A Inspection on Sentiment Research of Big Data: Techniques, Open Challenges

L.Sudha Rani, S.Zahoor-Ul-Huq, C.Shoba Bindu

Abstract:Due to the invention of Web 2.0, the users have become more interest to share their content day by day. The emergence of various social networking sites has added to a greater extent to these activities. These provide a very good platform for the users to share the opinions of the persons across the globe. The opinions shared by the customers on the web can have the prevalent over the service industry. Many industries such as educational institutions, researchers, business organizations are concentrating opinion mining which is also called as sentiment analysis (SA) to retrieve the views and opinions posted by the public. This paper made a survey on Sentiment Analysis (SA) which aims to discusses technical aspects of SA (techniques, types) .This paper further highlights the main challenges faced by SA. These challenges present a lot of scope for research work in the future.

Keywords:Big Data, Sentiment Analysis, Structured data, Unstructured data.

I. INTRODUCTION

From the past years, the growth of the data is increased rapidly and it leads to the development of big data [1]. The people from all over the world are practiced for the devices like digital sensors, social media applications and communication tools [2]. For instance, associations catch the developing volume of value-based information through which trillions of bytes of data are produced as far as viewpoints from providers to clients. The headway of computerized sensors and correspondence advances have prompted the improvement of the IoT technology [4]. With such improvement, person to person communication destinations and specialized gadgets like advanced cells, workstations, and PCs enable people to associate with one another to make gigantic measures of expansive information [3]. For example, Twitter is a tremendous system of 467 million clients which creates 175 million tweets once a day [5]. In 2011, the International Data Corporation (IDC) has indicated that the world is now produced around 1 zettabyte (ZB) of information and the rate at which the information is developing has been expanding. It is expected that by 2020, the sum information will achieve 44 ZB, with in any event some part of them is printed information [7] which is accessed from different social media sites lie Google messenger, WhatsApp, Facebook and Twitter. As per the survey, there is almost 450 million tweets sent for each day and 300 million tweets among them are shared. In the mean time, it is evaluated that 5 billion messages on Face book is posted with 6 billion likes once a day. In addition, it is expected that the measure of information will be constantly develop as a result of the deluge of computerized innovations that have just risen in the advanced era[1]. The broad utilization of advancements and quick stream of information throughout the years has likewise supported in the development of enormous information business investigation. This supposed to be called as big data analytics.

"Big data is defined as the data set whose computation time is more than the tolerable time in using the traditional software tools to store, manage and manipulate the data". In general, big data is composed of Volume Variety and Velocity. Volume represents the massive amount of data, i.e, it contains huge volumes and data sets and it requires huge computation for analysis. The volume of the data is ranges from TB to PB. Velocity represents the speed of the computation. It needs to perform the data computation in lightning speed. For instance, the video monitoring system continuously monitors the data and identifies the useful data in matter of seconds. This mechanism is different from traditional data mining approaches. The Variety represents the categories of data. The data sets in the big data are collected from different sources are of different formats such as unstructured, structured and semi-structured.

Big data presents many no of challenges to the organizations because of it's complexity in nature. At the basic dimension, associations confront the test of taking care of and putting away a colossal measure of information [8]. Analyzing meaningful information and visualization of the analyzed data in a variety of forms [9] is another important challenge. Furthermore, the distribution of information across different sites in the world is another challenge.

Before the appearance of the Internet, we depended on friends for item or administration suggestions, casting a ballot see amid nearby races. The Internet simplifies our efforts to get conclusions and encounters from the general population who are neither in expert system nor in our own contract. The measure of conclusions and remarks on the Internet has developed massively amid the most recent decade.

The remaining paper is described as follows. Section II deals with overview of text mining and analytics.Section III deals with overview of sentiment analysis. Section IV gives the detailed discussion of various methods used for Sentiment Analysis Section V discusses the open challenges of Sentiment Analysis over big data , finally the conclusion is drawn.

Revised Manuscript Received on April 12, 2019.

