International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.289-298
http://dx.doi.org/10.14257/ijdta.2016.9.12.26

Extracting the Sentiment Score of Customer Review from Unstructured Big Data Using Map Reduce Algorithm

Syed Imtiyaz Hassan

Department of Computer Science & Engineering, Hamdard University, New Delhi, India.
s.imtiyaz@gmail.com

Abstract

Big Data is a term used to identify the datasets that due to their large size, is very difficult to manage with traditional techniques. This data may be in the order of magnitude of petabytes. It can be found easily on web, especially on social media in the form of customer blogs, reviews and comments. Generally it is unstructured data or semi-structured data. One can use this big data to generate values by calculating sentiment score. Map Reduce is one of the most popular algorithm in Hadoop environment to perform such task. The objective of present research is to automate the process of extracting sentiments expressed about specific features of a product. For this purpose three datasets generated by Amazon for different types of electronics product reviews has been used. The data sets used consists of reviews of the products Nikon Coolpix 4300 Camera, Nokia 6601 mobile and the Canon G3camera. Map Reduce algorithm on Hadoop environment that is considered faster, reliable and fault-tolerant for processing big amounts of data in-parallel on large clusters, has been used to extract sentiment score.

Keywords*: Big Data, Opinion Mining, Map Reduce, HDFS, Unstructured Data.*

1. Introduction

Huge number of users are active globally on the web. They posts their reviews on products which help other users in taking decision through them. The numbers are increasing day by day with the increase in computer literacy in developing countries. As a large numbers of reviews are available for a single product which makes it difficult for a customer to see each and every reviews and make a decision according to them. Thus, mining this data, identifying the user opinions and classify them is an important task.

Opinion mining are very effective in such circumstances. It helps promoters to evaluate the success of an ad campaign or new product, to decide which versions of a product or service are popular and to identify which demographics like or dislike specific product features. For example, a review on a website might be broadly positive about a digital camera, but be particularly negative about how heavy it is. Being able to identify this kind of information in a systematic way gives the supplier a much clear picture of opinion of public than surveys or focus groups do, because the data is created by the customer.

Opinion mining (also called sentiment analysis), involves building a system to collect and categorize opinions about a product [1]. Opinion Mining aims to obtain feelings of the writer expressed in positive or negative comments by analysing a large number of documents [2]. It is a technique of computation to quantify sentiments of people by assigning appropriate scores to positive and negative opinion about the product [3]. Therefore, the main task of sentiment analysis is to classify the documents and specify its polarity. Polarity can be represented as positive, negative or neutral. Automated opinion mining often uses machine learning to mine sentiment of a document [4]. Map Reduce algorithm on Hadoop environment is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The objective of present research is to automate the process of finding sentiments expressed about specific features of a product stored in unstructured big data.

2. Literature Review

There are significant research in the direction of classification of unstructured data using machine learning. The fact that big data analytics is becoming an important tool to improve efficiency and quality in organizations and its importance is going to increase in the next few years were already predicted by [1]. In the year 2002, [5] considered the problem of classifying documents by overall sentiment to find whether a review is positive or negative. Further, a method for sentiment analysis was proposed by [6]. [6] has mined the product features based on customer's comment and then identified opinion sentences in each review to make a decision whether the opinion of the sentence is positive or negative. Finally it produced summarized result. A novel technique for sentiment classification with polarity shifting detection was proposed by [7]. This counting-based classifier significantly improves the performance of sentiment analysis across the different domains. Feature based opinion mining techniques were again explored by [8]. For this purpose an application interface "postbuk" was used for mining and authenticating the comments extraction from the Facebook user and server. The mining datasets were pre-processed by NLP and then classified by using support vector machine classification technique. The Cassandra, a big data's Hadoop framework were used to collect the large datasets from Facebook.

A number of researchers used Twitter for extracting comments. It has been analyzed by [4] that Naive Bayes and Maximum Entropy machine learning classification on Twitter has accuracy above 80% when trained with emoticon data. Further a work by [3] has explained how twitter data can be tested and classified into different classifications. Utility of linguistic features for detecting the sentiment of messages on Twitter was also investigated by [9]. Down the line Hadoop has been used as tool for sentiment classification. Hadoop were used by [10] to provide a way of sentiment analysis for processing the huge amount of data on a Hadoop cluster for faster in real time execution. Hadoop was again find its place in [11]. MapReduce were used to perform key-value pair level incremental processing in this work. This technique not only support one-step computation but also more sophisticated iterative computation, which is widely used in data mining applications.

