ISSN No:-2456-2165

Review on 'Big Data - Sentiment Analysis'

Yash R Karanje
Department of Computer Engineering
NBN Sinhgad School of Engineering, Pune, India

Abstract – We all are very well aware with today's world that at what extent it has been socialised and it doesn't need to be mentioned explicitly. People express their thoughts on various digital platforms and there comes the opportunity to know the opinion of society upon any concerned issue. Here, we'll be using Twitter as a social media platform and one's individual tweet as a characteristic property will act as the examinee since tweet is the sharp and shortest way to describe ourselves and so it is adopted by society. In this review paper, firstly it focuses on big data concepts and its handling techniques. Then it focuses on tweets mining from Twitter and their sentiment extraction. Sentiment Analysis is a construct of Data Mining & NLP. Also, it shows working of NLP libraries and how it uses Naïve Bayes Algorithm. In short, this paper is review about existing technologies to handle big data and so its sentiment analysis with their pros & cons.

Keywords:- *Big data, Sentiment Analysis, Opinion Mining, Data Mining, Natural Language Processing, Naïve Bayes Algorithm, Twitter, Social Media.*

I. INTRODUCTION

Social media has taken a drastic boom in last few years and it has become the need of time now. Social sites like Facebook, Twitter, etc. are the digital platforms where people can express their thoughts and could make things happen. Here, in this review paper, we'll be dealing with Twitter data. Tweet is a characteristic property and a short but descriptive way of representation for the user's thoughts. According to Twitter Usage Statistics [8]; every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. Just the number is enough explanation for amount of data being generated and this number is exponentially rising.

Most of the data available is in unstructured format and it is very difficult to operate such huge unstructured data. To deal with such scenario, use of Hadoop technology is appreciated. It's a framework which performs computations over large datasets. Its working is based upon concept of distributed system. Hadoop internally works with one special functionality called as Map Reduce. Again, Map Reduce comes with 3 sub-functionalities including Query, Map & Reduce. 'Query' does the filtering task. 'Map' functionality does the mapping of key with their values. Then comes the 'Reduce' operation in which any required function is applied over the set of values so that they combined to give output. Once we're done with the excursion of expected required data, next stage is Sentiment Analysis aka Opinion Mining. Here, it is supposed to get the state of opinion whether it is Positive or Negative or Neutral. Sentiment feature extraction is done by assigning polarity label to individual document unit so that complete tweet's polarity and subjectivity can be calculated. It shows the influence that surroundings have on user.

Opinions matter a lot in order to take any decision for the sake of any organisation or somewhere else. Before taking any crucial decision, taking opinions is must but analysing the result is also of the same importance. Twitter is just a platform and it's in our hand how do we make use of it. One for sure that we use it to express ourselves but it could be used for marketing and trading kind of purposes as well. For example, political parties can get review of their curriculum whether it is accepted by people or not and its extent as well. Same goes for any movie review to predict how much it could earn based upon people's reviews again. Same we can implement for some sort of service provided by any organisation. Currently, drawing out conclusion from opinions to take the fair decision is very important. This main moto is discussed in this review paper by doing Sentiment Analysis over Twitter data.

II. METHODOLOGY

A. Big data

Big data the word is enough to give idea about size of data but there's a lot related with it. Data is growing at tremendous amount of speed and that too in unstructured manner. Moreover, to obtain availability, Horizontal architecture is adopted. Because of this, the major issue of computing resources has been arose.

➢ *Defining Big data: [5]*

- *Variety*

Different classes of data are present including raw, semi structured, structured and unstructured data. Since different resources are available including sensor data, social media, web pages, web logs, email, documents and etc. so, it is obvious to have multi class data.

- *Velocity*

The term 'Velocity' deals with the amount of data being generated from different resources per unit time. In short, it tells the rate at which data flows.

