International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Sentimental Analysis and Detection of Rumours for Social Media Data using Logistic Regression

Asha R, Rahul Jain, Gourav Das, Pranjnay Bharadwaj

Abstract: ***Over the last decade ,the Internet has become an ubiquitous and enormous suffuse medium of the user generated content and self-opinionated knowledge. Users currently have the facility to specify their views, opinions and ideas publically. Victimizing social media platform is a place where people can express their mindsets and feelings in a well associated manner and hence is productive and economical . These ever-growing subjective knowledge are doubtless, an especially made for supply of data of any reasonably method process. The Sentiment Analysis aims at distinctive self-opinionated knowledge during an Internet and classifying them in line with their polarity whether or not they contain positive ,negative or neutralizing references. Sentiment Analysis could be a drawback of text based mostly analysis however there are difficulties which are needed to be pondered upon that would create a tough parameter as compared to ancient text based analysis. It depicts the state where it has a desire of trial to figure out these issues and it's spread out many chances for further analysis for handling negative sentences, hidden emotions , slangs and sentence sarcasm. The project also proposes additional features compared to other previous model projects by enabling the detection of rumor , identifying and analyzing whether message given via user belongs to rumor category or not using Logistic Regression process in Machine Learning domain.***

KEYWORDS—Machine Learning , Sentiment Analysis , Rumor Detection, Slang , Social Media Data

I. INTRODUCTION

Sentimental Analysis describe process of computational identification and categorizing the data into thoughts, ideas or opinions expressed in medium of data written or Sentimental Analysis describe process of computational identification and categorizing the data into thoughts, ideas or opinions expressed in medium of data written or text to obtain the mindset track of public opinions in terms of positive, negative or neutral. However, identifying and analyzing thousands of data over social media can all go wrong especially words expressing emotions, sarcasm and contradiction. This system enable to classify basic tweets whether that is positive or negative, complex or ironic statement classification is not possible using the existing system .

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Ms .Asha R*, Assistant Professor in Computer science Department in SRM Institute of Science and Technology, Chennai
Rahul Jain, currently pursuing B tech degree in Computer Science and Engineering at SRM Institute of Science and Technology
Gourav Das, currently pursuing B tech degree in Computer Science and Engineering at SRM Institute of Science and Technology
Pranjnay Bharadwaj, currently pursuing B tech degree in Computer Science and Engineering at SRM Institute of Science and Technology

It is the basis of which further steps to classify the system and modification could be done. Large data storage is possible using database management easily.
Classification and distribution can be enhanced to a larger extent using techniques like Naïve Bayes, Logistic Regression. Disadvantages of this Existing System are numerous like Sentence Sarcasm means the sentence which contains contradicting statements or opinions is difficult to depict by this system .So new method is needed to enable identification. Statement Contradiction refers to different approaches taken by user to describe the emotions explicitly.

- Expressing Emotions is unavailable here due to understanding the form of writing text which is either in support or hatred on a particular subject/topic .
- Emoji Identification is not possible by existing method because of its usage can be in sarcastic manner. This cannot be predicted by classical method.
- Statement Vulnerability is another issue to be pondered upon and should be identified properly.

Various algorithms are developed which is used to analyze the data ,with the goal to extract information .The objective of the paper are to give detailed explanation about the Logistic Regression algorithm .Verification of data online is reviewed and analyzed using Logistic Regression.
To identify sentence sarcasm ,emojis and emotions of user and compare the data to gain a bigger picture of the issue.

II. DEFINITIONS

A .SENTIMENTAL APPROACH AND ANALYSIS

- Enable identification of sentences as positive and negative opinions ,emotions which enable checking
- Computation and Analysis of the data or text study of opinions ,sentimental and emotions expressed .
- **Opinions**: Opinions refers to conclusion open to dispute(each user have different ideas and thinking ability on a particular opinion.)
- **View**: Refers to a subjective opinion of a user over certain issues.

B. SENTIMENT

- Corroborative measure to be taken for hidden justification or certainty .
- **Belief**: Deliberate infallible acceptance and intellectual assent of credulous belief.

Pre Processing of Datasets
Preprocessing is the initial step or founding steps for analyzing the tweets /messages online of the user. Datasets generated are to be surveyed is collected and particular analyzing of each set of data is done .

