations to external sources. We evaluate our methods using a significant number of human assessments about the credibility of items on a recent sample of Twitter postings. Our results shows that there are measurable differences in the way messages propagate, that can be used to classify them automatically as credible or not credible, with precision and recall in the range of 70% to 80%.

Mendoza, M., Poblete, B, and Castillo, C et. al. [5] had published a "Twitter under Crisis: Can we Trust. In this article we explore the behaviour of Twitter users under an emergency situation. In particular, we analyse the activity related to the 2010 earthquake in Chile and characterize Twitter in the hours and days following this disaster. Furthermore, we perform a preliminary study of certain social phenomenon's, such as the dissemination of false rumours and confirmed news. We analyse how this information propagated through the Twitter network, with the purpose of assessing the reliability of Twitter as an information source under extreme circumstances. Our analysis shows that the propagation of tweets that correspond to rumours differs from tweets that spread news because rumours tend to be questioned more than news by the Twitter community. This result shows that it is possible to detect rumours by using aggregate analysis on tweets.

III. EXISTING SYSTEM:

Sentiment analysis inspires corporations to define clients' preferences about products, services, and brands. Further, it plays an important role in interpreting information about industries and corporations to reserve them in making entity review. Sarlan established a sentiment analysis through extracting number of tweets with the help of prototyping and the results organized customers' views via tweets into positive and negative. Their research divided into two phrases. The first part is based on literature study which involves the Sentiment analysis techniques and methods that nowadays are used. In the second part, the application necessities and operations are described preceding to its development. The distinct approaches and conclusions of algorithm performance were compared. Methods were used which were supervised ML based, lexicon-based, ensemble methods. Authors used four methods that were Twitter sentiment Analysis using Supervised ML Approaches; Twitter sentiment Analysis using Ensemble Approaches. Twitter sentiment Analysis is using lexicon-based Approaches.

Disadvantage:

- Features extraction techniques not only convert textual features into vector form but also helps to find significant features necessary to make predictions.

IV. PROPOSED SYSTEM

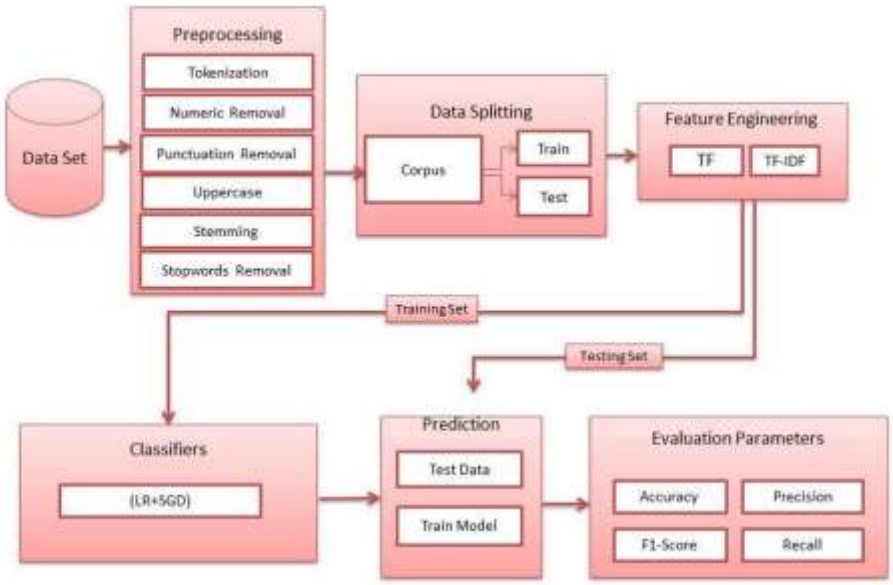
- Automatic emotion recognition, pattern recognition and computer vision have become significantly important in Artificial Intelligence lately with applications is a wide range of areas.
- Efficient methods are important to automatically label text data due to its noisy nature. In the past many studies have been performed on Twitter sentiment classification.
- Twitter is a highly demanding app in the world and is a successful platform in social media.
- One can improve the performance of models by recognizing patterns efficiently and through effective averaging combination of models.
- Twitter dataset used in this experiment is scrapped from Kaggle repository. First the dataset is pre-processed by removing unwanted data.

