DOI 10.29042/2018-3797-3801

Helix Vol. 8(5): 3797- 3801

Sentiment Analysis of Tweets at Sentence Level Using Hadoop

*[1]Yazala Ritika Siril Paul, [2]Dilipkumar A. Borikar
[1, 2] Shri Ramdeobaba College of Engineering and Management, Nagpur, India
*Email: *paulys@rknec.edu, borikarda@rknec.edu*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract
In the last couple of decades Sentiment Analysis has attracted considerable amount of interest from the research community across the globe. Sentiment Analysis brings to light the underlying point of view(s) in a text; for example classifying a review as positive, negative or neutral. The main aim is to extract a set of potential confident features from the review and then classify them into emotions which they strongly depict. In this paper we dig deeper into analysing the sentiments and classify them into Ekman's six basic emotions i.e. Anger, Fear, Disgust, Sadness, Happiness and Surprise using Hadoop with the assistance of Apache Flume, Apache Hive.

Keywords: Apache Hadoop, Apache Flume, Apache Hive, Sentiment/Emotion Analysis, Unstructured Data

Introduction

Starting from the advent of 1990s, which is marked as the beginning of transmission of modern Internet, the usage of internet has been increased in various forms. In today's world since textual data is increasing in an unimaginable range, there are many organizations that are trying to use this large amount of data to extract people's opinions towards their products. Social Networking Sites (SNS) proves to the best source for obtaining the data for analysis. However since the data collected from sources is huge, it is highly unfeasible to manually analyze them.

Data is critical to organizations for the immense value that it offers. In the past organizations have considered data uneconomical and the rest have used solutions which limited their capabilities to derive the value from data. In modern times technologies have advanced, businesses have become more dynamic and organizations now want to derive value from the data they always had. It is important for them to identify the data that needs to be captured and stored. With the increase in data, changes in the approach to capture, manage, process and visualize data is highly required. The need to aggregate and analyze this data to derive hidden insights has grown with exponential data growth at an exponential rate. The capability to do this has become the basis of competition and the success that follows.

Sentiment Analysis or opinion mining deals with the computation and classification of opinions, sentiment and the subjectivity that is depicted by the text. The opinion of people is kept in high regards during the decision making process of any organization. This is the reason why opinion mining/sentiment analysis has become a very important issue for researchers. It is indeed a challenge to process a large mass of data which is available on the Web. We consider Twitter as a source for data collection since it is a very popular microblogging site, a means of communication and a collaborative system that allows people to share short text messages. Although, the tweets does not exceed 140 characters; for a small group of users it has been increased to 280 characters. In this paper we aim to analyze the tweets and classify them into Ekman's six basic emotions.

Hadoop: Apache Hadoop is an open-source software framework from the open source community. Various vendor's specific distributions also exists in the market, such as Cloudera distribution of Hadoop, Hortonworks data-platform, IBM Big-Insight etc. Apache Hadoop, popularly known as Hadoop allows distributed storage, horizontal scalability, distributed storage, distributed and parallel processing of very large datasets on commodity machines which form a cluster.

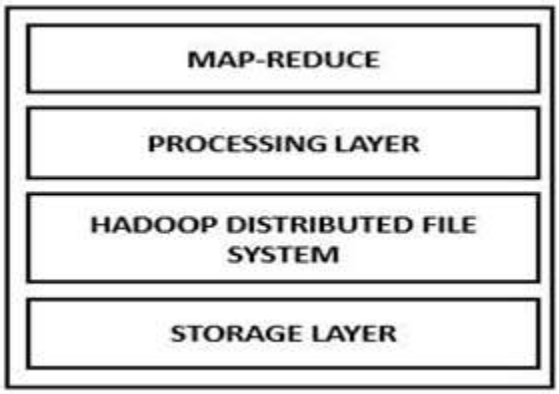

Figure 1. Core of Apache Hadoop

Hadoop Distributed File System (HDFS): HDFS follows the architecture of Master-Slave and has following components. The Name node, which acts as

master, is a commodity hardware which consists of LINUX/UBUNTU/GNU operating system and a name-node software.

