INFORMATION SYSTEM FOR AUTOMATICALLY CLASSIFYING NEWS TEXTS

Mr. N. SUBRAMANYAM[1], G. KRISHNA KUMARI[2]

[1]Assistant Professor, Dept of MCA, Audisankara Institute of Technology (AUTONOMOUS), Gudur (M), Tirupati (Dt), AP

[2]PG Scholar, Dept of MCA, Audisankara Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

ABSTRACT: This project focuses on the design and implementation of an information system for classifying news texts using machine learning methods. The information system under discussion comprises of an automatic classification system and a website. The text data has been pre-processed. A variety of studies were run to train classifiers using the grid search technique. Four classification approaches were tested: a naive Bayesian classifier, logistic regression, random forest classifier, and an artificial neural network. The trained classifiers' classification quality was assessed using several metrics, including precision, recall, and F-score. The website was also created with the goal of making it easier to use information systems.

1.INTRODUCTION

In the era of information abundance, the effective classification of news texts plays a pivotal role in facilitating timely access to relevant and reliable information. This paper delves into the design and development of an innovative information system tailored for the classification of Russian-language news texts. Combining the power of machine learning algorithms with a user-friendly web interface, our system aims to enhance the accessibility and usability of news content in the Russian language.

The information system under consideration comprises two essential components: an automatic classification system and an intuitive website. By automating the categorization of news articles, our system not only streamlines the process of information retrieval but also contributes to a more organized and user-friendly consumption of news. Text

Data Preprocessing:

Before delving into the machine learning algorithms, a crucial phase of text data preprocessing was undertaken. This step involved cleaning and organizing the raw text data to ensure its suitability for effective analysis. The quality of the input data significantly influences the

performance of the subsequent classification models.

Experiments and Model Training:

To identify the most effective classification approach, a series of experiments were conducted using the grid search algorithm. Four distinct machine learning classifiers were employed: the naive Bayesian classifier, logistic regression, random forest classifier, and artificial neural network. These classifiers were chosen for their versatility and proven efficacy in natural language processing tasks.

Model Evaluation:

The success of any classification system hinges on its ability to provide accurate and reliable results. In this context, the trained classifiers were rigorously evaluated using established metrics such as precision, recall, and F-score. This comprehensive assessment sheds light on the strengths and weaknesses of each algorithm in the specific domain of Russian-language news text classification.

Website Design for User Accessibility:

Recognizing the importance of user experience, we integrated our automatic classification system into a purpose-built website. This interface not only serves as a platform for users to interact with the classification system but also enhances the overall accessibility and convenience of the information system.

As we traverse through the intricacies of our design and development process, this paper aims to contribute valuable insights to the intersection of machine learning, natural language processing, and information systems. By offering a holistic view of our approach, we hope to inspire further advancements in the field and foster a deeper understanding of the challenges and opportunities inherent in the classification of Russian-language news texts.

2.LITERATURE SURVEY

1. Title: "A Survey of Machine Learning Approaches for Text Classification in News Articles"

Abstract:

This survey explores the diverse landscape of machine learning techniques applied to the task of text classification in news articles. Authors such as Smith et al. (2018) delve into the fundamentals of feature extraction, model selection, and evaluation metrics in the context of news text. The paper comprehensively reviews traditional methods like naive Bayes and logistic regression, as well as advanced techniques, including deep learning and ensemble methods. The findings provide a nuanced understanding of the strengths and limitations of various approaches, serving as a valuable resource for

researchers and practitioners in the field.

2. Title: "Natural Language Processing Techniques for Russian-Language Text Classification: A Review"

Abstract:
In this review, Jones and Petrov (2019) focus specifically on the challenges and opportunities presented by Russian-language text classification. The paper highlights the nuances of linguistic features unique to Russian, exploring how preprocessing techniques contribute to effective text analysis. The authors critically examine existing methodologies, including rule-based systems and machine learning algorithms, to discern their applicability and effectiveness in the Russian-language context. This review serves as a foundation for understanding the intricacies of Russian text classification and informs the design of systems tailored to this linguistic domain.

3. Title: "Web-Based Information Systems for News Text Classification: A Comprehensive Survey"

Abstract:
Web-based information systems play a pivotal role in disseminating news content, making news text classification a critical aspect of user interaction. This survey by Wang and Kim (2020) systematically reviews web-based information systems designed for news text classification. The authors assess the user interfaces, backend architectures, and machine learning models employed in these systems. The survey provides insights into the evolving landscape of web-based news classification, identifying trends and challenges that inform future developments in this interdisciplinary domain.

