International Journal of Engineering & Technology, 7 (4.5) (2018) 374-378

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An Approach To Twitter Sentiment Analysis Over Hadoop

Yazala Ritika Siril Paul[1]*, Dilipkumar A. Borikar[2]*

[1,2] *Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, INDIA*
*Corresponding author E-mail:[1]*paulys@rknec.edu*, [2]*borikarda@rknec.edu*

Abstract

Sentiment analysis is the process of identifying people's attitude and emotional state from the language they use via any social websites or other sources. The main aim is to identify a set of potential features in the review and extract the opinion expressions of those features by making full use of their associations. The Twitter has now become a routine for the people around the world to post thousands of reactions and opinions on every topic, every second of every single day. It's like one big psychological database that's constantly being updated and which can be used to analyze the sentiments of the people. Hadoop is one of the best options available for twitter data sentiment analysis and which also works for the distributed big data, streaming data, text data etc. This paper provides an efficient mechanism to perform sentiment analysis/ opinion mining on Twitter data over Hortonworks Data platform, which provides Hadoop on Windows, with the assistance of Apache Flume, Apache HDFS and Apache Hive.

Keywords: *Apache Flume; Apache Hadoop; Apache Hive; Sentiment Analysis; Twitter*

1. Introduction

We live in an era where the textual data on the Internet is growing at a very rapid pace and many companies are trying to use this huge amount of data to extract people's views towards their product. The best source of unstructured text information is included in social networking sites, where it is unfeasible to manually analyze such huge amount of data. There are large number of social networking websites that enables the users to contribute, modify and also grade the content. The users can express their personal opinions about any specific topic. Some examples are blogs, forums, product reviews sites and social networking sites, like Twitter, Facebook, etc.

Sentiment Analysis: Sentiment analysis is a technique for processing natural language text so as to evaluate the position, sensitivity or assessment of the people about a specific subject, merchandise or topic.

Why Sentiment Analysis? Everyday enormous amount of data is created from social networks, blogs and other media and diffused in to the World Wide Web (WWW). This huge data contains very crucial opinion related information that can be used to benefit businesses and other aspects of commercial and scientific industries. Manual tracking and extracting this useful information is not possible, thus, Sentiment analysis is required. Sentiment Analysis is the phenomenon of extracting sentiments or opinions from reviews expressed by users over a particular subject, area or product online. It clubs the sentiments orientation in categories like "positive" or "negative". Hence, it determines the general attitude of a speaker or a writer with respect to the topic in context.

Analysis of Sentiments on Twitter data provides an effective way to bring forth the users opinions that is necessary for decision making process in various sectors. Twitter allows the user to post short text messages called tweets of 140 characters but has doubled to 280 characters for a small group of users. There are 240+ millions of active users and 500+ million tweets are generated every day. The twitter audience varies from common man to celebrities hence resulting in unambiguity of tweets. Twitter also allows the users to use hashtags which are used to mark topics. So for each hashtag, there may be a lot of tweets and many new tweets are generated every minute. We are using Hadoop framework, in order to handle so many tweets. By using Hadoop, we analyze twitter data and classify them in 6 basic emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise. This data can be further represented using pie charts and can also be used for finding country based sentiment analysis.

Since only 20% of data in organizations is structured, and the rest is all unstructured, so it is very crucial to manage the unstructured data which goes unattended. Hadoop is able to manage different types of Big Data, whether it is structured or unstructured, or any other type of data and then makes it useful during any decision making process.

Hadoop is simple, relevant and schema-less. Though Hadoop generally supports Java programming language, any other programming language can be used in Hadoop benefit from its Map Reduce technique. Even though Hadoop works best on Windows and Linux, it can also work on other operating systems like BSD and OS X.

Hadoop: Hadoop is an Apache open source framework which is used for the distributed storage and for the processing of large datasets of Big Data using the Map-Reduce programming model. It consists of computer clusters which are built from commodity hardware. Hadoop is designed to scale up from single server to thousands of machines. Hadoop framework also allows the user to quickly write and test distributed systems. It is not only efficient but also automatically distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU scores. It does not rely on hardware to provide fault-tolerance and high availability (FTHA). Apart from being an open source, it is compatible on all the platforms since it is Java based.

