Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques

Satya Mohan Chowdary G[1], K. Pavanasupriya[2], V. Deva Manojna[3],A. Madhava Sai[4], B. Ganesh[5]
[1,2,3,4,5]*Pragati Engineering College*

***Abstract* - Social media is growing trend now a days. Every day millions of user review and rate tourist places on tourism websites. Sentiment analysis can be performed over these reviews which will be helpful to find tourist place popularity. Based on sentiment analysis result, tourist can easily decide tour destination to be visited. In this paper sentiment analysis has been implemented using machine learning approach. The Dataset has been collected from various tourism review websites. Here we have performed comparative study of feature extraction algorithms i.e. Count Vectorization, TFIDF Vectorization. Along with classification algorithms Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF). Performance of algorithms has been compared using various parameters like accuracy, recall, precision and f1-score. From experiment we found that TFIDF Vectorization feature extraction algorithm has improved accuracy of classification algorithm as compare to Count Vectorization for given review dataset. In sentiment classification of tourist place reviews TFIDF Vectorization+RF has given highest accuracy 86% for a research dataset used.**

I.INTRODUCTION

Social media is rapidly growing now a days. Millions of users post reviews and rate tourist places on a daily basis over tourism websites. For analyzing this review sentiment analysis can be performed. Proper analysis of reviews will be able to find a trend of tourist place popularity. Summarized results from sentiment analysis will help tourists to decide the tour destination and tour planning. In this research paper two feature extraction algorithms have been used i.e. Count Vectorization and TFIDF Vectorization algorithm. Also three classification algorithms Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) have been used for sentiment classification. Comparison of performance has been performed for combination of feature extraction and classification algorithms on the basis of parameters like execution time, accuracy, recall, precision and f1-score. The content of this paper is structured as follows. Literature surveys on sentiment analysis are reviewed in Section II. Section III defines the Basic concept of Machine Learning. Section IV describes our Methodology of sentiment analysis for tourist place review classification, its visualization and performance evaluation. Section V presents the experimental implementation using machine learning algorithms for tourist place popularity distribution calculation. Section VI contains the results of the experiment executed. Section VII presents the comparative analysis of sentiment analysis using machine learning algorithms used in research study. Section VIII concludes this research paper. Section IX describes the future scope of the research paper.

II. LITERATURE STUDY

1. Sentiment Analysis: A Comparative Study on Different Approaches AUTHORS: M.D. Devika. Sunitha Amal Ganesh
Sentiment analysis (SA) is an intellectual process of extricating a user's feelings and emotions. It is one of the pursued fields of Natural Language Processing (NLP). The evolution of Internet based applications has steered a massive amount of personalized reviews for various related information on the Web. These reviews exist in different forms like social Medias, blogs, Wiki or forum websites. Both travelers and customers find the information in these reviews to be beneficial for their understanding and planning processes. The boom of search engines like Yahoo and Google has flooded users with copious amounts of relevant reviews about specific destinations, which is still beyond human comprehension. Sentiment Analysis poses as a powerful tool for users to extract the needful information, as well as to aggregate the

collective sentiments of the reviews. Several methods have come to the limelight in recent years for accomplishing this task. In this paper we compare the various techniques used for Sentiment Analysis by analyzing various methodologies.

2. Comparative analysis of Twitter data using supervised classifiers AUTHORS: Rohit Joshi, Rajkumar Tekchandani
Online Microblogging on social networks have been used for indicating opinions about certain entity in very short messages. Existing some popular microblogs like Twitter, facebook etc, in which Twitter attains maximum amount of attention in the field of research areas related to product, movie reviews, stock exchange etc. We had extracted data from Twitter i.e. movie reviews for sentiment prediction using machine-learning algorithms. We applied supervised machine-learning algorithms like support vector machines (SVM), maximum entropy and Naive Bayes to classify data using unigram, bigram and hybrid i.e. unigram + bigram features. Result shows that SVM surpassed other classifiers with remarkable accuracy of 84% for movie reviews.

3. A Survey on sentiment analysis challenges AUTHORS: Doaa Mohey El-Din Mohamed Hussein
With accelerated evolution of the internet as websites, social networks, blogs, online portals, reviews, opinions, recommendations, ratings, and feedback are generated by writers. This writer generated sentiment content can be about books, people, hotels, products, research, events, etc. These sentiments become very beneficial for businesses, governments, and individuals. While this content is meant to be useful, a bulk of this writer generated content require using text mining techniques and sentiment analysis. But there are several challenges facing the sentiment analysis and evaluation process. These challenges become obstacles in analyzing the accurate meaning of sentiments and detecting the suitable sentiment polarity. Sentiment analysis is the practice of applying natural language processing and text analysis techniques to identify and extract subjective information from text. This paper presents a survey on the sentiment analysis challenges relevant to their approaches and techniques.

4. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques
AUTHORS: Timor Kadir, Fergus Gleeson
The amount of text that is generated every day is increasing dramatically. This tremendous volume of mostly unstructured text cannot be simply processed and perceived by computers. Therefore, efficient and effective techniques and algorithms are required to discover useful patterns. Text mining is the task of extracting meaningful information from text, which has gained significant attention in recent years. In this paper, we describe several of the most fundamental text mining tasks and techniques including text pre-processing, classification and clustering. Additionally, we briefly explain text mining in biomedical and health care domains.

5. Text Classification using Different Feature Extraction Approaches Text Classification using Different Feature Extraction Approaches. AUTHORS: Hong Liang, Xiao Sun, Yunlei Sun & Yuan Gao Erdelyi.
Selection of text feature item is a basic and important matter for text mining and information retrieval. Traditional methods of feature extraction require handcrafted features. To hand-design, an effective feature is a lengthy process, but aiming at new applications, deep learning enables to acquire new effective feature representation from training data. As a new feature extraction method, deep learning has made achievements in text mining. The major difference between deep learning and conventional methods is that deep learning automatically learns features from big data, instead of adopting handcrafted features, which mainly depends on priori knowledge of designers and is highly impossible to take the advantage of big data. Deep learning can automatically learn feature representation from big data, including millions of parameters. This thesis outlines the common methods used in text feature extraction first, and then expands frequently used deep learning methods in text feature extraction and its applications and forecasts the application of deep learning in feature extraction.5.9 hours of the CME arrival time, with 54% of the predictions having absolute errors less than5.9 hours. Comparison with other models reveals that CAT-PUMA has a more accurate prediction for 77% of the events investigated; and can be carried out very fast, i.e. within minutes

after providing the necessary input parameters of a CME. A practical guide containing the CAT-PUMA engine and the source code of two examples are available in the Appendix, allowing the community to perform their own applications for prediction using CAT-PUMA

III. EXISTING SYSTEM

A customer can become an active user by giving reviews about different products/services which may be useful to other potential customers. But there are hundreds, thousands or even more product/service related reviews available on the web and reading all those available reviews is a very tedious and taxing task for the customer. Therefore, there is a need gap for apt techniques which automatically summarize these reviews into a positive or a negative category to give useful information to the user.

Disadvantages of Existing System

- To identify their sentiments but comparatively very less work has been done in the domain of tourism reviews.
- Xing Fang and Justin Zhan proposed a new feature vector generation algorithm to perform sentiment polarity categorization of product reviews, but it not accurate for the result

IV. PROPOSED SYSTEM

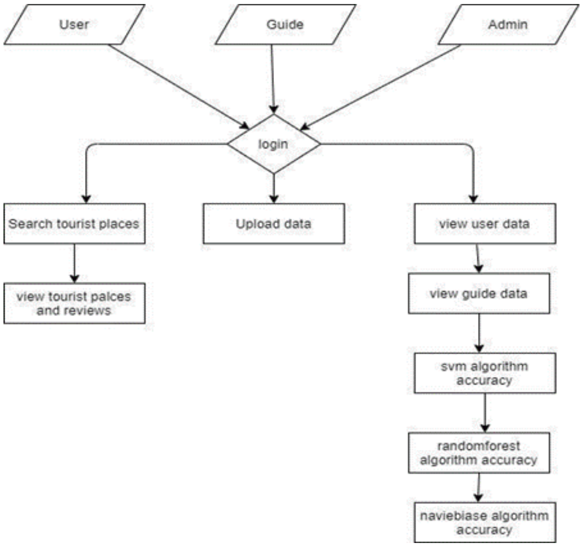
In the proposed method various techniques of sentiment analysis has been studied and compared. Different levels of sentiments are document level, sentence level, aspect level which has been elaborated Approaches used for sentiment analysis in this paper are machine learning based,
Rule based and lexical based. Inside machine learning approaches various techniques are SVM (Support Vector Machine), NB (Naive Bayes), also feature driven sentiment analysis has been described in detail. Various approaches to sentiment analysis have been compared; its corresponding advantages and disadvantages are described in detail. From Various parameters of comparison like performance, efficiency, and accuracy it has been found that machine learning approach gives the best result.

Advantages of Proposed System:

- It is observed that significant work has been done in the domain of product reviews, movie reviews, restaurant reviews, blog posts etc.
- The sentimental analysis in the tourist domain Researchers have explored various sentimental analysis techniques such as Naive Bayes and Support Vector Machine.

V. DATA FLOW DIAGRAM

- The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
- DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



VI. MODULES

- User
- Tourist guide
- Admin
- Machine learning

MODULES DESCRIPTION:
User:
The User can register the first. While registering he required a valid User email and mobile for further communications. Once the User registers, then the admin can activate the User. Once the admin activates the User then the User can login into our system. After login he can search the particular tourism place to view the reviews of the tourism place. User also can search the details of tourism place like information about tourism places and packages etc...

