Understanding Tourist Experiences Sentiment Analysis of Destination Reviews Using Machine Learning

Sk. Shabbir Basha 1, L. Venkata Roshini 2, A. S. L. Yasaswini 3, G. Divya Tulasi 4, V. Thanush 5, P. Venkata Sai Siddesh 6

#1 Associate Professor in Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali.

#2#3#4#5B.Tech with Specialization of Computer Science and Engineering- Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

Abstract_ In today's society, using social media is very common. Tourism websites get millions of reviews and ratings from users. These testimonials can lead to best analysis that leads to reveal a destination's general level of popularity among tourists. Tourists can get the vacation spot from the data in websites. In this work, a machine-learning strategy for sentiment analysis is presented. The information in the dataset comes from a wide range of travel review websites. We have analyzed the best method in uprooting techniques, Count Vectorization, TFIDF-Vectorization and compared their performance. Besides the well-known NB (Naive Bayes), SVM (Support Vector Machine), and RF (Random Forest) classification techniques. Several metrics including accuracy, recall, precision, and f1-score, have been used in evaluating the algorithm's relative performances. In our experiments, in terms of classification accuracy for the test dataset, we found that the TFIDF Vectorization feature extraction method performed better than the count Vectorization methodology.

INDEX TERMS: *Classification, TFIDV (Term Frequency-Inverse Document Frequency), Sentiment Analysis, SVM (Support Vector Machine), Random Forest, Machine learning.*

1. INTRODUCTION

Social networking websites are becoming increasingly popular. Millions of individuals use travel review websites every day to express their thoughts and experiences about various tourist spots. All of these testimonies can be understood using sentiment analysis. Through rigorous research and reviews, a pattern in a location's appeal among tourists can be found. The compiled findings of sentiment analysis will help travelers make destination decisions and plan their ensuing itineraries. The Count Vectorization algorithm and the TFIDF Vectorization approach are two

different feature extraction strategies used in this work to accomplish its objectives. The three classification techniques used for sentiment analysis are Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). Only a few of the measures used to assess the effectiveness of various feature extraction and classification algorithm combinations include execution time, accuracy, recall, precision, and f1-score.

2. LITERATURE SURVEY

M. D. Devika, C. Sunitha, Amal Ganesh "Sentiment Analysis: A Comparative Study on Different Approaches" Scinece Direct Fourth International Conference on Recent Trends in Computer Science Engineering

Sentiment Analysis (SA) is crucial in extracting user emotions from vast online reviews. This paper explores various SA techniques, aiming to efficiently categorize reviews. By focusing on analyzing social media reviews, the research aims to provide valuable insights for decision-making processes. It discusses machine learning methods such as SVM, NB, and Maximum Entropy, alongside semantic analysis and rule-based approaches. Research emphasizes analyzing social media reviews for better decision-making. By considering diverse methodologies, the study seeks to enhance sentiment analysis accuracy and usability in navigating the wealth of online information.

Rohit Joshi, Rajkumar Tekchandani "Comparative analysis of Twitter data using supervised classifiers" 2016 International Conference on Inventive Computation Technologies (ICICT)

Presented at the 2016 International Conference on Inventive Computation Technologies (ICICT), the paper authored by Rohit Joshi and Rajkumar Tekchandani offers a comparative analysis of Twitter data utilizing supervised classifiers. The study focuses on employing machine learning techniques to analyze Twitter data, a rich source of real-time information and opinions. By utilizing supervised classifiers, the authors aim to categorize tweets into different classes or sentiments, enabling insights into public opinions, trends, and sentiments on various topics. The paper likely explores the performance of different supervised classification algorithms in handling Twitter data, assessing factors such as accuracy, precision, and recall. Through this comparative analysis, Joshi and Tekchandani contribute to the understanding of how machine learning techniques can be effectively applied to social media data, particularly Twitter, for tasks such as

sentiment analysis and trend detection.

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)

The paper explores a comprehensive overview of sentiment analysis techniques. Presented at the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud), the paper delves into the burgeoning field of sentiment analysis, which aims to computationally assess the emotional tone behind text data.

Through this survey, the paper provides valuable insights for researchers and practitioners interested in understanding the state-of-the-art in sentiment analysis and its applications across diverse domains.