Fig. 1: Hadoop Architecture

Hadoop Distributed File System (HDFS): HDFS is based on the Google File System and is suitable for the distributed storage and processing. When a file is placed in HDF it is broken down into blocks minimum of 64 MB block size. Each of these blocks are then saved on some nodes in the cluster called Data Nodes. Meta data of these files is stored on a specific node called Name Node in the cluster. To prevent the failure at data nodes, these blocks are replicated across different nodes in the cluster. The default replication value is 3. A Hadoop cluster can comprise of a single node or thousands or millions of nodes. HDFS provides file permissions and authentication.

Namenode is a commodity hardware that works on the GNU or LINUX operating system and also the namenode software. The system that has been assigned a namenode acts as the master server and it does the following tasks:

- Managing the file system namespace
- Regulating client's access to files
- Executing the file system operations such as renaming, closing and opening files and directories.

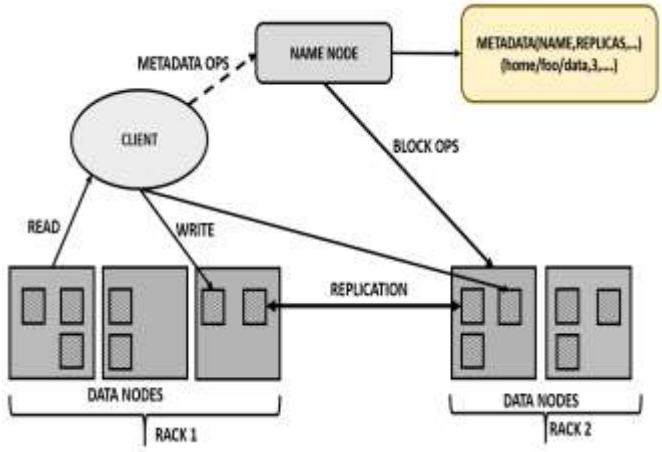

Fig. 2: HDFS Architecture

Datanode which is a commodity hardware works on operating systems like GNU or LINUX and a datanode software. For every node (commodity hardware/ system) in a cluster, there will be a datanode. These nodes manage the data storage of their system and performs the following tasks:

- A datanode performs read-write operations on the file systems, as per client request.
- A datanode performs operations such as block creation, deletion and replication according to the instructions of the namenode

MapReduce: MapReduce is a parallel programming model used by Hadoop for writing distributed applications for efficient processing of large amounts of data (multi-terabyte datasets), on large clusters (thousands of nodes) of commodity hardware in a reliable and fault-tolerant manner. The MapReduce algorithm contains two important tasks, namely map and reduce. The key reason to perform mapping and reducing is for speeding-up the execution of a specific process by splitting the process into multiple tasks, thus enabling parallel work. The map takes a set of data and converts the data into another set of data, where individual elements are broken down into various tuples (key-value pairs). Reduce has the task to take the output of map as an input and then combine those data tuples into a smaller set of tuples. Reduce task is always performed after the map task just as the name suggests.

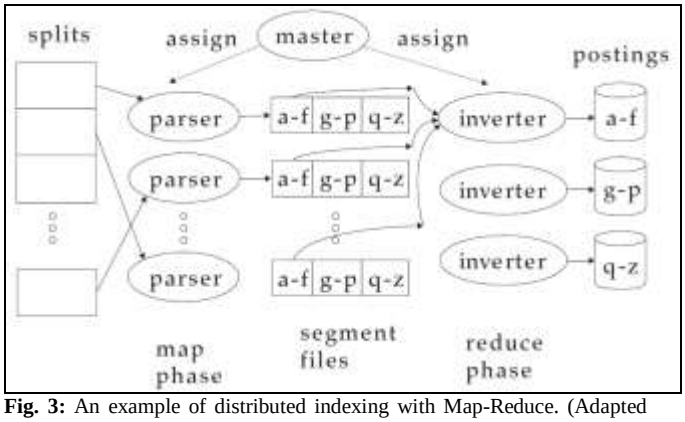

Fig. 3: An example of distributed indexing with Map-Reduce. (Adapted from [22])

Apache Flume: Apache flume is a tool/ service/ data ingestion mechanism which is highly reliable, distributed and configurable tool. Flume was principally designed to copy streaming data (log data) from various servers to centralized stores, for example HDFS.