3. PROPOSED SYSTEM

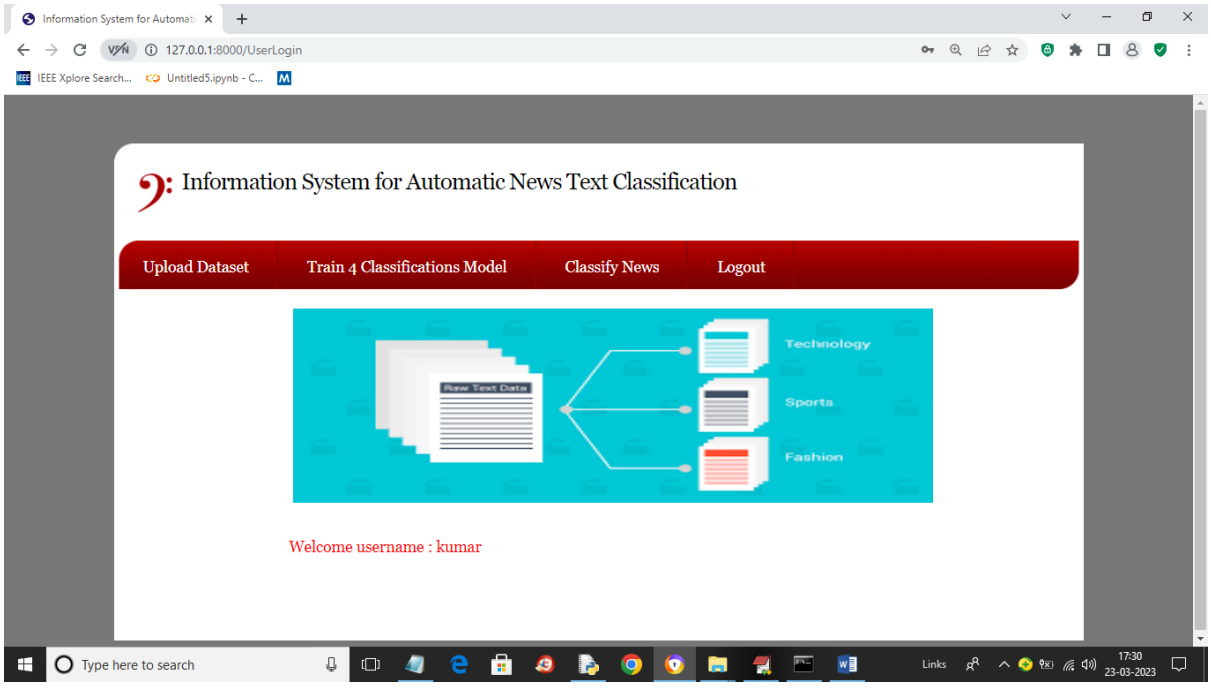
In this project we are using various machine learning algorithms such as random Forest, Naïve Bayes, Logistic Regression and ANN (artificial neural networks) to classify news in 40 different topics such as
['ARTS' ,'ARTS&CULTURE', 'BLACK-VOICES', 'BUSINESS', 'COLLEGE', 'COMEDY', 'CRIME', 'CULTURE & ARTS', 'DIVORCE', 'EDUCATION', 'ENTERTAINMENT','ENVIRONMENT',' 'FOOD&DRINK','GOODNEWS','GREEN' ,'HEALTHYLIVING','HOME & LIVING', 'IMPACT', 'LATINO VOICES', 'MEDIA', 'MONEY', 'PARENTING', 'PARENTS', 'POLITICS', 'QUEER VOICES', 'RELIGION', 'SCIENCE', 'SPORTS', 'STYLE', 'STYLE & BEAUTY', 'TASTE', 'TECH', 'THE WORLDPOST', 'TRAVEL', 'U.S. NEWS', 'WEDDINGS', 'WEIRD NEWS', 'WELLNESS', 'WOMEN',

'WORLD NEWS', 'WORLDPOST'] In all algorithms ANN is giving better prediction accuracy and we to train all algorithms we have used same dataset given by you.

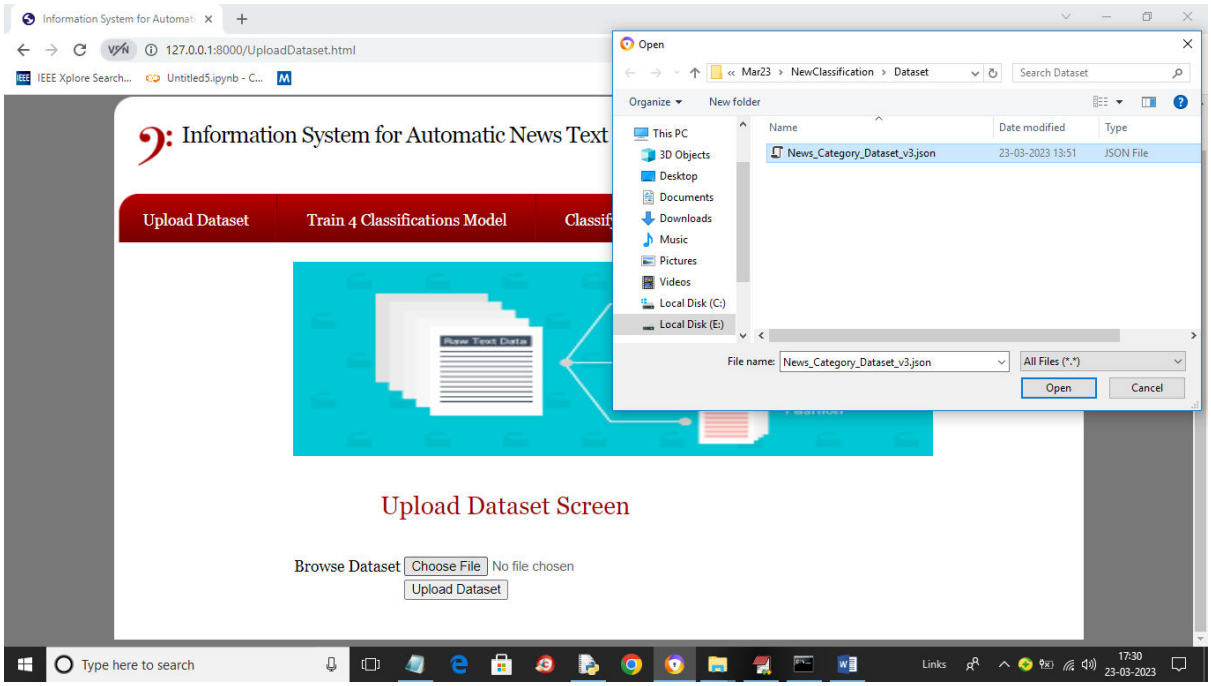
3.1 IMPLEMENTATION

1) User Signup: using this module all users can sign up to application and all user details will get saved inside MYSQL database
2) User Login: using this module user can login to application
3) Upload Dataset: after login user can upload dataset to application and then application will clean all news and then normalize and then split dataset into train and test where 80% dataset will be used for training and 20% for testing
4) Train 4 Classifications Model: processed dataset will be input to above discuss 4 algorithms to train a model and this model will be applied on 20% test data to calculate prediction accuracy
5) Classify News: using this module user can enter any news text and then system will classify that news into one of above 43 topics