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"

The paper authored by Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut offers a concise exploration of text mining methodologies. Presented as "A Brief Survey of Text Mining: Classification, Clustering, and Extraction Techniques," the paper was likely showcased at a conference or published in a journal. It provides an overview of text mining techniques. Text mining involves extracting valuable information and knowledge from unstructured text data, and the authors cover a range of approaches utilized in this process. From traditional methods like classification and clustering to advanced techniques such as information extraction, the paper outlines the landscape of text mining research. This survey serves as a valuable resource for researchers, practitioners, and enthusiasts seeking to understand the fundamentals and advancements in text mining techniques.

3. PROPOSED SYSTEM

In this research, machine learning algorithms including Support Vector Machine (SVM), Naïve Bayes, and Random Forest are utilized alongside two feature extraction algorithms: TFIDFVectorizer, CountVectorizer. These algorithms play a pivotal role in extracting meaningful features from dataset. TFIDFVectorizer prioritizes words based on both the frequency and the semantic importance, while CountVectorizer solely considers the word frequency. By training these algorithms with extracted features, the

project aims to determine the optimal combination of algorithm and feature extraction method to accurately predict sentiment from text reviews.

4. SYSTEM ARCHITECTURE

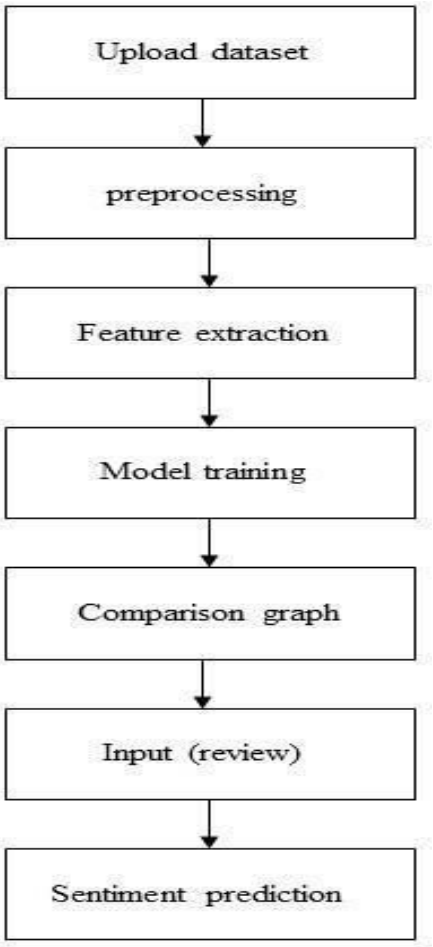


Fig 4.1: Architecture

5. MODULES DESCRIPTION

5.1 Upload Dataset

Using this module we will upload reviews dataset to the application. The dataset is selected from Kaggle website.

5.2 Data Preprocessing

Using this module we will read all reviews and then eliminate stop words and special symbols and apply NGRAM techniques to review clean text.

5.3 Count Vectorization

Using this module clean text will be converted to a count vector where each word count will be calculated and then a features vector will be generated.

5.4 TFIDF Vectorization

Using this module Term-frequency (TF) and IDF (Inverse Document-Frequency) will be calculated and then generate a features vector.

5.5 RUN SVM, Naïve Bayes, and Random Forest with TFIDF

Using this module we will train all 3 algorithms with TFIDF features and then calculate execution time, accuracy, precision, Recall, and, F1SCORE.

5.6 RUN SVM, Naïve Bayes, and Random Forest with Count Vector

Using this module we will train all 3 algorithms with Count Vector features and then calculate execution time, accuracy, precision, Recall, and F1SCORE.

5.7 Comparison Graph

Using this we will visualize the performance graph of both feature extraction algorithms with various machine learning algorithms

