International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Sentiment Analysis for Tweets using Patterns and Strategies to Detect the Genuineness of Tweets.

Deepti Kulkarni, Nagaraju Bogiri

Abstract*: Sentiment analysis related with distinguishing and classifying opinions or sentiments expressed in given text. Social media is completing a colossal quality of wealthy knowledge within the type of tweets, standing updates, diary posts etc. Sentiment analysis of this user generated knowledge is incredibly helpful in knowing the opinion of the gang. Twitter sentiment analysis is troublesome compared to general sentiment analysis thanks to the presence of slang words and misspellings. Knowledge base approach and Machine learning approach square measure the 2 methods used for analyzing sentiments from the text. Public and private opinion a few wide range of subjects' square measure expressed and unfold frequently via numerous social media. Twitter offers organizations quick and effective thanks to analyze customers' view toward the crucial to success within the market place. Developing a program for sentiment analysis is an approach to be accustomed computationally live customers' perceptions. This project cognitive content together with varied patterns for tweets along a side multiple strategies to discover the sentiment class expressed in a very tweet and if a tweet is real or not. We proposed work to classify sentiments of tweets from people to determine if people are happy, sad, angry, etc. about particular topic. Also the purpose of the work is to check genuineness of tweets so that rumors about any topics can be detected and mitigated. This approach can be used in various fields further as like detecting people sentiment about particular social issues. Also fake tweets and rumors which may further exploit to social issues like riots, religion complexities can be removed. To achieve these goals and fetch tweets we used Twit4j API and various techniques such as NLP, TF-IDF, and Sentiment Classifications are applied to get results accordingly. We had maintained our own database of words for dictionary purpose as well as have been used OpenApache NLP with their predefined dictionaries.

Keywords—NLP Sentiment analysis, machine learning, influence of tweets, POS

I. INTRODUCTION

There are more than 100 million peoples who daily used Twitter and they tweets more than 500 million tweets every day [2]. With large number of audience, Twitter systematically attracts users to express their opinions and perspective about any issues, brand, company or the other interested or trending topic. According to this, Twitter is employed as an informative supply by several organizations, establishments and firms. Using Twitter, users can post or share their opinions in the form of tweets using only 140 characters [3][4]. This results in folks compacting their statements by exploitation slang, abbreviations, emoticons, short forms etc. Along with this, people convey their opinions by using sarcasm and polyse my [6]. Hence it' seven to term the Twitter language as unstructured. In order to explain sentiment from tweets, sentiment analysis is used [1]. This will be employed in several areas like analyzing and monitoring changes of sentiment with a happening, sentiments relating to a specific complete or unleash of particular product, analyzing public read of state of government policies etc.

A number of analysis has been done on Twitter knowledge so as to classify the tweets and analyze the results obtained. In this project we have a tendency to aim to predict the emotions from tweets by checking the polarity of tweets as positive, negative or irrelevant. Sentiment analysis could be a method of explanation sentiment of a specific statement or sentence in order to get emotions of users toward it. Sentiments are according to the topic of interest [8]. The area unit needed to formulate that what reasonably options can 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. We have proposed the seven different classes of sentiment for classification. The dimension of the sentiment category is crucial consider in deciding the potency of the model. For example, we can have two-class sentiment classification (positive and negative) or three class sentiment classification (positive, negative and irrelevant) [10]. Sentiment analysis approaches will be loosely categorized in 2 categories – lexicon based mostly and machine learning based. Lexicon based approach is unsupervised because it proposes to perform analysis exploitation lexicons and a rating methodology to gauge opinions. Whereas in machine learning approach it uses feature extraction and training the model using feature sets defined and some dataset [1].

We need to classify tweets in multiple types of sentiments by going further that just finding the polarity. We propose some classes such as sadness, happy, love to classify tweets according to their features extracted into these classes. Here as used in paper [1] we will not be using fixed dataset yet will be fetching live tweets of users whom the system follows.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Deepti Kulkarni, Computer Engineering Department, K. J.College of Engineering & Management, Savitribai Phule Pune University, Pune, India
Nagaraju Bogiri, Computer Engineering Department, K. J.College of Engineering & Management, Savitribai Phule Pune University, Pune, India.

II. LITERATURE SURVEY

Beakcheol Jang and Jungwon Yoon [2] conducted comprehensive measurements to understand the characteristics, including similarities and differences, of data from the news and SNSs. The resulted variations square measure as follows: it's difficult to get constant topics within the news and SNS. The news covers to official events and SNSs covers to personal interests. The news mentions a specific topic continually, and the transition from one topic to another in SNSs is fast. The issues mentioned on SNSs square measure completely different each day. The news can be well describing specific events with just a single keyword, but many keywords are required to find the required data or interest in SNSs.