Data-node is also a commodity hardware having LINUX/UBUNTU/GNU as the operating system and a data-node software. The responsibility of data-nodes is to manage the data storage of their system. Reading and writing operations on the file systems and also performance of the operations such as deletion, block creation and replication according to the instructions given by the name-node are done by the data-nodes.

Map-Reduce: It is not only a processing technique but also a programming model for distributed computing. The Map-Reduce algorithm consists of two important tasks, viz. Map Task and Reduce Task.

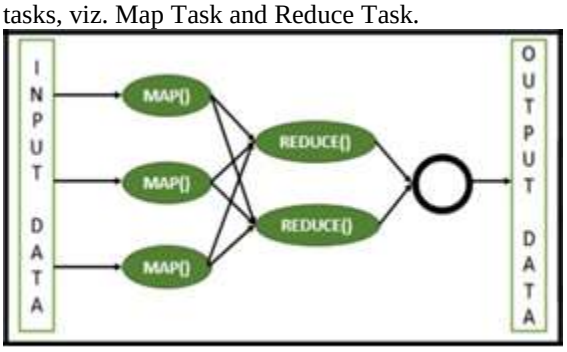

Figure 2. Map-Reduce Framework

The Mapper stage processes the input data which generally is in the form of file or dictionary which is stored in Hadoop File System (HDFS). The input file is processed sequentially by the Mapper function and creates several small chunks of data. The Reducer stage is a combination of shuffling and reducing. The job of Reducer function is to process the data that comes from Mapper function. After processing a new set of output is generated which will be stored in HDFS.

Apache Flume: Flume is an Apache open source framework which is used as a data ingestion mechanism tool. It is a mechanism of moving large amount of data into Hadoop cluster.

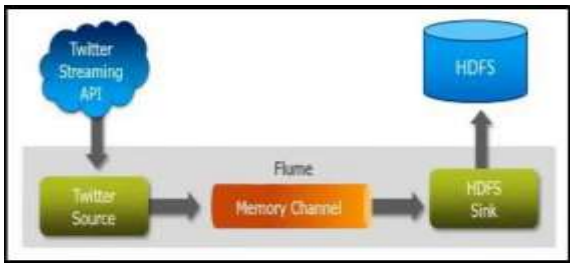

Figure 3. Apache Flume Architecture

Related Work

A lot of research work is done on analyzing the sentiments, rule-based techniques, bag-of-words and on various machine learning methods. The two main directions for research in opinion mining is either document level [1][2] or sentence level analysis [3][4]. Most methods for classification of the documents at the level of sentences are usually based on the identification of terms or phrases of opinions. To do so there are basically two methods:

(1) Rule-based methods and (2) Lexicon-based methods.

Turney [5] introduced a new unsupervised learning algorithm in order to rate a review as thumbs up (recommended) or thumbs down (not recommended). The algorithm consists of three steps: (a) the first step is extracting the phrases which contains adjectives or adverbs. (b) The second step is the core of the algorithm where semantic orientation of each and every phrase is estimated (c) Lastly classifying the reviews based on the average semantic orientation of the phrases. However this algorithm do not extract features on which the opinions have been expressed. It was further processed at sentence level where the sentiment of each document was determined by Hu and Liu [6].

Wilson, Weibe and Hoffmann [7] proposed a new approach for sentiment analysis at phrase level which first determines the expression as neutral or polar and then disambiguates the polarity of the polar expressions. With this approach they were able to classify the polarity of the context for a large subset of sentiment expression automatically.