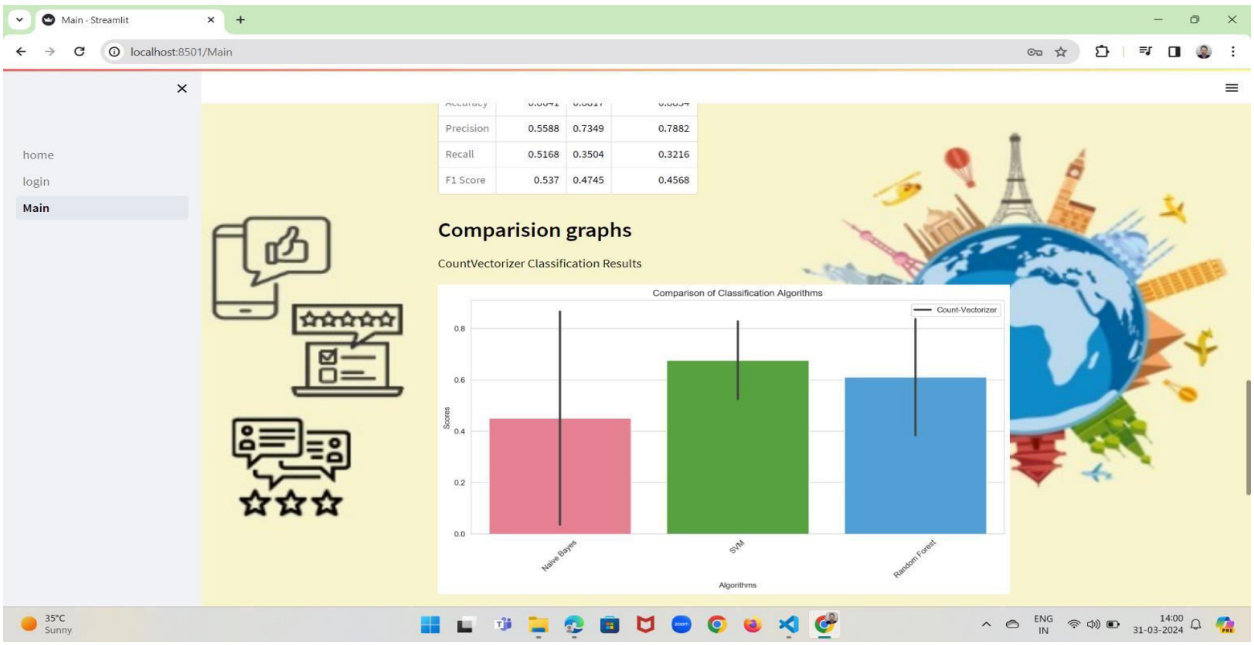
5.8 Predict Sentiments from Review

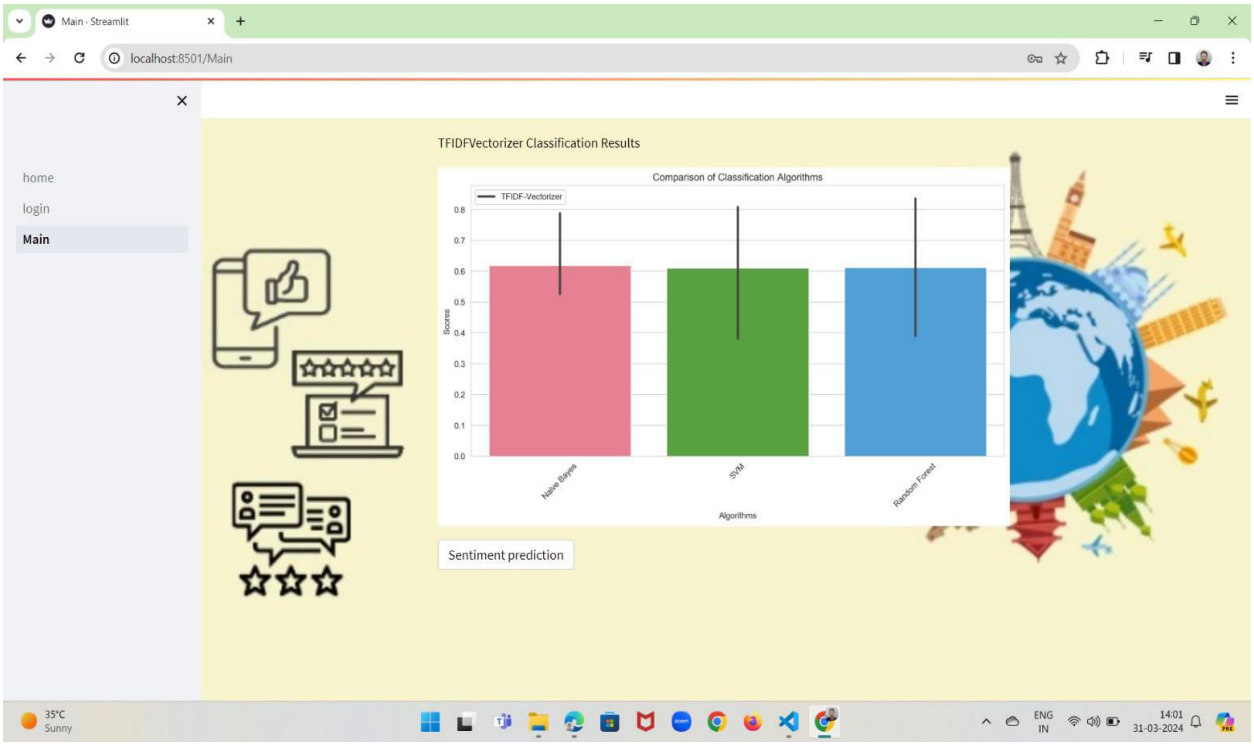
Using this module user can enter his review

