iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62610

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Text Classification for News Article

Sanket Pawar[1], Onkar Kudage[2], Shreyas Dalavi[4], Mustkeem Baraskar[5], Mihir Vanave[6], Sudarshan Bhosal[7], Mrs. S.S. Jadhav[7]

[1, 2, 3, 4, 5, 6]*Electronics & Tele-communication Engineering, Sveris College of Engineering, Pandharpur, Solapur, Maharashtra, India*

Abstract: ***In the era of information overload, effective organization and categorization of news articles are essential for providing users with relevant and timely information. This project focuses on the development and implementation of a text classification system for news articles. The primary goal is to automatically categorize news articles into predefined topics or classes, enhancing the user experience by enabling efficient content discovery and navigation. The project begins with a comprehensive collection of a diverse dataset of news articles spanning various domains such as politics, sports, technology, entertainment, and more. Preprocessing techniques are employed to clean and tokenize the text, followed by feature extraction methods that capture meaningful patterns within the text data. Various machine learning algorithms, including but not limited to, Naive Bayes, Support Vector Machines, and neural networks, are explored and evaluated to determine the optimal model for text classification. To enhance the performance of the classification system, advanced techniques such as word embeddings and transfer learning are investigated. Word embeddings like Word2Vec, Fast Text, or Glove capture semantic relationships between words, improving the model's ability to understand context. Transfer learning, particularly using pre-trained language models like BERT or GPT-3, leverages large-scale language understanding, enabling the model to generalize better even on limited labeled data. The evaluation of the text classification models involves metrics like accuracy, precision, recall, and F1-score, ensuring a comprehensive understanding of their performance across different classes. Hyperparameter tuning and model optimization are conducted to achieve the best possible results.***

Keywords: Classification model, News classification, Text classifiers, Text Mining

I. INTRODUCTION

In an era characterized by an overwhelming influx of digital information, the task of efficiently categorizing and organizing news articles has become increasingly challenging. As news consumption shifts from traditional print media to online platforms, the need for automated systems that can accurately classify articles into appropriate categories has never been more pronounced. Text classification, a subset of natural language processing (NLP), has emerged as a powerful tool to address this challenge by enabling machines to automatically assign relevant labels or categories to news articles based on their content. The sheer volume of news articles generated daily, spanning diverse topics, sources, and writing styles, underscores the necessity for efficient and accurate classification. Manual categorization of these articles is not only labor-intensive but also susceptible to human error and bias. Hence, there is a growing demand for automated methods that can streamline the process, enabling news organizations to offer their readers curated content and enhancing user experience. This project aims to contribute to the field of text classification by developing a robust and adaptable system specifically tailored for news articles. By harnessing the capabilities of machine learning and NLP techniques, the project seeks to build a model capable of analyzing the textual content of news articles and assigning them to predefined categories accurately. The system's potential applications span from news aggregators and recommendation systems to sentiment analysis and trend tracking, demonstrating its relevance in both information organization and user engagement.

II. PROPOSED METHODOLOGY

1) Text Classification of News Articles Using Machine Learning on Low-resourced Language: Tigrigna by Author Awet Fesseha; Shengwu Xiong; Eshete Derb Emiru; Abdelghani Dahou.2021 In this paper the author has proposed Text categorization or Textual document is a method that becomes more significant in tagging a textual document to their most relevant label. However, not all languages have parallel textual growth; without free and absences of a dataset, text categorization becomes interesting for Tigrigna language, i.e., low-resourced language.

2) A Study of Text Classification for Indonesian News Article. By Grelly Lucia Yovellia Londo; Dwiky Hutomo. 2015 This paper aims to do a study of text classification for Indonesian News Article. This research will explore the use of some machine learning algorithms like support vector machine, multinomial naive Bayes, and decision tree to classify news article in the Indonesian Language. From the experiment, it showed that support vector machine algorithm achieves high performance with 93% in f1 score.

3) Newspaper Article Classification using Machine Learning Techniques by J Sree Devi, M. Rama Bai, Chandrashekar Reddy.2019 About the Paper: Newspaper articles offer us insights on several news. They can be one of many categories like sports, politics, Science and Technology etc. Text classification is a need of the day as large uncategorized data is the problem everywhere. Through this study, we intend to compare several algorithms along with data preprocessing approaches to classify the newspaper articles into their respective categories. Convolutional Neural Networks (CNN) is a deep learning approach which is currently a strong competitor to other classification algorithms like SVM, Naive Bayes and KNN. We hence intend to implement Convolutional Neural Networks - a deep learning approach to classify our newspaper articles, develop an understanding of all the algorithms implemented and compare their results. We also attempt to compare the training time, prediction time and accuracies of all the algorithms.