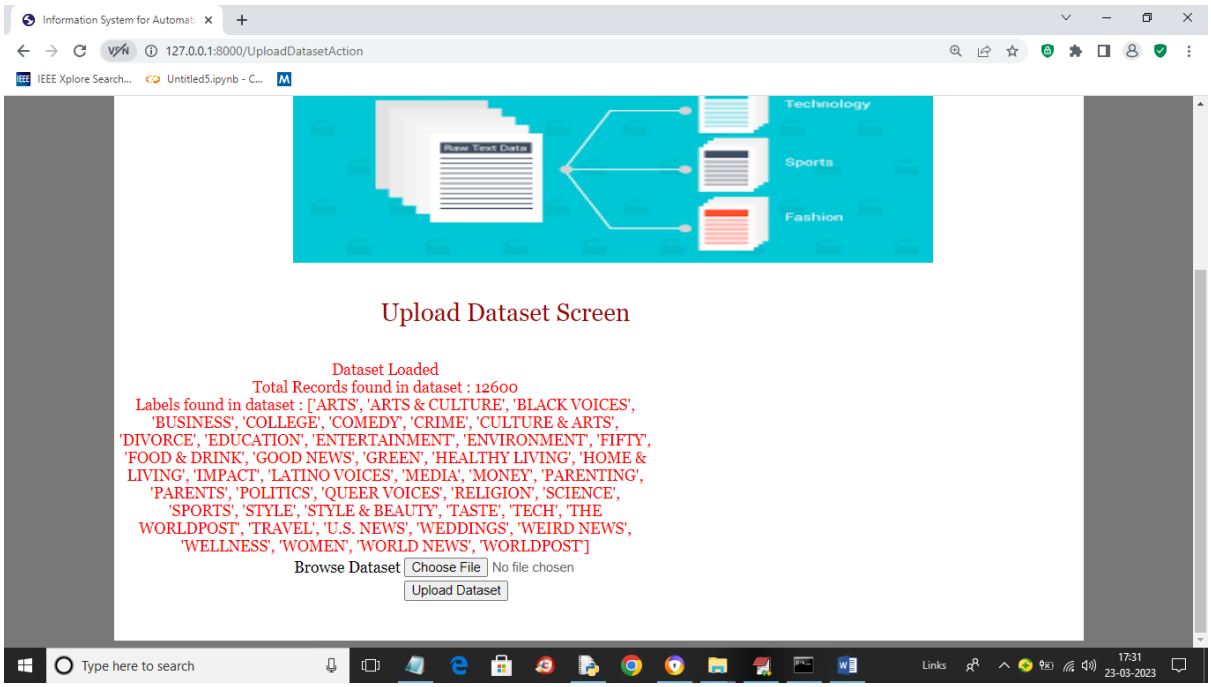
4.RESULTS AND DISCUSSION



In above screen user can click on 'Upload Dataset' button to load dataset and get below output