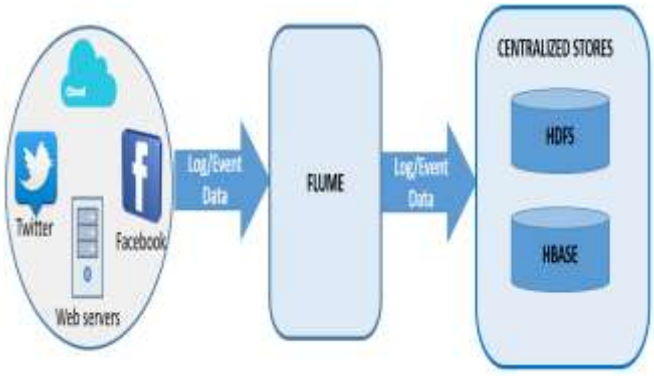

Fig. 4: Apache Flume working design

Apache Hive: It is data warehouse software project. For providing data summarization, query and analysis, Apache Hive is built on top of the Apache Hadoop. Hive provides an SQL-like interface to connect with the data stored in various databases and file systems that integrates with Hadoop using queries. Apache Hive is used for analytical purposes.

2. Problem Statement

Social media is one of the popular media right now to share opinions or variety of topics and twitter is very popular social site to share everything related to opinion on variety of topics and the discussions on current affairs. These tweets generates a huge amount of information related to different areas like government, election, etc. millions of tweets is generated every-day and which

is very useful in decision making process because everyone is sharing their views and opinions on various issues and topics.
Our aim is to analyze and classify the sentiments of the given textual messages into Ekman's 6 basic emotion viz anger, disgust, fear, happiness, sadness, surprise over Hadoop

Fig. 5: Ekman's 6 basic Emotions

3. Literature Review

Sentiment analysis is the most popular trend in today's world. A lot of work has been done in this sector. Following are some of the approaches which are most popular in today's world.
Starting from document level, Turney [1], in this paper a simple unsupervised learning algorithm for rating a review as thumbs up or thumbs down is introduced. The algorithm comprises of three steps: (a) Extraction of phrases containing adjectives or adverbs. (b) Estimation of semantic orientation of each and every phase, which is also the core step of the algorithm (c) Classification of the reviews which are based on the average semantic orientation of the phrases. This algorithm attained an average accuracy of 74%.
Pang and Lee [2], studied the relationship between the subjectivity detection and polarity classification with the result showing that the subjectivity detection can compress reviews into much shorter extracts that still contain the polarity information at a level equal to that of the full review. The results show that for the Naïve Bayes polarity classifier inputs are more effective than the originating document which only means that they are not only shorter but also cleaner representation of the intended polarity.
It is observed that the later papers [1, 2] calculate the summation of polarity of the adjectives and adverbs contained within text. Given a limitation of 140 characters on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis.
In order to detect sentiments of the users, researchers have used different machine learning techniques, such as, Maximum Entropy, Support Vector Machine and Naïve Bayes because they proved to be more accurate than other algorithms [3, 4]. A comparative study in terms of relative performance shows that Naïve Bayes tends to do worst and SVM's tend to do the best, even though the difference aren't very large [21].
Wilson, Wiebe and Hoffmann [5] have proposed a new approach to phrase-level sentiment analysis. Firstly determination of whether an expression is neutral or polar is done and after that disambiguating the polarity of the polar expressions is performed. Thus making it possible to automatically identifying the contextual polarity for a large subset of sentiment expressions while achieving results that are significantly better than baseline by using this approach.
Batool, R. *et al.* [6] analyzed tweets to classify the data and sentiments or opinions from twitter data more precisely. The information from tweets were extracted using the keyword based knowledge extraction. The extracted knowledge are further enhanced by using domain specific seed based enrichment technique. The methodology that is proposed here facilitates the extraction of entities, keywords, synonyms, and parts of speech (POS) from tweets which are then used for tweets classification and for sentiment analysis. The system proposed was tested on a data set with collection of 40,000 tweets. Increase in information gain in the range of 0.1% to 55% was achieved when knowledge enhancer and synonym binder module was applied on the extracted information.
Kumar Singh, P. et al. [8] aimed to automate the process of gathering online, end user reviews for any given product or service and analyzing those reviews in terms of the sentiments expressed about specific features. This involves the filtration of irrelevant and unhelpful reviews, quantification of the sentiments of thousands of (useful) reviews. And finally, providing the end user (business/manufacturer) summarized data about the expressed sentiments in way of intuitive and easy to understand graphs, charts and other visualization. This data can then be used to improve business outcomes and ensure a very high level of customer satisfaction.
Lei Huang et al. [9] explored augmenting different n-gram features in conjunction with POS tags into the training of supervised classifiers including Naive Bayes (NB), Support Vector Machines (SVM) and Maximum Entropy (MaxEnt). It is observed that by almost 3% the MaxEnt trained from a combination of unigrams and bigrams outperforms other models trained from a combination of POS tags and unigrams. Subramaniyaswamy et al. [14] in this paper deals with Hash Algorithm, MRAP (Map-Reduce Access Patterns), Hadoop Map reduce and Collaborative Filtering. In this research work, firstly the unstructured data is structured and then processed by using Map Reduce technique and after that the automatic prediction of user's taste is done through collaborative filtering. Map reduce is the most efficient technique for processing large volume of data and the application of collaborative filtering and the sentiment analysis provides recommendation generation for any number of data provided as input. The MRAP performed multiple sequential reads per map task, whereas in map reduce reads performs only single read for a single map task. In Collaborative filtering restructuring the predicted is based on the user's suggestions to generate a recommendation system. It is observed that collaborative filtering takes less time for predicting the tweets whereas MRAP consumes more time. Also collaborative filtering consumes less space as compared to MRAP.