3. Unstructured Big Data

Some well-known companies like LinkedIn, Yahoo, Google, and Amazon are generating large quantity of structured, semi-structured and unstructured data every day [5]. Among these, unstructured data covers a large portion of data. Unstructured (information) data or unstructured information is a term used for information that either did not have a pre-defined model of data or it is not well organized in a pre-defined manner. The some of the sources of these unstructured data are: sensors which are used to gather climate information, data from social media sites, digital pictures and videos, records cell phone GPS signals, and , purchase transaction to name a few. This data is termed as big data [8]. IDC – EMC Digital Universe study published that data will grow to 44 trillion gigabytes by 2020 among them 70-80% might consists of unstructured information [12].Unstructured information is typically text-heavy, but may contain data such as numbers, dates and other facts as well. It is very difficult to comprehend using traditional programs as compared to data stored in fielded form in databases or annotated or semantically tagged in the documents. That is why it is difficult to utilize unstructured data properly. So it is very important to utilize such data as it can give useful information after analysis [9].

4. Sentiment Classification

Sentiment Classification broadly refers to binary categorization, multi-class categorization, regression and ranking. Sentiment Classification mainly consists of two important tasks: sentiment polarity assignment and sentiment intensity assignment [2]. Sentiment polarity assignment deals with analyzing, whether a text has a positive, negative, or neutral semantic orientation. Sentiment intensity assignment deals with analyzing, whether the positive or negative sentiments are mild or strong. The words which have positive impact on the opinion of the analyzer are counted as positivity contributing word. The words which have negative impact on the opinion of the analyzer are counted as negativity contributing word. The words which are having no impact on the opinion of the analyzer are counted as neutral and have no use in analysis. Sentiment classification can be performed on three basic levels: [13]

- *Document level:* It classify the whole document as positive, negative or neutral. Classification is performed for the complete document and then decision is taken whether the document shows positive or negative sentiment. The basic unit of information is always a single document of opinionated text.
- *Sentence level:* It classify the complete sentences as positive, negative or neutral. Classification is performed for the complete sentence instead of document.
- *Aspect & Feature level:* It classify sentences/documents as positive, negative or neutral based on the aspects of those sentences/documents commonly termed as aspect-level sentiment classification.

5. Distributed Processing of Huge Datasets on Hadoop

Hadoop is an open source framework written in Java by Apache that allows distributed processing of huge datasets across clusters of computers using simple programming models. It is one of the processing tools that is used to analyze and process large data sets [14]. A Hadoop application works in an environment that provides distributed storage and evaluation across clusters of node or computers. Hadoop is designed to scale up from single server to thousands of machines, each giving local computation and storage [10]. In short, Hadoop framework is able enough to develop applications and capable for running on clusters of computers and they can perform statistical analysis for large amounts of data set.

It consists of two main components: MapReduce, which is used to process the data; and Hadoop Distributed File System (HDFS) for storage of data set. MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce framework operates on <key, value> pairs. It views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types. The term MapReduce actually refers to the following two different tasks that Hadoop programs perform:

- *The Map Task:* This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples.
- *The Reduce Task:* This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

The map-reduce methodology used first by Google to perform web crawling at a faster rate.

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. It is a Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers. It has demonstrated production scalability of up to 200 PB of storage and a single cluster of 4500 servers, supporting close to a billion files and blocks [15].

Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.

6. Extracting the Sentiment Score of Customer Review using Map Reduce

For processing and analyzing large data sets Map Reduce (on Hadoop) is considered one of the most powerful tool for opinion mining [13]. A method to calculate and analyse comments or review given by the customers or user is implemented in Java (on Hadoop). The implemented system works in two phases: The Mapper phase and the Reducer phase. A positive and negative word dictionary is used to identify positive and negative words and also a Stop word dictionary is used to identify stop words from the reviewed product [16]. The system can extract useful information out of unstructured files which go unused most of the time but contain useful information. Processing unstructured file make the results more refined and accurate. The major modules and subsystems used involved are:

a. Mapper

It mapper input key/value pairs are set to intermediate key/value pairs. Mapper will prepare the data for processing inside the Reducer. The sub-components of Mapper are:

File parser: It will split the data into records and records in turn will split into individual key-value pair.