The following steps are necessary to ensure the preprocessing. These are as follows:

- Attenuate URL
- Remove hash tags ,targets symbols

TABLE 1: Radom data taken from various sources over online platform.

A. HASH	*Tweets*	*http://demeter.inf.ed.ac.uk*	*31,681 pos tweets,62,567neg tweets,128,859 neutral tweets*	*Total 223,107 Tweets taken*
B. EMOT	*Tweets and Emotions*	http://twittersentimentappspot.com	*230,911 pos tweets,150,070neg tweets*	*Total* 380,981 tweets taken
C. ISIEVE	*Tweets*	*www.i-sieve.com*	*1520 Pos tweets,200Negative tweets,2,295* Neutral tweets	*Total 4015 tweets taken*
D. Sample	*Tweets*	*http://goo.gl/UQvdx*	*667 Tweets*	*Total 667 tweets taken*
E. Stanford Data set	*Movie Review*	*http://ai.stanford.edu/~amaas/data/sentiment*	*50000 movie reviews*	*Total 5000 reviews taken*
E. Spam Data set	*Spam Rerviews*	*http://myleott.com/op_spam*	*400 deception and 400 truthful reviews in positive and negative reviews*	*Total 1200 reviews taken*
F. Soe data set	*Sarcasm and Nasty Review*	*http://nlds.soe.uc.sc.edu/iac*	*1000 discussions,~39,000* Pots ,and some~73,00,000 words	*Total 1000 discussion*

- Identify data and classify emotions .
- Ignore and change irrevocable punctuations,symbols numbers.
- Elucidating the short forms and understanding their meaning.
- Remove and ignore data present in the different languages(consider a particular language for analyzing)

Various subject or topics of the reviews taken are for example movie reviews, sarcastic and nasty reviews and also spam reviews .Total data is further classified into positive ,negative or neutral statements and thus ,total count is recorded. The table below clearly justifies the standard of particular site reference and domain of survey conducted from various sources like tweets ,movie review, sarcastic reviews .For example, hash tweets are considered and total tweets used of survey is decided .The classification of those hash tweets is done in terms of positive ,negative or neutral .The public opinion is recorded and analyzed further upon.

Abbreviations and Acronyms:

Neg refers to Negative
Pos refers to Positive
Neu refers to Neutral

ADVANTAGES

- Classification of Sentences is enabled via this method.
- Identification of positive ,negative and neutral statements is done regardless of the contextual text.
- Analyzing the trends and giving the real report of survey over a particular domain.
- Determination of speakers attitude and opinions is possible.

III. PROPOSED SYSTEM

Sentiment are the words or sentence that represent opinion in a positive ,negative or neutral way .Here the proposal made is a new hybrid approach involving both corpus based and dictionary based analysis. The method or technique we are using is Logistic Regression using Machine Learning. We consider the emotions ,sarcasm subjective criticism of the user and classify them accordingly .We also consider to detect the rumors involved in the process.The project later involves detection of rumors to identify whether the message is a rumor or not by using mathematical formulae and comparison.

Two main part of division involves:

Data Extraction and Pre Processing of Extracted Data.

Other important features involved are:

- Retrieval of Tweets
- Pre Processing of Extracted Data
- Parallel Processing
- Removal of Stop words
- Scoring Sentiments
- Output Showcasing

IV. OBJECTIVE

The objective of project is to perform Sentimental Analysis and detection of rumors for social media data using Logistic Regression.Classification of sentences and detection of sentence sarcasm is possible in the proposed system using the data as input .The objective is also to enable to analyze the trends and identification of maximum flow of public opinion over the specific issues.

Enable to identify the speakers attitude or sentiments and classify them as positive, negative or neutral statements and detection of rumor is analyzed to check the availability of rumor in a particular text using mathematical formulas.

V. APPROACHES FOR SENTIMENT ANALYSIS

The approaches related to sentiment process require couple of techniques for sentiment analysis for data available online:

A.MACHINE LEARNING TECHNIQUES:

Machine learning is forced upon assertion upon approaches classification technique in order to classify text to classes.

There are typically two kinds of ML techniques :

1. UNSUPERVISED LEARNING:

- List or data which is non categorical and could not be ubiquitous like with exact targets at all and therefore depends on various processes.

2 .SUPERVISED LEARNING :

- Depending upon the data provided to the main source or center during the process. These are referred as sets which are trained to get meaningful outputs when encountered during decision making process.

