Sentimental Analysis on Twitter using Pig and Hive

Ajit Noonia, Vikas Verma, Ruchika Khandelwal, Kushagra Gautam

Abstract*: Data science is the analytic process to explore new prediction and pattern when to process the collected data. Data analysis is done using large sets of databases and due to them we can easily form patterns and then they could be recognized. This will helpful for prediction of new challenges and circumstances. From the perspective of statistics data analysis of large observational databases has very challenges which made it a research area in abroad as well as in India. Different tools are available in market to process and analyze the large set of data for prediction of future trends and due to which knowledgeable decision should be created. Bigdata and hadoop are one of them. In this paper we have collected 5000 above tweets and then we have done pre-processing over it and then done sentimental analysis so as to get negative and positive tweets and then done prediction over it so as to get the people's sentiments over a particular person.

***Keywords* : Big Data, Sentimental Analysis, tokenizer, hadoop, Apache hive, Apache pig, pre-processing.**

I. INTRODUCTION

Bigdata is a term used for large data sets and helps in analyzing that datasets so that we can extract useful information from it as it is very difficult for the traditional database system to deal with large datasets. And for this we use Hadoop, which is a programming paradigm which helps in processing of large sets of databases in the computing environment. The Hadoop ecosystem consists of many components (which are used for analyzing) like apache pig, apache hive, zookeeper, spark, etc.

Apache Hive, an open-source data warehouse system, is used with Apache Pig for loading and transforming unstructured, semi-structured, or structured data for data analysis and getting better business insights. Apache Pig, a standard ETL (Extract, Transform, Load) scripting language, is used to export and import data into Apache Hive and to process large number of datasets.

As we know nowadays, social media plays a vital role in everyone's life. It plays a role of networking in the life of contemporary society. And mostly twitter which provides a service for people to communicate ad stay connected through the exchange of quick, frequent messages. It's an online micro blogging service that enables users to send and read tweets or we can say comments over a particular topic. Twitter is a social networking service on which people interacts with each other by posting their views on a particular topic recently happening in the world and from famous personalities to middle class persons. Its world limit is 280 words and these short messages are known as tweets in accordance to twitter and from people are able to know the people's sentiments with those tweets and more than a crore tweets are posting in a day. Twitter, contains nearly 500 million users and more than 200 million messages per day, has quickly became an important for organizations to monitor their reputation and brands by analyzing the sentiment of the people by their Tweets posted on the twitter related about them, their markets, and competitors. Performing Sentiment Analysis on Twitter is simple in comparison to any other method. This is because the tweets are very short (only about 140 characters) and usually contain slangs, hash tags, etc. Sentiment Analysis may be described as a text mining technique for analyzing the sentiment of a text message that is a tweet. Twitter sentiment or opinion expressed through it may be negative, positive or neutral. However, no algorithm can give you complete accuracy or prediction on sentiment analysis. Twitter Sentiment Analysis has various applications like in business we can access business strategies, we can also know the customer's feelings about a particular brand or product that is either its positive aspects getting from public or why that product is not getting shelled that is to find out its limitation by checking people's response that they are write in the form of tweets. It is useful in politics too as it keeps track on political views to check that how much difference between the actions and statements at the government level or we can analyze election results and many more things related too that. We can also check public actions like monitoring social phenomenon and know what is happening in the world.

Methods like, negative and positive words to find the sentence is however inappropriate, because the sense of the text block depends a lot on the context. This may be done by looking at the POS (Part of Speech) Tagging. By doing this project we are trying to find the sentiments of the people over a particular type for example demonetization which would be helpful to may type of people who are attached to financial institutions or we can say politicians and it may be normal people also as from higher class to middle class to lower class either politicians actors and big fame personalities everyone is there so it is easier to know the people's sentiments in every form

Revised Manuscript Received on January 30, 2020.

* Correspondence Author

Ajit Noonia*, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.

Vikas Verma, Assistant Professor, Department of Computer Science and Engineering, Jaipur National University, Jaipur, India.

Ruchika Khandelwal, Department of Computer Science and Engineering, Jaipur National University, Jaipur, India.

Kushagra Gautam, Department of Computer Science and Engineering, Jaipur National University, Jaipur, India.

Retrieval Number: B7051129219/2020©BEIESP
DOI: 10.35940/ijitee.B7051.019320
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

so it could be beneficial in any purpose.