- *Volume*

The name 'Big data' itself represents the Big size of data. Today, total amount of data generated per day is 2.5 exabytes [6] out of which social media fuels up with max contribution.

➢ *Big data challenges*

- Privacy & Security
- Data Sharing, Storage and Processing
- Control over data volume
- Does all data need to be stored & How to decide data to be analysed
- Algorithm efficiency over processing data & resources used
- Defining test cases

➢ *Tools and Techniques – Hadoop: [5]*

Hadoop is an open source project hosted by Apache Software Foundation. It is infrastructure for distributed system. This tool has mainly two techniques:

➢ *HDFS (Hadoop Distributed File System)*

The problem of data failure is handled by HDFS. As it is clear that we're dealing with tremendous amount of data; so, retrieval time from single source would be definitely large enough and also the problem of failure of storage devices can't even be neglected. In such cases, Horizontal Computing architecture is implemented. Replication or Redundancy is created by means of creating copies of same data at different devices so that in case of failure, it's another copy will be available that too in considerable time.

Working of HDFS consists of two nodes – namenode and datanode which are also known as master node and slave node respectively. The task of namenode is to manage file system namespace, metadata for directories while datanode stores and retrieves the blocks as per instruction. Here, namenode controls the data retrieval process from datanodes and gets the list of blocks as return value.

➢ *Map Reduce*

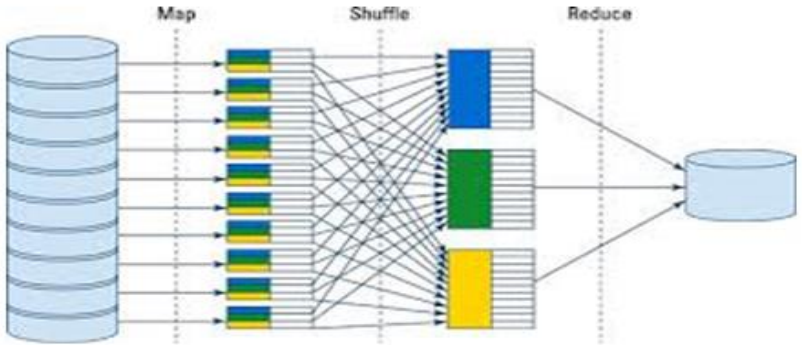

Fig 1:- Map Reduce Programming Paradigm [7]

The issue of combining data being read from multiple sources with different structure possibly in simultaneous manner is solved by Map Reduce technique.

Hadoop internally works with one special functionality called as Map Reduce. Again, Map Reduce comes with 3 sub-functionalities including Query, Map & Reduce. 'Query' does the filtering task. Only the required data which follows some specific phenomenon gets filtered out from bulk amount of data. 'Map' functionality does the mapping of key with their values. Key is nothing but an entity which is a representation of set of values following same criteria. Then comes the 'Reduce' operation in which specifically required function is applied over the set of values so that they combined to give output. This is one of the fastest ways for result provision.

B. Natural Language Processing (NLP):

Natural Language Processing is ability of a computer program which avails the machine to interpret and analyse human spoken language. This isn't that simple because understanding human slangs with regional dialects and social context is very complex. In our case, tweets are nothing but human spoken language. In order to draw out something from them, they need to be analysed and for that, at first they must be processed. This is where NLP comes in Sentiment Analysis.

To perform NLP tasks, different libraries are available in Python. NLTK can be considered as foremost library. TextBlob is another one built upon NLTK and best for beginners and easy to understand. Spacy is highly powerful and suitable for professional level.

ISSN No:-2456-2165

➢ *Sentiment Analysis*

The process of deriving whether the people's opinion bend towards positive side or negative side or neutral side against any particular topic/issue is known as the Sentiment Analysis.