Pros: Focused on news and SNS and prediction is good.

Cons: Keywords database is must.

Fang and Zhang [3] proposed approach for the calculating polarities and strengths of Chinese sentiments phrases in this study, which could be used to analyze semantic fuzziness of Chinese language. It uses a value of probability, than hexed value for the polarity and strengths detection of sentiment phrases of Chinese language, compared with the conventional methods.

Pros: New approach with fuzzy logic

Cons: Works with only Chinese language which is not useful for other countries.

Mondher and Tomoaki [4] proposed an approach for sentiment classification, where a set of tweets is classified into 7 different types of classes. The obtained results show some potential: the accuracy obtained for multi-class sentiment analysis within the information set used was 60.2%. Though, they tend to believe that a lot of optimized coaching dataset would give higher performance and accuracy.

Pros: Multiple sentiment class support is included.

Cons: Accuracy is not much achieved from available dataset

Aldo Hernández, Victor Sanchez [5] had proposed a sentiment analysis method of tweets developed upon a linear regression model. The approach deploys natural language processing analysis on a collected corpus or data and determines negative sentiments within a particular context. The objective is to predict the response of specific teams concerned in hacking policy among completely different Twitter users once the sentiment is negative enough prediction of hacking can be done.

Pros: Very effective to get opinions about specific issue using dataset for particular.

Cons: Focuses on particular issues and need dataset for each issue to predict sentiments.

Manju Venugopalan and Deepa Gupta [6] has combined tweet sentiment classification model covering domain oriented lexicons, unigrams and specific features of tweets using machine learning techniques has been developed. Classification accuracies by this method are found to increase by a median of around two points across completely different domains.

The incorporating effectiveness ideas of domain specificity within the polarity lexicon and therefore the capacities of specific tweet options to extract sentiment have been valid. Pruning the unigrams supported their important presence in a very category has simplified the model to an outsized extent.

Pros: Lexicons and unigrams are focused and machine learning techniques applied.

Cons: Techniques are more focused than results for sentiment.

Rincy Jose and Varghese S Chooralil [7] have implemented a real time, domain independent twitter sentiment analyzer using sentiment dictionaries like SentiWordNet and WordNet. It compared political sentiment towards two politicians by plotting graphs exploring results of sentiment analysis on real-time extracted twitter data. This was done by applying WSD and negation handling for increasing accuracy of sentiment analysis. Negation handling results in 1% improvement in classification accuracy and WSD ends up 2.6% improvement in classification accuracy.

Pros: Online dictionary like WorldNet is used for enhanced dataset.

Cons: Focused on particular issue of election prediction

III. PROPOSED METHODOLOGY

Tweets with the specific words or with search query should be collected, for example e.g. “#PulwamaAttacks” or “Modiji”. Such tweets collections should allow by The Twitter API only for accessing in our application. According to this reason, this step must start as soon as an event arises or a fake tweet begins. First of all tweets are fetched from twitter by using Twitt4j API, also particular tweets can be searched. Tweets are then classified according to sentiment classes such as happy, sadness, disgust, anger, etc. Then the tweets are preprocessed such as tokenization, removal of symbols and other things will take place here. Then the features of tweets will be extracted based on keywords to detect the sentiment class. Polarity of tweet will be found based on Stanford NLP algorithm or Open Apache NLP algorithm. After NLP pattern matching as shown in Table 1 will be applied to it and then strategies will be applied. In this method we will collect all three results and if any both of them return true then that tweet will not be genuine.

Sentiment classes for tweet classification are as follows,

1. Love/ Happy
2. Neutral
3. Anger
4. Hate
5. Sadness
6. Surprise
7. Disgust

To achieve classification we will use database of respective words of each classes and compare tweet texts with database to detect class. TF-IDF algorithm along with Decision tree will be used for this purpose. For achieving sentiment classification following feature extractions will be used,

1. Features related with Sentiment.
2. Features related with Punctuation.
3. Features related with Syntactic and stylistic.
4. Features related with Semantic.
5. Top words present in tweet.
6. Features related with Pattern.

Features read from tweets are all textual features. The NLP algorithms as prescribed above will return a sentiment score as -1, 0, 1 which we assume as Senti Score[1]. After sentiment score it will be compared with data dictionary manually minted for sentiment class.

