Real-time Sentiment Analysis of Big Data Applications Using Twitter Data with Hadoop Framework

Divya Sehgal and Ambuj Kumar Agarwal

Abstract Twitter and other social networking sites generate huge amount of data on a daily basis. Frequent interactions can be witnessed with data generated through Linked-in, facebook, and gmail. Twitter being the largest social networking site generates data of very large amount because of millions of tweets and followers which are increasing per day. This imposes a big problem of processing and analyzing the data. As it is a case of handling big data, the technology of Hadoop comes into picture. Using Hadoop eases the process of analyzing the data. The work of analyzing twitter data is undertaken in the paper.

Keywords Big data · Hadoop · HDFS · Map reduce

1 Introduction

Today, the data generated through social networking sites is increasing day by day. Twitter one of the most popular social networking platforms deals with both structured and unstructured format of data. However, Twitter data is mostly in unstructured format like followers, tweets, likes, and expressions, etc. It is very difficult to process the data easily. All kinds of industries and companies are using this type of data for the future development and advertising work.

D. Sehgal (✉) · A.K. Agarwal
CCSIT, Teerthanker Mahaveer University, Moradabad, India
e-mail: sonasahgal199@gmail.com

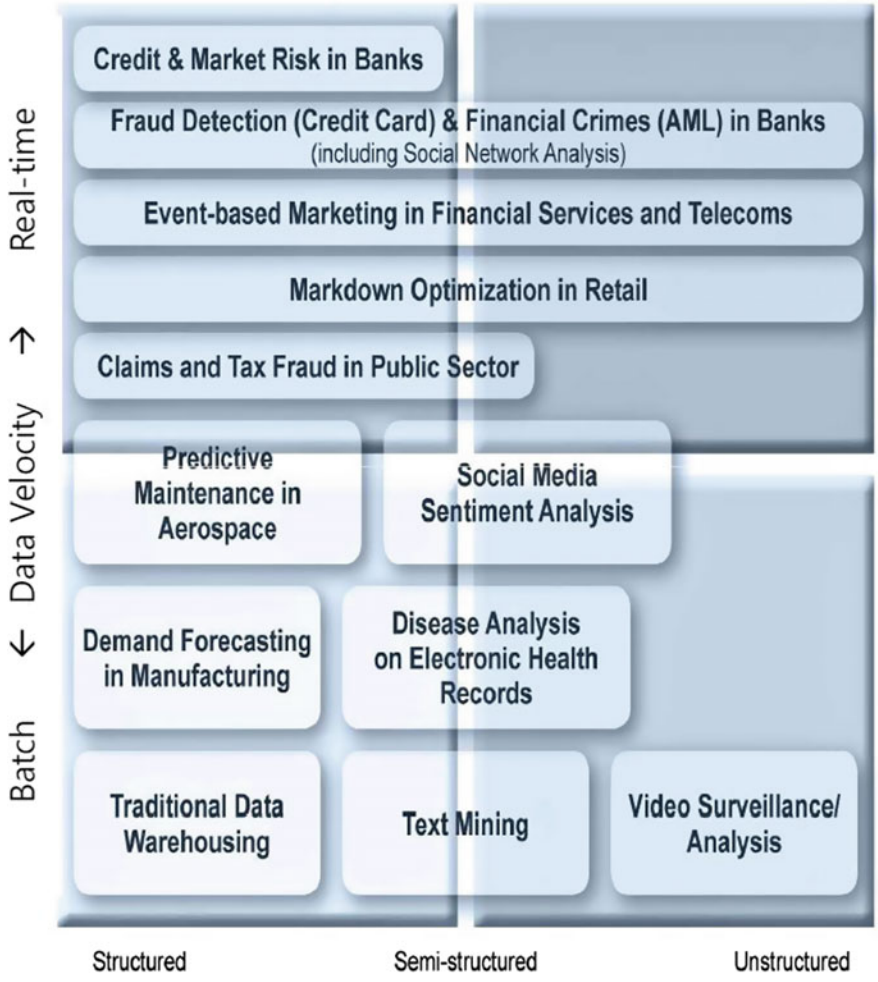
A.K. Agarwal
e-mail: ambuj4u@gmail.com

M. Pant et al. (eds.), *Soft Computing: Theories and Applications*,
Advances in Intelligent Systems and Computing 584,
https://doi.org/10.1007/978-981-10-5699-4_72

