IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

Sentiment Analysis on Huge Streaming Data using Hadoop

S. Sugandhi[1], M. Mohamed Surputheen[2]

Research Scholar, Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Tamilnadu, India[1]

Associate Professor, Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Tamilnadu, India[2]

Abstract: Sentiment analysis has been a major focus in all consumer oriented industries due to the availability of huge amount of customer opinions in the Internet. This paper presents a sentiment analysis framework that utilizes the processing efficiency of the Hadoop ecosystem to provide real time sentiment analysis. The framework divides the process of sentiment analysis to two major sections; content pre-processing and evaluation. Experiments shows that our application has the ability to scale and handle huge amounts of data.

Keywords: Sentiment Analysis; Polarity Identification; Hadoop; Tokenization; Stemming

I. INTRODUCTION

Sentiment analysis has become one of the major necessities of today's business industry. Business, both online and offline require major inputs from consumers in terms of feedbacks. Not many consumers provide direct feedback. In order to formulate effective business strategies, consumer requirements/ feedbacks are needed. The advent of social networks and mass adoption of them has presented a major opportunity for these organizations to leverage the interests of the user and convert it to business strategies.

Though user opinions are reflected in the messages posted in social networking sites, these messages are textual and do not exhibit direct correspondence to the products that are dealt with by the organizations. Correlation between opinions exhibited by public and products manufactured by the companies is the missing link in this scenario 1. The strategies proposed for constructing this missing link is called text mining. In order to be specific, the methods concentrating on mining social networking data tends to identify the major player or contributor in the sentence, along with identifying the polarity of the sentence, which helps identify the sentiment levels of the sentence.

Appropriate analysis and usage of this information would provide an upper hand to organizations in terms of marketing appropriate products to the customers [4,5]. Sentiment analysis helps classify documents into opinions or polarities. Due to a huge number of useful applications [1, 6, 7] involved in this, this study has received considerable attention in the research community. Some major applications include product review classification [2], product opinion mining [3], stock movement prediction [8], currently trending topics along with their polarities etc.

II. RELATED WORKS

Several classification based methods for sentiment identification have been proposed. Silva et al., proposed a classifier ensemble based sentiment identification method [9]. This method considers the query term and classifies the tweets as positive or negative. This method has its major concern in product or organization based analysis. It uses Multinomial Naive Bayes, SVM, Random Forest, and Logistic Regression for classification. The advantage of this approach is that it provides best results due to the combination of several algorithms. The downside of this approach is that it tends to increase the processing overhead in manifolds.

Baecchi et al., proposed a feature based learning model that provides effective sentiment analysis [10]. Balahur et al., proposed a multi-lingual based sentiment analysis [11]. This method has its major concerns in operating with multiple language support for effective results. This paper is based on two languages, English and Spanish. This method analyzes the hybrid features and explore multi-lingual data, and it boasts of improved accuracies. Saif et al., proposed a context based semantic analysis of twitter data [12]. This method operates on both entity level and tweet level, providing enhance accuracy. Katz et al., and Korenek et al., proposed similar context based sentiment analysis methods [13],[14]. This method proposes to exhibit high robustness to noise, by eliminating them, hence providing better analysis.

VithiyaRuba et al. presented a similar twitter based sentiment analysis method [28]. Zol et al. presented a sentiment based opinion generation technique [29]. Behavior analysis methods [15],[30] are also prevalent, that takes the sentiment analysis to next level. Emotion tracking and time series emotion tracking provides an

DOI 10.17148/IJARCCE.2016.5948

enhanced way to track social/ product sentiments. Zhu et al., proposed a time series based emotion tracking [16] that models emotions on the basis of time. Several such techniques tracking emotions on the basis of time has been proposed [17], [18], [19]. User specific emotion analysis also exist in literature that tracks the emotions of a single user [20]. Irony in text proves to be the most complex structure to be evaluated. Reyes et al., Kreuz et al., Glucksberg et al. and Lucariello et al., proposed some effective techniques working on detecting irony[21],[22],[23],[24].

