International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

Fake News Detection and Sentiment Analysis in Twitter

Priyanka V. Tatipamal[1], Prof. Sanjay S. Badhe[2]

P G. Student, Department of Electronics & Telecommunication Engineering,D. Y. Patil College of Engineering, Ambi, Talegaon, Pune, India[1]

Professor, Department of Electronics & Telecommunication Engineering,D. Y. Patil College of Engineering, Ambi, Talegaon, Pune, India[2]

ABSTRACT: Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in the source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user-generated data is very useful in knowing the opinion of the crowd. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. Knowledge base approach and Machine learning approach are the two strategies used for analysing sentiments from the text. Public and private opinion about a wide variety of subject is expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. Twitter offers organizations a fast and effective way to analyse customers' perspectives toward the critical to success in the market place. Developing a program for sentiment analysis is an approach to be used to computationally measure customers' perceptions. This project uses a knowledge base including various patterns for tweets along with multiple strategies to detect the sentiment expressed in a tweet and if a tweet is genuine or not. Various machine learning and knowledge base approaches are used to compare patterns and apply strategies and NLP for sentiment analysis.

KEYWORDS: NLP (Natural Language Processing), Sentiment Analysis, Machine Learning, Pattern Matching, Twitter Data, POS (Part of Speech)

I. INTRODUCTION

Twitter has emerged as a major micro-blogging website, having over 100 million users generating over 500 million tweets every day. With such large audience, Twitter has consistently attracted users to convey their opinions and perspective about any issue, brand, company or any other topic of interest. Due to this reason, Twitter is used as an informative source by many organizations, institutions and companies. On Twitter, users are allowed to share their opinions in the form of tweets, using only 140 characters. This leads to people compacting their statements by using slang, abbreviations, emoticons, short forms etc. Along with this, people convey their opinions by using sarcasm and polysemy. Hence it is justified to term the Twitter language as unstructured.

In order to extract sentiment from tweets, sentiment analysis is used. The results from this can be used in many areas like analysing and monitoring changes of sentiment with an event, sentiments regarding a particular brand or release of a particular product, analysing public view of government policies etc. A lot of research has been done on Twitter data in order to classify the tweets and analyse the results. In this project we aim to predict the sentiments from tweets by checking the polarity of tweets as positive, negative or irrelevant. Sentiment analysis is a process of deriving sentiment of a particular statement or sentence. It's a classification technique which derives opinion from the tweets and formulates a sentiment and on the basis of which, sentiment classification is performed. Sentiments are subjective to the topic of interest. We are required to formulate that what kind of features will decide for the sentiment it embodies. In the programming model, sentiment we refer to, is class of entities that the person performing sentiment analysis wants to find in the tweets. The dimension of the sentiment class is crucial factor in deciding the efficiency of the model. For example, we can have two-class tweet sentiment classification (positive and negative) or three class tweet sentiment classification (positive, negative and irrelevant). Sentiment analysis approaches can be broadly categorized in two classes – lexicon based and machine learning based. Lexicon based approach is unsupervised as it proposes to perform analysis using lexicons and a scoring method to evaluate opinions. Whereas machine learning approach involves use of feature extraction and training the model using feature set and some dataset.

The basic steps for performing sentiment analysis includes data collection, pre-processing of data, feature extraction, selecting baseline features, sentiment detection and performing classification either using simple computation or else machine learning approaches. Features are like Parts-of speech features i.e. nouns, adverbs, adjectives, etc. in

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

each tweet are tagged. We for the accuracy purpose of polarity will be detecting sentiments along with various knowledge base patterns and multiple machine learning strategies to evaluate the sentiments. Alongside we will be checking if the tweet is genuine or not or if has influenced by other tweets which can be very useful in rumours mitigation on social medias. This approach will produce the higher accuracy for polarity by considering POS factor and genuineness as well as can be used in various sectors such as analysing product reviews or government policies, etc. where it can be found if negative influence is spread and if it affects people.

II. LITERATURE SURVEY

[1] Characteristics Analysis of Data from News and Social Network Services, "Beakcheol Jang", 2018. In this study, we conducted comprehensive measurements to understand the characteristics, including similarities and differences, of data from the news and SNSs. The observed differences are as follows: It is challenging to find the same topic in the news and SNS. The news responds to official events whereas SNSs respond to personal interests. The news mentions a specific topic continually, whereas the transition from one topic to another in SNSs is fast. The issues discussed on SNSs are different every day. The news can identify specific events with a single keyword, but many keywords are required to find the required data in SNSs.