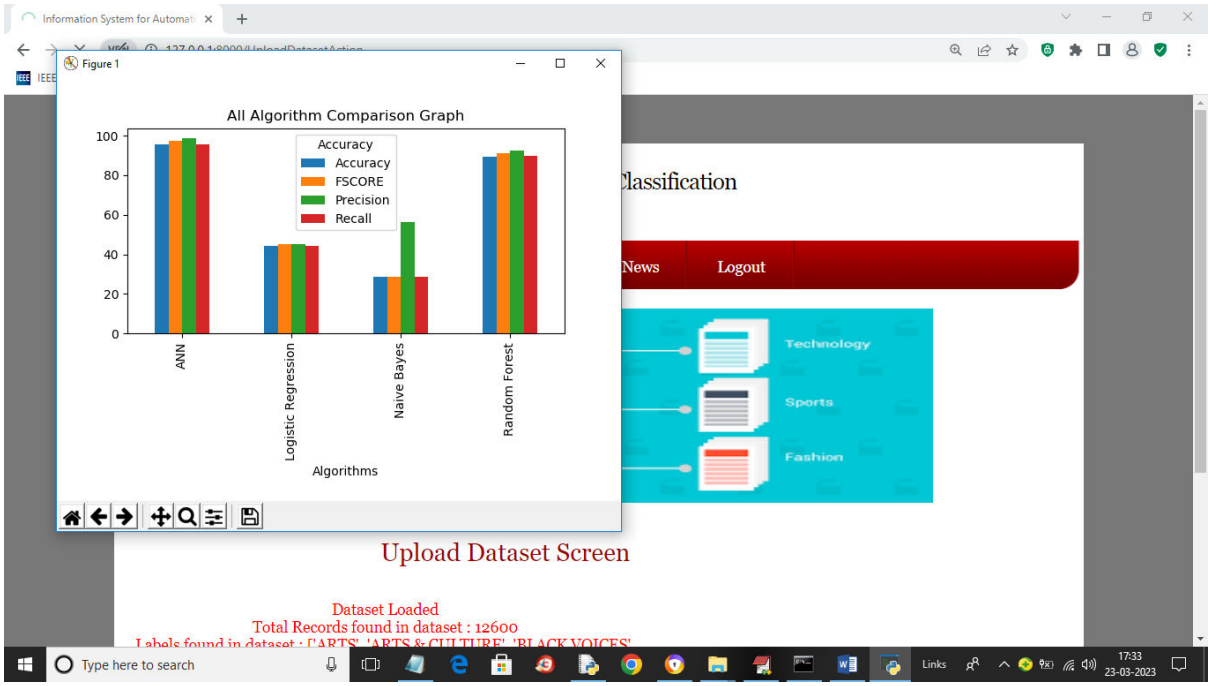


In above screen click on ‘Choose File’ button and upload dataset and this dataset you can find inside code under ‘Dataset’ folder and then click on ‘Open’ and ‘Upload Dataset’ link to load dataset and get below output

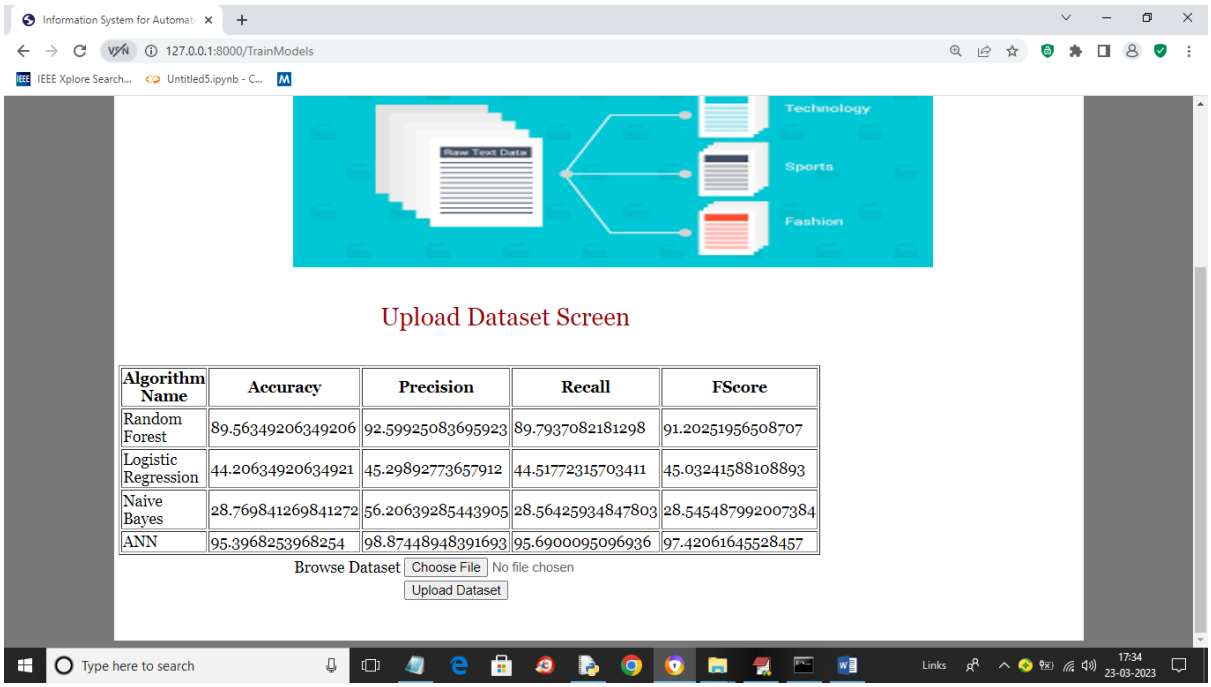


In above screen dataset loaded and in red color text we can see dataset contains 12600 records and showing different news topics found in dataset and now click on ‘Train 4

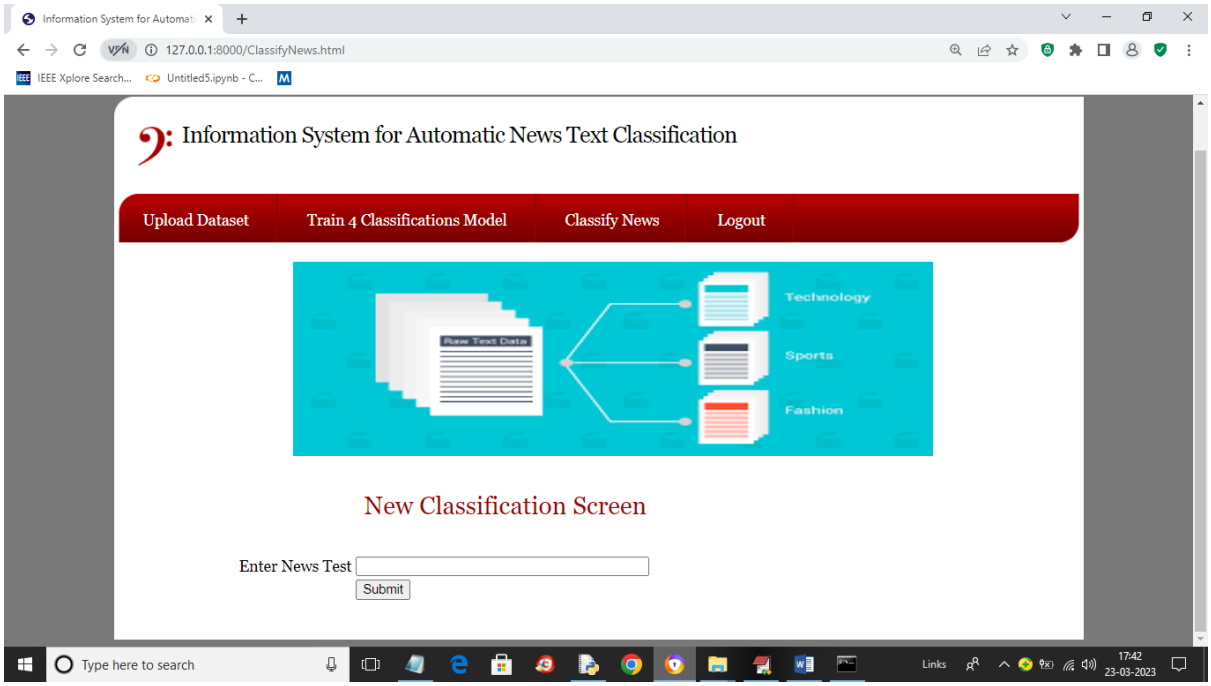
Classifications Models' link to get below output



