SANGAM International Journal of Multidisciplinary Research

Vol. 02, Issue 02, March (2025) E-ISSN: 2584-086X

EFFICIENT BINARY CLASSIFICATION OF HINDI NEWS USING CONTEXT-BASED RETRIEVAL

Subhashini Spurjeon Kashi[1*], Dr. Ani Thomas[2]

[1*] Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg (Chhattisgarh), India
k.subhashini@bitdurg.ac.in
[2] Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg (Chhattisgarh), India
ani.thomas@bitdurg.ac.in
*Corresponding Author: *k.subhashini@bitdurg.ac.in*

Available online at: www.sijmr.org

Abstract— One of the most commonly spoken languages in India is Hindi, and as digital platforms have grown, an enormous quantity of Hindi-language documents is being created for news portals, government websites, and the public and commercial sectors. Text classification research has been stimulated by this growth in textual data, which is essential for news categorization, content tagging, sentiment analysis, and spam detection. However, limited research studies have focused on the classification of Hindi news articles. The purpose of the current research aims to bridge that gap by analyzing machine-learning approaches to Hindi news article classification. Hindi news articles for this work were collected from different news portals and assigned specific news labels according to news content. The classification process was divided into two stages: primary and secondary. The primary classification involved binary categorization of the articles into "India" and "International" labels. The secondary classification is multi class classification according to their content. Various ML and NLP models were prepared and analyzed by training the hindi news dataset, including SVM, LR, TF-IDF with NB, LSTM, CNN, Bidirectional LSTM, RNN, and BERT. The models were evaluated by performance criteria like recall, accuracy, precision, and F1 score. SVM performed well out of all the models on hindi dataset and achieved the highest testing accuracy of 85.83%, with an F1 score of 85.74% and of 86%. The results indicate that the models will perform better as the dataset size improved, highlighting the importance of data preparation and consistency between training sets and classes. This classification approach can be adapted for other Indian languages as well.

Keywords*— SVM, RNN, CNN, BiLSTM, LR, news articles*

I. INTRODUCTION

Hindi evolved to take on its current form around a millennium ago. Hindi has become an essential global language nowadays, and it needs to be shown online as an expressive language for news portals, government websites, and the commercial and public sectors [1]. Unicode has made it possible to read and write on the Web, but a few essential problems still need to be considered carefully, such as Information Extraction and text classification. Text classification is a necessary task today since there is an issue with vast amounts of uncategorized data everywhere. Natural language processing and machine learning approaches provide intelligent and practical solutions for analyzing, comprehending, and deriving meaning from Hindi news articles. Hindi News articles are vital for everyone to learn about world affairs, current happenings, sports, politics, economy, business, etc. These News articles must be categorized into predefined, mutually exclusive categories. The first classification level is binary, i.e., the data is classified into two categories – Indian and International. The secondary classification includes dividing the data into multi-label classes. They fall under one of the multiple categories, including political, crime, science and technology, sports, etc. The main challenge with using the text in Hindi to the classifying model is that, in most of the situations, the context of the text may need to be clarified. Therefore, an efficient algorithm for classifying Devanagari texts must still be developed [2]. News stories in India are still manually processed and categorized by editors. As a result, a system that can automatically assign articles to the appropriate class based on their context is required [3]. For the Classification of Hindi Newswire article for the first level, the system needs two predefined classes देश, विदेश shown in Figure 1.

The news classification system requires applying various pre-processing techniques to the text before preparing the classifying model. The pre-processing involves tokenizing, stop word removal, feature extraction, and POS tagging the words to clean up the dataset and prepare it for a machine-learning model, improving its accuracy and effectiveness. Tokenization is breaking up a written document into smaller pieces, known as tokens. The Characters, subwords, or words can all be used as tokens. Stop word removal eliminates tokens and punctuation irrelevant to the classification. The process of reducing a word to its root form is called stemming. After that, these root words are given to classifiers and represent the sentence in the document to which they belong. This article examines and summarizes the effectiveness of the SVM, TF-IDF with Naïve Bayes, LR, CNN, RNN, LSTM, Bidirectional LSTM, and Bert models on the dataset of 1200 Hindi news articles. This research article represents an analysis of various classifier models where the SVM and regression model outperform in binary classification in terms of accuracy, precision, recall, and F-score, respectively.