VI . Equations

The performance can be classified and can be defined or described using four parameter furthermore could be calculated by the following below formulae :

Measurement of Infallible:

(P)= (a+b) /(a+b+c+d)

P = a / (a+c)

R= a / (a+d)

P1 = (2×P×R) / (P +R)

In which

- a, total cases of true positive sentences
- d, total cases of false negative sentences
- c , total cases of false positive sentences
- b ,total cases of true negative sentences
- P1 ,performance instances
- P , precision calculation
- R , Recall Instruction

B.)Lexicon-Based Approaches

Lexicon Based Approaches involves the fundamental process method requiring the sentimental dictionary along with opinionative words which is then matched with the data to determine polarity of data as positive, negative or neutral. The sentiment scores (0-10) is given to the data containing similar ideas and the particular opinions . Lexicon-based enables relation about classification upon sentiment lexicon which is dossier and documents connected contains sarcasm, phrases and are collected for especially the bucolic genre .

Table 2:

Machine Learning	Method	Data Set	Acc	Author
	SVM	Movie Review	86.40%	Pang Lee
	Co Training SVM	Twitter	82.50%	Liu
	Deep Learning	Stanford Sentiment Treebank	80.70%	Richard

The table above depicts the performance based on Sentimental Analysis on Machine Learning domain.

The methods in machine learning domain used are Support Vector Machine or SVM ,Co Training SVM and Deep Learning describing about movie reviews ,social media data and their accuracy levels .These above stats are given by different authors while working on the same domain.

VI. LOGISTIC REGRESSION

A technique used to describe and predict different analytical approaches via graphs .To explain the relations among various parameters and describe about the behavioral and characteristic change of concerned data .It helps in providing bigger approach of the survey and enable modification and improvement in the same.

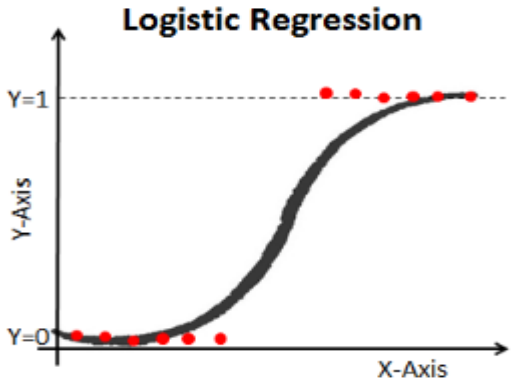


Fig 1: Describing Logistical Regression Graph and Analyzing the Trends

VII. DETECTION OF RUMOUR AND CORROBORATIVE SYSTEM

The ratio of a particular system or text is compared with threshold parameter to deceit whether it is a rumor or not which is feigning in nature or not and corroboration is done using below formulae.

The threshold of a post is:

- rumor Threshold >=0.1
- non rumor Threshold <0.1

Viola due to this parameter ,we can judge the detection or presence of rumor and identify the apocryphal nature of a text and even explain whether text contain stupefaction or not .

Rumor Substantiation System

We extracted from various social media platforms and analyzed the trend that predominantly the two prerequisites classification of the rumors are as described; dependableness, involving parameters like “true” and “false”, however other contains ambivalent or equivocal mixture also containing uncanny, incongruous and enigmatic sentences . A rumor was taper off in true or false as justified suffuse or douse medium like snopes.com confirmed it per se. It contained minimum of around five hundred posts having quite five comments.. The rumor verification system could be a methodology wherever upon many inputs over a specific issue is taken and a survey to search out the general public poll and their take over that issue is recorded and therefore calculated .

Since, folks of each generation is actively victimized of social media so these survey will be worthy in analysis purpose and showing real photos of all possible domain in life and even regarding the future events.

Common Used Abbreviations and Their Expansion

Table 3:

ABBREVIATIONS	EXPANSION	ABBREVIATIONS	EXPANSION
Btw	By the way	Ngt	Night
U	You	Neg	Negative
Ri8	Right	Admin	Administration
Rn	Right now	Fyi	For your information
M	Am	Pos	Positive
Tbh	To be honest	Ttyl	Talk to you later
Congo	Congratulations	Luv	Love

MODULE

Sentiment Analysis module works as follows:

- Preprocesses texts of comments, reviews, posts, or tweets received from social networks; removes stop words; extracts features based on vector representation of words using Senti Word Net;
- Returns probability of Positive/Negative sentiment associated with the event or notion under analysis;
- If required, provides a full package of quality metrics, such us accuracy, f1 score, ROC-AUC, etc.
- The module was trained on millions reviews from different sites, and reached high quality of analysis.