2 Big Data

Big data comprises of both structured and unstructured data, which includes video, audio, data generated through emails, etc. Social sites generate vast amount of data through daily activities. It should be understood that the historical data maintained in industries and business sector help them to thrive in this competitive world. This can be termed as big data. It is very difficult to analyze and process the Big data. However, with the help of Hadoop technology, the complexity in analyzing and processing of data can be reduced substantially. Big data is a heterogeneous collection of complex data sets and produced by varied sources like television, mobile, etc.

There are three characteristics of big data—



1. volume
2. velocity
3. variety

Structure of the big data—

1. **Volume**—Volume mainly defined in terms of amount of data. Volume of data is growing exponentially. Management of high volume of data in reference with processing and storing is always complex.
2. **Velocity**—Social sites are constantly generating complex data in unstructured and semi-structured form. Increasing the collection of big data with the help of mobile, televisions, and more advance technologies as Internet mainly influences high velocity.
3. **Variety**—Variety of big data exists in structured and unstructured format. Structured data are always of fixed format and there is no possibility of changes of this data like tabular data, ERP, etc.

3 Related Work

Sentiment analysis is very popular technology in today's world. Vast amount of work has been done in this field. Mostly, work in this area relates to storage of the data.

(i) Semantic analysis assumes much importance: also it deals with document and word type of the data and is mostly dependent on NLP processing techniques.
(ii) It deals with point-wise data and information and is mathematical in nature.

4 Hadoop

Hadoop is an open source framework which is freely available for every user. It is based on the Java programming framework. Hadoop is a project of apache. Hadoop is a framework which is available to support for the reliable and scalable distributed computing system. Hadoop framework was designed for solving the problems like processing the data and analysis the big data (Fig. 1).

1. Execution Engine (Map Reduce)
2. Hadoop distributed file system (HDFS) (Fig. 2)

Fig. 1 Data replication [1]

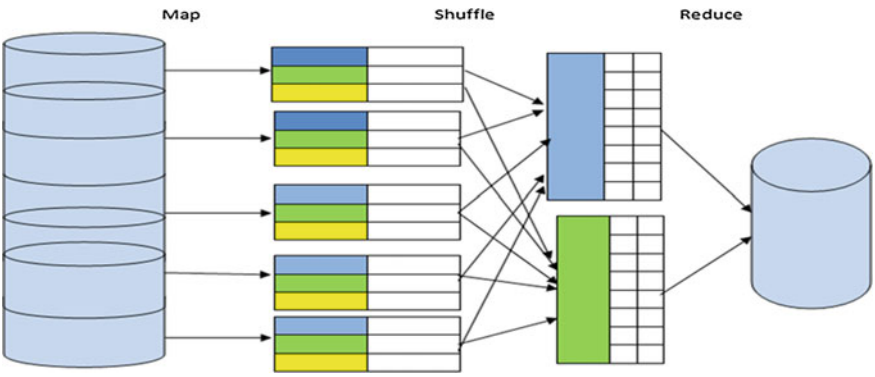


Fig. 2 Map reduce [2]

5 Hadoop Distributed File System (HDFS)

It is mainly handling large amount of big data. Big data stores the blocks in HDFS. It is client server architecture. HDFS comprises of name node and many data nodes. It stores the data for the named nodes that is known as name node. The Name node searches to track the data node positions. Also, it is responsible to support the file system operations. If the name node fails in running the operation, then Hadoop does not support and recovers any data node state. Data replication is done for achieving targets of fault tolerance. In HDFS, the large sized data cluster is stored as a parallel sequence of blocks.

Using Our Approach

In this paper, we focus on mainly speed performing analysis and also accuracy. What removes the various problems in big data technology? Like part of speech and

tagging using opennlp, it is easy to solve the problem. It is mainly known as tagging and used for following purposes.

