A Novel Approach for Sentiment Classification on Train Reviews

Rashmi Thakur
Ph.D. Research Scholar, MPSTME,
NMIMS University, Mumbai, India
thakurrashmik@gmail.com

M. V. Deshpande
Professor & Dean, School of CSIT,
Symbiosis University of Applied Sciences, Indore, India

Abstract— Sentiment Analysis is the active area of research which focuses on analyzing the opinions or emotions of users and classify them into positive or negative reviews. In this paper, we propose a new approach for sentiment classification of train reviews using the map reduce concept. As we are aware that in this era of big data, tremendous data/reviews are gathered via social media sites at different locations which are distributed. Existing systems of Indian railways don't classify and analyze the reviews into positive and negative sentiments. Also there is no automatic classification of departments depending upon the complaints or reviews received for further action. We address this issue by developing the novel approach for sentiment classification using map reduce framework.

Keywords— Sentiment Analysis, Sentiment Classification, Natural Language Processing

I. INTRODUCTION

Public transportation is a necessity for creating sustainable communities where, the people usually book transportation tickets together, leave for the same destination, and especially have the same purpose such as doing business, sightseeing, or visiting relatives [2]. People usually have special needs when they travel with different purposes. Hence, travel purposes of the group of passengers can be inferred and may help carriers or airports provide precise and personalized services or recommendations for passengers. Moreover, the experienced journey of the passengers can be analyzed in detail with respect to waiting time, in-vehicle time, and number of transfers, together resulting in a measure of passenger service [3]. Currently, people utilize many social sites to share their opinions on different issues associated with transportation (e.g., rockslides, jams in traffic, traffic collisions or landslides. New clients can see the reviews which other people have given on the same category of subject and react accordingly on the same subject (e.g., roads or city streets jammed, street-side organizations, and associations). Conversely, a large volume of tweets or reviews can puzzle web surfers trying to determine immediate and safe routes [12]. Therefore, sentiment analysis plays a vital role in responding to the reviews and to meet the user satisfaction [6].

Sentiment analysis (also known as opinion mining [13]) that analyzes people's opinions/sentiments/emotions from texts is an active research field in natural language processing [14]. It has become popular research area which is drawing attentions from both research and industry communities in current era. [1]. Sentiment Analysis is beneficial in various fields like education, e-commerce etc. [17] [9]. With the help of Sentiment analysis analyze people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text [18]. A sentiment lexicon consists of some words and phrases which can express positive or negative sentiments, but using only a sentiment lexicon for sentiment classification is not enough due to the opposite orientations of these words in different domains [9]. In sentiment analysis, sentiment classification which can be treated as branch of text classification has become popular research area as from 2000 there is increase in public opinions on social sites and blogs [15].

Some of the researchers prefer different terms for sentiment classification such as opinion mining, sentiment analysis, subjectivity analysis, review mining, and opinion extraction [16]. Finding a solution for the accurate and timely classification of emotion is a challenging task [4].There are lots of classical feature extraction methods in the traditional text classification [20], such as Chi-square test, DF, etc., but these methods cannot be applied onto sentiment classification directly [8]. Sentence-level sentiment classification is a fundamental and extensively studied area in sentiment analysis. Lexicon-based approaches [11], typically utilize a lexicon of sentiment words, each of which is annotated with its sentiment strength. Text categorization problem is focused by sentiment classification which can be treated by Corpus-based methods [18]. They mostly build sentiment classifier from sentences with annotated sentiment polarity. The sentiment supervision can be manually annotated, or automatically collected by sentiment signals like emoticons in tweets [19] or human ratings in reviews [21] [1].

II. LITERATURE REVIEW

TABLE I. ANALYSIS ON VARIOUS METHODS

Author	**Method**	**Advantages**	**Disadvantages**
Duyu Tang *et al*. [1]	Sentence-level sentiment classification	It does not require any syntactic or polarity annotations in segmentation level.	In this method, not every expression can be composed by the meaning of its constituents.
Cagatay Catal and Mehmet Nangir [9]	Vote algorithm with Naive Bayes, Support Vector Machine (SVM), and Bagging.	Better performance.	Expensive and time consuming when dealing with huge datasets.
Chihli Hung and Hao-Kai Lin [10]	support vector machines (SVMs)	Enables better decision-making process. It maintains same classification accuracy as that of method which uses full-length documents.	Word sense disambiguation is not considered during the extraction process.
Vo Ngoc Phu *et al*. [4]	Fuzzy C-Means (FCM) method for English sentiment classification with Hadoop MAP (M) /REDUCE (R) in Cloudera.	It processes big data involving millions of English documents and the execution time of this model to conduct sentiment analysis on big data is short.	It takes a long time to implement and it is costly to build the algorithms of the model in the distributed system.
Mohammad Salehan and Dan J Kim [5]	Online Customer reviews (OCR) using a sentiment mining approach for big	It creates scalable automated systems for sorting and classification of OCR.	It did not consider the differences in expression of emotions between the real life and the virtual space.

