A survey to Identify Gap in Sentiment Analysis Using Big Data Analytics

Ritu Patidar[1], Sachin Patel[2]
[1] Department of Computer Science Engineering, SAGE University, Indore
[2] Department of Computer Science Engineering, SAGE University, Indore

Abstract

Many researchers have addressed the different issues, challenges and problems associated with the sentiment analysis and emotion mining over the last decade. The advent of e-commerce over the last decade and changed the complete paradigm of the business environment. Instead of going to the shops personally, the users can directly access the details of the products they want to buy, compare them of different ecommerce websites, and order them. The online shopping has saved so much of time, money and energy of the users as well as the vendors. This research paper performs a survey on different sentiment analysis research work and explore gap in different solutions. Its also investigate the contribution of machine learning and artificial intelligence in the analysis of user sentiments. The complete study has been performed on small and big data size work to evaluate the feasibility and accuracy of the existing solutions. The complete research paper starts with introduction of sentiment analysis and keeps forward with comparison chart of different existing solutions. It also addresses problems and expected solutions.

Keywords: Sentiment Analysis, Support Vector Machine [SVM], Naive Bayes Algorithm

I. INTRODUCTION

Sentiment Analysis is the technique to examine or understand the feelings and thoughts of users. It can be defined us: "Sentiment Analysis is the process for computationally identifying and categorizing the opinion and expressions from piece of text. It also help to address the opinion of user in terms of positive, negative and neutral. Natural Language Processing is used for real time understanding and getting more accurate evaluation.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy". The field of research work carried on this thesis is the big data analytics applied to the detailed sentiment analysis. The required analysis is performed on the reviews and feedbacks of users on e-commerce companies. A deep study of the big data environment is required to be performed to identify the potential of the field. The capability of big data analytics to process a huge amount of data of the range of terabytes in considerably small amount of time is needed to be understood. The basic framework used to cope up with this huge amount of data is required to be gone through so as to match the problem domain with the field. The challenges associated with the sentiment analysis of the text data for large scale of data with a heavy diversity are also the need of the research. The major objective of including this chapter is to get the depth of the big data analytics through the detailed analysis of the architecture of environment. The major factors enabling the framework to process this large amount of data are also discussed in this chapter so that the advantages could be utilized in the further research. The drawbacks also lead to the constraints needed to be considered throughout the process. The rationale behind the inclusion of this chapter in the architecture of thesis is the development of the insight towards the broader aspects of data handling in a very large space. The utility of the framework used for dealing with the big data for sentiment analysis has also been explored in this chapter.

II . HADOOP ECO SYSTEM

Big data is characterized by three V's in broader sense, Volume, Velocity and Variety. These characteristics govern the nature of the data which helps in deriving the strategy and algorithms to deal with big data. The vastness, diversity, heterogeneity and speed of evolution are encompassed in these three V's as shown in fig 3.1. The characteristics are discussed in brief below.

A. Volume

It reflects the amount of data which is generated, processed, stored and operated within the system. Generally the data of size of around terabytes comes into

the category of big data. The range however may go to Zetabytes considering the ever increasing growth of internet technologies and the number of users all around the globe.

B. *Variety:*

This characteristic of big data represents the diversity and heterogeneity of the data generated from different source, internal or external. The type of data may be any kind of multimedia data like text, images, video, live streams etc. The sources of these data also are from large range of equipments of different specifications. The data may represent any kind of physical quantity with varying sensitivity like Electronic Medical Records (EMR), meteorological data, security data, satellite data, navigation data, etc. This large range of diversity is presented through Variety characteristic. The inter-relation of this data also leads to a complexity of links between them. These link types add the complexity of the data further. The variety also relates to the possible uses associated with a raw data and the associated features of the datasets received from various points of data collection.

C. *Velocity*

This reflects the speed of generation of data in big data environment. With the hugeness and diversity of the data, the velocity with which the data is generated is also of utmost importance. It is because of the capability of the available processing hardware and software. The time taken by the available resources to deal with this big data plays an important role in reflecting the performance of big data analysis, therefore velocity of generation of data is also considered as a very important characteristic of big data.

