ISSN 2395-1621

Mining Social Media Data Using NLP and Hadoop Clustering

[#1]Devika Girme, [#2]Madhuri Godage, [#3]Snehal Gajare, [#4]Saloni Shekatkar, [#5]Prof.V.S.Rajput

[1]girmedevika007@gmail.com
[2]madhurigodage3@gmail.com
[3]sgajare077@gmail.com
[4]salonishekatkar@ymail.com
[5]vnkalyankar_sits@sinhgad.edu

[#12345]Information technology Engineering, Sinhgad Institute of Technology and Science (Affiliated to Savitribai Phule Pune University) Pune, India.

ABSTRACT

Social media like Facebook today are not only just a website .They are now become much popular communication tool for internet users. It is a medium through which users belonging to any of category, profession can make their comments. These all comments have contained some features along with it.These comments or status are really useful which are actually viewed as their OPINIONS.Opinions are really important while we need to analyze any of product, topic, discussion and whatever which will require some user opinions to draw some inferences and conclusions from them. Social media plays an important role for this intention.We focused on facebook statuses comments which we can view as opinions of users or their reaction on concern we want to analyze.We are usingNatural language processing and hadoop clustering.In which Hadoop clustering is specially designed for storing and analyzing huge amount of unstructured data.We are classifying posts of Facebook into POSITIVE,NEGATIVE and NEUTRAL.Its pure new and unique technique proposed in the field of opinion mining.

ARTICLE INFO

Article History

Received 25th March 2016

Received in revised form :

27th March 2016

Accepted : 29th March 2016

Published online :

1st April 2016

Keywords:–Natural language processing,hadoop,Hadoop cluster.

I. INTRODUCTION

The dramatic and exponential growth of content available on web and its classification has now become an efficient methodology to make the contents of large repository in an organized manner.Social networking websites are the newera of expressing views .Today every fifth person put their opinions,views,comments on these micro-blogging and social sites like TWITTER,FACEBOOK and many more.Authors of those comments, views and opinions write their point of perception on any of discussion topic. It may include any political issue, religious issue, technology, product, movie review and much more daily gossiping issues coded in their surroundings. Now people are using internet as a communication tool among their social network including friends, family, friends of friends to these micro-blogging and social network sites. In this we gradually put and share their opinions among their friends on these sites which finally becomes huge and relevant repository for any of particular entity or organization. Such dataset collected from all these sites can be efficiently used for marketing, case study and social studies. Organizations that required can easily draw inferences and conclusions regarding their product, technology or political point whatever they all are concerning with by going through opinions comes from facebook posts.

II. LITERATURE SURVEY

[1] Mining Social Media Data for Understanding Students Learning .This paper is beneficial in learning analytics, educational data mining, and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. [2] Opinion Extraction and Classification of Real Time Facebook Status Social media like Facebook today are not only just a website. They are now become much popular

communication tool for internet users. It is a medium through which users belonging to any of category, profession can make their comments. These all comments have contained some features along with it. These comments or status are really useful which are actually viewed as their OPINIONS. Our classifier is able to extract three features GOOD, BAD and AVERGAE from that statuses respectively. As per classifier results we perform evaluations experiments which further can be work for feature mining of user opinions on facebook.

III.PROPOSED SYSTEM

The Graph *API* is the primary way to get data in and out of *Facebook's* social graph more about the different operations that can be performed *using* the *API*. We can generate positive, negative and neutral opinion from facebook posts, messages or comments. Firstly, the system is accessible to only developer of facebook. FacebookappID is to be created by developer. After login by developer to facebook's account, posts can be accessed. The textual data is listed into a table in which their category is mentioned. The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured or, perhaps through a limited number of clearly-enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context. All posts which are accessed using nlp are copied to a text file. The text file to be considered is an appendable file. Hadoop cluster is given an input of that text file. The text file is analyzed and according to nlp posts are categorized into positive, negative or neutral. To simplify the output we are creating a graph using applet. Different colours in graph specify different categories accordingly.

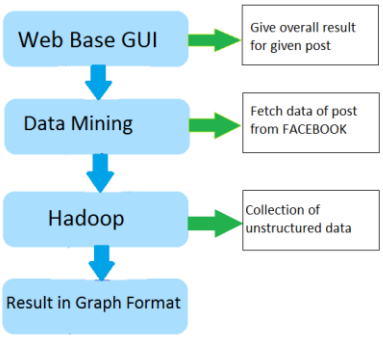


Figure 1: General Architecture

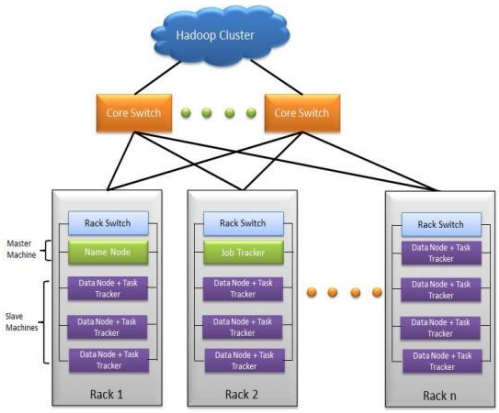


Figure 4.2: Hadoop cluster architecture

The graph shown below states the percentage of positive ,negative and neutral posts of facebook.