L.Sudha Rani,Asst. Prof, CSE Dept, GPREC,Kurnool GPREC,KURNOOL JNTUA,Anantapur, A.P, India (Email: sudha1021@gmail.com)

Dr.S.Zahoor-Ul-Huq,Professor,CSE Dept, GPREC,Kurnool GPREC,KURNOOL JNTUA,Anantapur, A.P, India(Email: szahoor@gmail.com)

Dr.C.ShobaBindu,Professor,CSE Dept, GPREC,Kurnool GPREC,KURNOOL JNTUA,Anantapur,A.P, India (Email: shobabindhu@gmail.com)

II. TEXT MINING AND ANALYTICS

From the past few years, huge amount of information has risen at an inconceivable rate; there is a essential to fuse some kind of investigation to increase significant insights from content, pictures, and recordings which is for the most part in unstructured arrangement. Text mining uses information mining, natural language processing and machine learning strategies to process content documents [10]. Text Analytics, while like Text mining as far as strategy, for the most part manage a lot of information to separate and produce helpful, important data and knowledge [11].Text mining/examination are initially led for two purposes. Breaking down individuals' feeling on an issue or wonder is the first and imperative reason for Sentiment Analysis. For this reason, Text Analytics experiences a gigantic measure of literary information to distinguish individuals' mentalities, thoughts, decisions, and feelings on an issue. Surveying individuals' assessment on an item, individual, occasion, association, or subject from a client or gathering of client's points of view is its second purpose [12]. Sentiment examination which utilizes an algorithmic system to perceive opinionated content [13].

III. AN OVERVIEW OF OPINION MINING AND SENTIMENT ANALYSIS

Sentiment Analysis is a method of examining thee unstructured text to find the suitable information and converting it into useful business information [13]. It determines if an expression is positive, negative or neutral and to what extent.

Sentiment Analysis is also known as opinion mining as it includes identifying opinions, emotions, attitudes of a company's product ,brand or service.

Sentiment analysis is divided into two learning techniques such as supervised learning and unsupervised learning.. Supervised Learning is used for subjectivity or objectivity identification in a piece of text. Unsupervised Learning is used for determination of different opinions or sentiments in relation to different aspects.

The earlier work on sentiment analysis mainly focused on unimodality related to text [16, 17], but recent research on sentiment analysis has developed their momentum like audio and video etc.

IV. TECHNIQUES

The following subsections explains various methods for Sentiment Analysis

A. KEYWORD-BASED CLASSIFICATION

This approach deals with the keywords such as happy, delight, sad, joy, terrified and miserable etc. The polarities considers both positive and negative. The limitations of this approach are it depends on only polarities. However some time, it takes the advantage over polarities and conveys the information [21].

B. LEXICON-BASED CLASSIFICATION

This approach builds the list of words and indexed them accordingly based on the positive and negative polarities. Based on the weight of the polarities the overall sentiment score of the post or text the constructed lexicon is used. The advantage of this method is it doesn't require the supervised methods and trained data. This approach is widely used in the blogs, forums and text reviews [22-23]. This method not have major impact in social media [23], as this data is in unstructured format and also contains new slang, abbreviations, new expressions etc..

C. MACHINE LEARNING-BASED APPROACH

This approach is used in numerous application areas and has become a significant research area. The machine learning approaches are the advancement of normal approaches, but they are smarter and have the self-learning capacity based on the inputs given by the user. These algorithms are majorly having the modules of training phase and testing phase.

Machine Learning Algorithms are categorized into following three types.

- Supervised
  - Regression
    - Linear
    - Polynomial
  - Classification
    - KNN
    - SVM
    - Decision Trees
    - Logistic Regression
    - Naive Bayes
- Unsupervised
  - Clustering
    - K-means
    - SVD
    - K-means
  - Association analysis
    - FP-Growth
    - Apriori
  - Hidden Markov Model
- Reinforcement

The following are some of the sentiment analysis algorithms used in the machine learning approach.

(1) RANDOM FOREST

Random Forests is an adaptable algorithm capable of performing both classification and regression. It is a type of Ensemble learning method which is commonly used as predictive modelling and machine learning technique [24].