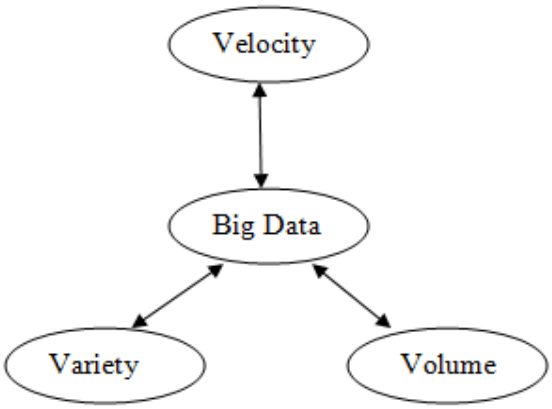

Figure 1-V Relation with Big Data

3. DATA CLASSIFICATION APPROACH

Data Classification is the approach to categories labeled or unlabeled data based on their pattern and category for most effective and efficient use. It can be broadly classify as:

1. Supervised Learning
2. Unsupervised Learning

Supervised learning is the approach that categorized labeled data based on common pattern. It uses training data ser to examine the data pattern and classify data in testing phase. The best possible use of such approach is to predict and observe trend of data movement. It is use to analyze the trained data and produce inferred function to examine and generate data classes for whole data sample .Subsequently, unsupervised learning is used to categorize unlabeled data from hidden labels. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This work considers two most significant classification algorithms for sentiment analysis known as Support Vector Machine and Naïve Bayes Classifier.

Support Vector Machine (SVM) model is often used as baselines for other methods in text categorization and sentiment analysis research. SVM was introduced in COLT-92 by Boser, Guyon & Vapnik, and it became rather popular since. SVMs can be used to solve various real world problems of Uncertainty in Knowledge-Based Systems.

Naive Bayes is a approach to classify model based on assign labels it is the most successful for large dataset. Simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

III. LITERATURE REVIEW

The detailed literature survey conducted in the earlier sections has shown that the area of feature extraction has been explored very deeply by the researchers. Many techniques have been presented for the feature extraction from text data received from different platforms. Some researchers have also explored the images and video data for the sentiment analysis, but the accuracy has been a major concern for them. Various machine learning techniques have

been proposed for the sentiment analysis but the scale of sentiments was narrow in these works. The amount of data used for this analysis was also not considered large. The strategies presented by the researchers have not considered the big data environment. The sentiment analysis for e-commerce companies using the reviews and feedback has also not been performed by many researchers. The rationale behind this research work is the absence of an intelligent sentiment analyzing framework for the e-commerce companies by considering huge amount of data with great diversity and velocity.

This section presented the detailed literature survey encompassing the complete spectrum of big data analytics and its applications in sentiment analysis. It started with the discussion on the background of the field of sentiment analysis to present the insight of the problem area of emotion mining. It also addressed the challenges and opportunities associated with this field. This chapter presents the objectives of this literature survey and that of the present study followed by the correlation among them to clarify the process and objectives of research in depth. More than 35 research papers have been thoroughly reviewed to present the research gap through Tabular format which provides a direction to proceed with the research work and to identify the required outcome.

Sr. No.	Name of the Author	Publishing Year	Title of the Paper	Parameter Covered
1	Hailong Zhang, Wenyan Gan, Bo Jiang	2014	Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey	Lexicon based Methods, Survey
2	Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo	2015	Approaches, Tools and Applications for Sentiment Analysis Implementation	Statistical tools, Machine learning approaches, sentiment analysis
3	T. Hai, K. Shirai, and J. Velcin	2015	Sentiment analysis on social media for stock movement prediction	Stock movement prediction, social media, Sentiment analysis
4	Y. Yu and X. Wang	2015	in the Twitter world: A big data analysis of sentiments in US sports fans' tweets	Win prediction, Twitter data analysis, World cup 2014
5	R. R. B. López, S. Sánchez-Alonso, and M. A. Sicilia-Urban	2015	Evaluating hotels rating prediction based on sentiment analysis services	Sentiment analysis, Hospitality business, TripAdvisor portal