III. NEWS CLASSIFICATION PROCESS WORKFLOW

There are different steps involved in news classification. Classification is a difficult activity as it requires pre-processing steps to covert the textual data into structured form from the un-structured form. Text classification process involves following main steps for classification of news article. These steps are data collection, pre-processing, feature selection, classification techniques application, and evaluating performance measures.

A. News Collection

The first step of news classification is accumulating news from various sources. This data may be available from various sources like newspapers, press, magazines, radio, television and World Wide Web and many more. But with the widespread network and information technology growth internet has emerged as the major source for obtaining news. Data may be in available in any format i.e. it may in .pdf, .doc, or in .html format.

B. News Pre-processing

After the collection of news text pre-processing is done. As this data comes from variety of data gathering sources and its cleaning is required so that it could be free from all corrupt and futile data. Data now needs to be discriminated from unrelated words like semicolon, commas, double quotes, full stop, and brackets, special characters etc. Data is made free from those words which appear customarily in text and are known as stop word.

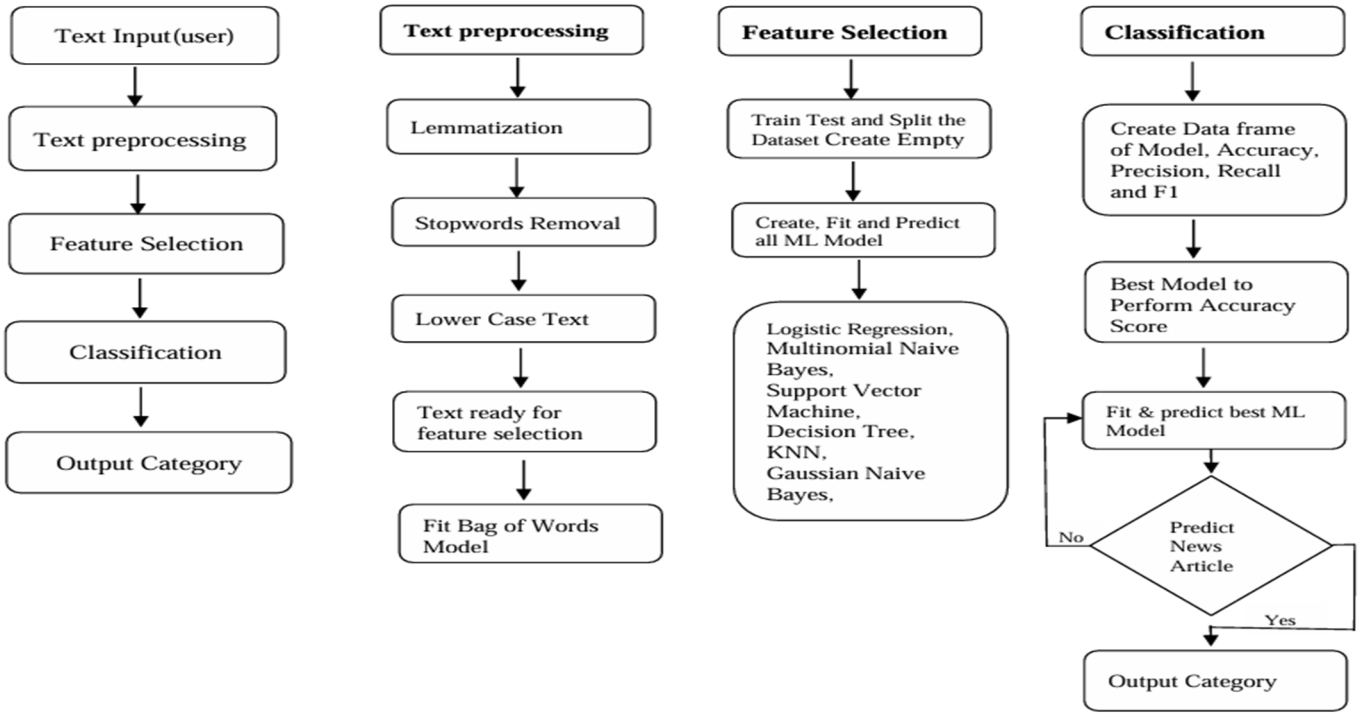


Fig: News Processing

1) Stop Word Removal

The stop words language specific and does not carry any information. It generally includes conjunctions, pronoun and prepositions. They are contemplated of low worth and are removed eventually. Thesewords need to be percolate before the processing of data.
Stop words can be removed from data in many ways. There removal can be on the basis of concepts i.e.the removal will be of the words which provide very fewer information about classification. Another way of removal of stop words is the removal of the words that are present in the list of Englishstop words. The list is made up of approx 545 stop words and is provided by Journal of Machine Learning Research. Stop words can also be abolished depending upon the frequency of their occurrence. In this method frequency of occurrence of words is computed and then weights are assigned to words. Then depending on theseweights the stop words are dropped.