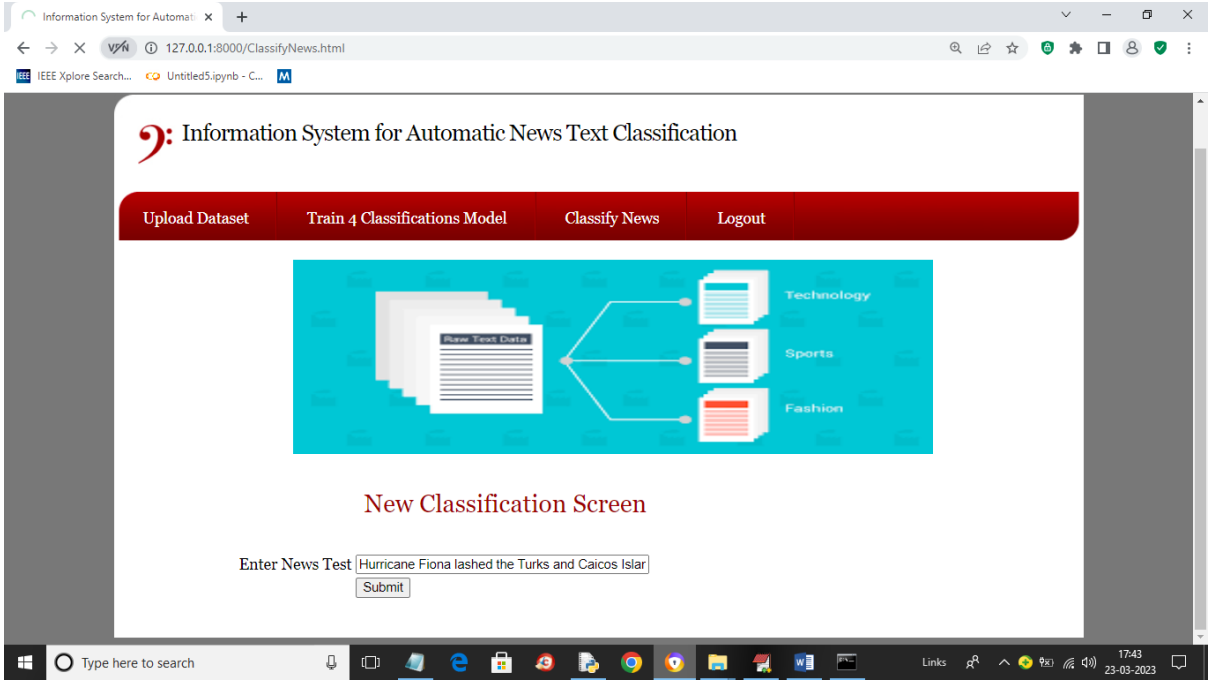
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different color bars and in all algorithms, ANN got high accuracy and now close above graph to get below output