	data analytics		
Farman Ali *et al*. [6]	Fuzzy ontology-based sentiment analysis and semantic web rule language (SWRL) rule-based decision-making	It automatically extract related sentiments from online consumer tweets and reviews and it successfully categorizes extremely obscure reviews, and intelligently determines transportation and city feature polarity.	It executed the irrelevant reviews and considered the noun, verb, adjective, and adverb as sentiment words, which decrease the precision rate of sentiment analysis.
Tao Chen *et al*. [7]	neural network based sequence model	Different opinion targets boosts the performance of sentence level sentiment analysis	Failed to explore the other sequence learning models for target expression detection.
Jinyan Li *et al*. [8]	hierarchical classification along with three filtering schemes	High prediction accuracy.	High performance is hard to achieve. In the process of operation it is not sensitive to missing data because many of the words in the text are low-frequency words

III. Challenges in Sentiment Classification

1) The major problem associated with the sentimental classification is that the online reviews consists of the user opinions in the form of abbreviations, shortenings and conjoined words, which are frequently used by the users in expressing their feelings about a point. The existing methods failed to have the full coverage over the online reviews as they are composed of reviews in abbreviations, shortenings and conjoined words [1]. In addition, they suffer from the polarity inconsistency.

2) Major issue faced by the classification methods is regarding the selection of the best feature that would yield the best result. The various features include the n-grams, synthetic n-grams of various types, words), or a combination of these features. Moreover, the size of the feature dimension is another major problem faced by the existing methods [4].

3) Precision is another major factor to be considered while performing the classification. The precision may be affected when the classifiers use the nouns, verbs, adjectives, and so on instead of considering only the sentimental words in the online reviews [6]. Moreover, most of the existing methods concentrated only on the polarity of the sentence rather than concentrating on the type of the sentence that carries different expressions of the reviews [7].

4) Some of the sentiment classification methods used only the numerical rating of the review and the word count of the review to validate the performance of the system mainly in validating the performance of the product. The existing methods failed to concentrate on the number of users, the positive and negative reviews [5].

5) In [7], they utilized linear SVM for sentiment classification after extracting the segment level features. The linear SVM is a traditional technique which has the limitation on over fitting problem and the convergence issue within bound of optimization problems.

6) In [4], the FCM clustering with MapReduce framework was used for sentiment classification. This method did not make use of any supervised learning mechanism for classification. Also, the manual level categorization is very challengeable and requires much cost and time if the data is big.

IV. Research Areas of Data Analytics and Data Mining in Indian Railways

1) **Railways have disparate Passenger data across five databases**, so the benefit lies in bringing this together to help build a more detailed, individual profile of each passenger. The data which is spread across the databases can be in different formats or can be unstructured data which can be combined to give best results. In current scenario when passenger books the ticket there is no permanent registration id generated which can be used in future to verify the historical data of the passenger and apply predictive analytics. Aadhar card no can be used as a permanent registration no so that once the profile of the passenger is generated there is no need to fill the complete information again and again while booking. If this no is linked then we can get all the past historical information of the traveler in a glance.

2) Customer satisfaction can be improved by reducing less delays and cancellations which increases customer loyalty and as a result increasing the bookings. By analyzing customer booking patterns, railways can also identify new routes to add and other services that will benefit both customers and the railway's bottom line. If the traveler misses the train, one can respond immediately by paying instant compensation offers like additional points, real-time re-booking, and customer service on his re-booked ticket.

3) **Seat Availability Forecast** – It can suggests the passenger which is the best train from one destination to another on various parameters like time and travel day. It gives users with smart predictions such as: which routes are busier, which are the areas where maximum people are travelling, information regarding delays in trains. Both the type of trains like long distance and local trains can include the information regarding the good food restaurants which are nearby the station. Logical insights from data can be used for saving consumptions in fuel, Shipment prediction as a result of which trains can run on time and keep the trains running on time.

4) **Recommender System for confused passenger:-** A person is in a dicey situation of when to book a ticket for journey date X. He cannot predict availability of ticket if booking date= A. Solution would be to ask a website visitor " a booking date" to Predict availability for journey date if he books on that booking date. Predictive analytics can be used based on the previous historical data and the festivals in the month to predict availability of tickets.