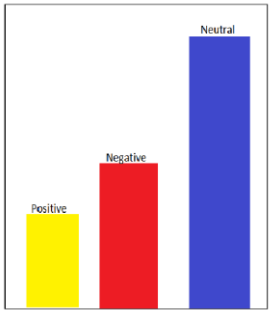


Figure 4.3: Graph

IV.ALGORITHMS

- **NATURAL LANGUAGEPROCESSING**

Natural Language Processing (NLP) will be briefly presented together with an overview of some sub-areas especially relevant to evolutionary computation. NLP is a research field concerned with the interaction between computers and natural (human) language, as spoken and written language bodies are being processed for various purposes. The field is situated between Computer Science and Linguistics, and deals with problems ranging from ambiguity resolution both on lexical and syntax level, part-of-speech-tagging (POS-), speech and text segmentation, to syntactic and semantic parsing. The problems have traditionally been solved with either rule-based or data-driven approaches, or in later times combinations of the two .Major application areas are spelling and grammar checking, machine translation, text summarization, question answering systems, and dialogue systems.The following are the steps used in NLP:

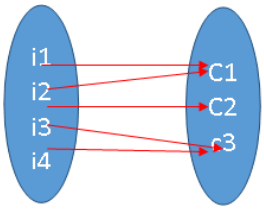
1.DEGRADATION
In degradation the whole string is separated into different parts.
2.STREAMING In Streaming,"ing,es,ion," etc words are separated.
3.STOP WORD FILTERING In stop word filtering,"is,am,was,had,are",etc words are considered.

4.WORD CLUSTERING In word clustering, k-means algorithm is used for clustering of comments or statuses into POSITIVE and NEGATIVE clusters.

- **MATHEMATICAL MODEL**

- Set Theory Analysis: Let S be the | Facebook Post Classification as the final set S = I, F, O Where, I = i | set of all input as a post for analysis i.e. I= i1,i2,i3

- O = o | set of all outputs as a result for analysis i.e. O= o1,o2,o3 F = f | set of all functions used in post analysis i.e. F= f1,f2,f3,f4 f1=fetchposts (i (n)) f2=classify (i(n)) f3=categorise (i (n), c) f4=generategraph (c (n), G) f3=categorise(i(n),c) ,categorized into category c f4=generategraph(c(n),G),all categories are mapped into a graph to give final output as graph 14



V. CONCLUSION

The final conclusion drawn from the work is we have developed very efficient and time saving method to classify millions of comments posted on facebook.These classified opinions will then become required data to judge the reviews of users regarding any concern belong to any issue. It reduces the manual survey work that had been done for drawing conclusions on opinion posted on facebook. This work could further extended for twitter tweets or any of frequently access social websites containing several reviews from different people. Using the most well-known machine learning algorithms, we conducted a comparative experimental procedure between the K-means and the NLP algorithms by combining different feature extractors. Those algorithms can achieve high accuracy for classifying sentiment when combining different features. Although Facebook statuses have unique characteristics compared to other corpuses (Reviews, News, etc), machine learning algorithms are shown to classify statuses with similar performance.Finally, the overall performance of the proposed methodology is satisfactory, however, we would like to further improve by tracking changes within peoples sentiment on a particular topic, explore the time dependency of our data and analyze their trendy topics dynamically. It would be very interesting to involve the temporal feature on this kind of analysis and not to focus solely on previous posts or discussions.

Future Scope:

Many other social sites such as Twitter,LinkdIn,etc can be classified by using such algorithms for making decisions .

VI. REFERENCES

[1]J. Kleinberg, “Bursty and Hierarchical Structure in Streams,” Data Mining Knowledge Discovery, vol. 7, no. 4, pp. 373-397, 2003.

[2]Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, “Real-Time Change-Point Detection Using Sequentially Discounting Normal- ized Maximum Likelihood Coding,” Proc. 15th Pacific-Asia Conf. Advances in Knowledge Discovery & Data Mining (PAKDD’ 11), 2011.

[3]S. Morinaga and K. Yamanishi, “Tracking Dynamics of Topic Trends Using Finite Mixture Model,” Proc. 10th ACM SIGKD Int’l Conf. Knowledge Discovery and Data Mining, pp. 811-816, 2004.

[4] Q. Mei and C. Zhai, “Discovering Evolutionary Theme Patterns from Text:An Exploration of Temporal Text Mining,” Proc. 11Th ACM SIGKDD Int’l Oonf. Knowledge Discovery in Data Mining, pp. 198-207, 2005

[5]J. Allan et al., “Topic Detection and Tracking Pilot Study: Final Reporrt,” P Proc. DARPA Broadcast News Transcription and Under-standing Workshop, 1998.