Tokenization is the process of dividing whole sentence into words where every individual word will refer a token. This one is the most basic & prior job under NLP. [9]

Here is a basic example of tokenization done using TextBlob library-

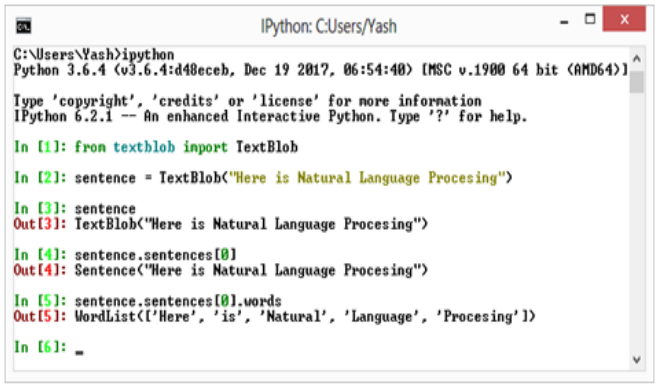

Fig 2:- Tokenization example

Sentence level Analysis [3] will be implemented here for individual tweet since every tweet is a set of sentences itself. The result expected is to get the final sentiment of that particular tweet. Polarity and Subjectivity are the two things expected here in which Polarity states whether opinion is positive or negative or neutral and subjectivity states the power of opinion i.e. how much intense your opinion is. Sentiment analysis is done by different lexicons [3]. For example, SentiWordNet which includes sentiment score (polarity) & sentiment strength (subjectivity) for tokens in its possible contexts. Libraries like TextBlob makes use of such lexicons to replace every individual token in tweet sentence by their respective polarity values so that the overall polarity of that document in its context can be calculated. This is where we get our actual result behind any tweet. Speech Recognition and Text to Speech are itself part of NLP.

- *Limitations*

✓ Use of slangs [3] – slangs are those words which quite deviate from native language and changes evolutionary.
✓ Making taunts – Remarks are made in a different way than what is actually supposed. Drastic change is seen in one's gist.
✓ Character limitation [3] – Limited number of characters can be processed at a time so sometimes it does effect on ideal result.
✓ User's own perspective of language use – Use of multiple languages that too in their own way (acronyms, short spells, etc.). Use of textual symbols & emoticons; rather emoticons are now recognized widely with their uniqueness.

➢ *Classification – Naïve Bayes Classifier:*

The Naïve Bayes Classifier is a supervised learning model which makes use of statistical method for classification. Since it's a probabilistic model, it allows to capture the uncertainties about the model by calculating probabilities [4]. The word 'naïve' means something which is simple, newbie and unaffected. So, this algorithm does the classification among classes which follow same kind of naïve features for whatever is the data set. Naïve Bayes algorithm works on Bayes theorem of conditional probability. Conditional probability is where happening of one event is conditional over another event. It gives the probability of an event based upon prior information of events that might be related to the current event. It is useful learning algorithm for observed data and past knowledge if existed. As this algorithm performs with independent features, works very fast & efficient for large data sets, includes noise and considers all possible cases; this is used for Twitter Sentiment Analysis and to classify tweets among all possible classes viz. Positive, Negative & Neutral. Its main use is in text classification and problems with multiple classes.

$$P(A).P(B|A) = P(B).P(A|B)$$

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

P(x|y) = Probability of event x provided event y is True/already happened.

opinion	to_do
Positive	yes
Positive	no
Neutral	yes
Positive	yes
Negative	yes
Positive	no
Neutral	yes
Positive	yes
Positive	no
Negative	no
Negative	yes
Positive	yes
Neutral	yes
Positive	yes
Neutral	no
Negative	no
Neutral	no
Negative	no
Neutral	no
Neutral	yes
Total	20

Frequency Count Table				
Opinion	yes	no	Grand Total	Probability
Positive	5	3	8	0.40
Neutral	4	3	7	0.35
Negative	2	3	5	0.25
Grand Total	11	9		
Probability	0.55	0.45		