Tourist guide:
The guide can register the first. While registering he required a valid User email and mobile for further communications. Once the guide registers, then the admin can activate the guide .Once the admin activates the guide then the guide can login into our system. Here guide can upload the tourism places.

Admin:
Admin can login with his credentials. Once he logs in he can activate the users and tourist guide. The activated users and guide only login in our applications. We can implement naïve bayes algorithms and svm and random forest to predict sentimental analysis .

Machine learning:
Machine learning refers to the computer's acquisition of a kind of ability to make predictive judgments and make the best decisions by analyzing and learning a large number of existing data. The representation algorithms include deep learning, artificial neural networks, decision trees, enhancement algorithms and so on. The key way for computers to acquire artificial intelligence is machine learning. Nowadays, machine learning plays an important role in various fields of artificial intelligence. Whether in aspects of internet search, biometric identification, auto driving, Mars robot, or in American presidential election, military decision assistants and so on, basically, as long as there is a need for data analysis, machine learning can be used to play a role.

VII. CONCLUSION

From research study, we can infer that TFIDFVectorization has outperformed over CountVectorization feature extraction algorithm by increasing accuracy of classification. But feature extraction using TFIDFVectorization requires more execution time than CountVectorization algorithm. In research, classification algorithms Support Vector Machine(SVM), Naive Bayes(NB), Random Forest(RF) have been used. It has found that TFIDFVectorization+RF outperformed other algorithms used on bases of several evaluation parameters like accuracy, precision, recall and f1-score.

REFERENCES

[1] M.D.Devika, C.Sunitha, Amal Ganesh "Sentiment Analysis: A Comparative Study on Different Approaches" ScienceDirect Fourth International Conference on Recent Trends in Computer Science Engineering https://doi.org/10.1016/j.procs.2016.05.124

[2] Rohit Joshi, Rajkumar Tekchandani" Comparative analysis of Twitter data using supervised classifiers" 2016 International Conference on Inventive Computation Technologies (ICICT)DOI: 10.1109/ INVENTIVE.2016.7830089

[3] Harpreet Kaur, Veenu Mangat, Nidhi "A Survey of Sentiment Analysis techniques " 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) DOI: 10.1109/ISMAC.2017.8058315

[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919 [cs.CL], July 2017

[5] Robert Dzisevicˇ , Dmitrij Sˇesˇok "Text Classification using Different Feature Extraction Approaches Text Classification using Different Feature Extraction Approaches" 2019 Open

Conference of Electrical, Electronic and Information Sciences (eStream)

[6] Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, Morteza Mastery Farahani ”A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification” 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th 18thMarch 2016, Coimbatore, TN, India.

[7] Rasika Wankhede, Prof. A.N.Thakare ”Design Approach for Accuracy in Movies Reviews Using Sentiment Analysis”. International Conference on Electronics, Communication and Aerospace Technology ICECA 2017

[8] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan ”Sentiment Classifi- cation using Machine Learning Techniques ” Proceedings of the Confer- ence on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86. Association for Computational Lin- guistics.

[9] Muhammad Afzaal, Muhammad Usman ”Novel Framework for Aspect- based Opinion Classification for Tourist Places” The Tenth International Conference on Digital Information Management (ICDIM 2015)

[10] Upma kumari, Dr. Arvind K Sharma, Dinesh Soni ”Sentiment analysis of smart phone product reviews using SVM classification techniques” 2017 International Conference onEnergy, Communication, Data Analytics and Soft Computing (ICECDS)

[11] Xing Fang and Justin Zhan ”Sentiment analysis using product review data ” Springer an Journal of Big Data (2015) 2:5 DOI 10.1186/s40537- 015-0015-2

[12] C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol. 2, pp. 121–167, 1998.

[13] Leo Breiman ”RANDOM FORESTS” Statistics Department University of California, Berkeley, CA 94720

[14] C. Sheppard, Tree-based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting. CreateSpace Independent Publishing Platform, 2017.

[15] Kamal Sarkar ”Using Character N-gram Features and Multinomial Naive Bayes for Sentiment Polarity Detection in Bengali Tweets” 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)

[16] Text Classification and Naive Bayes https://web.stanford.edu/ jurafsky/slp3/slides/7NB.pdf

[17] Dixa Saxena, S. K. Saritha, PhD , K. N. S. S. V. Prasad ”Survey Paper on Feature Extraction Methods in Text Categorization” International Journal of Computer Applications (0975 – 8887) Volume 166 – No.11, May 2017

[18] https://www.tripadvisor.in/

[19] https://www.mouthshut.com.