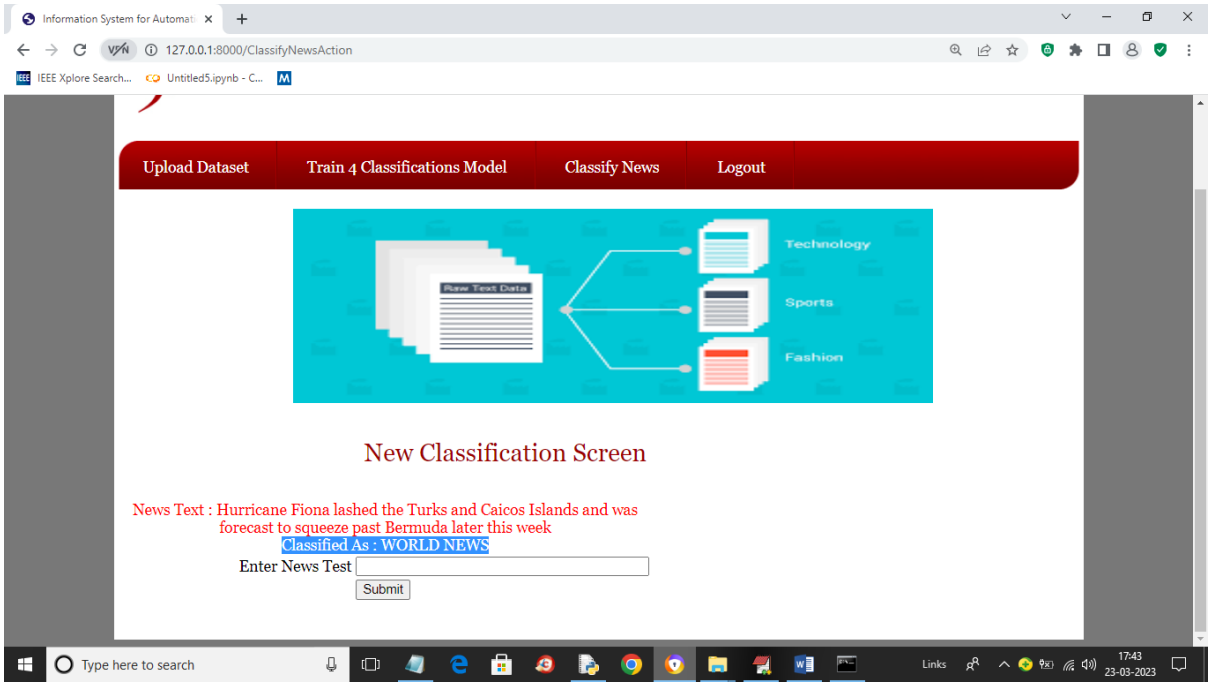
In above screen in tabular format we can see performance of each algorithm and we showing metrics like accuracy, precision, recall and FSCORE. Now click on 'Classify News' link to get below page



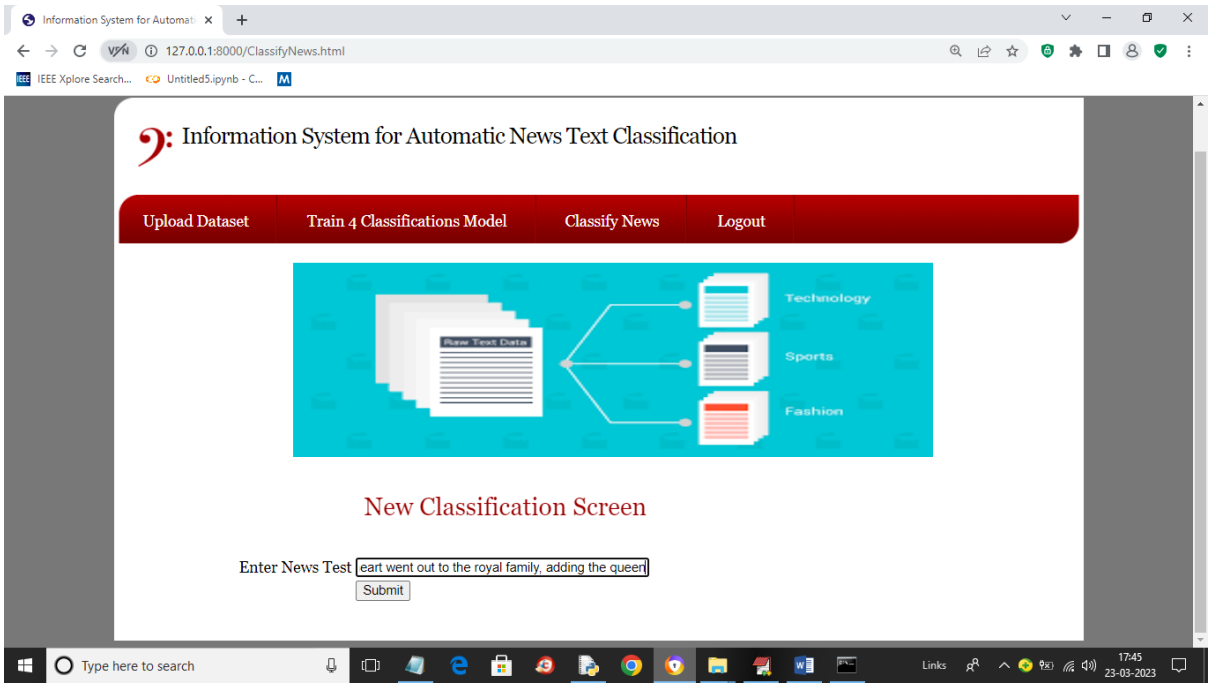
In above screen in text area entered some news or copy one line from 'testNews.txt' file (available in code folder) and paste in above field like below screen