2) Word Stemming

After the removal of stop words the next activity that is performed is stemming. This step reduces a word to its root. The motive behind using stemming is to remove the suffixes so that the number of words wouldbe brought down. For example the words like user, users, used, using all can be reduced to the word "USE". This will reduce the required time and space.
For stemming there exists many stemmers like S-Stemmers, Lovins Stemmer , Porter Stemmer , Porter Stemmer, Paice/Husk Stemmer. Among these stemmers M.F. Porter is mostly used.

C. Feature Selection

When there exist a large number of features and each of the features is a well known descriptive word for each class, a lot of time may be required in classification and it may be possible that expected accuracy may not be achieved and to overcome these issues, a process named as feature selection is adopted in which only those relevant and highly effective features are chosen, which may prove more noticeable for better news classification. A large number of techniques exists in literature for selecting appropriate features like Boolean weighting, Class Frequency Thresh holding, Term Frequency Inverse Class Frequency, Information Gain.

D. News Classification

After feature selection the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories. The most common news classification methods are Naive Bayes, Artificial Neural Networks, and Decision Trees, Support Vector Machines, Support Vector Machines, K-Nearest Neighbours

1) Naive Bayes

Naive Bayes is a problistic classifier based on text features. It calculates class labels and probability of classes. Naive bayes isn't made up of a single algorithm for classification but it includes a large number of algorithms that work on a single principal for training classifiers and the principal states that the value of a particular feature is autonomous of value of any other feature specified in a class. In the past classification of news article naive bayes were used. But due to its incorrect parameter assessment revamped accuracy was reported. The best thing about Naive bayes algorithm is that it works equally well on both textual as well as numeric data and it is easy to implement and calculate. But it shows poor performance when the features are short text classification.

2) Support Vector Machines

SVM has been used a lot for news text classification.SVM has a unique feature that it includes both negative and positive training sets which is generally not preferred by other algorithms.

3) Artificial Neural Networks

This network drew its concepts from neurons in which huge calculations are performed very easily by providing sufficient input and are used to estimate functions which are based on large number of inputs. Neural network when used with Naive Bayes presented a new approach known as Knowledge based neural network which is efficient in managing noisy data as well as outliers. Artificial neural network yields good results on complex domains and allows performing fast testing. But the training process is very slow.

4) Decision Tree

Decision tree is a classifier for text categorization represented in form of a tree in which each node can act as leaf or decision node. Decision tree can make appropriate decisions in situations where decisions are to betaken quickly and allowing slight delay may lead to significant difficulties.

Decision Trees are quite easily perceived and rules can be easily produced through them. Decision Trees can be used to solve intricate problems very easily. It comes with a clause that training decision tree is an expensive task. Besides this one news can be connected to one branch only. If there occurs a mistake at the higher upper level it can cause the whole subtree go invalid.

5) K-nearest Neighbors

K-nearest neighbors is a simple algorithm and a non-parameterized way of classification and regression in case of pattern recognition. For using this algorithm we need to refer K-similar text documents. It reckons the similarity against all documents that exists in the training set and uses it for making decisions about presence of class in the desired category. Neighbour that have same class are the most probable ones for that class and the neighbours with highest probability are assigned to the specific class. K-nearest neighbours is effective and non-parameterized algorithm. The biggest pitfall is that it requires a lot of classification time and it is also difficult tofind a optimal value of K.

K-nearest neighbour is a type of lazy learning where function Generalization beyond the data is delayed until a query is made to the system. K-nearest neighbour is one of the simplest machine learning algorithms.

IV. CONCLUSION

A review of news classification is bestowed in this paper. All the steps i.e. pre-processing, document indexing, feature selection, and news headlines classification are examined in detail. In addition, stop words filtering using frequency based stop words removal approach is also discussed. In future these algorithms can betested on larger corpora. Moreover these algorithms can be improved so that efficiency of categorisation could be improved. A combination of algorithm can be used in order to achieve clustering in a faster way.