II. RELATED WORK

Kharde and Sonawane (2016) [1] has proposed a survey over sentimental analysis over twitter and constitutes a comparative analysis of some techniques like machine learning and lexicon-based approaches to find out the better result. It also contains evaluation metrics using various machine learning algorithms such as SVM and navie bayes and lexicon-based method and they both are correct at their places like machine learning methods can be considered as the baseline learning methods, while lexicon-based methods which requires few efforts in human-labeled document. They are also working on the effects of classifier so that they can classify them easliy. As they are trying to get better accuracy, they have used the bigram model which provides better sentiment accuracy as compared to other models. So, for this they are trying to combine both the techniques to get the better sentiment accuracy for betterment of results.

Apoorv agarwal et.al. (2017) [2] has introduced some concepts to increase the feature engineering those concepts are POS-specific prior polarity features and the tree kernel. They have tried to increase the overall gain and accuracy in sentiment analysis and they have succeeded to that is 4% more accuracy and proposed state-of-the-art unigram model. They studied both the tree kernel and feature based models and observes that both these models work outstandingly over the unigram baseline.

Kumar and Sebestian (2015) [3] has proposed a hybrid approach of doing sentiment analysis over tweets using both corpus based and dictionary-based methods. Basically, in this paper the researchers are comparing both these methods of doing sentiment analysis and presented a case study to illustrate the use and effectiveness of both. For doing this, they extracted the opinion words from the tweets. The corpus-based method was used for the adjectives and the dictionary-based method used for the verbs and adverbs in the process of findin sentiments.and this was all about done using linear equation and at last it was tried to improve efficiency only.

Shubhangi D Patil (2014) [4] has tried to give review over sentimental analysis. As Twitter sentiment analysis plays a very somewhat important role for most of the decision-making situations where public opinion is necessary to be considered. This paper is providing the research-oriented review and analysis of different twitter sentiment analysis tasks which are Feature selection methods and Sentiment classification methods. The researcher has tried to explain the various methods for the feature selection as well as sentiment classification task. The various twitter sentiment analysis datasets which are freely available for research purpose are listed with their available tools of twitter analysis which are available online. Though a lot a work has already been done in this area, many issues are still to be investigated.

Shubham Goyal (2016) [5] tried to do the sentiment analysis by utilizing the data only and take a proper dataset and worked upon it to get the better results. Here the researcher has taken a particular topic as a case study that is Food price crisis and public opinion is considered as by doing sentiment analysis. The tweets are extracted using twitter API. The proper database has been created to save the tweets and over which the pre-processing is done so as to remove the special characters, short words, Spam, url, etc. and then the tweets are then stemmed and tokens are provided to them and TF-IDF score is calculated for the keywords used for getting the results. And then for performing Feature selection, the methods are used are Chi-Square and information gain. Analysis has been done using KNN and Naïve Baye's and by using them a hybrid structure has been created. The results of the classifiers were satisfactorily but the hybrid-KNN is better in comparison to Naïve Baye's in terms of accuracy.

III. RESEARCH MEHODOLOGY

1. Data Description: - There is a data of twitter tweets which includes up to 5000 tweets in which it includes all the data related to that tweet so that sentimental analysis can be performed properly and it is in csv format.

It has many fields like first is the text that is the tweets that are done and then either it is favorite or not that is either the tweets are recorded to enable quick access in future and then if it favorites then calculated the no of favorites.