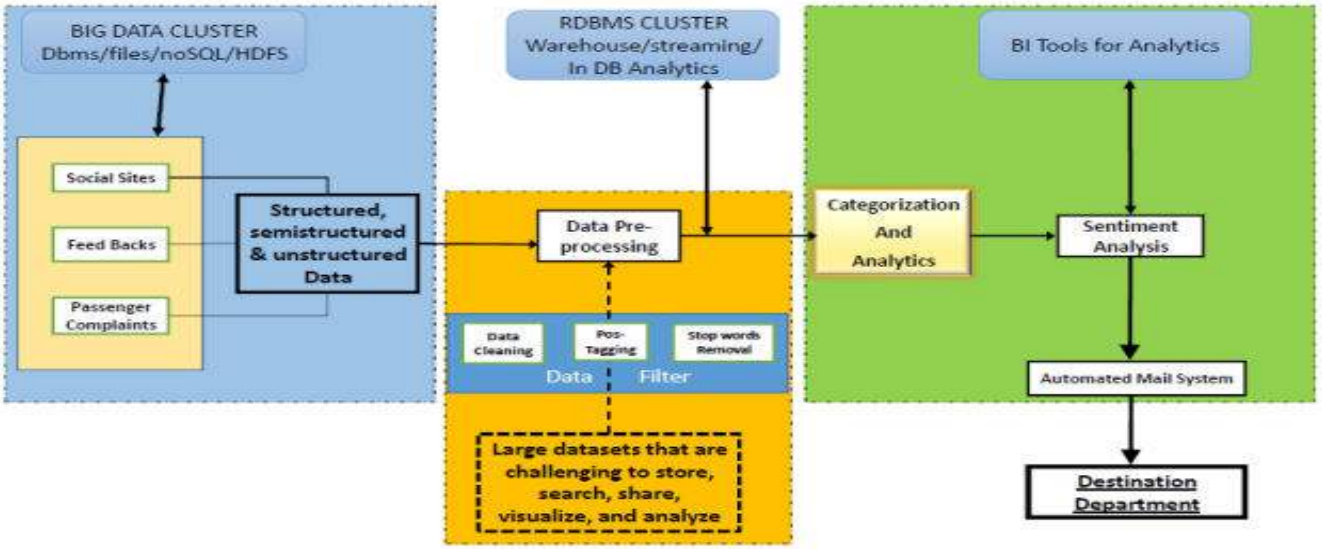
5) **Customer Complaints: -** Currently on Indian railways site whenever the Passenger logs a complaint the complaints are classified by the passenger. Railways dept. needs to classify the complaints automatically and divert it to respective dept. Text analytics can be used to solve the above problem.
6) Baggage tags can be gathered and scanned so that customers can collect their baggage via mobile apps.
7) A quick response can be given via compensation offers to the missed flight of the customers though additional points, real time rebooking.
8) Travel apps that track weather and deliver alternative itineraries based on lifestyle preferences.
9) Customer experiences can be collected for mining the data and then by applying analytical solutions to it one can get a view regarding what your customers think about you. This can help to respond immediately and take decisions accordingly.

V. Existing Complaint Management System of Indian Railways

Above diagram describes the complaint Management System in which customers can log their complaints on web portal. With the advancement in social media users very rarely post their complaints in the above portal. Also when the users the posting their complaints in the web portals, the classification regarding the type of complaint is done by customers itself. As shown in diagram the various types of complaint classifications are non-availability of water, passenger booking etc. Then this complaints are diverted to respective department according to the type selected by the users.

Drawback of above System:- Classification is done manually by the end user which needs to be done automatically by using data mining and analytics. The system don't classify the reviews/complaints on the basis of positive and negative sentiments. Also it does not take into consideration the reviews / Complaints from the social media sites like twitter etc.

VI. PROPOSED METHODOLOGY



Phase 1: The primary intention of this research is to design and develop an approach for sentiment classification on train reviews and MapReduce framework. Here, a new classifier will be developed for classification and the map reduce framework will be adapted for handling the big data. In order to adapt the MapReduce framework, two process of sentiment analysis such as, feature extraction and classification will be performed by devising the mapper and reducer function. The mapper can able to read the review database from different data centers. Also, the mapper function converts the reviews which are stored as text document to the feature vector. To find the out the feature, the features explained in [1] like, All-caps, Emoticon, Hashtag, Elongated units, sentiment lexicon, negation, bag of units, punctuation and the statistical features based on frequency will be utilized. Then, the reducer will have the new Classifier which will classify the reviews into positive class and negative class.

Phase 2: In the second phase of work, the classified reviews will be again applied to second level hierarchy which contains the K-Entropy based decision tree. This method will classify the train reviews into a respective departmental category so that the corresponding authorities can take a final solution on the reviews to further improve their customer satisfaction.