Retrieval Number: H7397068819/19©BEIESP
DOI: 10.35940/ijitee.H7397.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The calculation for sentiments will be as follows,

- Positive words present in tweet.
- Negative words present in tweet.
- Highly sensitive positive words present in tweet (i.e., words having score returned by Senti Score greater or equal to 1)
- Highly sensitive negative words present in tweet (i.e., words having score returned by SentiScore less or equal to -1)
- Ratio of emotional words p(*t*) defined

$$\rho(t) = \frac{PW(t) - NW(t)}{PW(t) + NW(t)}$$

Where, PW is the total number of positive words as returned by SentiScore [1] present in tweet and NW is the total score of negative words as per returned by SentiScore [1] present in tweet. If in tweet, Emotional words are not present p is set to 0.

(a) Analysis of tweets and searching of fake tweets were done manually. In this step, all the irrelevant tweets were sort out. For example, we ignore tweets talking about holidays in Brussels because unfortunate "Brussels attacks".

(b) Collection of more tweets applicable to the story with keywords which are missed in the beginning of Step 1 (this step is optional). For example, at the time of searching for fake tweets we might come across tweets talking about another fake tweet. We add the keyword that describes this new fake tweet in our tweet collection.

(c) Categories tweets to fake tweets. Group all tweets which refer to the same fake tweet.

(d) Finding all the unique users present in a fake tweet. In Steps 1to 2 this set of users will be used. Before staring of fake tweet, we should collect most recent 400 tweets of user. This step is necessary to examine the users' past behavior and sentiment, for e.g. during the fake tweet if there is change in users' writing style or sentiment changes and whether these features are important for the model. For our best knowledge, in academic literature in building fake tweet classifier, this set of features is considered.

1. For propagation graph, collection of users' followers (friends) is necessary.
2. Users' information, such as registration date and time, description, whether account is verified or not etc is collected.

Table 1 Patterns for tweets analysis

Pattern	Description
like_count	No of users who have liked the post
comment_count	No of users who have commented on post
repost_count	No of users who have reposted the post
senti_score	Sentiment score of post.
picture_count	No of pictures posted in post.
tags_count	No of #tag in post
mention_count	No of @mentioned in post
smiley_count	No of smiley's in post
question-marks_count	No of question marks in post
firstperson_count	No of first person in the post
length_post	Length of the post.
is_repost	Whether the post is repost
timestamp	Hour the post was posted
source	How the post was posted

The system operates on tweets and tweets are fetched by using Twitt4j API using Java as programming platform. The steps involved in working are as follows,

1. Authenticate
2. Fetch Tweets
3. Preprocess (Tokenization, removal of special symbols and URLs)
4. Read Tweet features such as mention counts, tag counts, smiley, URL, etc.
5. Perform NLP to detect polarity of preprocessed tweet.
6. Match patterns from read features of tweets as shown in table 1.
7. Apply strategies such as time of tweet, username, location and compare this info in the know database.
8. If polarity is negative and any of the pattern matches or polarity is positive but patterns are matching and strategies returns true then tweet is not genuine

A. MATHEMATICAL MODEL

Let S be the closed system defined as,

S = {Ip, Op, A, Ss, Su, Fi}

Where, Ip=Set of Input, Op=Set of Output, Su= Success State, Fi= Failure State and A= Set of actions, Ss= Set of user's states.

- Set of input=Ip={username, password}
- Set of actions =A={F1,F2,F3,F4,F5,F6} Where,
  - F1= Authentication of user
  - F2 =Fetching Tweets from twitter
  - F3 = Compare all Tweets
  - F4 = Influence of negativity spread by tweets
  - F5= Classification on tweets
  - F6= Show Result
- Set of user's states=Ss={registration state, login state, selection of tweets, feature selection, classification, logout}
- Set of output=Op={Show twits analysis results}
- Su=Success state={Registration Success, Login Success, Fetch/Search tweets success}
- Fi=Failure State={Registration failed, Login failed, API failure}

Set of Exceptions= Ex ={Null Pointer Exception while various states, Record Not Found (Invalid Password) state , Null Values Exception while fetching tweets, Limit exhaust while fetching tweets}

The presented mathematical model is based on set theory of modeling a system where full set of system is considered as set 'S'. This is the simplest way of presenting system in mathematical forms and is very easy to understand.