Information Extractor: It gives the key/value pairs that are used to do the sentiment analysis.

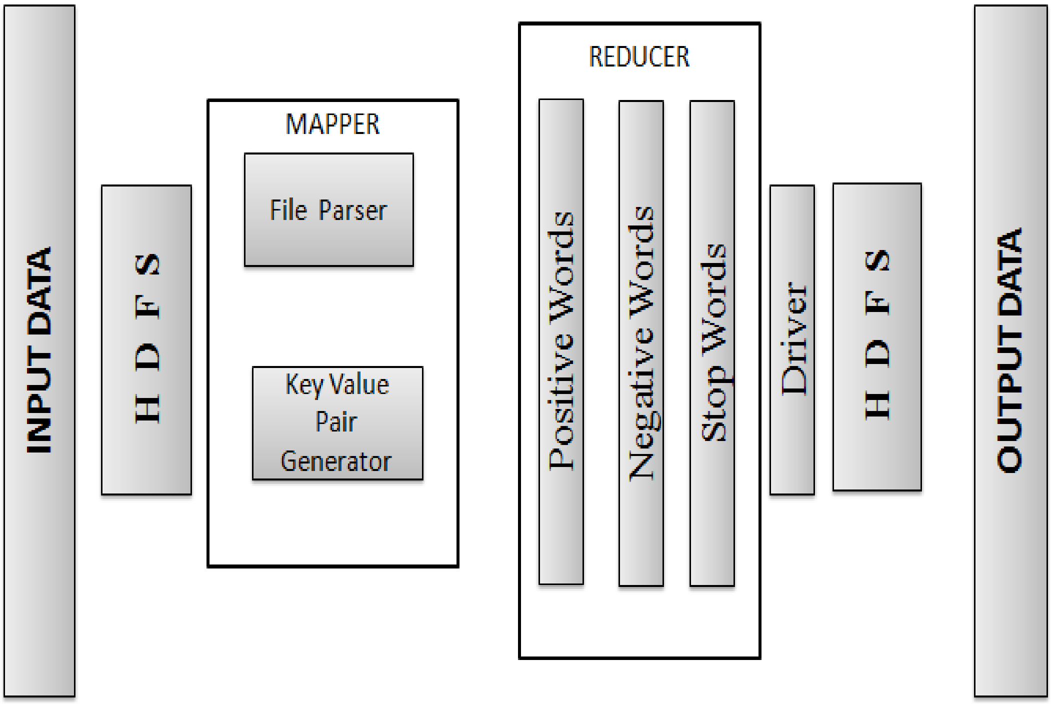


Figure 1. The Major Modules and Subsystems

b. Sorter

It will sort the key/value pairs generated by the mapper in intermediate steps. This sorted key/value pairs acts as the input to the Reducer phase.

c. Reducers

Input files are analyzed by Reducer using following subsystems:

Positivity and Negativity Calculator: For instance, "*this xyz product is good*." – in this sentence good is positive and hence results in a positive sentence. So here we make the polarity of a word positive. We increase the polarity strength when a word is positive. Thus positive word = +1. For instance, "*this xyz product is bad*" – in this sentence bad is negative and hence results in a negative sentence [17]. So here we reverse the polarity of a word. We decrease the polarity strength when a word is negative. Thus negative word = -1.

Stop word extractor: Stop words are the words which do not participate in the sentiment analysis, so removal of these words will not affect the experimental results.

d. *Polarity calculator:* The score calculator is the main part of proposed work. It uses positive, negative and stop-word dictionary to extract the words which change the opinion of analyzer from each sentence of the document. The Polarity Calculator identifies whether a document is positive or negative. It will calculate the positivity and negativity of a document, if we have more negativity then overall reviews are negative otherwise positive. After calculating polarity the output will be stored in HDFS.

