ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

Review on Twitter Data Analysis using R Language

Virendra Yadav[1], Shivani Nagarikar[2], Srushti Chahande[3], Chetna Sahu[4], Ankita Wandre[5], Harshini Urmale[6]

Assistant Professor, Department of Computer Science and Engineering,

Priyadarshini Institute of Engineering and Technology, Nagpur, India[1]

Dept of Computer Science & Engineering, Priyadarshini Institute of Engineering & Technology, Nagpur, India[2,3,4,5,6]

Abstract: Rich source of information is present in social media. Social media provides meaningful information of gleaning emotions of people and understanding public attitude and mood more deeply. Internet provides a huge platform for exchanging ideas, online learning, sharing ideas on social networking sites such as Google plus, Instagram, Twitter, Facebook. There is huge volume of data in the web with the advancement of web technologies and its growth. Social networking sites give exposure, allow people to express and share their view about the topic discussion with various communities and can post message across the world. Lot of work is been progressed in the field of sentimental analysis of twitter data. The emphasis of this survey is to analyse the in the tweets where opinions are highly unstructured, heterogeneous opinions are either positive, negative or neutral. In this paper we present a survey and comparative analysis of existing techniques for opinion mining. there are two main parts in framework initially, ensemble classifier schema, combines knowledge based, generic domain independent with machine learning methods, used to perceive contents of emotions in user generated data. To model the emotional level of the topic based on emotional recognition, a graph-based method is used to create the topic emotional graph visualizing public moods and emotion on topic. Encouraging results are observed.

Keywords: Sentiment analysis, Twitter, Predictive Analysis, Affinity analysis, Pre-processing, R Language, Social Network

I. INTRODUCTION

The way of communication has drastically changed due to social media and provides people new means to connect in real time globally with the information, news and events. A significant change in the role of users has changed from simple passive information seekers to active producers. We get an interactive media through online communities where consumers influence and inform other through forums. The large amount data of sentimental data is in the form of tweets, status updates, blogs, comments, reviews, posts, etc. Opportunity for business is provided through social medias as a platform which connects them directly to the customers. Decision making of people are mostly dependent upon the user generated contents over online data over a great extent. The occurrences of social media and web 2.0 technologies are eagerness can be seen among the people to express their opinions on web regarding to plethora of aspects and express their attitude on entities, person, activities, events, and products. Every day, social media generates a vast amount of heterogeneous bid social data which necessitates automated methods to extract and analyse knowledge. From the user generated data on social networking sites a important piece of information could be extracted which is an underlying emotional content on social sites and emotions of people is very important procedure in micro blogging. Since, emotions can give very indicative aspects of his/her status, personality of person, behaviour, thoughts, thinking process. The recognition of emotions for big social data can enhance to understand the public mood and attitude towards various events, understanding of people's status and also provides various clues to determine his/her personality.

Emotional modules are employed to specify how people feel or think about the given entity such as events, topic or other. Mining these voluminous data of reviews and opinion can provide indicative information for understanding collective human nature and can be extremely beneficial to many domains like political stance, market campaigns, product review, company feedbacks and many other. Example: If a person wants to purchase a product or intents to use any service then they scroll through the reviews online, discussion about the product on the social media before coming a decision. It is difficult to analyse the content generated by users and their approaches accordingly. So, there we need am automation various sentiment analysis techniques and us.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

II. LITERATURE SURVEY

A. A framework for analyzing big social data and modelling emotions in social media (2018). In this work, they presented a generic framework for the analysis of big social data and the modelling of public emotions and mood. Mining these big volumes of opinions can provide indicative information for understanding collective human behavior and can be extremely valuable to many domains. The framework consists of two main parts. The first part concerns the analysis of the textual social data and the recognition of their emotional content. For this purpose, a generic, domain-independent knowledge-based method is illustrated and the development of a corresponding to that performs the analysis of the text and determines its emotional contents is presented. Initially an ensemble classifier schema, which combines a generic domain independent, knowledge-based tool with machine learning methods, is used to recognize emotional content in user generated social data. The goal of ensemble classifier schema is to efficiently leverage the advantages of base classifier to increase efficiency and accuracy of the emotion recognition process. Then the second part that concerns a graph-based method is utilized to model the emotional level of a topic based on the emotions recognized and then it creates the topic's emotional graph visualizing public emotions and mood on the topic. The results from the case study are quite encouraging.