4. Proposed Approach

In this section a method to analyze and classify the sentiments on the basis of sentiment score is highlighted. The platform used here is Hortonworks Data Platform (HDP) which gives a platform to operate Hadoop on Windows. The preprocessing stage is divided into two phases.

1. In the first phase of preprocessing the subject of the tweets to be retrieved from the source data is finalized and then 'URL links' and '@' are removed while ingesting the data into HDFS via Apache Flume so as to reduce the load in the HDFS. Flume provides end to end connectivity, hence removing the risk of losing data during the data ingestion process.
2. After completion of the first step, the HDFS will contain the Raw Tweets, Time-zone Map and Dictionaries which are not schematized. Since there are so many field in the raw tweets we schematize only those parts that we need. For example, ID, retweet count, status, location, year,

month, etc. The same will be done for Dictionaries and time-zone map which will be a one-time process.

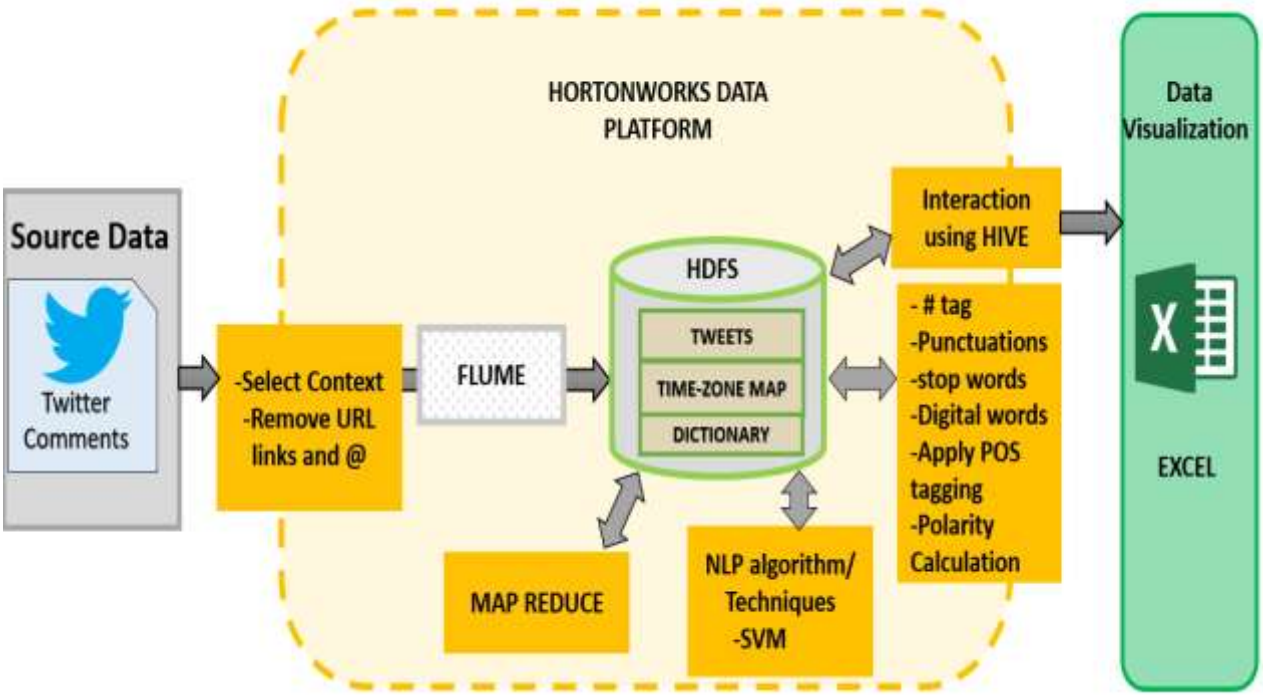

Fig. 5: Proposed System

3. The second phase of preprocessing will commence when the data is ingested in HDFS.

   a. *Stop word removal*: Stop words are those words which generally do not carry any potential useful information but are added to get the grammar of the sentence.

   b. *Repeated letters removal:* If a letter is repeated more than twice consecutively, the number of its occurrence is reduced to two.

   c. *POS tagging:* Once the data is cleansed, POS (Parts of Speech) tagging is done. POS tagging helps in identifying the part of the word in a sentence.

   d. *Digital Words Removal:* Digital words such as dates (28-03-2017) which do not contribute in calculating sentiments will be removed.

4. After applying POS tagging (using Stanford tagger), we now have the part-of-speech that a particular word belongs to. Combining it with information in SentiWordNet dictionary yields sentiment score for each word.