6	T. P. Liang, X. Li, C.-T. Yang, and M. Wang	2015	What in consumer reviews affects the sales of mobile apps: A multifaceted sentiment analysis approach	Consumer reviews, sales of mobile apps, Sentiment analysis
7	E. D. Avanzo and G. Pilato	2015	Mining social network users opinions to aid buyers' shopping decisions	Buyers' shopping decisions, social network
8	Divya Sehgal and Ambuj Kumar Agarwal	2016	Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework	Sentiment Analysis, Big Data, Social media, Hadoop
9	Suchita V Wawre, Sachin N Deshmukh	2016	Sentiment Classification using Machine Learning Techniques	Sentiment Classification, Machine Learning Techniques
10	I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez	2016	Analysis of the effect of sentiment analysis on extracting from tweets and forum posts	Sentiment analysis, Adverse drug reactions, Social media data
11	E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song	2016	Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news	Topic-based content, Ebola virus, Twitter, news

12	J. Li and P. Meesad	2016	Combining sentiment analysis with socialization bias in social networks for stock market trend prediction	Sentiment analysis, socialization bias, stock market trend prediction
13	W. Chen, Y. Cai, K. Lai, and H. Xie	2016	A topic-based sentiment analysis model to predict stock market price movement using Weibo mood	Topic-based sentiment analysis, stock market price movement, Weibo
14	R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, Jr.	2016	Predicting wins and spread in the Premier League using a sentiment analysis of Twitter	Win prediction, Twitter, Sentiment analysis
15	C. Alfaro, J. Cano-Montero, J. Gómez, J. M. Moguerza, and F. Ortega	2016	A multi-stage method for content classification and opinion mining on weblog comments	Multi-stage method, opinion mining
16	K. Philander and Y. Y. Zhong	2016	Twitter sentiment analysis: Capturing sentiment from integrated resort tweets	Sentiment analysis, Hospitality business, social media
17	W. Chung and D. Zeng	2016	Social-media-based public policy informatics: of US Immigration and border security	Sentiment analysis, public policy informatics, US Immigration and border security
18	Q. Zhou, R. Xia, and C. Zhang	2016	Online shopping behavior study based on multi-granularity opinion mining: China versus America	Online shopping behavior, Opinion mining, China versus America

19	Speriosu M, Sudan N, Upadhyay S, Baldridge J.	2016	Twitter polarity classification with label propagation over lexical links and the follower graph	Polarity classification, label propagation, lexical links, follower graph, Twitter
20	Wong FMF, Tan CW, Sen S, Chiang M.	2016	Quantifying political leaning from tweets, retweets, and retweeters	Political leaning, Quantification, Twitter, retweets.
21	E. Cambria	2016	Effective computing and sentiment analysis	Sentiment analysis, Computation complexity, Time complexity
22	M. Mazhar Rathore, Anand Paul, Awais Ahmad	2017	Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions	Sentiment Analysis, Big Data, Geo-Social media, Real time decision

Table 1. Comparision Table

IV. RATIONALE

The amount of data generated through these feedbacks is too huge for a person to evaluate and come to a conclusion manually. Considering the business base of very big giants of the online market like, Wallmart, Amazon, Flipkart, Reliance, etc, who have millions of users, it is near to impossible to handle this huge amount of data. The speed at which this data is generated also make it a near to impossible challenge to handle manually. The problem becomes more severe, when the heterogeneous nature of the reviews came into account. The users are from very different backgrounds, having varying expectations and range of products is also huge. Incorporating all these challenges associated with the amount of data with a huge velocity and variety, the assessment of the data is a big challenge for the service providers. Recently, big data analytics has emerged as boon in the field of computing technology which is capable of dealing with the very huge amount of data of varying velocity and variety in a very efficient way. The parallel distributed computing architecture with high computing capabilities has made it possible to access, process and analyze this big data. This thesis proposed an implementation of this big data analytics technology to analyze the text data received from reviews and the feedbacks of the users of various e commerce companies and identifying the sentiments of the users about the specific product or service. The concept of natural language processing is used in this work to identify the orientation of the users. The sentiments are classified into four categories in this work as positive, negative, neutral and unsure. A dictionary is prepared on the basis of keywords projecting the sentiments and the emotions of the users which is preprocessed as per the requirement.