B. Some of the earlier and recent results on sentiment analysis of Twitter data analysis and visualizations using the R language on top of the Hadoop platform (2017) and Visualization of big data analysis on social media (2017) here the focus was to leverage existing big data processing frameworks with its storage and computational capabilities to support the analytical functions implemented in R language, they developed a platform that would consider peoples activities in social media, and display results. This was done by grouping together like-minded people in a series of interactive data visualizations that will allow trends to be found. The work showed how the sentiment and mindset of people varies with location i.e. which areas in a country and in the entire world have people with many thoughts.

C. Visualization of Big Data Analysis on Social Media required to collect data from the desired source (here twitter). This data is made more machine sensible than its previous form due to various steps of pre-processing. In this, the home page requests for the topic of interest and then the related tweets is fetched and this is put into any of the user's preferred visualization tools. Through Coffee and JavaScript languages API was accessed and it is linked to MongoDB. The tweets were stored in MongoDB's cache till it is being processed and then we are able to remove those tweets and just store the Sentiments and Locations. The client-side processor takes in the raw tweets as an input and outputs and this information is then sent to the server and a copy is stored in DB. The Visualization is the final stage of that work which helped the system to understand and the sentiments better. All visualizations commonly display the sentiment score with regards to the topic. Some visualizations add more data such as comparisons or change in Sentiment with respect to timeline.

III. METHODOLOGY

A. Extract tweets from twitter: Extraction of tweets from twitter about the particular area of interest. The collection of tweets is processed using mining tools or can be done using API for analysis of result in desired time period. We need to collect the data from the source (twitter) for performing sentimental analysis.

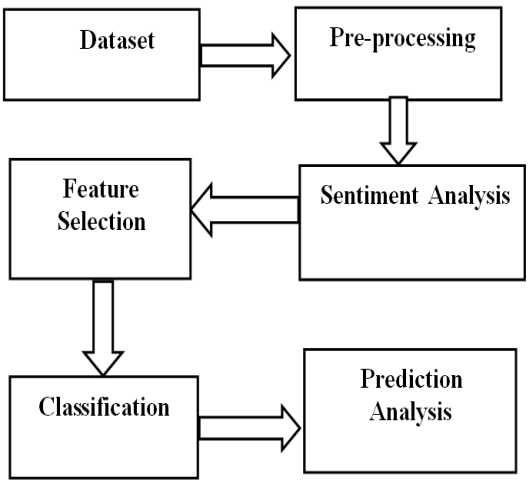

Figure 1. Methodology

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

Due to various steps of pre-processing, this data is made more machine sensible than its previous forms. In this, the home page requests for the topic of interest and then the related tweets are fetched and this is put into any of the user's preferred visualization tools. Through Coffee and JavaScript languages API was accessed and it is linked to MongoDB. The tweets were stored in MongoDB's cache till it is being processed and then we are able to remove those tweets and just store the Sentiments and Locations. The client-side processor takes in the raw tweets as an input and outputs and this information is then sent to the server and a copy is stored in DB. The Visualization is the final stage of that work which helped the system to understand and the sentiments better. All visualizations commonly display the sentiment score with regards to the topic. Some visualizations add more data such as comparisons or change in Sentiment with respect to timeline.

B. Pre-Processing: The efficiency of other steps is decided on the basis of pre-processing. The aim should be involved for making syntactical corrections of desired tweets, the steps should be involved to make data more machine readable in order to reduce the ambiguity in feature extraction. Below are some steps that are used for pre-processing of tweets extracted from the source(twitter).