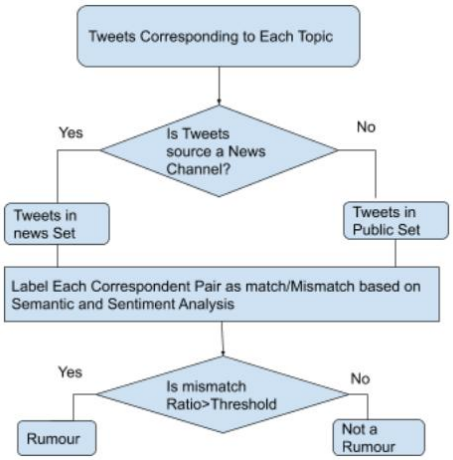


Fig 2: Flow Diagram representing Online Social media data(Twitter) describing Sentimental Analysis trends and Detection of Rumor

VIII. CONCLUSION

The Sentiment Analysis aims at distinctive self-opinionated knowledge during an Internet and classifying them in line with their polarity whether or not they contain positive ,negative or neutralizing references. It depicts the state where it has a desire of trial to figure out these issues and it's spread out many chances for further analysis for handling negative sentences, hidden emotions , slangs and sentence sarcasm. The project also proposes additional features compared to other previous model projects by enabling the detection of rumor , identifying and analyzing whether message given via user belongs to rumor category or not using Logistic Regression process in Machine Learning domain.

The detection of rumors are now done successfully and we are able to classify them in further sub division categories whether they are positive, negative or neutral statements.

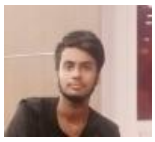
We can also identify the slangs and abbreviations used in a particular text and understand its meaning easily and efficiently.

REFERENCES

1. Rehana Moin ,Zahoor- ur- Rehman, Khalid Mahmood,"Framework for Rumors Detection in Social Data",[2018]
2. Walaa Mehdat, Ahmed Hasan ,Hoda Korashy "Sentimental Analysis algorithms and applications",[2014] .
3. Oscar Araque,Ignacio Corcuera-Platas, "Enhancing Deep Learning Sentimental Analysis with ensemble techniques in social applications" ,[2012].
4. Z, "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts," [2015].
5. Liu B. Sentiment Analysis and Opinion mining. Synth Lect Human Lang Technology [2012].
6. Tao Chen, Improving Sentimental Analysis via Sentence Classification".[2017]

AUTHORS PROFILE

Ms .Asha R is an Assistant Professor in Computer science Department in SRM Institute of Science and Technology ,Chennai .She has completed her Masters degree and is helping young students to grow in their respective fields .She has already published journal on International platforms .Her interest research lies in Web services ,Cloud Services and Software Engineering. Email : ashasekar02@gmail.com



Rahul Jain is currently pursuing B tech degree in Computer Science and Engineering at SRM Institute of Science and Technology and will be graduated in 2021.He has strong interest in Data Analytics and is learning Machine Learning and Artificial Intelligence.
He is a great server of society and is officially a member of Lions International Club and Bhumi NGO. His main ambition is to complete his Masters Degree and Ph D. degree by getting specialization in Machine Learning and Artificial Intelligence domain .He wants to combine the knowledge of machine learning and data analytics field to solve real life problems and looks forward to research.
E-mail : jain40054@gmail.com

Gourav Das is currently pursuing B tech degree in Computer Science and Engineering at SRM Institute of Science and Technology and will be graduated in 2021.His interest lies in project development and strong domain includes Machine learning and Cloud Services. His aim is to solve real life problem and contribute to help others. Email : gouravdasrahul@gmail.com

Pranjnay Bharadwaj is currently pursuing B tech degree in Computer Science and Engineering at SRM Institute of Science and Technology and will be graduated in 2021.His interest lies in Data Science and Machine Learning domains .His long term aim is to create a start up based upon web services and to help in solving real life problem. Email: pranjnaybharadwaj@gmail.com