Fig 3:- Sample example

Let us consider a problem statement in which we've set of scenarios and corresponding target variable 'to_do' which states about a particular task to be done or not, depending upon opinion. Task will be done in positive scenario. So, to perform that task, we'll consider first event as to_do - 'yes' provided opinion/scenario is positive so, second event as 'positive'. [10]

P(yes | positive) = (P(positive | yes) * P(yes)) / P(positive)

P(positive | yes) = 5/11 = 0.45
P(yes) = 11/20 = 0.55 P(positive) = 8/20 = 0.40
So, => ((0.45) * (0.55))/0.40
=> 0.62

Similarly,
P(yes | neutral) = (P(neutral | yes) * P(yes)) / P(neutral)
P(yes | negative) = (P(negative | yes) * P(yes)) / P(negative)

Here, we get probability with the help of Bayes theorem. Polarity is considered as a vector quantity. Probability obtained from Bayes theorem is treated as scalar value of the polarity and its vector property will be defined by kind of sentiment it shows which is either optimistic or pessimistic. So now, depending upon polarity, tweets get classified among different classes viz. positive, negative & neutral and this is how Naïve Bayes Classifier works.

C. Workflow Diagram

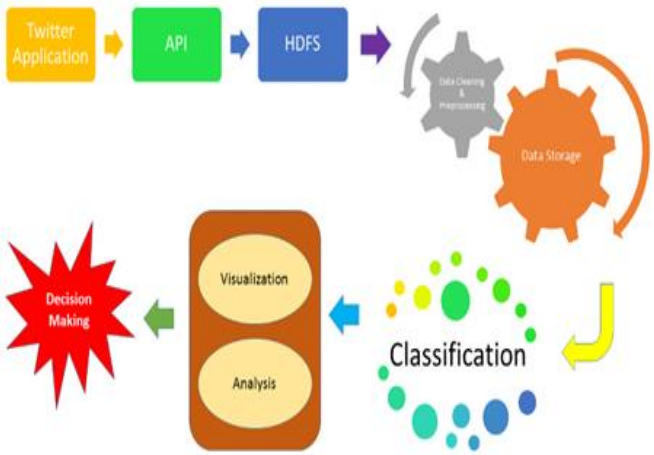

Fig. 4: Workflow diagrams

➢ *API*

API stands for Application Programming Interface. It is a bifrost between source & sink. An API is established with the help of certain authentication keys and tokens of twitter handle which does the communication in between twitter application as source & HDFS as sink. Once the API is set, then only tweets can be fetched. Here, we've used Tweepy Streaming API to access the source Twitter application. Tweepy Streaming API is based upon REST API. It makes use Request-Response communication model. Tweepy is faster than REST and can fetch more data than REST in single hit. So, we've used Tweepy here.

- How API works - API takes the required parameters from user and generates a URL. Then it will hit the browser as a request and will get the response.
- URL generation process – To get the output/response, there must be a request made to server application. Here, request is URL which takes parameters including credentials to access twitter application, #tag, no. of tweets to be fetched & some optional parameters like location, time frame, etc.

➢ *Data Cleansing & Pre-Processing*

Data Cleansing shortly means removing out unnecessary parts from data to get it exactly in required format so that sentiment generated will be accurate.

So, this includes Removal of [4]

- Hashtags – These are usually used to refer trending things. It makes no effect on sentiment.
- Handle name – Every twitter handle has a unique username. With the help of '@' symbol, it is possible to mention another user but it has nothing to do with sentiments.
- Emoticons & Emojis – They allow user to express their feelings and emotions in a better way of pictorial representation. So it is necessary to replace them with their corresponding values while processing.
- Special Symbols – These are usually punctuation marks which doesn't contain any sentiment.
- Unnecessary texts – Texts like URLs and unnecessary white spaces should be removed.
- Retweets – Template format showing Retweets should be removed.

Once the data gets cleaned, it must be tokenized so that sequence of tokens can be sent for analysis process.