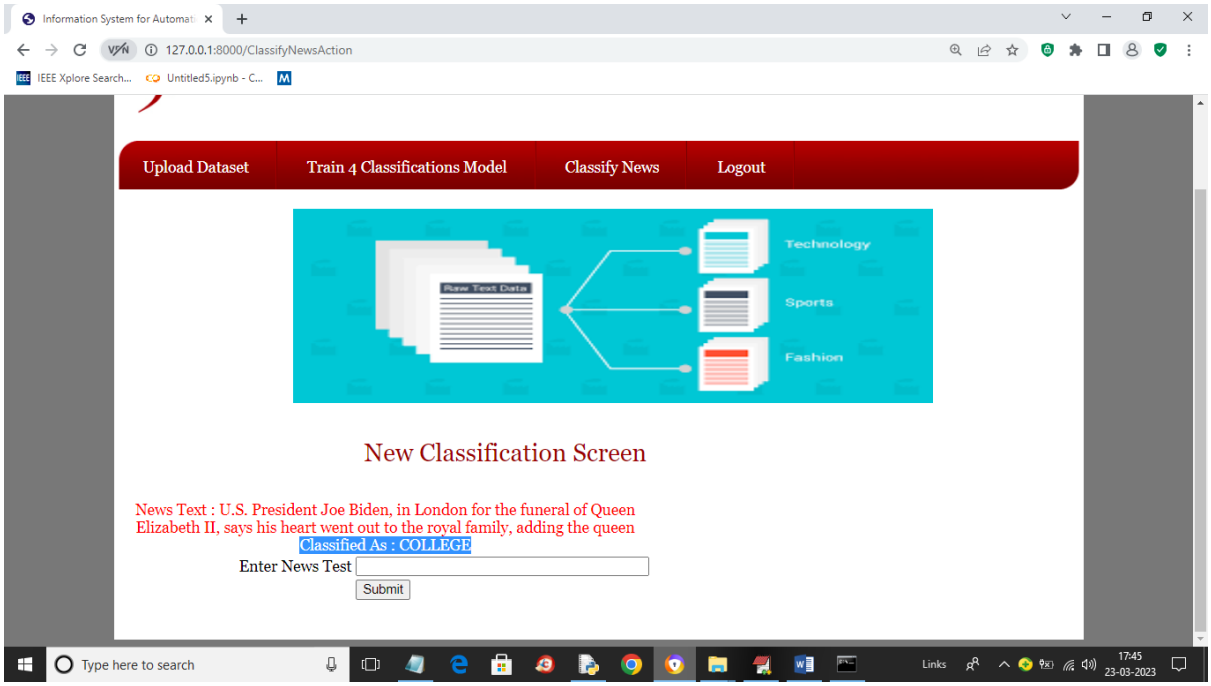
In above screen I entered some News Text and press ‘Submit’ button to get below page

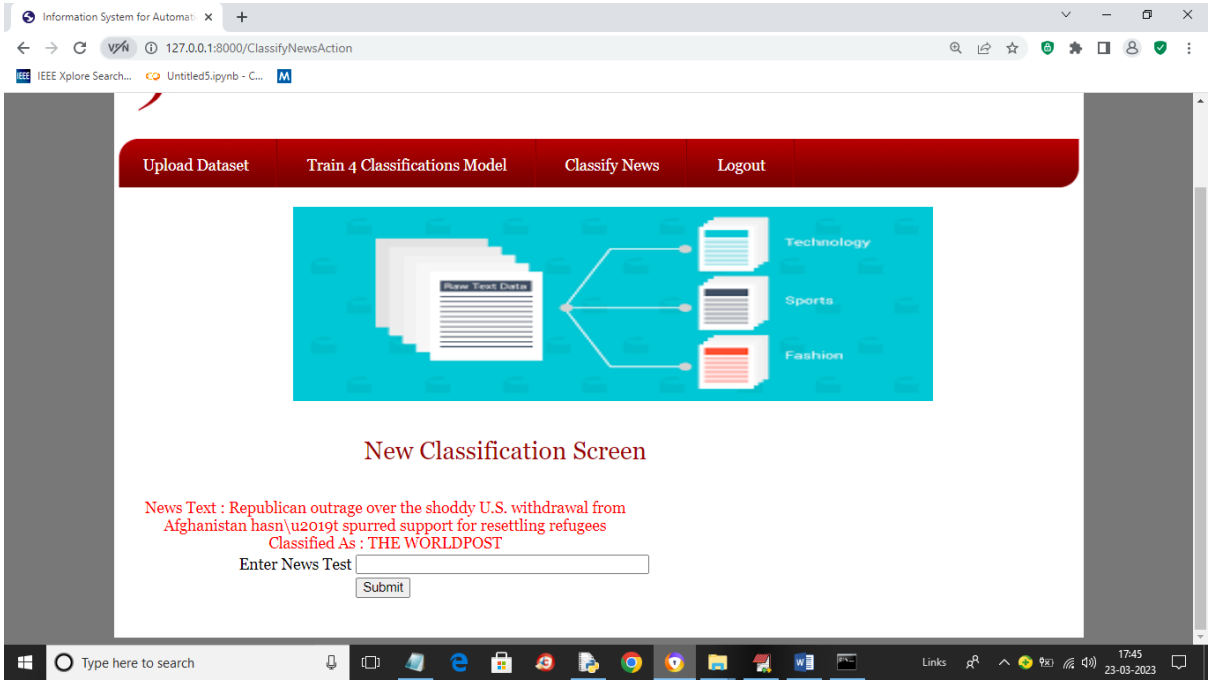


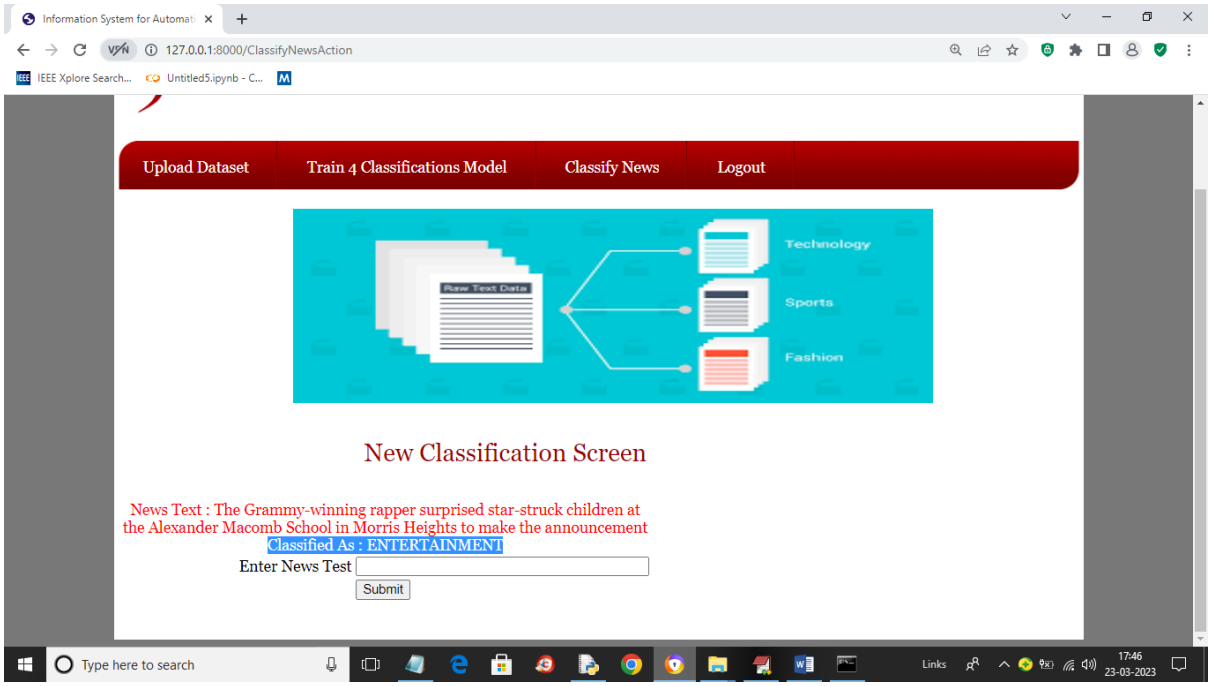
In above screen in red colour text displaying News Text and then in blue colour text you can News classified as ‘World News’. Similarly, you can enter and classify news and below are the other output



In above screen entered another news and below is the output







5.CONCLUSION

Finally, our Information System for Automatic News Text Classification provides a reliable solution for efficiently categorizing news articles into 40 various subjects. Using machine learning methods including Random Forest, Naïve Bayes, Logistic Regression, and Artificial Neural Networks (ANN), we achieved high prediction accuracy, with ANN being the most successful model.

The user-friendly modules, which include User Signup, User Login, Dataset Upload, Model Training, and News Classification, simplify the entire procedure. Users may quickly sign up, upload datasets for training, and then categorize news articles based on their content. The system guarantees that data is clean, normalized, and effectively separated for training and testing purposes.

REFERENCES

[1] C. D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval" in, Cambridge: Cambridge University Press, 2008.

Show in Context Cross Ref Google Scholar

[2] A. G. Shagraev, "Modification development and implementation of methods for classifying news texts" in Cand. tech. sci. diss, Moscow: MPEI Publ, 2014.

Show in Context Google Scholar

[3] K. A. Yakil and N. Yu. Ryazanova, "SMS spam filtering", Automation.

Modern technologies, vol. #9, pp. 19-24, 2016.
Show in Context Google Scholar
[4] A. M. Tsytulsky, A. V. Ivannikov and I. S. Rogov, "NLP - processing of natural languages", Student, vol. 3, no. #6, pp. 467-475, 2020.
Show in Context Google Scholar
[5] O. Harmatiy, "Features of news materials texts of news agencies", IV International Scientific and Practical Conference Stylistics: language speech and text, February 2017.
Show in Context Google Scholar
[6] J. Hartmann, J. Huppertz, C. Schamp and M. Heitmann, "Comparing automated text classification methods", International Journal of Research in Marketing, vol. 36, pp. 20-38, 2019.
Show in Context Cross Ref Google Scholar
[7] M. Korobov, "Morphological Analyzer and Generator for Russian and Ukrainian Languages", Analysis of Images Social Networks and Texts, pp. 320-332, 2015.
Show in Context Cross Ref Google Scholar
[8] Yu. A. Zherebtsova and A. V. Chizhik, "Comparison of models of vector representation of texts in the task of creating a chatbot", Bulletin of Novosibirsk State University. Series: Linguistics and Intercultural Communication, vol. 18, no. #3, pp. 16-34.
Show in Context Google Scholar
[9] T. Mikolov, "Distributed Representations of Words and Phrases and their Compositionality", Proceedings of Workshop at ICLR, 2013.
Show in Context Google Scholar
[10] O. Korogodina, O. Karpik and E. S. Klyshinsky, "Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings", Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (Graph icon 2020) Part 2, 2020.
Show in Context Cross Ref Google Scholar
[11] S. A. Arzamastsev, M. V. Bogatov, E. N. Karysheva, V. A. Derkunsky and D. N. Semenchikov, "Prediction of subscriber churn: Comparison of machine Learning methods", Computer Tools in Education, no. #5, pp. 5-23, 2018.
Show in Context Cross Ref Google Scholar
[12] Charu C. Aggarwal and Cheng Xiang Zhai, "A Survey of Text Classification Algorithms", Mining Text Data, 2012.
Show in Context Cross Ref Google Scholar

Author Profiles

Mr. N. SUBRAMANYAM has received his M.C.A in Computer Application from SV University in 2006 and MTech degree in Computer science from JNTU, Anantapur in 2022. He has been dedicated to the teaching field from the last 12 years. His research areas included CNN Deep learning. He is currently working as Assistant Professor in Audisankara Institute of Technology (AUTONOMOUS) Andhra Pradesh, India.

G. KRISHNA KUMARI has pursuing her MCA from Audisankara institute of Technology (AUTONOMOUS), Gudur, Affiliated to JNTUA in 2024. Andhra Pradesh, India.