III.OUR APPROACH

Sentiment analysis of user comments or feedbacks have played a vital role in successful marketing of products. Though the initial analysis were conducted on the text retrieved from organization specific web sites, it was later identified that social networking sites provide much better and unbiased results. Current business analysis are all conducted using social networking data. The social networking data tends to be streaming data, generated at large. This necessitates the use of Big Data techniques. The current method uses HDFS to store the data and MapReduce Paradigm is used to process the data returning the sentiment indicators. In the current social networking scenario, data tends to be not just text but also a group of symbols known as emoticons [25],[27]. The emoticons also play a vital role in identifying the polarity of the current document under scrutiny [26]. Hence our method uses both text and emoticons as the base indicators to identify the sentiment value. The document processing of tweets eliminates symbols and processes the other constituents of the tweet to identify the polarity of the tweet under analysis.

A. Tokenization

Tokenization is the process of dividing a string into stream of text; words, symbols, phrases and other meaningful elements. The entire contents of the tweet are taken and several filters are applied to them sequentially to eliminate special characters, symbols and spaces from the text. The output returned from the tokenizer is usually in the form of a vector, containing all the tokens corresponding to the tweet under analysis. Composition of these output tokens are limited to alphabets alone.

B. Stemming

Though tokenization breaks down the document into its basic units, the words are usually in their inflected forms or in a form that describes the tense of the sentence. Seed words are required in order to identify the polarity. This process is carried out in stemming. Stemming is the process of eliminating prefix and suffix from a word, to obtain the basic structure of the word, also called the seed word. Several algorithms have been proposed for stemming. The first stemming algorithm was proposed by Lovins[31]. The most popular and the standard stemming algorithm is the Porter’s Stemming algorithm, proposed by Martin Porter [32]. The current approach uses the contemporary version of the Porter’s algorithm, called the Porter 2.

C. Normalization/ Lemmatization

Stemming reduces inflection in words, however some words are incomplete and ripped after stemming. Normalization converts such words to their basic form. It reduces the text to its canonical format such that uniformity is achieved in the corpus, making the corpus consistent. Normalization is usually domain dependent. Hence a domain based ontology is required to perform normalization. The canonical words are chosen based on the domain and a domain-based ontology is maintained for this phase. Any word that does not obey the rules described in this phase is eliminated. Normalization and lemmatization almost similar, hence they are combined and performed as a single process. WordNet 3.0 database is used as the reference for the process of normalization and lemmatization.

D. Polarity Prediction

After the completion of tokenization, stemming, normalization and lemmatization, the returned tokens are of the root/ base form of the words that were used in the actual text. Polarity of terms is obtained from WordNet 3.0, which is a human curated dataset for polarity analysis. This dataset is the base for all operations, hence to parallelize the entire operation, the data is stored in the distributed cache in Hadoop Distributed File System (HDFS). This operation distributes the polarity data to all the data nodes in the Hadoop system, hence reducing the need for communications and references. The obtained polarity values are aggregated. This aggregated polarity identifies the sentiment of the current tweet. Since specific values are used to identify the sentiment of the text, the resultant value not only depicts the polarity, but also the intensity of the polarity. This intensity defines the impact the tweet makes on a user. Eg. A tweet with an intensity of -0.02 has low negative impact, while a tweet with intensity -0.78 has a high negativity associated with it.
Several methods tend to identify the polarity of the text alone, without identifying the intensity of the polarity. The proposed approach provides intensity based calculations, using which the polarity along with the sentiment intensity can be obtained.

IV.RESULTS AND DISCUSSION

Experiments were conducted to analyze and identify the accuracy of the proposed method. Inputs were passed from a client node connected to the cluster. Map Reduce programs were executed in six phases, each phase performing a single task in map and reduce phases.
The STS Gold Sentiment Corpus is a human annotated dataset containing 2034 tweets. Classification levels exhibited by the algorithm is presented in Fig. 1. It could be observed that an accuracy of 98% is exhibited by the classifier, representing high levels of predictions.