and then the application calculates sentiments from that review.

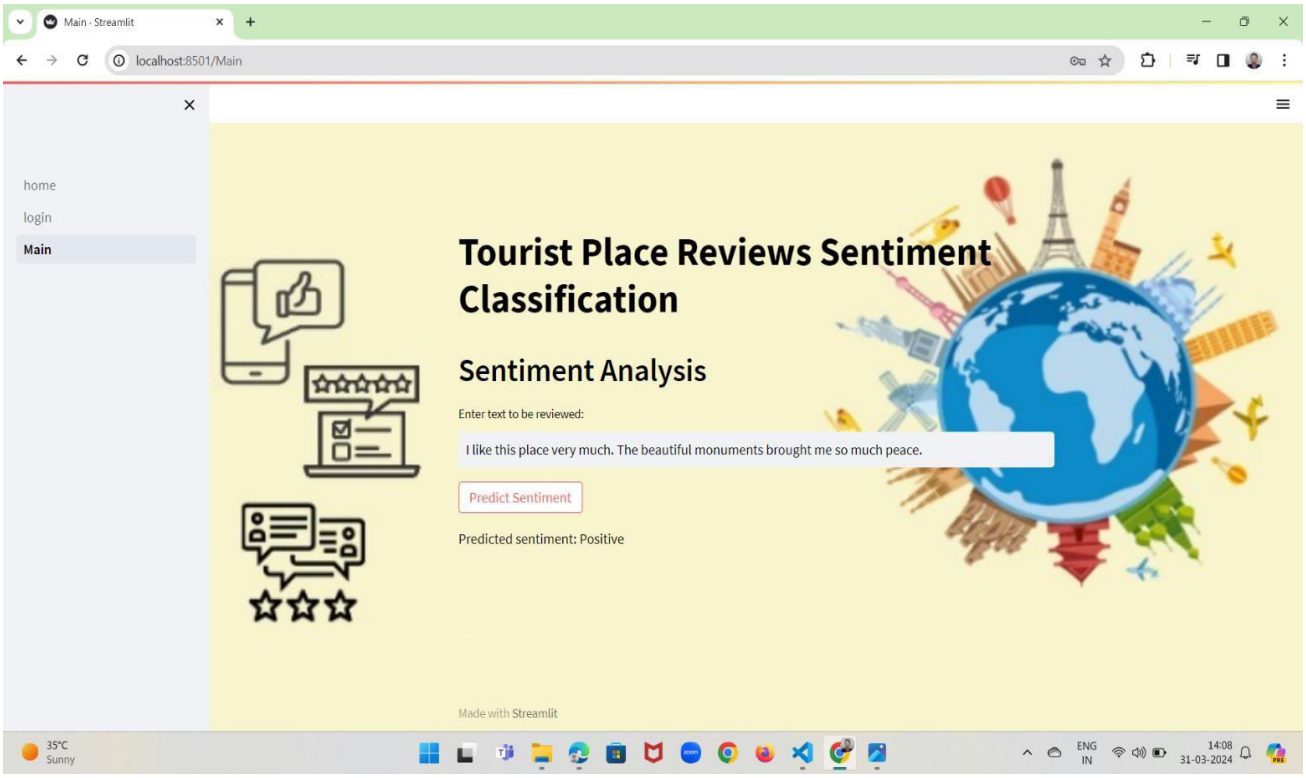