B.1 Removal of Re-Tweets

- **Converting upper case to lower case:** we are using case sensitive analysis that might take two occurrences of same words as different due to their sentence case.
- **Stop word removal:** Stops words that don't affect meaning of tweet are removed.
- **Twitter feature removal:** The presence of usernames and URL's is futile because it is not important from the perspective of future processing. All usernames and URL's are converted into generic tags or removed.
- **Stemming:** Reducing different types of word with similar meanings helps in reducing the dimensionality of the feature set.
- **Special character and digit removal:** Sometimes the digits and special characters are fixed with words that don't convey any sentiment, hence their removal can help in associating two words that were otherwise considered different.
- **Creating a dictionary to remove unwanted words and punctuation marks forms the text**
- **Expansion of slangs and abbreviations**
- **Spelling correction**
- **Generating a dictionary for words that are important or for emotions**
- **Part of speech (POS) tagging:** POS taggers are very efficient for explicit feature extraction as it assigns tag to each word in text and classify a word to a specific category like noun, verb, adjective, etc.

The transaction file is prepared that contains opinion indicators namely verb, adjective and adverbs. Some of the emotion identifiers also have to be identified such as the repeated sequence length, tweets percentage in caps and the no of exclamation marks. Thus, pre-processing of all the tweets is processed in the following ways:

Remove all URL's (e.g. www.webconfs.com),hash tags (e.g. #heading), targets(@unsaid), special Twitter words ("e.g. RT").

- Calculate the percentage of the tweets in Caps.
- Correct spellings – the sequence of characters that are repeated is tagged by a weight. This is to be done for differentiation between the regular usage and emphasized usage of a word.
- After counting the number of exclamations marks all the punctuations are removed.
- By using a POS tagger, The NL Processor linguistic parser, we tag the adjectives, verbs and adverbs.

C. Feature Selection: Feature is a piece of information that can be used as characteristic which can assist in solving a problem. The quality and quantity of features is very important as they are important for the result generated by the selected model. It is an important part of machine learning where it refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. Selection of useful words of tweets is feature extraction.

- Unigram feature: At a time only one word is considered and decided whether that word is capable of being a feature.
- N-gram feature: At a time, we can consider more than one word for deciding whether that word is capable of being a feature.
- External lexicon: The list of words is used to predefine negative and positive sentiments. Frequency analysis collect features with higher frequencies, further some of them are removed due to presence of similar sentiment words and group of these words are created. The affinity analysis is performed along with this which focuses on higher order n-grams in tweet feature representation. The pre-processed Dataset has many distinctive properties which has key

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

features that are considered as feature vectors for the classification task. Some features examples that have been reported in literature are:

1. Words and their frequencies: unigrams, bigrams and n-gram models are considered as features with frequency counts.
2. Parts of speech Tags: Parts of speech like adverbs, adjective and some group of verbs and nouns are good indicators of sentiments and subjectivity. Syntactic dependency is generated by dependency tree and parsing tree.
3. Opinion Words and Phrases: Apart from specific words, there are some phases and idioms which convey sentiments can also be used as features.
4. Position of terms: The position of text can signify on how much the term makes difference in overall sentiment text.
5. Negation: Negation is an important part but at same time difficult to interpret. Due to presence of negation the polarity of opinion also changes.
6. Syntax: Many researchers use syntactics pattern like collocations as feature to learn subjectivity patterns.

D. Classification: Classification in data mining is technique that assigns categories in order to aid in more accurate results and analysis to the collection of data. It predicts continuous valued functions and categorical class labels. For example, we can build the classification model to categorize either the bank loan application is risky or safe, or to predict the expenditure in dollars of potential customers on computer equipment given on their occupation and income using predictive model.

E. Predictive Analysis: It is an area of statistics that deals with extraction information from data and using it to predict trends and behaviour patterns Often the unknown event of the interest in the future ,but predictive analytics can be applied to many type of unknown whether it be in the past, present or future .the core of predicted analytics relies on capturing relationships between explanatory variables and the predicted variables from the past occurrence's, and exploiting them to predict the unknown outcome. The accuracy and usability of results will depend greatly on level of data analysis and the quality assumptions. It is also stated as predicting at a more detailed level of granularity i.e. generating predictive scores for each individual organizational element.