[2] Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness, "Ying Fang, Jun Zhang", 2018. A new method for the calculation of polarities and strengths of Chinese sentiment phrases is proposed in this study, which could be used to analyse semantic fuzziness of Chinese. It uses a probability value, rather than xed value for the polarity strengths of sentiment phrases, compared with the conventional methods.

[3] A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter, "Mondher Bouazizi, Tomoaki Ohtsuki", 2017. In this project, we have proposed a new approach for sentiment analysis, where a set of tweets is to be classified into 7 different classes. The obtained results show some potential: the accuracy obtained for multi-class sentiment analysis in the data set used was 60.2%. However, we believe that a more optimized training set would present better performances.

[4] Security Attack Prediction Based on User Sentiment Analysis of News Data, "Aldo Hernández, Victor Sanchez", 2016. This project proposed a sentiment analysis method for news based on a linear regression model. The method employs natural language processing analysis on a collected corpus and determines negative sentiments within a specific context. The objective is to predict the response of specific groups involved in hacking activism when the sentiment is negative enough among different News users.

[5] Exploring Sentiment Analysis on News Data, "Manju Venugopalan, Deepa Gupta", 2015. A hybrid news sentiment classification model incorporating domain-oriented lexicons, unigrams and news specific features using machine learning techniques has been developed and the classification accuracies have been found to improve by an average of around 2 points across different domains. The effectiveness of incorporating concepts of domain specificity in the polarity lexicon and the capacities of explicit news features to extract sentiment has been validated. Pruning the unigrams based on their significant presence in a class has simplified the model to a large extent.

[6] Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation, "Rincy Jose, Varghese S Chooralil", 2015. In this project, we have implemented a real time, domain independent twitter sentiment analyser using sentiment lexicons such as SentiWorldNet and WorldNet. It compared political sentiment towards two politicians by plotting graphs using results of sentiment analysis on real-time extracted twitter data. This was done by applying WSD and negation handling in order to increase accuracy of sentiment analysis. Negation handling results in 1% improvement in classification accuracy and WSD results in 2.6% improvement in classification accuracy.

III. EXISTING SYSTEM

In Existing System, to analyse the behaviour of news required maximum resources. To analyse the fake news, we required man power to deep down into it and check the authentication of news. We have to check all possible connection with news manually. It is time consuming and costly approach. Limitation of existing system: -

- Time Consuming Process
- Man-Power Required.
- Deep Knowledge required.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

- Cost driven approach.

IV. PROPOSED SYSTEM

In the proposed system, we will retrieve tweets from twitter using twitter API based on the query. The collected tweets will be subjected to pre-processing. We will then apply the various patterns and strategic algorithms including some of machine learning algorithms for NLP to supervise the data. The results of the algorithms i.e. the sentiment and influence will be represented in graphical manner (pie charts/bar charts). The proposed system is more effective than the existing one.

This is because we will be able to know how the statistics determined from the representation of the result can have an impact in a particular field as well as influence of negativity spread by rumours.

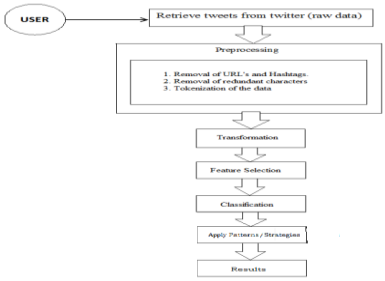

Fig.1: - System Architecture

V. METHODOLOGY

➢ Natural Language Processing:

✓ Fake News/Rumour Detection: -

A rumor can be described as a temporal communication network, where each node corresponds a communicating user, the edges correspond to communication between nodes and the temporal aspect captures the propagation of messages through the network. The intuition is that there are measurable differences between the temporal communication network corresponding to false and true rumours. In order to capture these differences, we need to identify characteristics of rumours. It makes sense that these characteristics would be related to either the nodes (i.e. users) in the network, the edges (i.e. messages) in the network or the temporal behaviour of the network (i.e. propagation).

✓ Linguistic: -

The linguistic features capture the characteristics of the text of the tweets in a rumour. A total of 4 linguistic features were found to significantly contribute to the outcome of our models. In the descending order of contribution these features are: ratio of tweet containing negations, average formality & sophistication of the tweets, ratio of tweets containing opinion & insight, and ratio of inferring & tentative tweets. We will now describe each of these features in detail.

• Vulgarity: The presence of vulgar words in the tweet.
• Abbreviations: The presence of abbreviations (such as b4 for before, jk for just kidding and irl for in real life) in the tweet.
• Emoticons: The presence of emoticons in the tweet.
• Average word complexity: Average length of words in the tweet.
• Sentence complexity: The grammatical complexity of the tweet.