VII CONCLUSION

Sentiment analysis is gaining its popularity due to increase in tremendous data on various social media sites like twitter where reviews are given by users. There is need to analyze this reviews in order to increase customer satisfaction. Various challenges arising in sentiment analysis are drawing the attention of researchers to work in this area. Indian Railways can be benefited by incorporating such text analytics techniques to handle the huge amount of reviews received for satisfying the customers.

REFERENCES

[1] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, and Ming Zhou, "A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, pp.1750 - 1761, NOVEMBER 2015

[2] Youfang Lin, Huaiyu Wan, Rui Jiang, Zhihao Wu, and Xuguang Jia, "Inferring the Travel Purposes of Passenger Groups for Better Understanding of Passengers", IEEE TRANSACTIONS ON INTELLIGENT

TRANSPORTATION SYSTEMS, vol. 16, no. 1, pp.235 - 243, February , 2015.

[3] Evelien van der Hurk, Leo Kroon, Gábor Maróti, and Peter Vervest, "Deduction of Passengers' Route Choices From Smart Card Data", IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 1, pp.430 - 440, FEBRUARY 2015.

[4] Vo Ngoc Phu, Nguyen Duy Dat, Vo Thi Ngoc Tran, Vo Thi Ngoc Chau, Tuan A. Nguyen, "Fuzzy C-means for English sentiment classification in a distributed system", Applied Intelligence, pp.1-22, 05 November 2016.

[5] Mohammad Salehan and Dan J Kim, "Predicting the Performance of Online Consumer Reviews: A Sentiment Mining Approach to Big Data Analytics", Decision Support Systems, vol.81, pp.30-40, January 2016.

[6] Farman Ali, Daehan Kwak, Pervez Khan, S.M. Riazul Islam, Kye Hyun Kim, and K.S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling", Transportation Research Part C: Emerging Technologies, vol.77, pp.33-48, April 2017.

[7] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN", Expert Systems with Applications, vol.72, pp.221-230, 15 April 2017.

[8] Jinyan Li, Simon Fong, Yan Zhuang, and Richard Khoury, "Hierarchical classification in text mining for sentiment analysis of online news", Soft Computing, vol.20, no.9, pp.3411–3420, September 2016.

[9] Cagatay Catal and Mehmet Nangir, "A sentiment classification model based on multiple classifiers", Applied Soft Computing, vol.50, pp.135-141, January 2017.

[10] Chihli Hung and Hao-Kai Lin, "Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification", IEEE Intelligent Systems, vol.28, no.2, pp. 47 - 54, 2013.

[11] 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.

[12] Ali, F., Kim, E.K., Kim, Y.G., " fuzzy ontology-based opinion mining and information extraction: a proposal to automate the hotel reservation system", Applied Intelligence, vol.42, no.3, pp.481–500, 2015.

[13] C. Havasi, E. Cambria, B. Schuller, B. Liu, and H. Wang, "Knowledge-based approaches to concept-level sentiment analysis," IEEE Intelligent System, vol. 28, no. 2, pp. 0012–14, Mar.-Apr. 2013.

[14] C. D. Manning and H. Schütze, " Foundations of Statistical Natural Language Processing" , Cambridge, MA, USA: MIT Press, 1999.

[15] Xia, R., Zong, C., and Li, S., (2011), "Ensemble of feature sets and classification algorithms for sentiment classification", Information Science, vol.181, no.6 , pp. 1138-1152, March 2011.

[16] Liu, B., (2012), "Sentiment analysis and opinion mining", Morgan & Claypool.

[17] Quan, C., Ren, F., "Unsupervised product feature extraction for feature-oriented opinion determination", Information Sciences, 272, pp. 16-28.

[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," In Proceedings of the EMNLP, pp. 79–86, 2002.

[19] J. Zhao, L. Dong, J. Wu, and K. Xu, "Moodlens: An emoticon-based sentiment analysis system for chinese tweets," In Proceedings of the SIGKDD, 2012.

[20] Yang Y, Pedersen JO , "A comparative study on feature selection in text categorization", In Proceedings of the ICML'97, pp.412–420, 1997.

[21] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," In Proceedings of the ACL, 2011.

[22] Vijay Mahadeo Mane, D.V. Jadhav, "Holoentropy enabled-decision tree for automatic classification of diabetic retinopathy using retinal fundus images",Biomedical Engineering / Biomedizinische Technik, 2016.

[23] B. Rajakumar, "The Lion′s Algorithm: a new nature-inspired search algorithm", Procedia, vol .6, pp. 126–135, 2012.