B. ALGORITHMS

We are using various standard algorithms such as NLP, TF-IDF, Naïve Bayes, algorithms. NLP algorithm id used to check the polarity of the tweets that either it is positive, negative or neutral. TF-IDF is used to check the pattern matching as well as to check that the term is present in the document or not. It is also used for ranking the document. Naïve Bayes classification is used to classify the user sentiments into multiple classes.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

C. OWN CONTRIBUTION

As discussed earlier in the paper many of the authors focused on only sentiment analysis of the tweets, some used online tools for that as Senta[1], and some used machine learning approaches as SVM classification[4]. Most of the approaches use offline dataset and it cannot be verified if those tweets are legitimate or has been removed by twitter. Hence we use online twitter API to fetch current tweets and process them. We not only find polarity of tweets but add sentiment classes to it, then by applying patterns and strategies we check if the tweets was genuine, if the tweet was rumor and the user who posted the tweet is legitimate user or not. Many times it may happen that some screen names of genuine users will match in our database but if their tweets is not matched in disrespectful sentiment class or not of negative polarity and pattern matching score is low then they will be ignored

D.System Architecture

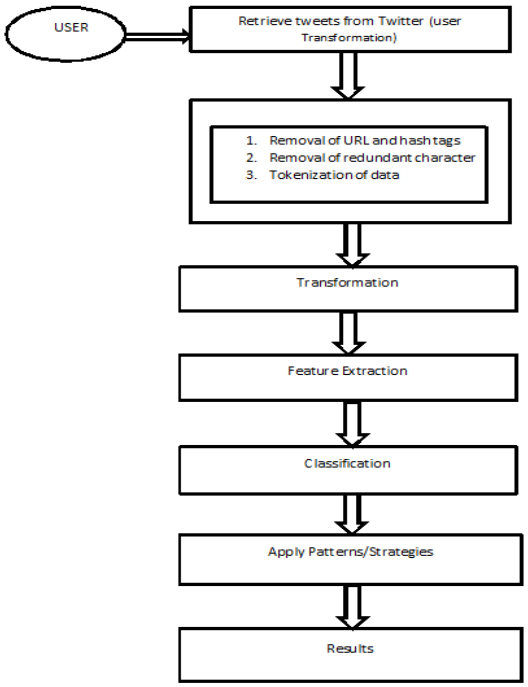

Fig 1 System architecture diagram

IV. PREDICTED RESULT AND DISCUSSIONS

According to the proposed system, Based on the query input, we will fetch tweets from twitter account using twitter API of that query. The fetched tweets will be subjected for preprocessing. We will then apply the various patterns and strategic algorithms as well as few machine learning algorithms for NLP for supervise the data. The algorithms result i.e. the sentiment and influence will be shown in graphical formats (pie charts/bar charts) for quick understanding. The system proposed is more practical than the existing one. This is as a result of we'll be ready to shrewdness the statistics determined from the illustration of the result will have a sway in a very specific field furthermore influence of negativity spread by fake tweets.

Many of the existing systems as discussed in literature focuses on only NLP algorithms and resulting in sentiment or polarity of the tweet. Our system extends these results with applying various pattern matching as shown in table 1 along with strategies to detect if the tweets were genuine or not so that other users will know how much to trust such tweets. The system also can be used in predicting anonymous or fake profiles on twitter which can be helpful further to mitigate such false tweets.

Comparative results of existing and proposed system is as following Table 2,

Parameters	Existing System	Proposed System
Sentiment Analysis	Yes	Yes
Polarity Detection	Somewhat	Yes
Classification	Somewhat	Yes
Pattern Matching	No	Yes
Fake Tweets	No	Yes
Graphical Analysis	No	Yes
User Alerts	No	Yes

Table 2: Comparative Results

With reference to table 2, it is clear that we overcome various problems in existing system and our approach works efficiently.

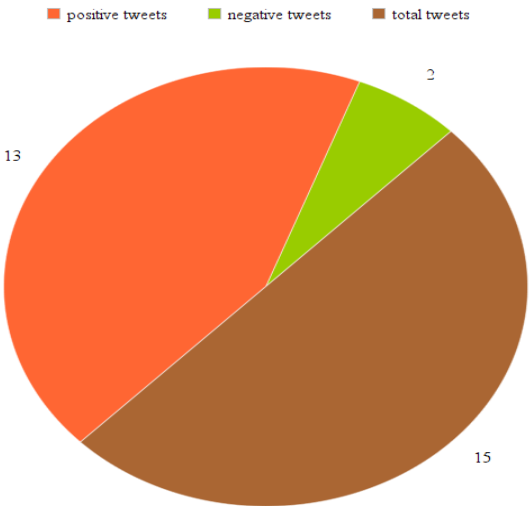

Fig.2: Pie Chart for Polarity of Tweets

As shown in Fig.2. The system will detect the polarity of tweets fetched from API. Polarity checking helps classification of sentiments. Also following Fig.3 shows the sentiment classification results in various classes,