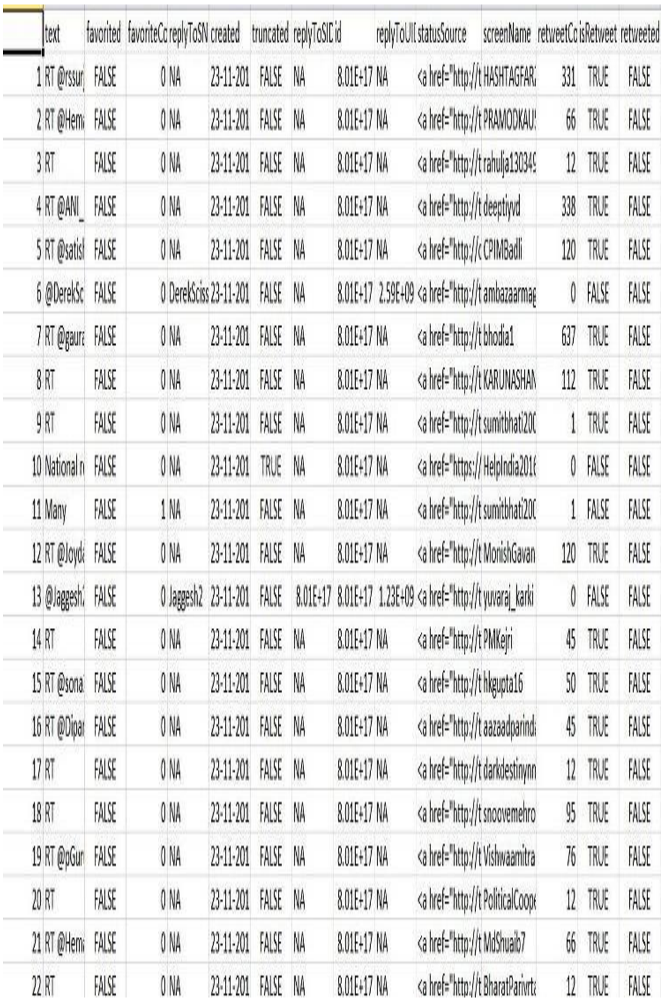
The next we have screen names that is the user chooses to identify them on that particular platform and in preference to that we have reply to screen name that is if someone has replied to screen names then they are saved in that particular column with their screen name only who have replied to the tweets and if anyone has not replied to screen names nil will be associated to it.

And other than that, it is saved that someone has retweeted over the tweets or not and then if someone has retweeted over that then its data is saved and preferably the retweet counts are also there. The created date is saved with the id of that tweet which is used in providing token to the tweets. It is checked either the tweets are truncated or not with their status source also. It is also checked that reply to SID and reply to UID is done or not if done then its id will b saved in that column only.

So, by using this kind of data we are trying to find out the sentimental analysis of that particular data so as to find the negative and positive tweets and neurtral too. The data dictionary we are using is the text file which contains the data which have positive and negative words which are used for the comparison of tweets after tokens get provided to them and then we are able to find out the negative and positive words.

The flowchart constitutes the algorithm over which the project works. It shows that firstly we have to collect the data as per out requirement that is tweet downloader and after that we will provide tokens to each and every cell so that we can easily provide pre-processing over it that is we will compare the tweets with dictionary words and then we will extract the unnecessary things the tweets that is feature extractor and then at last we will classify the negative and positive tweets so that we can easily do the prediction over the sentiments.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.



	text	favorited	favoriteCc	replyToSN	created	truncated	replyToSIC	id	replyToUI(	statusSource	screenName	retweetCo	isRetweet	retweeted
1	RT @rssur	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	HASHTAGFAR	331	TRUE	FALSE
2	RT @Hem	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	PRAMODKAU	66	TRUE	FALSE
3	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	rahulja130349	12	TRUE	FALSE
4	RT @ANI_	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	deeptiyvd	338	TRUE	FALSE
5	RT @satis	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://c	CPIMBadli	120	TRUE	FALSE
6	@DerekSc	FALSE	0	DerekSciss	23-11-201	FALSE	NA	8.01E+17	2.59E+09	<a href="http://t	ambazaarmag	0	FALSE	FALSE
7	RT @gaura	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	bhodia1	637	TRUE	FALSE
8	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	KARUNASHAN	112	TRUE	FALSE
9	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	sumitbhati20C	1	TRUE	FALSE
10	National r	FALSE	0	NA	23-11-201	TRUE	NA	8.01E+17	NA	<a href="https://	HelpIndia2016	0	FALSE	FALSE
11	Many	FALSE	1	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	sumitbhati20C	1	FALSE	FALSE
12	RT @Joyd	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	MonishGavan	120	TRUE	FALSE
13	@Jaggesh	FALSE	0	Jaggesh2	23-11-201	FALSE	8.01E+17	8.01E+17	1.23E+09	<a href="http://t	yuvaraj_karki	0	FALSE	FALSE
14	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	PMKejri	45	TRUE	FALSE
15	RT @sona	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	hkgupta16	50	TRUE	FALSE
16	RT @Dipa	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	aazaadparind	45	TRUE	FALSE
17	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	darkdestinynn	12	TRUE	FALSE
18	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	snoovemehro	95	TRUE	FALSE
19	RT @pGur	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	Vishwaamitra	76	TRUE	FALSE
20	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	PoliticalCoop	12	TRUE	FALSE
21	RT @Hem	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	MdShuaib7	66	TRUE	FALSE
22	RT	FALSE	0	NA	23-11-201	FALSE	NA	8.01E+17	NA	<a href="http://t	BharatParivrt	12	TRUE	FALSE