7. Experimental Setups

The data set used for experiment are different types of electronics product reviews available on Amazon [18]. Three data sets are used consisting of reviews of the products

Nikon Coolpix 4300 Camera, Nokia 6601 mobile and the Canon G3camera. The size of above three text files are depicted in Table 1.

Table 1. File Size

Record No.	File Name	Size of File (Approx)	Record No.	File Name
1	Nokia 6610.txt	54 KB	1	Nokia 6610.txt
2	Nikon Coolpix4300.txt	38 KB	2	Nikon Coolpix4300.txt
3	Canon G3.txt	64 KB	3	Canon G3.txt

The data available is of unstructured file format i.e. in simple text file format. The comments are all starting with ## in simple text file. The reviews contain positive, negative and stop words. Positive and negative word dictionaries are used to filter out positive and negative words [2] [19]. Skip words are left unused for the sake of simplicity in this experiment.

Eclipse is a very powerful IDE for Java development. That is why it is used as an experiment tool since Hadoop and Map Reduce programming is done in Java in general.

8. Execution Methodology and Findings

A data set of Amazon for a particular product is selected. The data is then stored into HDFS. The data can be in unstructured format consisting of reviews by the user. The input dictionaries are maintained for positive, negative and stop words including emoticons for the respective emotions. Positive words may be such as accomplish, achievable, achievement, abound, achievable, accurately etc. Negative words consist of negative words like absurdness, accept, abused, accident, abolish etc. A neutral word like a, is, to, but . , ! etc. Map Reduce programming techniques is used for mining of data fed into the system. For this purpose the content of input file is read inside mapper and converted into tokens of <Key, Value> pair e.g. <Good,9> and then reducer calculates the sum of all the words that have positive effect and negative effect on the opinion. Since analysis is done at document level, the sentence level classification of sentiments is not required and the decision is taken for the complete document. The result will consist of sentiment score that falls between 0 and 1. The formula used for calculating sentiment score is:

Sentiment Score = (Positive - Negative) /(Positive + Negative) ………………….. *(1)*

Positivity of a document is calculated as:

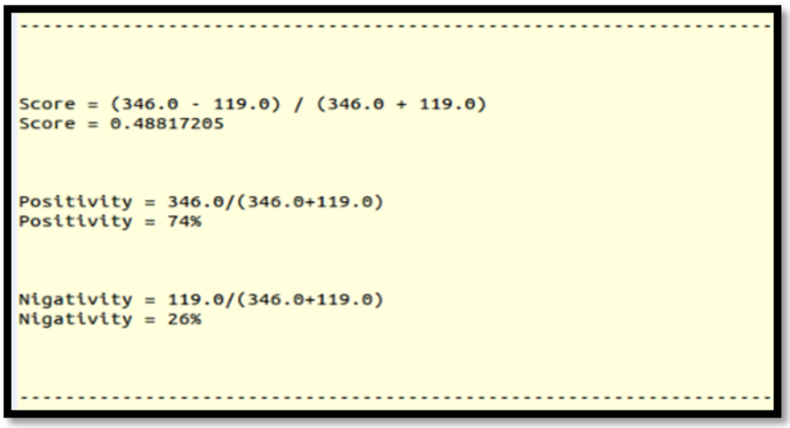
Positivity = Positive / (Positive + Negative) ……………………………………. (2)

Negativity inside a document is calculated as:

Negativity = Negative / (Positive + Negative) ………………………… ……….. (3)

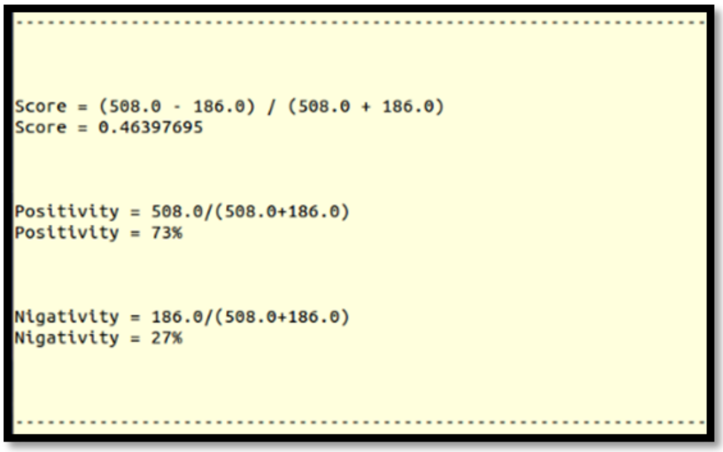
Here Positive consists of sum of all the positive words in the document. Negative words consist of sum of negative words in the document. Neutral words consist of the set of all the words which are to be skipped while evaluation.