Retrieval Number: H7397068819/19©BEIESP
DOI: 10.35940/ijitee.H7397.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

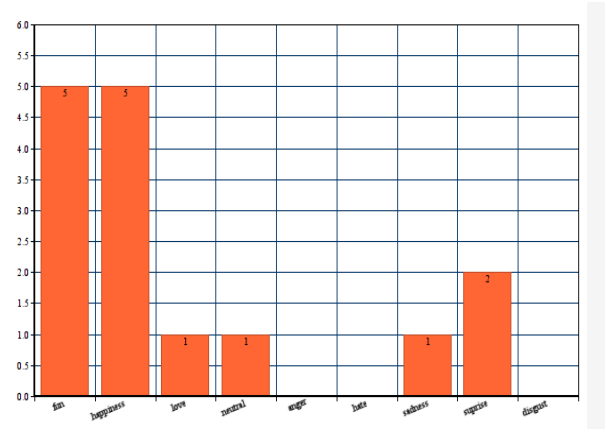

Fig.3: Sentiment Classification of tweets

Finally our system checks how many tweets along of fetched tweets were genuine and results achieved are shown in Fig.4.

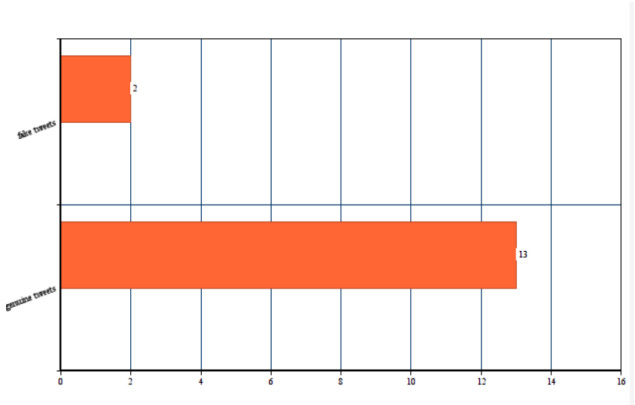

Fig.4: Tweet Genuienity

V. CONCLUSION

The proposed system set out to classify the tweets in different classification and genuinely check of Twitter posts. This system 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 was uninfluenced rumor 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 as well as querying and fetching of particular tweets from twitter are possible. Finding influence or negativity spread by users can be useful in various analytical tasks. From results obtained we conclude that proposed system is capable of classifying and recognizing sentiments of people about particular social issue also it is clear that fake or tweets are predicted and it can be useful for even government agencies to prevent social disaster. We obtained accuracies which varies between 90-95% and found to be satisfactory.

REFERENCES

1. MondherBouazizi and TomoakiOhtsuki, “A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter”, pp. 2169-3536 on August 2017
2. Beakcheol Jang and Jungwon Yoon “Characteristics Analysis of Data from News and Social Network Services” , pp. 2169-3536 on March 2018
3. Hai Tan & Jun Zhang, “Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness”, pp. 2169-3536onApril 2018
4. MondherBouazizi and TomoakiOhtsuki, “A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter”, 2169-3536 on August 2017
5. Aldo Hernández, Victor Sanchez “Security Attack Prediction Based on User Sentiment Analysis of Twitter Data”, May 2016
6. ManjuVenugopalan and Deepa Gupta “Exploring Sentiment Analysis on Twitter Data”, on Aug. 2015
7. Rincy Jose and Varghese S Chooralil “Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation”, on Nov. 2015
8. Anurag P. Jain and Mr. Vijay D. Katkar “Sentiments Analysis ofTwitter Data Using Data Mining”, pp.978-1-4673-7758-4 on Dec. 2015
9. Gaurav D Rajurkar and Rajeshwari M Goudar “A speedy data uploading approach for Twitter Trend and Sentiment Analysis using HADOOP” , pp. 978-1-4799-6892-3 on Feb. 2015
10. Ahmed TalalSuliman, Khaled Al Kaabi, “Event Identification and Assertion from Social Media Using Auto-Extendable Knowledge Base”, pp. 2161-4407 on July 2016
11. MohdFazil and Muhammad Abulaish, “A Hybrid Approach for Detecting Automated Spammers in Twitter”, on Nov. 2018.
12. Yan Zhang and Weiling Chen, “Detecting Rumors on Online Social Networks Using Multi-layer Autoencoder”, on June 2017.

AUTHORS PROFILE

Ms. Deepti Kulkarni studying in ME(Computer Engineering) in K. J. College of Engineering, Pune, India.

Mr. Nagaraju Bogiri ME(Computer Engineering) is professor in K. J. College of Engineering, Pune, India