5. RESULTS AND DISCUSSION

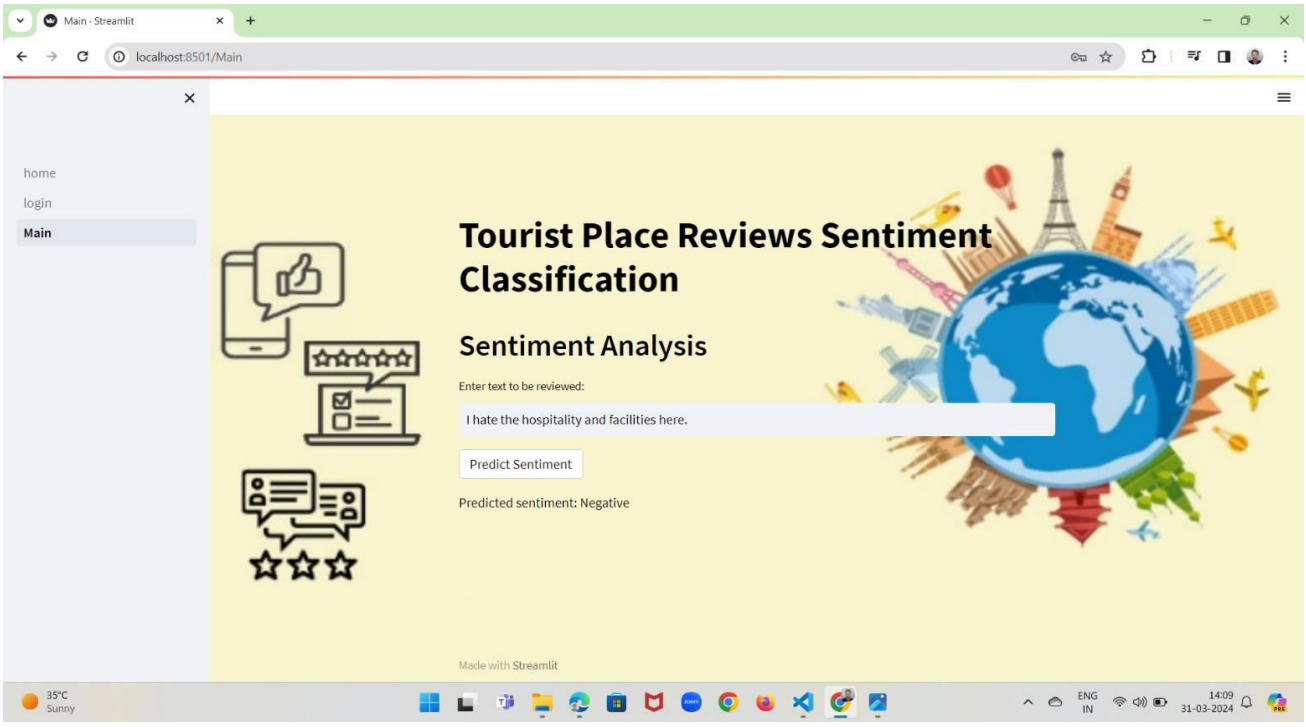
The comparison graphs are generated.