The output of the analysis is given as:



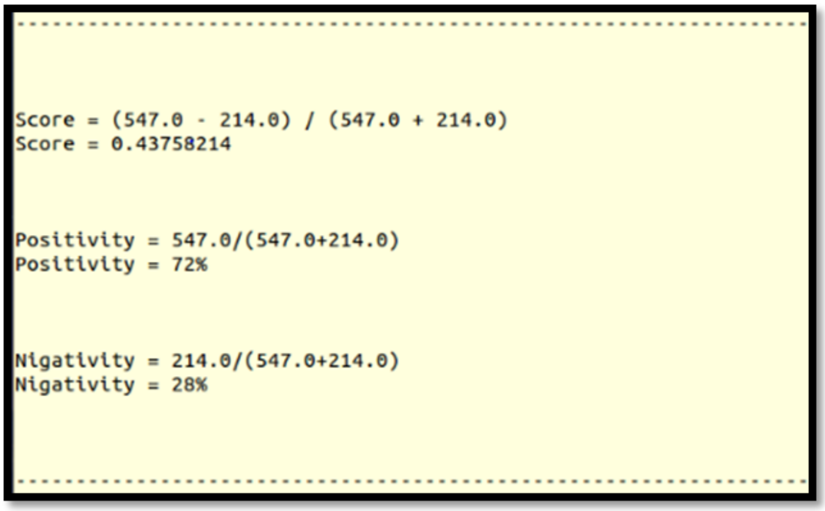
```
-----------------------------------------------------------------


Score = (346.0 - 119.0) / (346.0 + 119.0)
Score = 0.48817205


Positivity = 346.0/(346.0+119.0)
Positivity = 74%


Nigativity = 119.0/(346.0+119.0)
Nigativity = 26%


-----------------------------------------------------------------
```

Figure 2. Output for Nokia 6610.txt



```
-----------------------------------------------------------------


Score = (508.0 - 186.0) / (508.0 + 186.0)
Score = 0.46397695


Positivity = 508.0/(508.0+186.0)
Positivity = 73%


Nigativity = 186.0/(508.0+186.0)
Nigativity = 27%


-----------------------------------------------------------------
```

Figure 3. Output for Nikon Coolpix4300.txt



```
-----------------------------------------------------------------


Score = (547.0 - 214.0) / (547.0 + 214.0)
Score = 0.43758214


Positivity = 547.0/(547.0+214.0)
Positivity = 72%


Nigativity = 214.0/(547.0+214.0)
Nigativity = 28%


-----------------------------------------------------------------
```

Figure 4. Output for Canon G3.txt

The Summary of the findings has been depicted in Table 2.

Table 2. Summary of Findings

S No.	File Name	Sentiment Score	Program Output Positivity	Program Output Negativity
1.	Nokia 6610.txt	0.48817205	74%	26%
2.	NikonCoolpix4300.txt	0.46397695	73%	27%
3.	Canon G3.txt	0.43758214	72%	28%

9. Conclusion

The large amount of data generated on online shopping portals and social media sites can be used for mining of opinion of users for a specific product. As the data is growing at a very fast speed nowadays, it is required to process this huge amount of data at a faster speed. The generated data may be of any file format structured, unstructured or semi structured it can be processed very easily with Hadoop. This research tried to generate some values from unstructured big data using Map Reduce on Hadoop. For this purpose the gathered online end user reviews of different types of electronics products, Nikon Coolpix 4300 Camera, Nokia 6601 mobile and the Canon G3camera were analyzed. This involves the filtering of irrelevant and unhelpful reviews, measuring the sentiments of thousands of (useful) reviews and finally, providing the end user summarized data about the expressed sentiments in terms of sentiment scores, positivity and negativity.