➢ *Visualisation*

There is a need to conceptualize whole result with the help of its pictorial representation. It could be possible in many forms including charts, graphs or tabular one. Libraries like Matplotlib, Seaborn, Plotly in python allows to use already defined functions to draw out figures from existing data.

➢ *Decision Making*

This is the last but most crucial stage where decisions are taken in business perspective for any organisation. Once the analysis has done, expected result & predicted outcome can be compared and depending upon how much it deviates, next decisions can be taken. So this makes it a vital process.

III. EXPERIMENT & RESULT ANALYSIS

We're going to consider an experiment for movie review analysis. We'll need a hashtag for fetching tweets from twitter application. Expected workflow of this experiment is catching certain number of the fetched tweets and analysing them in order to find out overall people's response or review for that movie.

The code is written purely in python. TextBlob is used for textual data (tweets) processing. Pandas is used to have the classified tabular representation of tweets. Tweepy is used to establish twitter API. Matplotlib & Seaborn are used for visualisation purpose.

In experiment setup, a recent movie named 'Captain Marvel' is taken. We'll need a hashtag for fetching tweets from twitter application so, '#captainmarvel' is used. Here, just for the sake of convenience, we have taken lesser number of tweets but ideally, that number should be big enough. We've considered 100 latest tweets to be fetched present right before execution. [1]

```
The status of Positive Sentiment over #captainmarvel for 100 tweets is: Average 0.70 & Percentage 52.00%
The status of Negative Sentiment over #captainmarvel for 100 tweets is: Average -0.26 & Percentage 4.00%
The status of Neutral Sentiment over #captainmarvel for 100 tweets is: Average 0.00 & Percentage 44.00%
tot_tweets 100
pos_count 52
neg_count 4
neutral_count 44
pos_opinion 36.233910533910534
neg_opinion -1.0321428571428573
neutral_opinion 0.0
The final status of Sentiments over #captainmarvel for 100 tweets is:
Final Polarity:
POSITIVE with 0.3520176767676769
Final Subjectivity:
POSITIVE with 0.4348122294372294
```

Fig. 6: Execution Result

From the final result obtained; we can analyse that on the basis of 100 tweets, we got count of 52 positive, 4 negative & 44 neutral tweets. The final polarity value which we've obtained is 0.3520; it depicts positive opinion with strength of 0.4348 as subjectivity.

It's all dependent of number of tweets we're considering. More the tweets, more the precise result. So it becomes the crucial deciding factor.