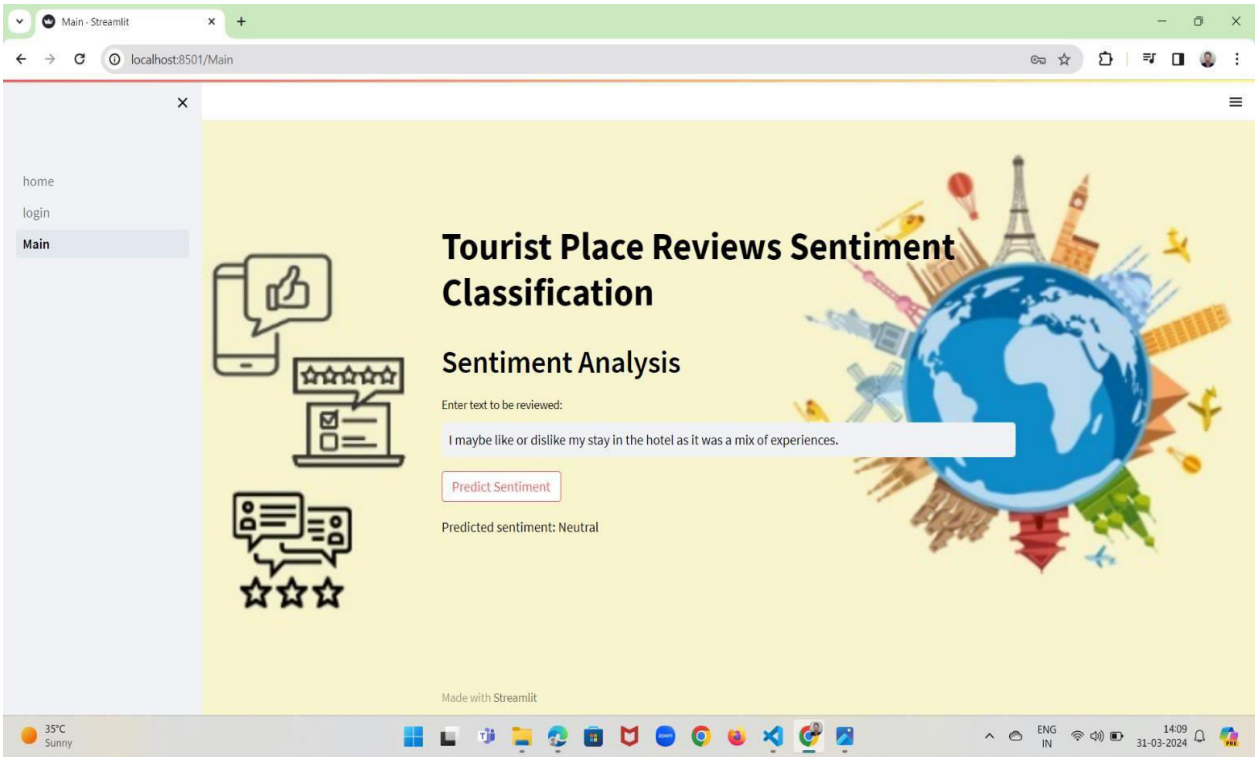




Sentiment Prediction:







6. CONCLUSION

Based on the results of the investigation, TFIDF-Vectorization appears to be a more effective feature extraction approach than Count Vectorization. However, the TFIDF Vectorization technique takes more time to execute than the Count Vectorization approach does when it comes to feature extraction. Some examples of classification algorithms used in studies include the Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF). Using metrics like accuracy, precision, recall, and f1-score,

The research study for machine learning-based review classification of tourist destinations has the potential to handle multilingual review classification in the future. Additionally, in an effort to increase classification accuracy, we will test using alternative feature selection techniques such as recursive feature elimination with cross-validation. For better performance in upcoming work, we will strive to apply deep learning-based algorithms for feature extraction and categorization.

7. FUTURE SCOPE

Automated cardiovascular disease diagnosis will be further improved using a variety of approaches, including ensemble learning for better performance, model design and hyperparameter tuning, and data

augmentation for model generalization. Real-time processing facilitates timely diagnosis, and the integration of extra clinical data improves interpretability and context. Validation and improvement are ensured by working with domain experts.

Trying out more complex CNN architectures, such as Transformer-based models and ResNet, can improve accuracy. Predictive powers are enhanced by adding lifestyle variables, medical history, and demographic data. Relevance is guaranteed by putting a feedback loop in place for ongoing model improvement. Creating intuitive mobile applications encourages proactive health management on the part of the user. Through increased clinical relevance, efficacy, and efficiency of automated diagnostic systems, patient outcomes and treatment quality will eventually be improved.

8. 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.

Author's Profiles

Sk. Shabbir Basha working as Associate Professor in Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali. He completed his M.Tech in Computer Science from Jawaharlal Nehru Technological University - Ananthpur. He has 19 years of Teaching experience.

L. Venkata Roshini, B.Tech with Specialization of Computer Science and Engineering- Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

A. S. L. Yasaswini, B.Tech with the Specialization of Computer Science and Engineering- Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

G. Divya Tulasi, B.Tech with Specialization of Computer Science and Engineering- Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

V. Thanush, B.Tech with Specialization of Computer Science and Engineering- Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

P. Venkata Sai Siddesh, B.Tech with Specialization of Computer Science and Engineering- Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.