DOI 10.17148/IJARCCE.2016.5948

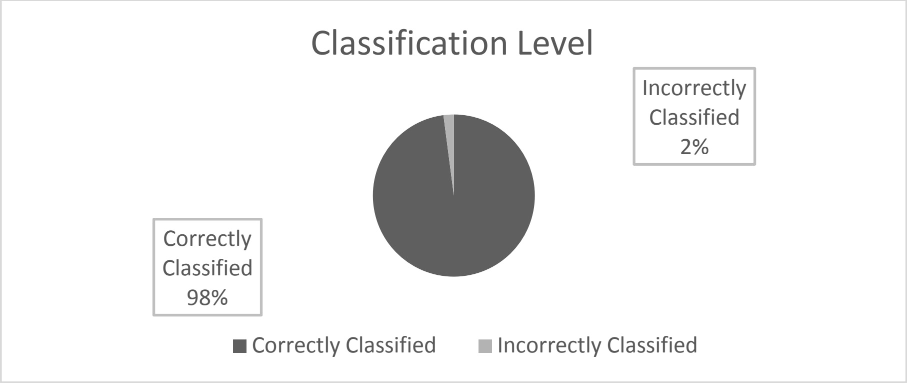

Fig. 1.Overall Accuracy

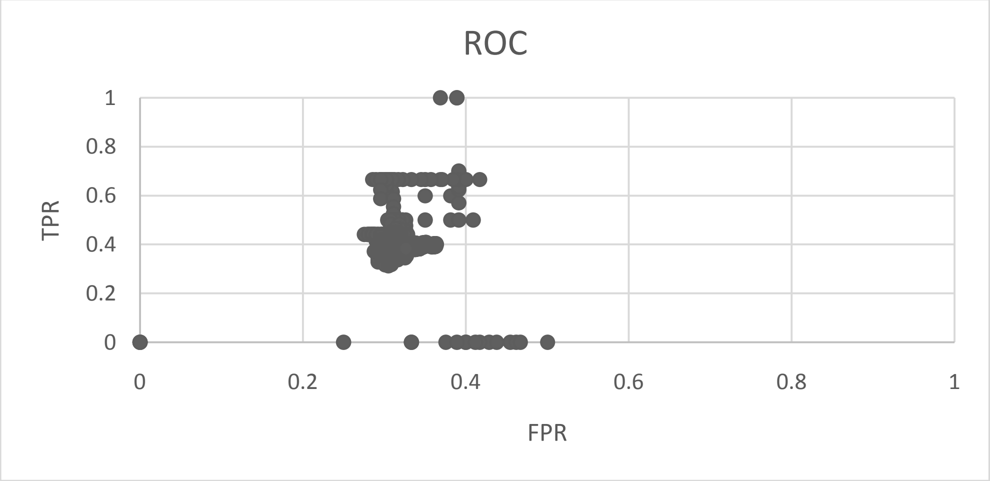

Fig. 2.ROC

Fig.2 shows the ROC curve plotted using the STC gold corpus data. It could be observed that most of the plotted coordinates are grouped in the top left corner of the graph, above the diagonal line. Some points even approach the top left (0,1) coordinate, which proves that the current method exhibits effective results.

It could be observed that the classifier exhibits a very high True Positive Rates (TPR) and low false positive rates <0.5. Some of the TPR values even reach 1, exhibiting 100% accurate true positive predictions. Fig. 3. shows the PR curve (Precision Recall Curve). It represents the accuracy of the results obtained and the accuracy of the selection mechanism.

The best classifier is the one exhibiting high precision and recall values. Precision represents the positive predictive value, while recall presents the rate of true positives obtained from the system. The current approach exhibits a high recall, with a moderate precision, exhibiting scope for improvements.