References

[1] A. Bifet, “Mining Big data in Real time”, Informatica, vol. 37, **(2013)**, pp. 15- 20.

[2] C. Cardie, “Empirical methods in information extraction”, AI Magazine, vol. 18, no. 4, **(1997)**, pp. 65-79.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, “Sentiment analysis of Twitter data”, Proceedings of ACL Workshop on Languages in Social Media, **(2011)**, pp. 30-38.

[4] A. Go, R. Bhayani and L. Huang, “Twitter sentiment classification using distant supervision”, Technical report, Stanford, **(2009)**.

[5] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques”, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, ACM, Stroudsburg, PA, USA, pp. 79-86. DOI=http://dx.doi.org/10.3115/1118693.1118704, vol. 10, **(2002)**.

[6] M. Hu and B. Liu, “Mining and summarizing customer reviews”, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), ACM, New York, NY, USA, DOI=http://dx.doi.org/10.1145/1014052.1014073, **(2004)**, pp. 168-177.

[7] S. Li, Z. Wang, S. Y. M. Lee and C. R. Huang, “Sentiment Classification with Polarity Shifting Detection”, Proceedings of International Conference on Asian Language Processing (IALP), **(2013)**, pp. 129,132.

[8] S. G. Grivas, M. Kaschesky and M. Schaaf, “Feature based Opinion mining - towards Performance Measure”, Proceedings of the IEEE International Journal of Advanced Computer Research, **(2013)**.

[9] K. Efthymios, W. Theresa and D. M. Johanna, “Twitter Sentiment Analysis: The Good the Bad and the OMG!”, Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, AAAI Press, **(2011)**, pp. 538-541.

[10] S. B. Mane, Y. Sawant, S. Kazi and V. Shinde, “Real Time Sentiment Analysis of Twitter Data Using Hadoop”, International Journal of Computer Science and Information Technologies, vol. 5, no.3, **(2014)**, pp. 3098-3100.

[11] Y. Zhang, S. Chen, Q. Wang and G. Yu, “i2MapReduce: Incremental MapReduce for Mining Evolving Big Data”, IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 7, **(2015)**.

[12] EMC – EDC Digital Universe [Online]. Available: http://www.emc.com/leadership/digital-universe/index.htm?pid=landing-digitaluniverse-131212

[13] A. Khan and B. Baharudin, "Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs", Int. J Comp Sci. Emerging Tech, vol. 2, no. 4, **(2011)**, pp. 539-552.

[14] Meg Cater (Apr 28, 2015), Move Large Files Fast: Overcoming the challenge of transferring huge unstructured data sets, [Online]. Available: http://www.signiant.com/blog/move-large-files-fast-overcoming-the-challenge-of-transferring-huge-unstructured-data-sets

[15] Apache Hadoop, [Online]. Available: http://hortonworks.com/apache/hdfs/

[16] C. Potts, "On the negativity of negation", in Proc. of SALT, **(2011)**, pp. 636-659.

[17] D. Davidov, O. Tsur and A. Rappoport, "Enhanced sentiment learning using Twitter hashtags and smileys", Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10), Association for Computational Linguistics, Stroudsburg, PA, USA, **(2010)**, pp. 241-249.

[18] Customer Review Datasets (5 products), [Online]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[19] C. C. Yang, Y. C. Wong and C. P. Wei, "Classifying web review opinions for consumer product analysis", Proceedings of the 11th International Conference on Electronic Commerce (ICEC '09), ACM, New York, NY, USA, DOI=10.1145/1593254.1593263, **(2009)**, pp. 57-63.

Author

Syed Imtiyaz Hassan is working as an Assistant Professor at Department of Computer Science & Engineering, Jamia Hamdard, New Delhi (India). He is having more than 15 years of professional experiences of teaching, research and project supervisions. He has supervised more than 100 students for inter disciplinary research and industrial projects. He has published more than 20 research papers in various national and international conferences and journals. He is the member of editorial boards of various international and national journals and also reviewed many research papers. His area of research are Smart & Sustainable City, Machine Learning, Smart Technologies and Software Engineering.