Zhao and Gui [9] discussed various textual pre-processing methods for classification of sentiments using Twitter dataset. The experiments show that when pre-processing methods like expanding the acronyms and replacement of negation are done, the F1 measure and the accuracy of Twitter sentiment classification classifier are improved, but it barely changes when URL's, stop words or numbers are removed. It was observed that after the application of various pre-processing methods, Naive Bayes and Random Forest classifiers are more sensitive as compared to Logistic Regression and support vector machine classifiers.

In the survey by Lee *et al* [10], they characterized the MapReduce framework and discussed its inherent pros and cons. They introduced its optimization strategies and discussed the open issues and challenges raised on parallel data analysis with MapReduce.

Jamoussi and Ameur [13] classified the sentiments of Facebook comments as positive or negative by using linguistic approach. They covered several sentiment words to create a lexicon since it plays a key role in sentiment analysis and also addressed the problem of

how to group and list the words which are present in the dataset into two separate dictionaries. The final classification into positivity and negativity is done using the dictionaries.

Proposed Approach

In this section we put forth the background and the method we proposed, along with the detailed description of our research work. The data is ingested in real time using Apache Flume in the HDFS from Twitter. In order to perform sentiment analysis, WordNet dictionary and a time-zone map to perform country-wise sentiment analysis is used. For opinion mining we have used EmoSenticNet dictionary which consists of 13190 words with their respective scores for the six basic emotions viz. anger, disgust, fear, happiness, sadness and surprise.

Considering only the subjective tweets, sentiment analysis is performed based on sentence level in the following three stages:

Stage 1: The first phase consists of pre-processing.

Stage 2: The second phase is by using the set of features extracted, create a feature vector.

Stage 3: Lastly machine learning is used for the classification of the tweets into sentiments/ emotions which they strongly depict.

Our pre-processing is divided into two phases. The first phase is applied during the ingestion of tweets in the HDFS using Apache FLUME. In phase 1, subject on which tweets are required for analysis is provided, for example, Hadoop, Bigdata etc. The usage of language filter enables only English tweets to be stored in the HDFS. Due to the first phase, unwanted tweet processing is avoided thus lowering the load on HDFS. The raw tweets loaded in the HDFS is first structured. The same is done with the dictionaries and the time-zone map.

The second phase of our pre-processing includes POS tagging. For POS tagging, we used Stanford tagger. The result obtained will be as follows:

INPUT	The concert was so amazing
OUTPUT	The/**DT** concert/**NN** was/**VBD** so/**RB** amazing/**JJ**

Where DT represents determiner, NN is noun, VBD is verb, RB is adverb and JJ is adjective.

The algorithm mentioned below will separate the required tag set.

ALGORITHM 1 Bag-of-Words (BOW) Task

1: **Function BOW** (String POS_Tag_File):