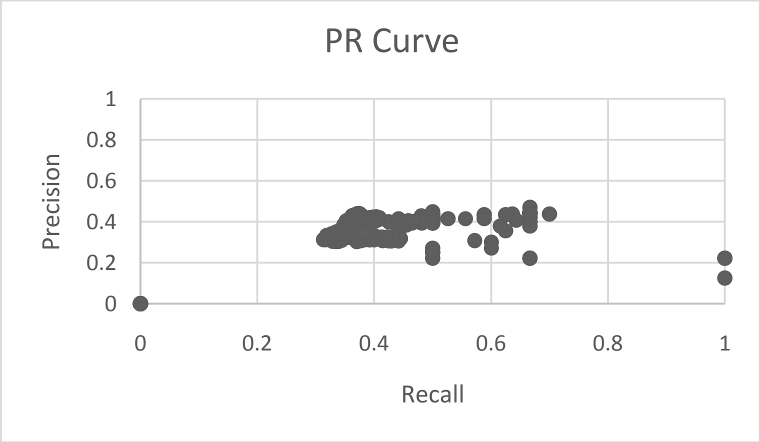

Fig. 3.PR Curve

DOI 10.17148/IJARCCE.2016.5948

V. CONCLUSION

Sentiment analysis is of huge prominence as the numbers of people using social networking are on a huge raise. Though converting this data to useful and analyzable information is a tedious task. This paper presents methods that inputs raw data provided by users in social networking sites and provides useful data that can be used in the evaluation phase to perform sentiment analysis. Future research directions include designing effective the data evaluation algorithms to minimize the time and maximize the scalability of the system

REFERENCES

[1] M.S. Hajmohammadi, R. Ibrahim, Z. Ali Othman, “Opinion mining and sentiment analysis: a survey,” Int. J. Comput. Technol. 2 171–178, 2012.

[2] H. Kang, S.J. Yoo, D. Han, “Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews”, Exp. Syst. Appl. 39- 6000–6010, 2012.

[3] L.W. Ku, Y.T. Liang, H.H. Chen, “Opinion extraction, summarization and tracking in news and blog corpora”, in: Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI, Palo Alto, California, pp. 100–107, 2006.

[4] M.I.Dascalu, C. N. Bodea, M. Lytras, P.O. de Pablos, A. Burlacu, “Improving e-learning communities through optimal composition of multidisciplinary learning groups”. Computers in Human Behavior, 30(0), 362–371, 2014.

[5] M. D. Lytras, and P. O. de Pablos, “The role of a ‘‘make’’ or internal human resource management system in spanish manufacturing companies: Empirical evidence”. Human Factors and Ergonomics in Manufacturing & Service Industries, 18(4), 464–479, 2008.

[6] M. Lytras, and P. O. de Pablos, “Software technologies in knowledge society”. Journal of Universal Computer Science, 17(9), 1219–1221, 2011.

[7] M. Lytras, M. A. Sicilia, A. Naeve, P. O. de Pablos, and M. D. Lytras, “Competencies and human resource management: Implications for organizational competitive advantage”. Journal of Knowledge Management, 12(6), 48–55, 2008.

[8] T. Nguyen, K. Shirai, J. Velcin, "Sentiment analysis on social media for stock movement prediction", Expert Systems with Applications, In Press, Corrected Proof, Available online 6 August 2015.

[9] N.F. da Silva, E.R. Hruschka, E.R. Hruschka. “Tweet sentiment analysis with classifier ensembles.” Decision Support Systems. 2014 Oct 31;66:170-9.

[10] C. Baecchi, T. Uricchio, M. Bertini, A. Del Bimbo. “A multimodal feature learning approach for sentiment analysis of social network multimedia.” Multimedia Tools and Applications. 2015:1-9.

[11] A. Balahur, J.M. Perea-Ortega. “Sentiment analysis system adaptation for multilingual processing: The case of tweets.” Information Processing & Management. 2015 Jul 31;51(4):547-56.

[12] H. Saif, Y. He, M.Fernandez, H.Alani. “Contextual semantics for sentiment analysis of Twitter.” Information Processing & Management. 2016 Jan 31;52(1):5-19.