Fig.1: Snapshot of the Data Used in the Project

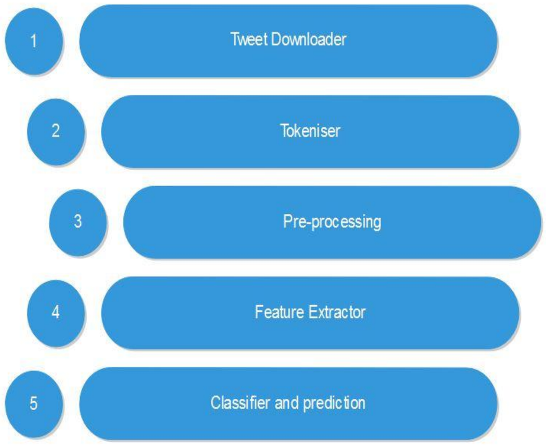


Fig. 2: Flowchart that Shows how the Algorithm Works

2. Prerequisites:

Hadoop Version 2.6.5
Apache Hive 1.2.1
Apache Pig version 0.17.0

3. Use Case:

- Load the data into Hive so as to make that structured.
- In Pig, process that data and then analyze it.
- Now load that data into different table in hive using pig.

4. Working:

a) Loading the csv(tweets.csv) file using pig.

b) Extracting the needed data:

$0 represents first field and $1 represents the second field

"id" is the alias name of $0 and "text" is the alias name of $1

c) The following characters are considered to be word

Separators: space, coma(,) parenthesis(()), star(*), double quote(").

TOKENIZE will split the records based on above separators and give the bag of words....

FLATTEN will remove the parenthesis like () and {} from the bag of words and tokens i.e. words from each record will get associate with each record as 3rd field in a iterative manner until the tokens will be finished.

d) For checking the effect of step 'tokens', use the commands

LIMIT AND DUMP.

e) Loading the dictionary which has rating for words.

f) 'Replicated' join is the special type of join in which second

Relation is small enough to fit into the main memory which will help in efficient join.... This will give sample data.

g) We are iteratively selecting the data from the relation

"word_rating" and selecting fields from respective relations To select the field which is the part of a certain relation weneed to use double colon "::"

h) Group relations on the basis of id and text combined.

i) Generate group which is from relation word_group and as all

the data will get group on the basis of (id,text) we can perform average of those tokens also as per their ratings.

j) Filter the positive tweets.

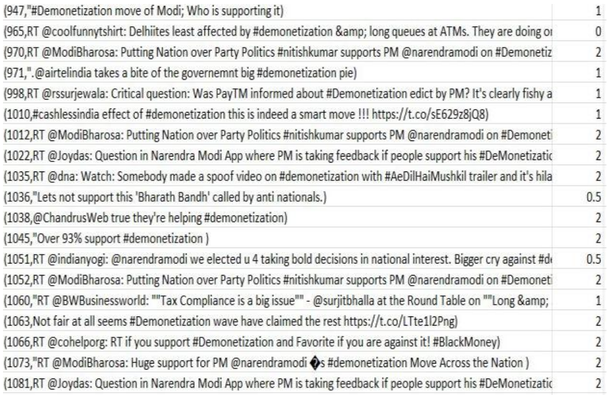
k) Filter the positive tweets.

l) Storing the positive and negative tweets output in any folder.