2: // POS_Tag_File = File containing the result of POS tag

3: **for each** W **AS** [adverb/adjective] **in Tweet do**

4: Bag-of-words-file = W

5: **end for**

6: **return** Bag-of-words-file

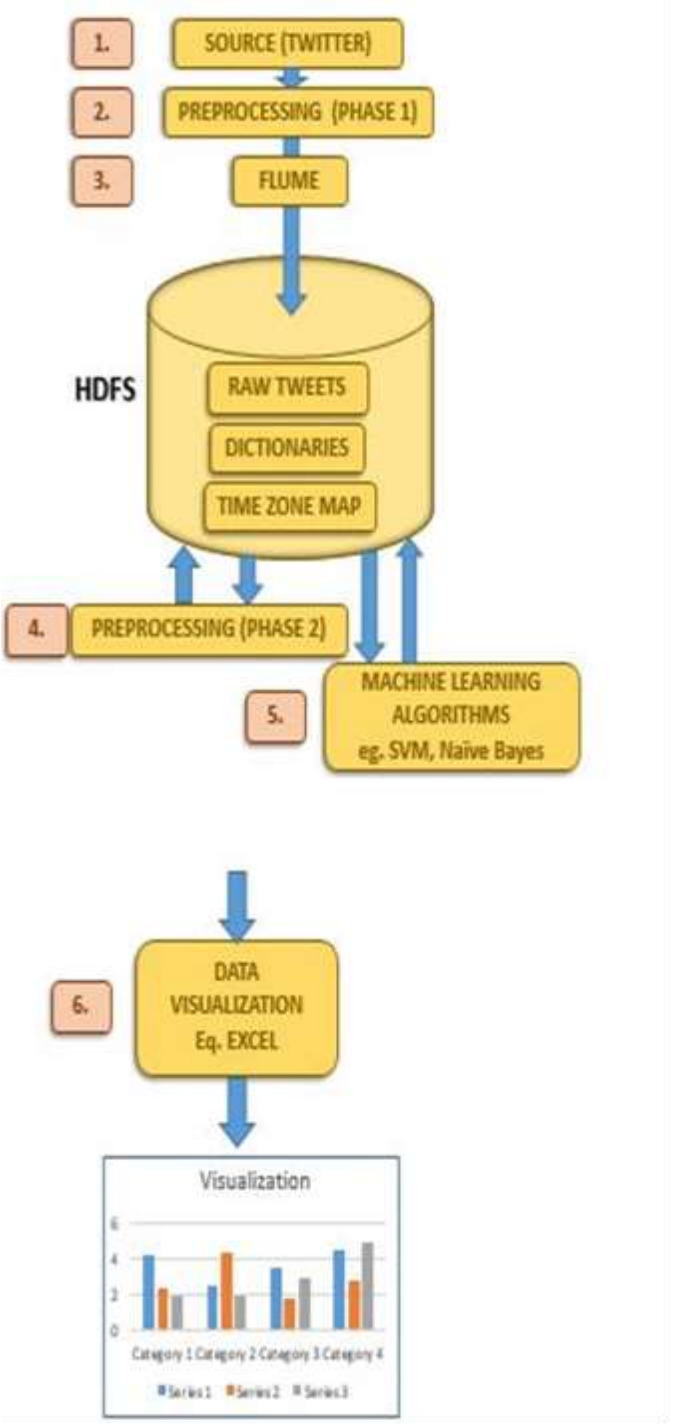

Figure 4. System Architecture

Hadoop uses Map-Reduce algorithm for parallel processing. However the algorithm differs for different analysis. The Map-Reduce algorithm we used is as follows:

ALGORITHM 2 MAPPING Task

1: **Function MAPPER** (String Bag-of-words-file, String file):

2: // Bag-of-words-file: File containing POS tagged tweets (adverbs and adjectives) stored in HDFS

3: //file: File containing sentiment scores stored in HDFS.

4: **for each** Tweet **in** Bag-of-words-file **do**

5: S = Sentiment_score (Tweet)

6: F= Feature (Tweet)

7: **return** (Tweet, F, S)

8: **end for**

Since only words tagged with adverbs and adjectives goes for mapping task, the stop words are excluded for further processing. The next step (REDUCE), works on the sentiment score and the feature for grouping.

ALGORITHM 3 REDUCING Task

1: **Function REDUCER** (int Sentiment_score, Iterator Wordlist):

2: //Sentiment: Sentiments generated by mapper

3: //Wordlist: List of words

4: **for each** feature F **in** Wordlist **do**

5: sum += 1

6: **end for**

7: mean = sum / 2

8: **return** (F, mean, Sentiment)

Experimental Results and Analysis

We consider a dataset which consists of sentences with their underlying sentiments representing Ekman's six basic emotions (anger, disgust, fear, sadness, happiness and surprise). The supervised machine learning techniques used for the classification are SVM (Support Vector Machine), Multinomial Naïve Bayes and Bernoulli Naïve Bayes.

The dataset used consists of 10000 tweets with 1000 tweets each for emotion anger and disgust and 2000 tweets for emotions fear, surprise, happiness and sadness.