(i) Firstly: Usage of words like a, an can be stopped. It is not useful for the real-time sentiment analysis.
(ii) Second approach is unstructured to structured: The twitter messages and the comments are mostly unstructured i.e. comment of "on bajrange bhai jaan" "favorite" is written "favorable," "God" is written "good," "aswm" is written for "awesome," "bd" is written for "bad."
(iii) Thirdly, emoticons: It is most expressive approach available on ideas and opinion. It is symbolic representation converted to words at this stage.

1. **Data and Real-time data features**

In this paper, the real time is very important. It is obtained from the data streaming API's available from twitter. It uses keywords like we are using in the movie bajrange bhai jaan. Objects that we use to perform the sentiment analysis are submitted to the twitter APIs. This provides Twitter, the tweets that are related to only that object.

Twitter data mostly used unstructured data. A tweet mostly consists of maximum 140 characters and likes. The messages' comments consist of a user name and timestamp. Mostly, timestamp is useful for the future development in our project. It is also helpful for different geographical regions (Fig. 3).

2. **Data defined part of speech**

The file contains the obtained tweets.

3. **Data in Root form**

It is widely used program to increase the overall efficiency and lowers the time access of the system. Root forms the word on twitter for the tweet are changed to their root form and split all that word which is unwanted and extra storage of the derived word's sentiment analysis.

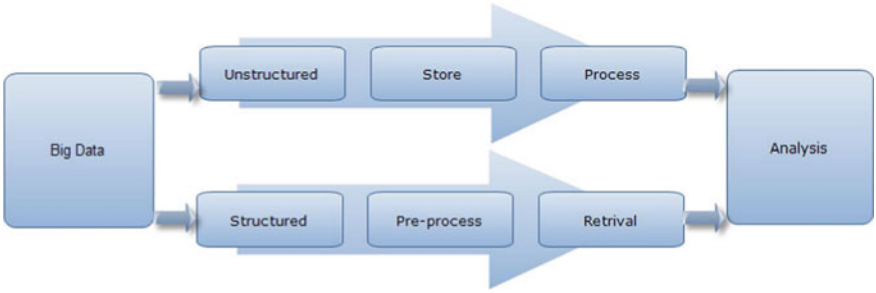


Fig. 3 Data processing [1]

4. **Sentiment Data Directory**

Now, the Real Time Data Directory is making and using standard directory for the big data sentiment, word net, and uses all condition for this word i.e. "good," "bad." The overall data is used to store the sentiment data in standard directories. It is local to the program, it is only primary memory. We can utilize our time in searching the main word in primary memory.

5. **Map Reduce Algorithm**

This is mapped reduce algorithm used for the tweets of the data. The sentiment values can be obtained from standard algorithm which we are using.

6 Final Data Accuracy

The complete overall accuracy in some twitters data i.e. accessing the data from sentiwordnet, opennlp, and wordnet is shown. The data are bajrange bhai jaan comments like negative word and positive word and neutral word. The data are compared by the help of movie bajrange bhai jaan, tweets like good, favorite, and negative words. The final special data available on web like following http://www.cs.tau.ac.il/~kfirbar/mlproject/twitter.data, now the checked data are as follows—

Sentiment	Count	Correct	%	Tolerance
Positive	739	520	72.22	−0.01
Negative	637	399	61.67	+0.05
Neutral	86	43	73.42	±0.003

The total accuracy of this project is 72.22. It is the mean of the total accuracy.

7 Total Time Efficiency

In our project, a necessary aspect is efficiency, that is why our project working well. Also reduces the time from hard disk that is only possible with the help of Hadoop and it is also using a lower time.

8 Conclusion

The Sentiment analysis is widely used at this time. The research is used for the analysis of the data. This project is also expanded to the social media platform and movie reviews like blogs and comment and likes per day. The accuracy is totally

value following. Also the use of hashtags and emoticons is very useful and necessary for social media data for this project. In our project, use of emoticons and hash tags like comment and tweets analyzed per day data. The total accuracy found is 72.22%.

References

1. Saleem, A., Agarwal, A.K.: Analysis and design of secure web services. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, Springer Singapore (2016)
2. Shukla, S., Lakhmani, A., Agarwal, A.K.: Approaches of artificial intelligence in biomedical image processing: a leading tool between computer vision & biological vision. In: 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Spring), Dehradun, 2016, pp. 1–6. doi:10.1109/ICACCA.2016.7578900
3. Beaver, D., Kumar, S., Li, H.C., Sobel, J., Vajgel, P.: Finding a needle in haystack: Facebook's photo storage. In: Proceedings of the nineth USENIX Conference on Operating Systems Design and Implementation, Berkeley, CA, USA, USENIX Association, pp. 1–8 (2010)
4. Rupanagunta, K., Zakkam, D., Rao, H.: How to mine unstructured data. Artic. Inf. Manag. (2012)
5. Marche, S.: Is Facebook making us lonely. Atlantic **309**(4), 60–69 (2012)
6. IBM What Is Big Data: Bring Big Data to the Enterprise, IBM (2012). Available: http://www.01.ibm.com/software/data/bigdata/
7. Duggal, R., Shukla, B., Khatri, S.K.: Big data analytics in Indian healthcare system—opportunities and challenges. In: Research Paper Accepted at National Conference on Computing, Communication and Information Processing (NCCCIP-2015), May 2015, pp. 92–104 (2015)
8. Chih-Wei, L., Chih-Ming, H., Chih-Hung, C., Chao-Tung, Y.: An improvement to data service in cloud computing with content sensitive transaction analysis and adaptation. In: Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual, 2013, pp. 463–468
9. Purcell, B.: The emergence of "big data" technology and analytics. J. Technol. Res. (2013). Garlasu, D., Sandulescu, V., Halcu, I., Neculoiu, G.: A big data implementation based on grid computing. Grid Comput. (2013)
10. Agarwal, A.: Implementation of cylomatrix complexity matrix. J. Nat. Inspir. Comput. **1** (2013)
11. Agrawal, T., Agarwal, A.K., Singh, S.K.: Study of cloud computing and its security approaches
12. Saxena, A.K., Agarwal, A.K., Ather, D.: How to secure design using threat modeling
13. Agarwal, A.K., Katiyar, V.: A study of software matrix systems: a comparative study of existing software matrix systems
14. Fatima, S., Agarwal, A., Gupta, P.: Different approaches to convert speech into sign language. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 180–183 (2016)
15. Shukla, S., Agarwal, A.K., Lakhmani, A.: MICROCHIPS: a leading innovation in medicine. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 205–210 (2016)
16. Grosso, P., de Laat, C., Membrey, P.: Addressing big data issues in scientific data infrastructure, 20–24 May 2013
17. Lin, J.: MapReduce is good enough? The control project. IEEE Comput. **32** (2013)

18. Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S., Stoica, I.: BlinkDB: Queries with bounded errors and bounded response times on very large data (2013)
19. Chih-Wei, L., Chih-Ming, H., Chih-Hung, C., Chao-Tung, Y.: An improvement to data service in cloud computing with content sensitive transaction analysis and adaptation. In: Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual, pp. 463–4681 (2013)
20. Sagiroglu, S., Sinance, D.: Big data: a review, 20–24 May 2013
21. Zhao, Y., Wu, J.: Dacha: a data aware caching for big-data applications using the Map Reduce framework. In: INFOCOM, 2013 Proceedings IEEE, Turin (2013)
22. Garlasu, D., Sandulescu, V., Halcu, I., Neculoiu, G.: A big data implementation, 17–19 Jan 2013
23. Zhang, X., Xu, F.: Survey of research on big data storage, 2–4 Sept 2013
24. Mukherjee, A., Datta, J., Jorapur, R., Singhvi, R., Haloi, S., Akram, W.: Shared disk big data analytics with Apache Hadoop, 18–22 Dec 2012
25. Bifet, A.: Mining big data in real time. Informatica **37**, 15–20 (2013)