[13] G. Katz, N. Ofek, B. Shapira. ConSent. Knowledge-Based Systems. 84(C):162-78, 2015.

[14] P. Korenek, M. Šimko. “Sentiment analysis on microblog utilizing appraisal theory.” World Wide Web. 2014 Jul 1;17(4):847-67.

[15] C. Mahajan, P. Mulay. “E3: Effective Emoticon Extractor for Behavior Analysis from Social Media.” Procedia Computer Science. 50:610-6, 2015.

[16] C. Zhu, H. Zhu, Y. Ge, E. Chen, Q. Liu, T. Xu, H. Xiong. “Tracking the evolution of social emotions with topic models.” Knowledge and Information Systems, 1-28,2015

[17] F.R. Chaumartin. “UPAR7: A knowledge-based system for headline sentiment tagging.” InProceedings of the 4th International Workshop on Semantic Evaluations 2007 Jun 23 (pp. 422-425). Association for Computational Linguistics, 2007.

[18] Z. Kozareva, B. Navarro, S. Vázquez, A. Montoyo. “UA-ZBSA: a headline emotion classification through web information.” InProceedings of the 4th International Workshop on Semantic Evaluations (pp. 334-337). Association for Computational Linguistics, 2007.

[19] K.H. Lin, C.Yang, H.H. Chen. “What emotions do news articles trigger in their readers?.” InProceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 733-734). ACM, 2007.

[20] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, Y. Yu. “Joint emotion-topic modeling for social affective text mining.” InData Mining, 2009. ICDM'09. Ninth IEEE International Conference on 2009 Dec 6 (pp. 699-704). IEEE.

[21] A. Reyes, P. Rosso, T. Veale. “A multidimensional approach for detecting irony in twitter.” Language resources and evaluation. 47(1):239-68, 2013.

[22] R. Kreuz. “Using figurative language to increase advertising effectiveness.” InOffice of Naval Research Military Personnel Research Science Workshop. University of Memphis, Memphis, 2001.

[23] S. Kumon-Nakamura, S. Glucksberg, M. Brown. “How about another piece of pie: The allusional pretense theory of discourse irony.” In Gibbs R, Colston H (Eds.). Irony in language and thought. London: Taylor and Francis Group. (pp. 57–96),2007.

[24] J. Lucariello. “Situational irony: A concept of events gone away.” Irony in language and thought, 467-98, 2007.

[25] A. Shukla, B.D. Chaudhary. “A study of usage of symbols and opinionated words in annotation for modeling literature survey experiences.” Education and Information Technologies, 91-111, 2015.

[26] S. Feng, K. Song, D.Wang, G. Yu. “A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs.” World Wide Web,18(4):949-67, 2015.

[27] X. Xiong, G.Zhou, Y. Huang, H. Chen, K. Xu. “Dynamic evolution of collective emotions in social networks: a case study of Sinaweibo.” Science China Information Sciences, 56(7):1-8, 2013.

[28] K. VithiyaRuba, D. Venkatesan, “Building a Custom Sentiment Analysis Tool based on an Ontology for Twitter Posts”, Indian Journal of Science and Technology,2015 July, 8(13), Doi no:10.17485/ijst/2015/v8i13/61464

[29] S. Zol, P. Mulay, “Analyzing Sentiments for Generating Opinions (ASGO)-A New Approach”,Indian Journal of Science and Technology, 8(S4), 2015.Doi no:10.17485/ijst/2015/v8iS4/62327

[30] T. A. Litvinova, P. V. Seredin, O. A. Litvinova ,“Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study”,Indian Journal of Science and Technology, 8(S9),2015, Doi no:10.17485/ijst/2015/v8iS9/51103

[31] J.B. Lovins, “Development of a stemming algorithm.” MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[32] M. F. Porter, “An algorithm for suffix stripping.” Program, 14(3), 130-137, 1980.

DOI 10.17148/IJARCCE.2016.5948