5. It is important to identify the name of the country where the tweet originated. This will ease localization of tweets to conduct country-wise sentiment analysis. It is done by joining time-zone map and the preprocessed view table.
6. The output of (5) which is in form of a key-value pair for each constituent word is then fed to Map-Reduce for calculating the final score in parallel.
7. After obtaining the final score, SVM algorithm is used to classify the sentiments into Ekman's 6 basic emotions: anger, disgust, fear, happiness, sadness and surprise.
8. To interact with the HDFS, we use Apache Hive which gives an SQL-like interface. Apache Hive facilitates data processing and summarization and provides for persistent, online storage.
9. And lastly, Microsoft Excel is used for visualizing the data in pie charts and 3D maps for displaying country-wise sentiment analysis. Besides Excel, the other proprietary BI visualization tools available in the market may also be used for better visualization.

5. Conclusion

A large amount of unattended unstructured data needs some structural arrangement in order for the data to be processed. The number of people interacting via social media are increasing in time and hence resulting in a large amount of data. Analyzing the sentiments from this voluminous data is challenging. The work of research in the field of sentiment analysis using Hadoop includes the analysis and classification of the sentiments into 3 groups, viz, positive, negative and neutral. This paper also attempts to classify sentiments into Ekman's six basic emotions, viz, anger, disgust, fear, happiness, sadness and surprise. Hadoop ensures the distributed processing and reduces the overall access time; resulting in a better response time. The paper presents an approach to sentiment analysis on Hadoop Framework and describes the activities in conducting sentiment classification and summarization. It is expected that the experimentation results will be encouraging and will provide a better insights into the working of the proposed approach.

References

[1] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification Of Reviews," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 417-424.