Application

Commerce: the companies can make use of research for collecting public opinion and perspective related to their products and brands. From the company, the survey to target customers is important for making out the ratings of their products. Hence Twitter can serve a good platform for collection of data and analysis of the particular data to determine customer satisfaction.

Politics: The major of tweets on Twitter are related to politics. Due to widespread use of twitter, many politicians are also aiming to connect to people though social media so that the people can know them and could recognise them during the political meets or elections. People post their disagreement or support towards government's policies, actions of elections, debates, rules and regulations etc. Hence analysing data from the social networking sites can help is in determining public views and opinion based on any on the particular topic.

Sports Events: Sports involve many events, championships, meets, gathering, cultural, controversies and events too. Many people follow their favourite sport players. These people tweet frequently about the performances performed by the sports player which helps them to analyse their performance during the game. We can use the data to gather public view of team performance, play action and official decisions etc.

IV. FUTURE SCOPE

Now a day's people empress their emotion and sentiments on social media in the form texts as well as emojis. Emojis are a small digital image processing or icon used to express an idea or idea in electronic communication. Many express their ideas, thoughts, emotions in the form of emojis on social media

CONCLUSION

In this work, we presented a framework for analysing social big data and modelling public emotions and mood. It was developed using R language and utilizes the big data processing technologies. Thus the essential knowledge for doing data and sentiment analysis of Twitter is been stated in this review paper. If we want to do sentiment analysis we need to know about twitter, the structure of extracted tweets and their meaning. So this paper will give the brief approach of tweets and the process which is to be followed for form extraction to analysis to be performed on real time twitter data and to analysis on the emotions and mood present of the real time twitter data.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

REFERENCES

[1]. Srinath Vijayaragavan. "Visualization of Big Data Analysis on social media", 2017 International Conference on Energy, Communication. Data analytics and Soft Computing (ICECDS).
[2]. Isidoros Perikos, Ioannis Hatzilygeroudis, "A framework for Analysing Big Social Data and Modelling Emotions in Social Media ", 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications.
[3]. Martin Sarnovsky, Peter Butka, Andrea Huzvarova , "Twitter Data Analysis and Visualisation Using the R Uanguage On Top of the Hadoop Platform",2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics.
[4]. Bravo-Marquez, F.,Mendoze, M., and Poblete, B.(2014)." Meta Level Sentiment Models for Big Social Data Analysis for Knowledge Based System", 69,86.-99.
[5]. Cambriya, E.,Living Stone, A.,&,A.(2012)," The R glass of emotions, cogitative behavioural system, 144-157.
[6]. Dietterich,"Assemble Methods in Machine Learning", proceedings of the first international workshop on multiple classifier system, MCS '00,Springer -Verlag, London, UK, 2000, PP.11-15
[7]. Stephan Gouws, Donald Metzler, Congxing Cai, Eduard Hovy, "Contextual Bearing on Linguistic Variation in Social Media", workshop on language in social media (LSM 2011), pp. 20-29, 23 june 2011
[8]. Abbott Rob, Walker Marilyn, Anand Pranav, E. Fox Tree Jean, Bowmani Robeson, Joseph King, "How can you say such things?!?: Recognizing Disagreement in Informal Political Argument", workshop on language in social media (LSM 2011), pp. 2-11, 23 june 2011.
[9]. J. P. Dijcks, "Big Data for the enterprise", Oracle White Paper, 2013.
[10]. J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", Proceedings of OSDI' 04: Sixth Symposium on Operating System Design and Implementation, pp. 107-113, 2004.
[11]. J. Ishwarappa-Anduranha, "A Brief Introduction on Big Data 5V Characteristics and Hadoop Technology", Procedia Computer Science, pp. 319-324, 2015