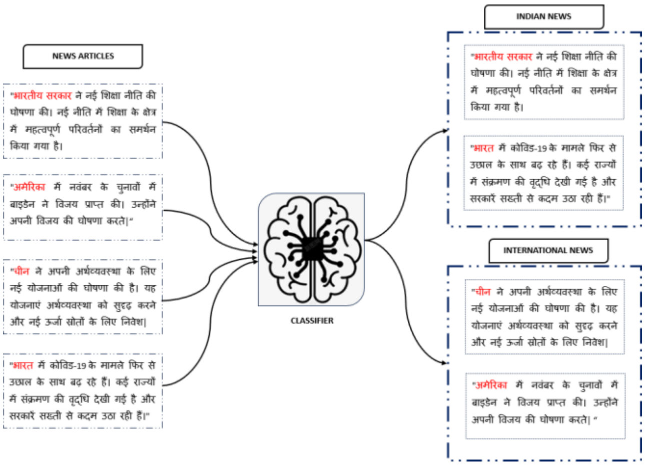

NEWS ARTICLES
"भारतीय सरकार ने नई शिक्षा नीति की घोषणा की। नई नीति में शिक्षा के क्षेत्र में महत्वपूर्ण परिवर्तनों का समर्थन किया गया है।
"अमेरिका में नवंबर के चुनावों में बाइडेन ने विजय प्राप्त की। उन्होंने अपनी विजय की घोषणा करते|"
"चीन ने अपनी अर्थव्यवस्था के लिए नई योजनाओं की घोषणा की है। यह योजनाएं अर्थव्यवस्था को सुदृढ़ करने और नई ऊर्जा स्रोतों के लिए निवेश|
"भारत में कोविड-19 के मामले फिर से उछाल के साथ बढ़ रहे हैं। कई राज्यों में संक्रमण की वृद्धि देखी गई है और सरकारें सख्ती से कदम उठा रही हैं।"
CLASSIFIER
INDIAN NEWS
"भारतीय सरकार ने नई शिक्षा नीति की घोषणा की। नई नीति में शिक्षा के क्षेत्र में महत्वपूर्ण परिवर्तनों का समर्थन किया गया है।
"भारत में कोविड-19 के मामले फिर से उछाल के साथ बढ़ रहे हैं। कई राज्यों में संक्रमण की वृद्धि देखी गई है और सरकारें सख्ती से कदम उठा रही हैं।"
INTERNATIONAL NEWS
"चीन ने अपनी अर्थव्यवस्था के लिए नई योजनाओं की घोषणा की है। यह योजनाएं अर्थव्यवस्था को सुदृढ़ करने और नई ऊर्जा स्रोतों के लिए निवेश|
"अमेरिका में नवंबर के चुनावों में बाइडेन ने विजय प्राप्त की। उन्होंने अपनी विजय की घोषणा करते| "

Figure. 1 Binary Classification system with Two classes

II. Related Works

This section represents text classification using different machine-learning algorithms and comprehensively summarizes related research work.

The previous research presents the development of a text classification system for BBC News. The authors independently selected and analyzed K-nearest neighbor, random forest, and logistic regression as classification algorithms in the classifier implementation part. A conclusion was reached after testing, analyzing, and cross-referencing these classifiers [4]. The authors mentioned that accuracy is an important parameter when applying machine learning algorithms to a specific data set. Thus, the results of the model demonstrates that the logistic regression classifier that uses the TF-IDF Vectorizer feature achieves the maximum accuracy. In an insufficient dataset, this algorithm has shown itself to be the most stable classifier. According to them, the random forest classifier performed well with reasonable accuracy. Out of the three algorithms, K-nearest neighbor algorithms had the lowest accuracy. For every parameter, the logistic regression classifier performed as predicted.

One more research work includes the classification process of five different procedures and carefully examines the results. The methods were used on raw and preprocessed data to explore the effects of data preparation. Next, the outcomes were compared using accuracy, training duration, F1-score, and recall values. The data were morphologically examined, standardized, and simplified using several techniques [5]. The approaches of NB, SVM, LTSM, LR, and RF were employed to train the data. Both before and after data preparation, the effectiveness of the different approaches was examined. Regarding training time and accuracy, the LTSM was shown to be the most efficient method. The results of this study indicate that coherence between the number of training sets and categories and normalized data are important factors affecting the approaches' performance. On the other hand, the method used in the feature extraction stage—TF-IDF—has the most influence on the outcomes. Furthermore, the techniques worked better when the data set's number of classes was decreased. The comparison also shows that the simplifying step has the most significant impact on the outcome out of all the preparation steps.

The advanced methods exist for choosing important characteristics like mutual information and information gain. Naive Bayes is one popular machine learning algorithm[6]. However, these models do have a data sparsity issue. Deep neural networks (DNNs) have a strong expressive capability and require less feature engineering than standard models, they have become prevalent in natural language processing (NLP) tasks. CNN and RNN are the two DNN variations that are utilized for text categorization [7]. The low complexity and ease of implementation of CNN model make it a helpful tool for classifying short and long texts. When the classes are relatively balanced, CNNs are a leading candidate because of their comparable performance and computational efficiency[8].