The following are the features of Random Forest Classifier:

- Most accurate leaning algorithm.
- Runs efficiently on large databases
- Performs implicit feature selection
- Requires no input preparation
- Works well for both classification and prediction problems.

(2) SUPPORT VECTOR MACHINE (SVM)

SVM is belongs to the supervised learning algorithms and it is used for classification of data [26]. SVM separates the different classes using the decision boundary. The hyper

planes are used in the SVM to divide the segments and these contain the common data.

The important features of SVM are given below:

1. The SVM is belongs to the class of supervised learning. The algorithm is trained with the set of labeled data and applies the classification based on the knowledge gained at the training phase.
2. The SVM has the advantage of classification and regression.
3. The SVM is used in the classification of non-linear data.

(3) NAÏVE BAYES

This classifier is a supervised machine learning algorithm that has long been used in most of the applications [27]. For classifying the data, it constructs a simple probabilistic model which is based on probability theorem. This method is used for classification of large data sets. This approach outperforms all the classification algorithms.

(4) DECISION TREE

These algorithms are another class of supervised learning which is used for classification [28]. This algorithms used for dependent and categorical variables which has the capacity to divide the data in homogeneous sets.

(5) K-NEAREST NEIGHBOUR

This algorithm is an unsupervised algorithm which uses the instance learning for classification. K-Nearest works through a nonparametric strategy of putting away all data sources and occurrences and arranges new inputs using the similarity measures.

V. OPEN CHALLENGES& RESULTS

(1) HETEROGENEOUS QUALITIES OF BIGDATA

As big data is collected from various sources, the data is in heterogeneous in nature and most of the data is in unstructured format. The sentiment classifier that is designed should work effectively to deal with such type of data. The traditional sentiment classifier can deal with the data that is collected from one single source such as online reviews about a particular entity which is collected from one single site or a company feedback records.

(2) INVESTIGATING SPARSE, UNCERTAIN, AND INADEQUATE DATA

The major feature of big data is that it contains the huge volumes of noisy data and has the data sparsity [23], This has an effect on the accuracy of the sentiment classification. The sentiment classifier should be designed which is not only able to classify the sentiment polarity of a message but also to predict the incomplete information to provide more accurate results.

(3) SEMANTIC RELATIONS IN MULTI-SOURCE DATA FUSION

Analysis of an event from different sources such as Twitter, Face book, YouTube, Instagram and constructing semantic associations between the data sources is very challenging task. Hence it becomes very important to construct such an efficient model which constructs semantic association between the data sources.

(4) THE IMPACT OF SOCIAL BOTS

Social bots are software programs which are designed to pretend as human users on social media websites. Recently social bots have become complicated as well as threatening. Politicians, marketers, and different firms use these automatic social bots in online environments in order to manipulate public opinion .The sentiment classifier which is designed have to answer the questions like 1) Is the output of sentiment analysis based on the authentic opinion of the user or is it a non authentic opinion which is generated by bots.?

2) To which extent, these social bots can influence the public opinion? The future research has to investigate these two questions.

VI. CONCLUSION

Sentiment Analysis has gained its significance due to the increase of social media usage, including reviews, forum blogs, micro-blogs, Face book, Twitter, and other social networks. Presently we approach an enormous measure of stubborn information which can additionally be utilized for various strategies for investigation. Typically more than 80% of web-based social networking information can be checked for examination purposes. Further the paper presents different open difficulties for Sentiment Analysis over huge information, where a great deal of research is required.