[2] Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." *Proceedings of the Association for Computational Linguistics (ACL)*, 2004

[3] Borikar D. A. and Chandak M. B. (2016), "An Approach to Sentiment Analysis on Unstructured Data in Big Data Environment." In: Unal A., Nayak M., Mishra D., Singh D., Joshi A. (eds) *Smart Trends in Information Technology and Computer Communications. SmartCom 2016. Communications in Computer and Information Science, vol 628. Springer, Singapore*

[4] A. C. E. S. Lima, L. N. de Castro and J. M. Corchado. "A polarity analysis framework for Twitter messages", in *Applied Mathematics and Computation, vol. 270, 2015, pp. 756–767.*

[5] Theresa Wilson, Janyce Wiebe and Paul Hoffmann, "Recognizing Contextual Polarity in Phrase-level Sentiment Analysis." *Proceedings of the conference on human language technology and empirical methods in natural language processing, ACL,2005*

[6] Batool, Khattak, Maqbool and Sungyoung Lee, "Precise tweet classification and sentiment analysis." *Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on , vol., no., pp.461,466, 16-20 June 2013*

[7] Xiaoqian Zhang, Shoushan Li, Guodong Zhou and Hongxia Zhao, "Polarity Shifting: Corpus Construction and Analysis." *Asian Language Processing (IALP), 2011 International Conference on , vol., no., pp.272,275, 15-17 Nov. 2011*

[8] Kumar Singh, Sachdeva, Mahajan, Pande and Sharma, "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites." *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference, vol., no., pp.329,335, 25-26 Sept. 2014*

[9] Go. A, Bhayani. R and Huang. L, "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford (2009)*

[10] Pak and Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." *Proceedings of LREC 2010 (2010)*

[11] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau , "Sentiment Analysis of Twitter Data" *Department of Computer Science Columbia University New York, NY 10027 USA* fapoorv@cs, xie@cs, iv2121@, rambow@ccls, becky@csg.columbia.edu 2011

[12] Sunil B. Mane, YashwantSawant, SaifKazi and VaibhavShinde, "Real Time Sentiment Analysis of Twitter Data Using Hadoop" *College of Engineering, Pune International Journal of Computer Science and Information Technologies 2014*

[13] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retrieval, vol.2, Nos. 1-2 (2008) 1-135, 2008 DOI: 10.1561/1500000001.*

[14] Subramaniyaswamy V, Vijayakumar V, Logesh R and Indragandhi V, "Unstructured Data Analysis on Big Data using Map Reduce", *2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), ScienceDirect 2015.*

[15] Tanvi Hardeniya and D. A. Borikar, "An Approach To Sentiment Analysis Using Lexicons With Comparative Analysis of Different Techniques", *IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 53-57*

[16] Piyush Gupta, Pardeep Kumar and Girdhar Gopal, "Sentiment Analysis on Hadoop with Hadoop Streaming", *International Journal of Computer Applications(0975-8887) Volume 121-No.11, July 2015*

[17] Deebha Mumtaz and Bindiya Ahuja, "Sentiment Analysis of Movie Review Data Using Senti-Lexicon Algorithm", *2nd International Conference on Applied and Theoretical Computing and Communication Technology, IEEE 2016*

[18] Kyong-Ha Lee, Yoon-Joon Lee,Hyunsik Choi, Yon Dhn Chung and Bongki Moon, "Parallel Data Processing with MapReduce: A Survey", *SIGMOD Record, December 2011, (Vol. 40, No.4)*

[19] E.Kouloumpis, T.Wilson and J.Moore, " Twitter Sentiment Analysis: The Good the Bad and the OMG! ", *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011*

[20] ZHAO JIANQIANG and GUI XIAOLIN, "Comparison Research on Text Pre-Processing Methods on Twitter Sentiment Analysis," *DOI: 10.1109/ACCESS.2017.2672677*

[21] Bo Pang and Lillian Lee and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Appears in Proc. 2002 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*

[22] Manning, Christopher D and Hinrich Schutze, "Introduction to information retrieval," *Cambridge University Press 2008.*

[23] "The Ultimate Hands-On Hadoop – Tame your Big Data!" *https://www.unanth.com.*