CLASSIFIER	EMOTION	PRECISION	RECALL	F-MEASURE	ACCURACY
SVM	ANGER	0.8549	**0.7783**	**0.8148**	**81.0431**
	FEAR	**0.7908**	0.8432	**0.8161**	
	SADNESS	0.7971	**0.8578**	**0.8264**	
	DISGUST	0.756	**0.6078**	0.6739	
	HAPPINESS	**0.8281**	**0.7718**	**0.7989**	
	SURPRISE	0.7924	**0.9545**	**0.8659**	
MULTINOMIAL NAÏVE BAYES	ANGER	**0.875**	0.6933	0.7736	74.7241
	FEAR	0.6692	**0.8954**	0.766	
	SADNESS	0.7244	0.8274	0.7725	
	DISGUST	**0.96**	0.4705	0.6315	
	HAPPINESS	0.7612	0.5728	0.6537	
	SURPRISE	**0.9**	0.8181	0.8571	
BERNOULLI NAÏVE BAYES	ANGER	0.6806	0.7641	0.72	74.8244
	FEAR	0.702	0.8292	0.7603	
	SADNESS	**0.8277**	0.7563	0.7904	
	DISGUST	0.775	**0.6078**	**0.6813**	
	HAPPINESS	0.8197	0.6844	0.746	
	SURPRISE	0.8928	0.5681	0.6944	

Table 1: shows the analysis results.
***Highest Precision, Recall, F-measure and Accuracy for Each Emotion are shown in Bold.**

The above results show that SVM yielded a total accuracy of 81.0431% whereas Multinomial Naïve Bayes and Bernoulli Naïve Bayes achieved an accuracy of 74.7241% and 74.8244% respectively.

Conclusion

In this research work, we have proposed a framework for further classifying the sentiments into Ekman's six basic emotions using Hadoop. Language filter was applied during the data ingestion in order to reduce load in HDFS. The unstructured ingested data that is stored in HDFS, is first structured and then processed by using MapReduce algorithm. The classification of emotions is performed by using machine learning techniques like SVM, Multinomial Naïve Bayes and Bernoulli Naïve Bayes earning maximum accuracy of 81.0431% from SVM.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up Sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 79–86, 2002.

[2] Kushal Dave and Steve Lawrence and David M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," In *WWW*, pages 519.528, 2003.

[3] M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining customer opinions from free text. IDA'2005.

[4] S. Kim and E. Hovy. Determining the Sentiment of Opinions. COLING'04, 2004.

[5] P. Turney. "Thumbs Up or Thumbs Down. Semantic Orientation Applied to Unsupervised Classification of Reviews,"ACL'02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Pages 417-424, 2002.

[6] Hu M. and Liu B. "Mining and Summarizing Customer Reviews," KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 168-177.

[7] Wilson T., Wiebe J. and Hoffmann P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In the Advanced Research and Development Activity (ARDA).

[8] 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

[9] Zhao Jianqiang and Gui Xiaolin. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. DOI:10.1109/ACCESS.2017.2672677

[10] 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)

[11] 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

[12] 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

[13] Jamoussi, S. Ameur, H. "Dynamic construction of dictionaries for sentiment classification," Cloud and Green Computing (CGC), 2013 Third International Conference on, vol., no., pp.418,425, Sept. 30 2013-Oct. 2 2013

[14] Manning, Christopher D, Hinrich Schutze, Introduction to information retrieval, Cambridge University Press 2008.

[15] Batool, Khattak, Maqbool, 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

[16] Yazala Ritika Siril Paul and Dilipkumar A. Borikar, "An Approach to Twitter Sentiment Analysis Over Hadoop," International Conference on Recent Trends in Engineering Sciences (ICRTES), February 20, 2018. [Presented]. To be published in coming issues Vol 8 of International Journal of Engineering Technology (UAE) [SCOPUS Indexed].