IV. RESULTS AND DISCUSSION

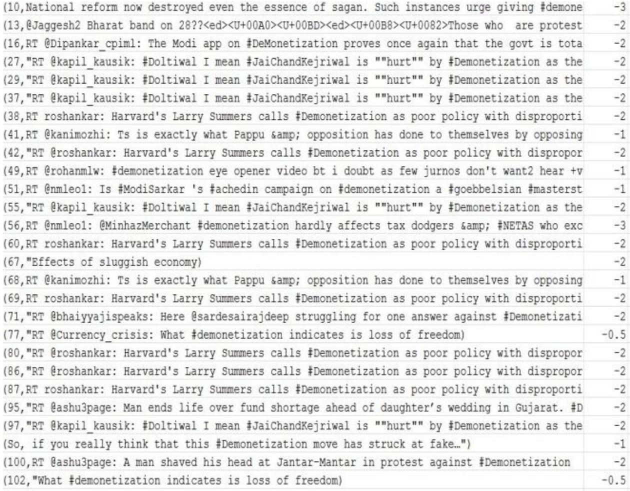
Now, after applying that algorithm we have processed data and analyzed it and make some patterns so that we can easily get the data we want by applying hive queries which are somewhat same as sql queries. We can get positive and negative tweets by comparing with them by the word dictionary for positive and negative words. The positive tweets will come in that format. As it will be compared to the data dictionary and then processed we are getting the output in the two columns one is the tweets we want with their serial no and who has tweeted that is screen name and then the tweet and in the second column the no of words which have been matched with word dictionary.

The second output we are getting is of negative tweets and they are also in the same form as we have in the positive tweets that is one column, we have tweets along with their id and screen name and in second column we have no of words matching to the data dictionary.



(947,"#Demonetization move of Modi; Who is supporting it)	1
(965,RT @coolfunnytshirt: Delhiites least affected by #demonetization &amp; long queues at ATMs. They are doing o	0
(970,RT @ModiBharosa: Putting Nation over Party Politics #nitishkumar supports PM @narendramodi on #Demonetiz	2
(971,".@airtelindia takes a bite of the governemnt big #demonetization pie)	1
(998,RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearly fishy a	1
(1010,#cashlessindia effect of #demonetization this is indeed a smart move !!! https://t.co/sE629z8jQ8)	1
(1012,RT @ModiBharosa: Putting Nation over Party Politics #nitishkumar supports PM @narendramodi on #Demoneti	2
(1022,RT @Joydas: Question in Narendra Modi App where PM is taking feedback if people support his #DeMonetizatic	2
(1035,RT @dna: Watch: Somebody made a spoof video on #demonetization with #AeDilHaiMushkil trailer and it's hila	2
(1036,"Lets not support this 'Bharath Bandh' called by anti nationals.)	0.5
(1038,@ChandrusWeb true they're helping #demonetization)	2
(1045,"Over 93% support #demonetization)	2
(1051,RT @indianyogi: @narendramodi we elected u 4 taking bold decisions in national interest. Bigger cry against #d	0.5
(1052,RT @ModiBharosa: Putting Nation over Party Politics #nitishkumar supports PM @narendramodi on #Demoneti	2
(1060,"RT @BWBusinessworld: ""Tax Compliance is a big issue"" - @surjitbhalla at the Round Table on ""Long &amp;	1
(1063,Not fair at all seems #Demonetization wave have claimed the rest https://t.co/LTte1l2Png)	2
(1066,RT @cohelporg: RT if you support #Demonetization and Favorite if you are against it! #BlackMoney)	2
(1073,"RT @ModiBharosa: Huge support for PM @narendramodi �s #demonetization Move Across the Nation)	2
(1081,RT @Joydas: Question in Narendra Modi App where PM is taking feedback if people support his #DeMonetizatic	2

Fig. 3: Snapshot of the Result that is Positive Tweets



(10,National reform now destroyed even the essence of sagan. Such instances urge giving #demone	-3
(13,@Jaggesh2 Bharat band on 28??<ed><U+00A0><U+00BD><ed><U+00B8><U+0082>Those who are protest	-2
(16,RT @Dipankar_cpiml: The Modi app on #DeMonetization proves once again that the govt is tota	-2
(27,"RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the	-2
(29,"RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the	-2
(37,"RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the	-2
(38,RT roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with disproporti	-2
(41,RT @kanimozhi: Ts is exactly what Pappu &amp; opposition has done to themselves by opposing	-1
(42,"RT @roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with dispropor	-2
(49,RT @rohanmlw: #demonetization eye opener video bt i doubt as few jurnos don't want2 hear +v	-1
(51,RT @nmleo1: Is #ModiSarkar 's #achedin campaign on #demonetization a #goebbelsian #masterst	-1
(55,"RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the	-2
(56,RT @nmleo1: @MinhazMerchant #demonetization hardly affects tax dodgers &amp; #NETAS who exc	-3
(60,RT roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with disproporti	-2
(67,"Effects of sluggish economy)	-2
(68,RT @kanimozhi: Ts is exactly what Pappu &amp; opposition has done to themselves by opposing	-1
(69,RT roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with disproporti	-2
(71,"RT @bhaiyyajispeaks: Here @sardesairajdeep struggling for one answer against #Demonetizati	-2
(77,"RT @Currency_crisis: What #demonetization indicates is loss of freedom)	-0.5
(80,"RT @roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with dispropor	-2
(86,"RT @roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with dispropor	-2
(87,RT roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with disproporti	-2
(95,"RT @ashu3page: Man ends life over fund shortage ahead of daughter’s wedding in Gujarat. #D	-2
(97,"RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the	-2
(So, if you really think that this #Demonetization move has struck at fake…")	-1
(100,RT @ashu3page: A man shaved his head at Jantar-Mantar in protest against #Demonetization	-2
(102,"What #demonetization indicates is loss of freedom)	-0.5