REFERENCES

Journal Papers

[1] Text Classification of News Articles Using Machine Learning on Low-resourced Language: Tigrigna by Author Awet Fesseha; Shengwu Xiong; Eshete Derb Emiru; Abdelghani Dahou.2021

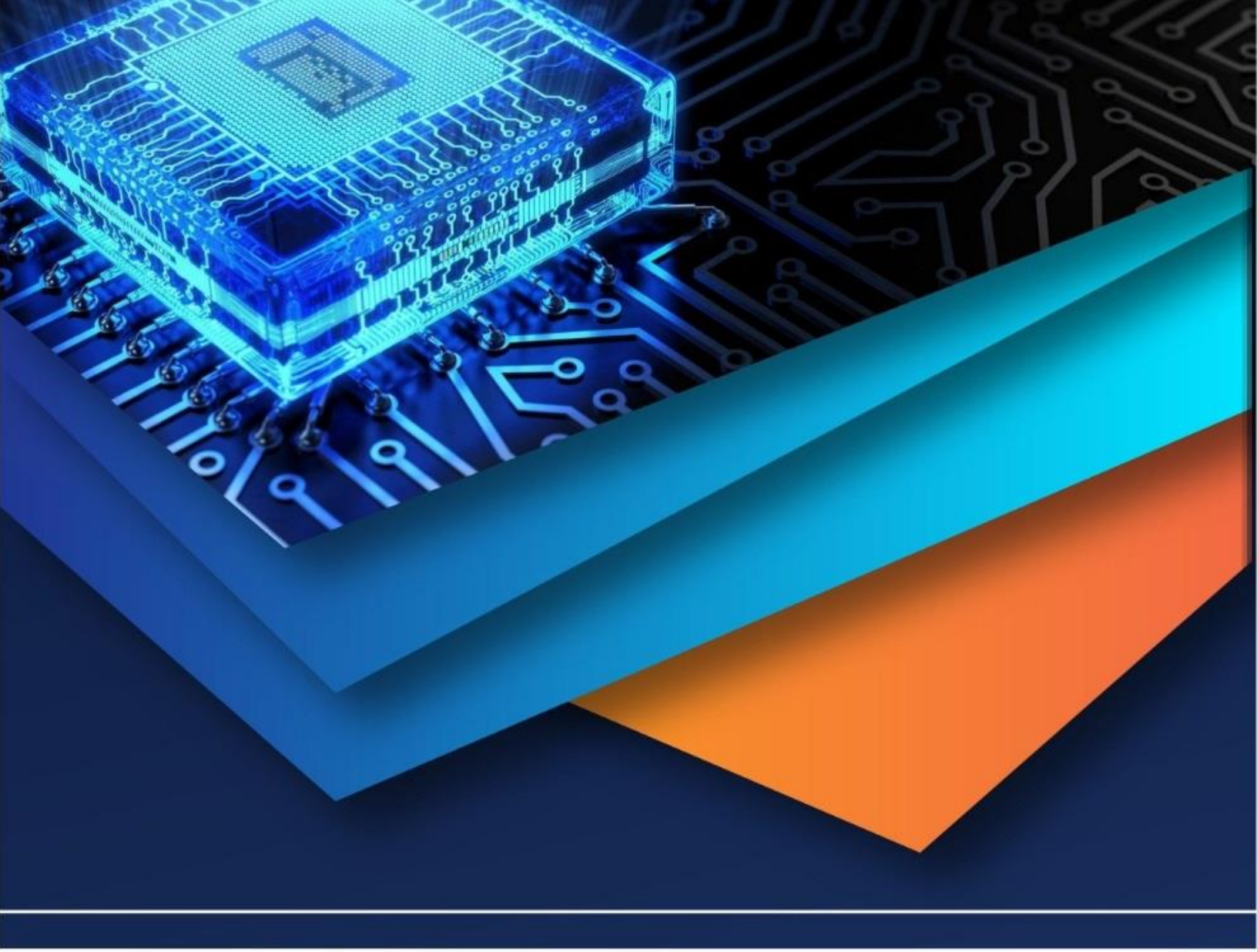
[2] A Study of Text Classification for Indonesian News Article. By Grelly Lucia Yovellia Londo; Dwiky Hutomo. 2015. http://library.unpar.ac.id/index.php?p=show_detail&id=204170#. 2015: p. 1-120.

[3] Newspaper Article Classification using Machine Learning Techniques by J Sree Devi, M. Rama Bai, Chandrashekar Reddy.2019

[4] News Text Classification Method and Simulation Based on the Hybrid Deep Learning Model by Ningfeng Sun and Chengye Du.2018

[5] Wang W, Carreira-Perpinan MA. The Role of Dimensionality Reduction in Classification. In Conference on Artificial Intelligence; 2015. California: AAAI. p. 2128-2134.

[6] Zareapoor M, K.R. S. Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection. New Delhi: 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY