Data Structure for the Automated Categorization of News Text

MR.SK.UDDANDU SAHEB[1], MISS.R LEELA MALEKYA[2]

#1 Assistant professor in the Department of DCA at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR (DT).

#2 MCA Student in Department of Computer Application (DCA) at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR (DT).

ABSTRACT

An information system for the categorization of news texts using machine learning algorithms is being planned and developed in this project. An online platform and an automated categorization system make up the data system in question. We have preprocessed the text data. In order to train classifiers using the grid search method, many experiments were carried out. We have tested four different categorization algorithms: naïve Bayesian, logistic regression, random forest, and artificial neural network. Several measures, including F-score, recall, and precision, have been used to assess the trained classifiers' classification quality. An additional goal in developing the website was to provide easy access to the information system.

INTRODUCTION

Classifying news texts effectively is crucial in this information abundance age for enabling rapid access to relevant and trustworthy information. The construction and implementation of a novel information system for the categorization of news texts written in Russian is the subject of this article. Our system's goal is to make Russian news material more accessible and usable by combining the power of machine learning algorithms with an intuitive online interface. An automated categorization system and an easy-to-navigate website are the two mainstays of the data system in question. Our technology automates the classification of news stories, which helps to make news consumption more structured and user-friendly while also simplifying information retrieval. Important text data pretreatment was carried out before diving into the machine learning techniques. In order to prepare the raw text data for proper analysis, this phase included cleaning and arranging it. The success of the following classification models is highly dependent on the accuracy

of the input data.Using the grid search method, a number of trials were carried out to determine the best categorization strategy. Neural networks, logistic regression, naïve Bayesian classifiers, and random forest classifiers were the four machine learning algorithms used. The adaptability and demonstrated usefulness of these classifiers in natural language processing tasks led to their selection. The capacity to provide trustworthy and precise findings is crucial for any classification system to be considered successful. Here, well-established measures including F-score, recall, and accuracy were used to assess the trained classifiers. In this detailed evaluation, we see how well each algorithm performs in the area of news text categorization in Russian and where it falls short. Incorporating our automated categorization system into a custom-built website allowed us to prioritize user experience. Not only does this interface facilitate user interaction with the categorization system, but it also makes the information system more accessible and convenient in general.
The goal of this study is to provide useful information for those working at the crossroads of information systems, machine learning, and natural language processing. We want to encourage further progress in the area and promote a better understanding of the difficulties and potentials of Russian-language news text categorization by providing a comprehensive overview of our methodology.

RELATED WORK

"A Survey of Machine Learning Approaches for Text Classification in News Articles"

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.

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

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.

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

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.

"Comparative Analysis of Machine Learning Algorithms for Multiclass Text Classification in News Domains"

This paper by Garcia et al. (2017) conducts a comparative analysis of machine learning algorithms in the specific context of multiclass text classification within news domains. The authors rigorously evaluate classifiers, including naive Bayesian, logistic regression, random forest, and support vector machines, using a diverse dataset of news articles. The study extends beyond binary classification, shedding light on the complexities of categorizing news texts into multiple classes. The findings contribute valuable insights into algorithm selection and performance evaluation for multiclass news text classification scenarios.

METHODOLOGY

New User Signup: Using this module, user is signing up by giving details like username, password, contact and email.

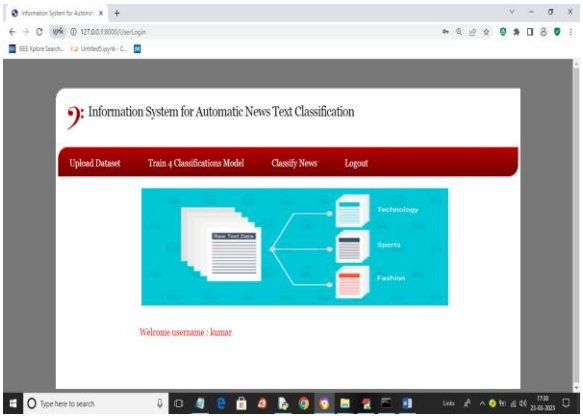
User Login: Using this module, user is login by giving username, password.

Upload Dataset: Using this module, dataset is loaded.

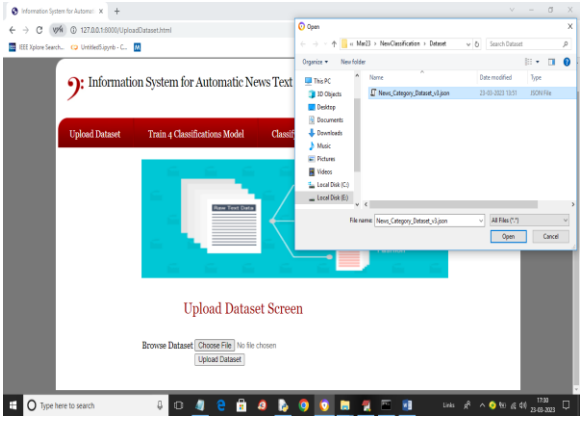
Train 4 Classifications Models: Using this module, classification models trained and ANN got high accuracy.

Classify News: Using this module, text area entered some news or copy one line from 'testNews.txt' file (available in code folder) and paste in above field

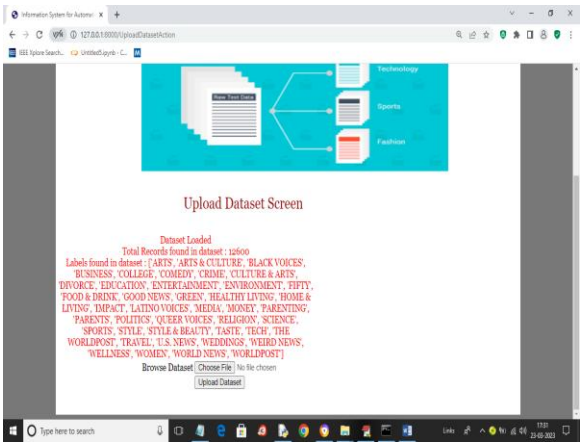
RESULT AND DISCUSSION



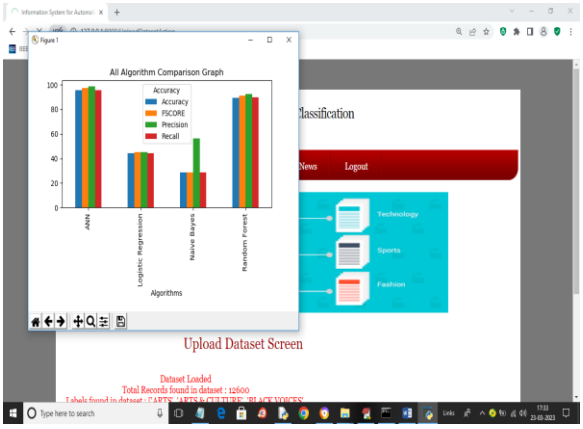
In above screen user can click on 'Upload Dataset' button to load dataset and get below output



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

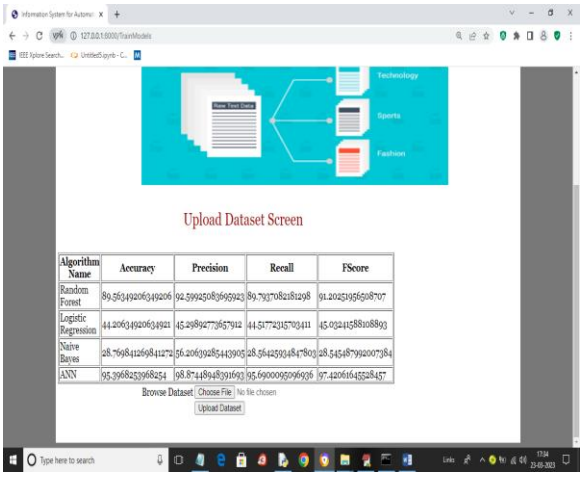


In above screen dataset loaded and in red colour text we can see dataset contains 12600 records and showing different news topics found in dataset and now click on 'Train 4 Classifications Models' link to get below output

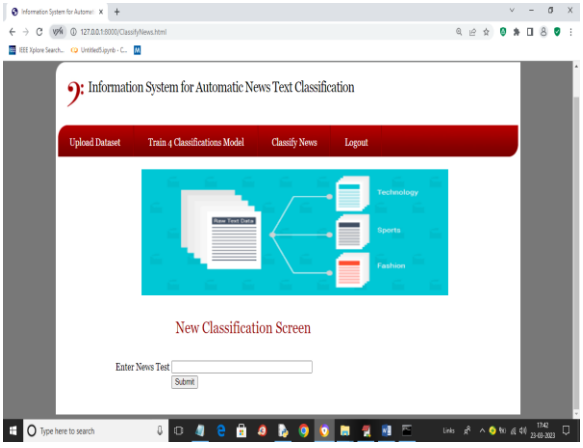


In above graph x-axis represents algorithm names and y-axis represents accuracy and