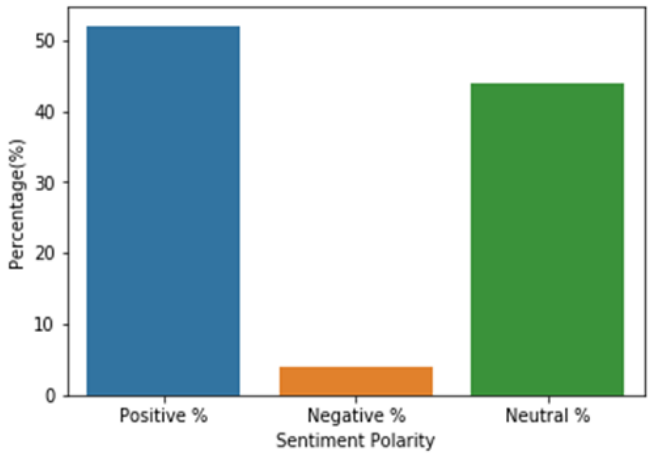

Fig. 7: Bar plot

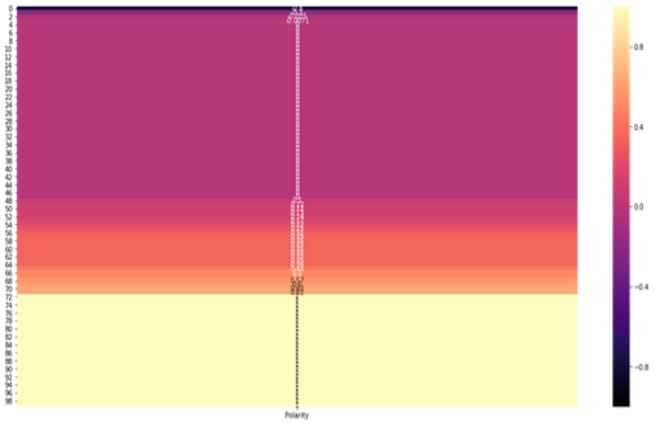

Fig. 8: Heat map

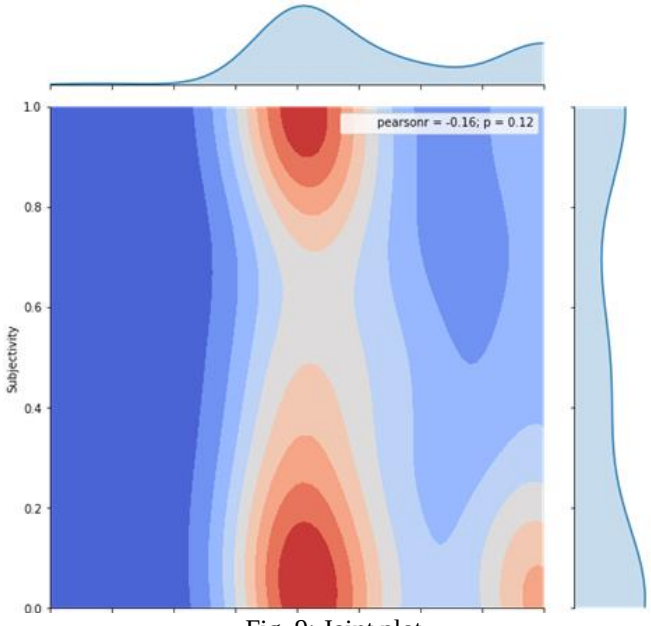

Fig. 9: Joint plot

Similarly, we can perform experiment for Traffic data analysis or Hospital data analysis, Feedback system, finding out what's trending by determining what the majority of the audience demands and Search Engine Optimization (SEO) as well by finding hot keywords to come up among the top results.

IV. CONCLUSION

This survey paper has explored different views of sentiment analysis. Starting right from Why & How sentiment analysis then the Methodology has explained very widely including Big data concepts & Natural Language Processing. We've discussed HDFS & Map Reduce as storage & programming architecture resp. Then we've seen NLP including mining & pre-processing of data, tokenization, sentiment analysis with its limitations & machine learning algorithm Naïve Bayes for classification. The workflow diagram shows sequential flow of complete process and at last we've seen one real time experiment.

Thus, all the essential information needed to perform sentiment analysis is mentioned in this paper with all its aspects. This clears that what study of literature matters in our life and its proper use in decision making will definitely lead us to higher level of accuracy & improve business productivity.

REFERENCES

[1]. ("https://github.com/Voidle/Twitter-Sentiment-Analysis-via-Py/blob/master/tejn.ipynb")
[2]. (Kumar & Bala, "Analyzing Twitter Sentiments Through Big Data", 2016)
[3]. (Wagh & Punde, "Survey on Sentiment Analysis using Twitter Dataset", 2018)
[4]. (Parveen & Pandey, "Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm", 2016)
[5]. (Katal, Mohammad Wazid, & R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices", 2013)
[6]. ("https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#5c6059060ba9")
[7]. ("https://encryptedtbn0.gstatic.com/images?q=tbn:ANd9GcTMRp6kPh3sZhRrI6UV6-Hh2r9f0hHFE1oVBRjpWgoNY-E0sLg")
[8]. ("http://www.internetlivestats.com/twitter-statistics/")
[9]. ("https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/")
[10]. ("https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/")