✓ User Identities: -

The user features capture the characteristics of the users involved in spreading a rumour. A total of 6 user features were found to significantly contribute to the outcome of our models. In the descending order of contribution these features are: controversialist, originality, credibility, influence, role, and engagement.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

✓ Propagation Dynamics: -

The propagation features capture the temporal diffusion dynamics of a rumour. A total of 7 propagation features were found to significantly contribute to the outcome of our models. In the descending order of contribution these features are: fraction of low-to-high diffusion, fraction of nodes in largest connected component (LCC), average depth to breadth ratio, ratio of new users, ratio of original tweets, fraction of tweets containing outside links, and the fraction of isolated nodes. All of these features are derived from a rumour's diffusion graph.

✓ Sentiment Analysis: -

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

✓ TextBlob: -

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks.

A good thing about TextBlob is that they are just like python strings. So, you can transform and play with it same like we did in python. Below, I have shown you below some basic tasks. Don't worry about the syntax, it is just to give you an intuition about how much-related TextBlob is to Python strings.

VI. RESULT & DISCUSSION

In proposed system we have created one web-based application using Python's Flask framework which is light weight. In proposed application we are fetching real time tweets from twitter data and applying algorithm on them to get result out of that. To access data from twitter, you need to have authenticate twitter developer account which allows you to access the data. After accessing the data, we are also storing that data into SQL database. Then we applied algorithms on that data. For sentiment analysis, we are using TextBlob and NLTK libraries. And for fake news detection, we have used TFIDF algorithm. It's taking approximately two to five minutes for execution. According to network complexity it fetches tweets from server and process them. Following screenshot shows the final output of project which has interpreted COVID19 keyword on twitter and which is accurate and efficient.

Fig 2. Fake News Detection

VII. MATHEMATICAL MODEL

Let 'S' be the system

Where,

S= {I, O, P, Fs, Ss}

Where,

I = Set of input

O = Set of output

P = Set of technical processes

Fs = Set of Failure state

Ss = Set of Success state

Identify the input data I1, I2, in

I = {(Twitter Data)}

Identify the output applications as O1, O2, on

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

O = {(Rumors Detection, Fake News Detection, Sentiment Detection)}
Identify the Process as P
P = {(Data Processing, Natural Language Processing, Sentiment Analysis, Pattern Recognition)}
Identify the Failure state as Fs
Fs = {(If fake news not predicted}
Identify the Success state as Ss
P = {(Fake news detected successfully)}

VIII. FUTURE SCOPE

In future we can integrate our system with another social media platform. We can optimize time complexity by using better resources. We can add more parameter to enhance the accuracy of system. We can implement this system with several other algorithm to compare the accuracy.

IX. CONCLUSION

The project set out to solve a practical problem of sentiment analysis and genuinely check of Twitter posts. We proposed a method using knowledge base patterns, strategies and machine learning approaches. These methods are proposed to increase the accuracy of sentiment check for tweets. Patterns can be used to evaluate if the tweets were a influenced rumour or a genuine post by any user. By using API of twitter, it is possible to work on live tweets than to work on offline data. Querying and fetching of particular tweets from twitter is possible by using its API. Finding influence or negativity spread by users can be useful in various analytical tasks.

REFERENCES

[1] Beakcheol Jang And Jungwon Yoon "Characteristics Analysis of Data from News and Social Network Services" Doi 10.1109/Access.2018.2818792, IEEE.
[2] Ying Fang, Jun Zhang "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness" 2169-3536 2018 IEEE.
[3] Mondher Bouazizi And Tomoaki Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter" 2169-3536 2017 IEEE.
[4] Aldo Hernández, Victor Sanchez "Security Attack Prediction Based on User Sentiment Analysis of Twitter Data" 978-1-4673-8075-1/16/$31.00 2016 IEEE
[5] Manju Venugopalan And Deepa Gupta "Exploring Sentiment Analysis on Twitter Data" 978-1-4673-7948-9/15/$31.00 ©2015 IEEE
[6] Rincy Jose And Varghese S Chooralil "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data Using Word Sense Disambiguation" 978-1-4673-7349-4/15/$31.00 ©2015 IEEE
[7] Anurag P. Jain And Mr. Vijay D. Katkar "Sentiments Analysis of Twitter Data Using Data Mining" 978-1-4673-7758-4/15/$31.00 ©2015 IEEE
[8] Gaurav D Rajurkar And Rajeshwari M Goudar "A Speedy Data Uploading Approach for Twitter Trend and Sentiment Analysis Using Hadoop" 978-1-4799-6892-3/15 $31.00 © 2015 IEEE