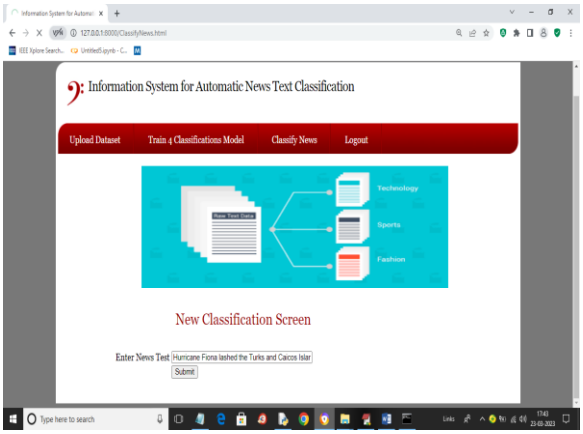
other metrics in different colour bars and in all algorithms ANN got high accuracy and now close above graph to get below output



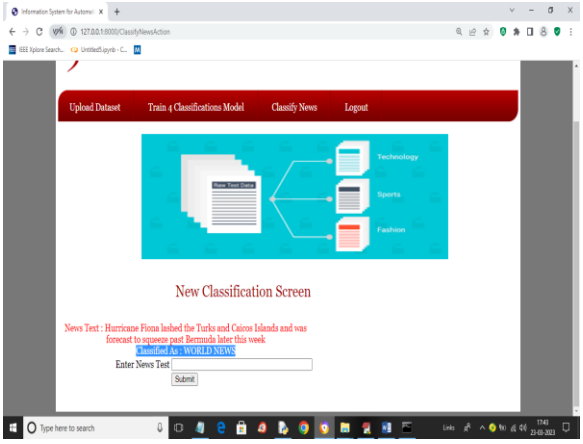
In above screen in tabular format we can see performance of each algorithm and we showing metrics like accuracy, precision, recall and FSCORE. Now click on 'Classify News' link to get below page



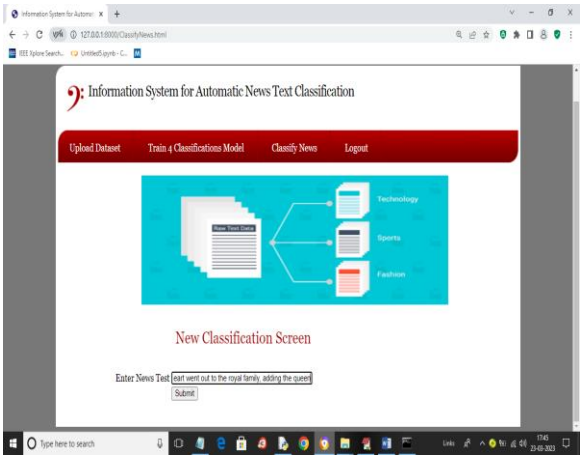
In above screen in text area entered some news or copy one line from 'testNews.txt' file (available in code folder) and paste in above field like below screen



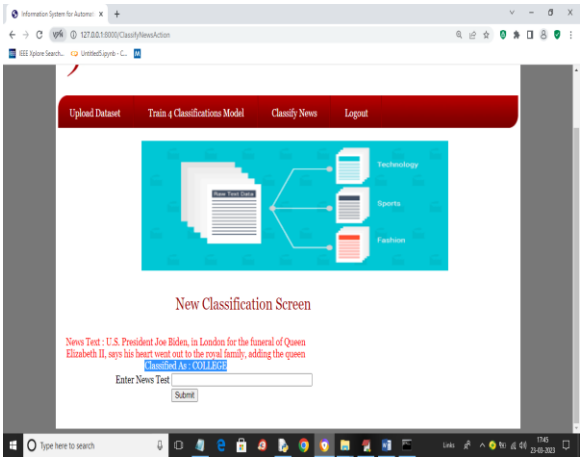
In above screen I entered some News Text and press 'Submit' button to get below page



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



CONCLUSION

The Information System for Automatic News Text Classification, a project of mine, provides a solid answer to the problem of effectively sorting news stories into forty distinct subject areas. Using ANN as the most effective model, I obtained higher prediction accuracy by using machine learning methods like Random Forest, Naïve Bayes, Logistic Regression, and others. Signup for Users, Login for Users, Upload Datasets, Train Models, and Classify News are some of the user-friendly modules that reduce friction throughout. Registration is simple, users may load training datasets, and then they can use the system to categorize news stories. Data standardization, sanitation, and efficient splitting between training and testing sets are all provided by the system.

REFERENCES

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

Show in Context CrossRef 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", StudNet, 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.

AUTHOR'S PROFILE

MR.SK UDDANDU SAHEB Completed his Master of computer Applications in Kakatiya University. He has published one paper in IJR Journal. Currently working as an Assistant professor in the department of DCA at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR District. His areas of interest are Data Structures, Machine learning, Java, and Web technologies.

MISS.R LEELA MALEKYA as MCA student in the Department of Computer Applications (DCA) at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR District. She has completed B. Sc (M. P. Cs) in Narayana Memorial Degree College Krishna University. Her areas of interest are java,Sql,pyton.