V. CONCLUSION

The complete survey analyzed that sentiment analysis is not only required to know about the user viewpoint but also help to observe the current trend of interest. This research paper attempts to explore the different gaps in sentiment analysis work and try to explore solution for same.

The proposed work is expected to present an accurate prediction and estimation framework for the sentiment analysis of the users through the text data retrieved from the e-commerce websites. The use of machine learning will add intelligence in the proposed work and thereby provide a justified and reasonable estimation of emotions of the users. The proposed work is implemented in the Big Data environment and therefore gives a fast, accurate and robust emotion detection mechanism.

REFRENCES

[1] *Divya Sehgal and Ambuj Kumar Agarwal, "Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework". 5th International Conference on System Modeling & Advancement in Research Trends, 2016, IEEE.*

[2] *M. Mazhar Rathore, Anand Paul, Awais Ahmad, "Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions". ICC SAC Symposium Big Data Networking Track, 2017 IEEE.*

[3] *C. Fellbaum, "Wordnet and wordnets," Encyclopedia of Language and Linguistics, pp. 665–670, 2005.*

[4] *B. Liu, "Sentiment analysis and subjectivity," Handbook of Natural Language Processing, pp. 627–666, 2010.*

[5] *H. Ji, S. Ploux, and E. Wehrli, "Lexical knowledge representation with contextonyms," in Proceedings of MT Summit IX, New Orleans, USA. Association for Machine Translation in the Americas, 2003.*

[6] *Chuanming Yu, "Mining Product Features from Free-Text Customer Reviews: An SVM-based Approach", 2009, Nanjing, China. ICISE 2009 December 26-28,*

[7] *Raisa Varghese and Jayasree M, "Aspect Based Sentiment Analysis using Support Vector Machine Classifier", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013.*

[8] *Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264.*

[9] *Suchita V Wawre, Sachin N Deshmukh, "Sentiment Classification using Machine Learning Techniques", International Journal of Science and Research (IJSR) Volume 5 Issue 4, April 2016.*

[10] *Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo, "Approaches, Tools and Applications for Sentiment Analysis Implementation",International Journal of Computer Applications (0975– 8887) Volume 125 – No.3, September 2015.*

[11] *Hailong Zhang, Wenyan Gan, Bo Jiang, "Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey", 11th Web Information System and Application Conference, 2014*

[12] *D. M. Rousseau, J. Manning, and D. Denyer, "11 evidence in management and .organizational science: Assembling the field's full weight of scientific knowledge through syntheses," Acad. Manage. Ann., vol. 2, no. 1, pp. 475_515, 2008.*

[13] *I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, ``Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts," J. Biomed. Inform.,vol. 62, pp. 148_158, Aug. 2016.*

[14] *E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," J. Inf. Sci., vol. 42, no. 6, pp. 763_781, 2016.*

[15] *T. Hai, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," Expert Syst. Appl., vol. 42, no. 24, pp. 9603_9611, 2015.*

[16] *J. Li and P. Meesad, "Combining sentiment analysis with socialization bias in social networks for stock market trend prediction," Int. J. Comput. Intell. Appl., vol. 15, no. 1, p. 1650003, 2016.*

[17] *W. Chen, Y. Cai, K. Lai, and H. Xie, "A topic-based sentiment analysis model to predict stock market price movement using Weibo mood," Web Intell., vol. 14, no. 4, pp. 287_300, 2016.*

[18] *R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, Jr., "Predicting wins and spread in the Premier League using a sentiment analysis of Twitter," Decis. Support Syst., vol. 88, pp. 76_84, Aug. 2016.*

[19] *Y. Yu and X. Wang, "World cup 2014 in the Twitter world: A big data analysis of sentiments in US sports fans' tweets," Comput. Hum. Behav.,vol. 48, pp. 392_400, Jul. 2015.*