REFERENCES

1. R. Addo-tenkorang and P. T. Helo, “Computers & Industrial Engineering Big data applications in perations supply-chain management : A literature review,” Comput. Ind. Eng., vol. 101,pp. 528–543, 2016.
2. A. Zaslavsky, C. Perera, and D. Georgakopoulos,, “Sensing as a Service and Big Data,” 2015.
3. J. Manyikaet al., “Big data: The next frontier for innovation, competition, and productivity,” 2011.
4. H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, “Vision and challenges for realising the Internet of Things,” Clust. Eur. Res. Proj. Internet Things, Eur. Commission, vol. 3, no. 3,pp. 34–36, 2010.
5. A. N. E. W. Approach and T. O. Analysis, "Unstructured data: A big deal in big data Deep dive : Analytics.”
6. R. L. Villars, C. W. Olofson and M. Eastwood, “Big Data: What It Is and Why You Should Care Information,” White Paper:IDC, June 2011 .
7. M. Khoso, “How much data is produced every day,” 2016.
8. A. Misra, A. Sharma, P. Gulia, and A. Bana, “Big Data : Challenges and Opportunities,” no. 2, pp.41–42, 2014.
9. M. Batty et al., “Smart cities of the future,” Eur.Phys. J. Spec. Top., vol. 214, no. 1, pp. 481–518,2012.
10. F. R. Luciniet al., “Text mining approach to predict hospital admissions using early medical records from the emergency department,” Int. J.Med. Inform., vol. 100, pp. 1–8, 2017.
11. Z. Khan, Z. Khan, T. Vorley, and T. Vorley, “Big data text analytics: an enabler of knowledge management,” J. Knowl. Manag., vol. 21, no. 1,pp. 18–34, 2017.
12. T. T. Thet, J.-C. Na, and C. S. G. Khoo, “Aspectbasedsentiment analysis of movie reviews on discussion boards,” J. Inf. Sci., vol. 36, no. 6, pp.823–848, 2010.

13. R. Piryani, D. Madhavi, and V. K. Singh,Analytical mapping of opinion mining and sentiment analysis research during 2000--2015," Inf. Process. Manag., vol. 53, no. 1, pp. 122–150, 2017.
14. N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in Proceedings ofthe 29th annual international ACM SIGIRconference on Research and development in information retrieval, 2006, pp. 244–251.
15. S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing, vol. 174, pp. 50–59,2016.
16. S. Poria, I. Chaturvedi, E. Cambria, and A.Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in Data Mining (ICDM), 2016 IEEE 16thInternational Conference on, 2016, pp. 439–448.
17. S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," Knowledge-BasedSyst., vol. 108, pp. 42–49, 2016.
18. I. Chaturvedi, E. Cambria, S. Poria, and R.Bajpai, "Bayesian Deep Convolution Belief Networks for Subjectivity Detection," in DataMining Workshops (ICDMW), 2016 IEEE 16thInternational Conference on, 2016, pp. 916–923.
19. S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," arXiv Prepr.arXiv1610.08815, 2016.
20. E. Cambria and B. White, "Jumping NLP curves:A review of natural language processing research," IEEE Comput. Intell. Mag., vol. 9, no.2, pp. 48–57, 2014.
21. E. Cambria, B. Schuller, Y. Xia, and C. Havasi,"New avenues in opinion mining and sentiment analysis," IEEE Intell. Syst., vol. 28, no. 2, pp.15–21, 2013.
22. M. Taboada, J. Brooke, M. Tofiloski, K. Voll,and M. Stede, "Lexicon-based methods for sentiment analysis," Comput. Linguist., vol. 37,no. 2, pp. 267–307, 2011.
23. A. Giachanou and F. Crestani, "Like it or not: Asurvey of twitter sentiment analysis methods,"ACMComput. Surv., vol. 49, no. 2, p. 28, 2016.
24. aLiaw and M. Wiener, "Classification and Regression by randomForest," R news, vol. 2, no. December, pp. 18–22, 2002.
25. L. Breiman, "Bagging predictors: Technical Report No. 421," Mach. Learn., vol. 140, no. 2, p. 19, 1994.
26. C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," Neural Networks, IEEE Trans., vol. 13, no. 2, pp. 415–425, 2002.
27. D. D. Lewis, "Naive (Bayes) at Forty : The Independence Assumption in Information Retrieval 2 The Naive Bayes Classier."
28. J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.
29. S. Sivanandam, SN and Deepa, Introduction to genetic algorithms, vol. 2, no. 6. 2007.