Fig. 4 Snapshot of the Result that is Negative Tweets

V. CONCLUSION

Bigdata is proved a greatest tool for decision making process using social media world and it is used for creating patterns in the data from the past data by doing processing over it. And the tools which are used that are the most important part of hadoop ecosystem that is pig and hive. And for processing data these are the bestest example till now. Hive helps in getting structured data from the unstructured form of data and also helps in getting the meaningful patterns from the data which helps in decision making. Whereas pig helps in processing the data to get the favorable outcomes which could be beneficial for our future use.

We have used the pig and hive which helped to change the unstructured form of data into the structured one and then pre-processing is applied so we can apply tokenizer to the tweets and then comparison is done using data dictionary and finally we can analyze the result and prediction should be done over a particular topic like demonetization or any other topic emerging on social media like twitter. So pig and hive would be the easier way to perform sentiment analysis.

FUTURE SCOPE

Finding the solution of: -

1. Twitter data crawled by third party as it will be the issue of security as well as no professional account are safe as on other social media business accounts are safe.
2. Limitation with Twitter APIs for crawling data that is standard API rate limits per window.
3. Noises included in randomly picked 5000 tweets as tweets are highly unstructured and no grammatical mistakes are very much as well as out of vocabulary words lexical variation and extensive use of acronyms.

Extending our work by: -

1. Using different other models and algorithms so as we can get statistical approach either by using bigdata and even machine learning too.
2. Temporal analysis can be added as future work in project so that we can create models as to do comparative analysis.
3. Consideration of Retweets as a factor so that we can do sentimental analysis over that we have taken that data but we are trying to do that also as retweets are very important factor in twitter.

REFERENCES

1. Kharde and Sonawane (2016), “Sentimental Analysis of twitter data- A survey of techniques”, International Journal of Computer Applications, 6(10), pp. 2321-9637.
2. Apoorv agarwal et.al. (2107), “Sentimental analysis of twitter data”.
3. Kumar and Sebestian (2015), “Sentiment Analysis on Twitter”, IJCSI International Journal of Computer Science.
4. Manikandan and Ravi (2014), “Big Data Analysis using Apache Hadoop”, IEEE, pp. 1- 4.
5. Dhawan and Rathee (2013), “Big Data Analytics using Hadoop Components like Pig and Hive”, American International Journal of Research in Science, Technology, Engineering & Mathematics, pp. 88-93.
6. Gadekar and Bhosle (2014),”A Review Paper on Bigadata & Hadoop”, International Journal of Scientific & Research publications, 4(10), pp. 1-7.
7. Smt. Shubhangi D Patil (2014), “Review of twitter Sentimental analysis”, International Journal of Scientific & Engineering Research, 5(10), pp. 2229-5518.
8. Shubham Goyal (2016), “Sentimental Analysis of Twitter Data using Text Mining and Hybrid Classification Approach”, International Journal of Advance Research, Ideas and Innovations in Technology, 2(5) , pp. 2454-132X.

AUTHORS PROFILE

Dr. Ajit Noonia is doctorate in Computer Science and Engineering from Jyoti Vidyapeeth Women’s University, Jaipur. He is currently Assistant Professor in Department of Computer Science in Chitkara University, Rajpura, Punjab, India. His research areas are VANET, Data Mining and Automation.

Mr. Vikash Verma is an Assistant Professor in the Department of Computer Science and Engineering, Jaipur National University, Jaipur, India. His research areas are VANET, Data Mining and Automation.