[20] *A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," New Media Soc., vol. 16, no. 2, pp. 340_358, 2014.*

[21] *C. Alfaro, J. Cano-Montero, J. Gómez, J. M. Moguerza, and F. Ortega, "A multi-stage method for content classification and opinion mining on weblog comments," Ann. Oper. Res., vol. 236, no. 1, pp. 197_213, 2016.*

[22] *R. E. Schapire, Y. Freund, P. Bartlett, andW. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," Ann. Stat., vol. 26, no. 5, pp. 1651_1686, 1998.*

[23] *K. Philander and Y. Y. Zhong, "Twitter sentiment analysis: Capturing sentiment from integrated resort tweets," Int. J. Hospitality Manage., vol. 55, pp. 16_24, May 2016.*

[24] *R. R. B. López, S. Sánchez-Alonso, and M. A. Sicilia-Urban, "Evaluating hotels rating prediction based on sentiment analysis services," Aslib J. Inf. Manage., vol. 67, no. 4, pp. 392_407, 2015.*

[25] *W. Chung and D. Zeng, "Social-media-based public policy informatics: Sentiment and network analyses of US Immigration and border security," J. Assoc. Inf. Sci. Technol., vol. 67, no. 7, pp. 1588_1606, 2016.*

[26] *T.-P. Liang, X. Li, C.-T. Yang, and M. Wang, "What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach," Int. J. Electron. Commerce, vol. 20, no. 2, pp. 236_260, 2015.*

[27] *D. Kim, D. Kim, E. Hwang, and H.-G. Choi, "A user opinion and metadata mining scheme for predicting box office performance of movies in the social network environment," New Rev. Hypermedia Multimedia, vol. 19, nos. 3_4, pp. 259_272, 2013.*

[28] *Q. Zhou, R. Xia, and C. Zhang, "Online shopping behavior study based on multi-granularity opinion mining: China versus America," Cogn. Comput.,vol. 8, no. 4, pp. 587_602, 2016.*

[29] *E. D. Avanzo and G. Pilato, "Mining social network users opinions`to aid buyers' shopping decisions," Comput. Hum. Behav., vol. 51, pp. 1284_1294, Oct. 2015.*

[30] *Speriosu M, Sudan N, Upadhyay S, Baldridge J., " Twitter polarity classification with label propagation over lexical links and the follower graph", In: Proceedings of the first workshop on unsupervised learning in NLP, EMNLP'11. Stroudsburg: Association for Computational Linguistics. p. 53–63., 2016*

[31] *Stavrianou A, Brun C, Silander T, Roux C., "NLP-based feature extraction for automated tweet classification," In: Proceedings of the 1st international conference on interactions between data mining and natural language processing, vol. 1202, DMNLP'14. Aachen: CEUR-WS.org; 2011. p. 145–146.,2017*

[32] *Wong FMF, Tan CW, Sen S, Chiang M., "Quantifying political leaning from tweets, retweets, and retweeters," IEEE Trans Knowl Data Eng. 2016;28(8):2158–72.*

[33] *Tumasjan A., "Predicting elections with Twitter: what 140 characters reveal about political sentiment," In: Fourth international AAAI conference on weblogs and social media. 2010.*

[34] *Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S., "A system for real-time Twitter sentiment analysis of 2012 US presidential election cycle," In: Proceedings of the ACL 2012 system demonstrations, ACL'12. Stroudsburg: Association for Computational Linguistics; 2012. p. 115–20.*

[35] *Wong FMF, Tan CW, Sen S, Chiang M., "Quantifying political leaning from tweets, retweets, and retweeters," IEEE Trans Knowledge Data Eng. 2016;28(8):2158–72.*

[36] *E. Cambria, "Effective computing and sentiment analysis," IEEE Intell. Syst., vol. 31, no. 2, pp. 102_107, Mar./Apr. 2016.*

AUHTOR

First Author –Mrs. Ritu Patidar, M.E (CSE), Sage University, Indore (India), ritupatidar89@gmail.com

Second Author – Dr. Sachin Patel, P.hd (CSE), Sage University, drsachinpatel.sage@gmail.com