- Then, the data was split into two sets: training set and testing set. The training set was given the percentage of 70% while the test set portion is 30%.
- After that feature engineering techniques are applied on the training set. Multiple machine learning classifiers are trained on the training set and tested using the test set.
- The evaluation parameters used in this experiment are: (a) Accuracy (b) Recall (c) Precision (d) F1-score.

Advantages:

- High security and more effective.
- This system gives more accuracy compared to another systems.
- One can improve the performance of models by recognizing patterns efficiently and through effective averaging combination of models.

V. SYSTEM ARCHITECTURE



Admin:

To provide following features-

1. Register into our system.
2. Login to our System.
3. Only Authorized Admin can access to our System.
4. Upload Related Dataset to Our System.
5. Data pre-processing process takes place like (stop-word removal,

   Stemming, Tokenization) to remove the unwanted data.
6. Classification of hashtags – Using Collaborative algorithm and CNN algorithm classification of the hashtag takes place.
7. Feature extraction - To predict the tweets keyword in which type of group based on the username.
8. Group Recommendation -To calculate Betweenness value which means the users who is in more than one group. Classify the tweets in datasets into the groups by using Collaborative Filtering algorithm.

- Remove all the redundant data.

- Clustering the group by using CNN algorithm.

9. Performance Analysis - To calculate the Algorithm accuracy by using Confusion matrix.

10. Result- Generate the graph based on the number of tweets in each group.

VI. CONCLUSION:

A novel combination of LR and SGD as a voting classifier for emotion recognition by classifying tweets as happy or unhappy. Our experiments showed that one can improve the performance of models by recognizing patterns efficiently and through effective averaging combination of models. The results showed that all models performed well on tweet dataset but our proposed voting classifier VC(LR-SGD) outperforms by using both TF and TF-IDF among all. Proposed model achieves the highest results using TF-IDF with 79% Accuracy, 84% Recall and 81% F1-score. The proposed model is further validated on two more dataset and achieved robust results.

VII. APPLICATION & FUTURE WORK

Seven Machine Learning models are implemented for emotion recognition by classifying tweets as happy or unhappy.

The future work will compare more feature engineering techniques and explore more combinations of ensemble models to improve the performance. In addition, new techniques will be investigated to deal with sarcastic comments.

REFERENCES

[1] J. Capdevila, J. Cerquides, J. Nin, and J. Torres, ‘‘Tweet-SCAN: An event discovery technique for geo-located tweets,’’ Pattern Recognit. Lett., vol. 93, pp. 58–68, Jul. 2017.

[2] T. Alsinet, J. Argelich, R. Béjar, C. Fernández, C. Mateu, and J. Planes, ‘‘An argumentative approach for discovering relevant opinions in Twitter with probabilistic valued relationships,’’ Pattern Recognit. Lett., vol. 105, pp. 191–199, Apr. 2018.

[3] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, ‘‘Unsupervised rumor detection based on users’ behaviors using neural networks,’’ Pattern Recognit. Lett., vol. 105, pp. 226–233, Apr. 2018.

[4] H. Hakh, I. Aljarah, and B. Al-Shboul, ‘‘Online social media-based sentiment analysis for us airline companies,’’ in New Trends in Information Technology. Amman, Jordan: Univ. of Jordan, Apr. 2017.

[5] R. Xia, C. Zong, and S. Li, ‘‘Ensemble of feature sets and classification algorithms for sentiment classification,’’ Inf. Sci., vol. 181, no. 6, pp. 1138–1152, Mar. 2011.

[6] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, ‘‘A novel stacked CNN for malarial parasite detection in thin blood smear images,’’ IEEE Access, vol. 8, pp. 93782–93792, 2020.