The researchers have mentioned three distinct feature engineering techniques—Count Vectorizer, TF-IDF, and Word2Vec—they preprocessed the data to extract the relevant characteristics and prepare it for machine learning algorithm training. Then, to identify the optimal combination, they applied three machine learning algorithms to each feature engineering technique: logistic regression, passive-aggressive, and multinomial naive Bayes. Consequently, the accuracy of Multinomial Naïve Bayes with count vectorizer was superior[9]. The authors also investigated and experimented with several deep-learning models for Marathi text classification. They showed that, on the available datasets, simple single-layer models based on CNN and LSTM combined with FastText embeddings outperform the BERT-based models[10].

One additional study of research classifies news headlines using a Long Short-Term Memory (LSTM) network, word embedding, cosine similarity index, and Bidirectional Encoder Representations from Transformers (BERT). This study concentrated on classification using pre-existing algorithms, such as LSTM and BiLSTM models, Word2Vec and GloVe embedding approaches, and BERT sentence vector labeling of the data. It is possible to see the accuracy of the classification for each class. Furthermore, since there was a decrease in accuracy due to a lack of training data in the politics and health categories, the precision, recall, and f1-score for these categories are extremely low compared to other types[11].

2.1 Motivation

Several authors have conducted text classification using various machine learning algorithms, but few have used multi-class classification. This venture necessitates integrating various natural language processing (NLP) techniques, combined with machine learning and a hybrid model, to evaluate the model's efficiency.

- ❖ Various classification techniques are available for English news text, but no technique has been developed for Devanagari text classification.
- ❖ A system capable of autonomously classifying news into predetermined categories is necessary.
- ❖ News classification should rely on extracting information at both sentence and paragraph levels rather than individual words.
- ❖ Automatic extracting of relevant information presents a significant challenge.
- ❖ The aim is to construct an Intelligent Information System for Hindi Newswire.
- ❖ The system should produce output aligned with the sentence's contextual understanding.
- ❖ Newswire article contents should be exclusively classified into their predetermined categories and not be reclassified into any other category.

III. RESEARCH APPROACH

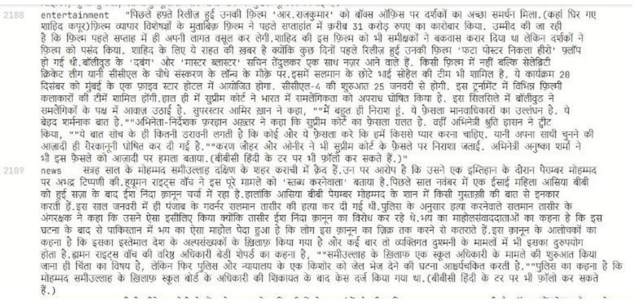
In this section, we discuss the methods implemented to carry out the various activities in the dataset cleaning and preprocessing stage and analyze the machine learning model for classifying Hindi newswire articles.

3.1 Dataset cleaning and preparation: One of the most essential elements of NLP and machine learning is data. The System requires News Articles from Various Categories to Implement the Classification Model. Thus, these data were gathered from GitHub and Kaggle's BBC Hindi News Dataset. Some of the data were also extracted from the IIT Patna Disaster dataset. This dataset is typically utilized for more than just training. A single training data set that has previously been processed is usually divided into many segments to assess how well the model was trained. Typically, a testing data set is kept apart from the data for this particular reason. The dataset contains about 3000 articles used to train and test the models.

Text Pre-processing and feature extraction: Pre-processing raw data is a crucial step in data preparation, improving the accuracy and efficacy of a machine learning model.

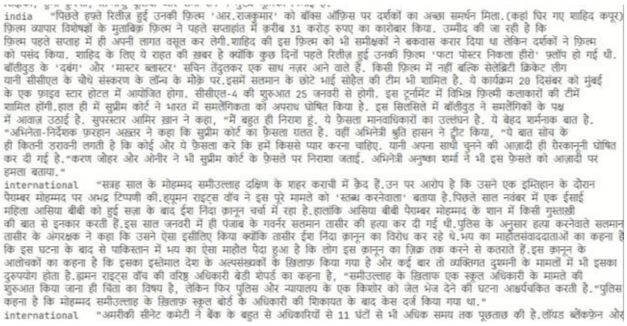
To get the dataset ready for the classification model to be applied, the following procedures were taken:

i) Dataset Cleaning: Cleaning a dataset is the process of eliminating extraneous and noisy data. The main objective is to create a uniform format for the dataset. The data was cleaned by following the steps: Cleaning the dataset started with deleting undesired, irrelevant, and empty ad lines. There are no longer any blank lines in the article; the advertisement lines have been removed, and multi-row articles have been converted into one row. The next step is to enclose each article within double quotes.



```
entertainment   "पिछले हफ़्ते रिलीज़ हुई उनकी फ़िल्म 'आर.राजकुमार' को बॉक्स ऑफ़िस पर दर्शकों का अच्छा समर्थन मिला.(कहां घिर गए
शाहिद कपूर)फ़िल्म व्यापार विशेषज्ञों के मुताबिक़ फ़िल्म ने पहले सप्ताहांत में क़रीब 31 करोड़ रुपए का कारोबार किया. उम्मीद की जा रही
है कि फ़िल्म पहले सप्ताह में ही अपनी लागत वसूल कर लेगी.शाहिद की इस फ़िल्म को भी समीक्षकों ने बकवास करार दिया था लेकिन दर्शकों ने
फ़िल्म को पसंद किया. शाहिद के लिए ये राहत की ख़बर है क्योंकि कुछ दिनों पहले रिलीज़ हुई उनकी फ़िल्म 'फटा पोस्टर निकला हीरो' फ़्लॉप
हो गई थी.बॉलीवुड के 'दबंग' और 'मास्टर ब्लास्टर' सचिन तेंदुलकर एक साथ नज़र आने वाले हैं. किसी फ़िल्म में नहीं बल्कि सेलेब्रिटी
क्रिकेट लीग यानी सीसीएल के चौथे संस्करण के लॉन्च के मौक़े पर.इसमें सलमान के छोटे भाई सोहैल की टीम भी शामिल है. ये कार्यक्रम 20
दिसंबर को मुंबई के एक फ़ाइव स्टार होटल में आयोजित होगा. सीसीएल-4 की शुरुआत 25 जनवरी से होगी. इस टूर्नामेंट में विभिन्न फ़िल्मी
कलाकारों की टीमें शामिल होंगी.हाल ही में सुप्रीम कोर्ट ने भारत में समलैंगिकता को अपराध घोषित किया है. इस सिलसिले में बॉलीवुड ने
समलैंगिकों के पक्ष में आवाज़ उठाई है. सुपरस्टार आमिर ख़ान ने कहा, ""मैं बहुत ही निराश हूं. ये फ़ैसला मानवाधिकारों का उल्लंघन है. ये
बेहद शर्मनाक बात है.""अभिनेता-निर्देशक फ़रहान अख़्तर ने कहा कि सुप्रीम कोर्ट का फ़ैसला ग़लत है. वहीं अभिनेत्री श्रुति हासन ने ट्वीट
किया, ""ये बात सोच के ही कितनी डरावनी लगती है कि कोई और ये फ़ैसला करे कि हमें किससे प्यार करना चाहिए. यानी अपना साथी चुनने की
आज़ादी ही ग़ैरकानूनी घोषित कर दी गई है.""करण जौहर और ओनीर ने भी सुप्रीम कोर्ट के फ़ैसले पर निराशा जताई. अभिनेत्री अनुष्का शर्मा ने
भी इस फ़ैसले को आज़ादी पर हमला बताया.(बीबीसी हिंदी के टर पर भी फ़ॉलो कर सकते हैं.)"
news    सत्रह साल के मोहम्मद समीउल्लाह दक्षिण के शहर कराची में क़ैद हैं.उन पर आरोप है कि उसने एक इम्तिहान के दौरान पैग़म्बर मोहम्मद
पर अभद्र टिप्पणी की.ह्यूमन राइट्स वॉच ने इस पूरे मामले को 'स्तब्ध करनेवाला' बताया है.पिछले साल नवंबर में एक ईसाई महिला आसिया बीबी
को हुई सज़ा के बाद ईश निंदा क़ानून चर्चा में रहा है.हालांकि आसिया बीबी पैग़म्बर मोहम्मद के शान में किसी गुस्ताख़ी की बात से इनकार
करती हैं.इस साल जनवरी में ही पंजाब के गवर्नर सलमान तासीर की हत्या कर दी गई थी.पुलिस के अनुसार हत्या करनेवाले सलमान तासीर के
अंगरक्षक ने कहा कि उसने ऐसा इसीलिए किया क्योंकि तासीर ईश निंदा क़ानून का विरोध कर रहे थे.भय का माहोलसंवाददाताओं का कहना है कि इस
घटना के बाद से पाकिस्तान में भय का ऐसा माहोल पैदा हुआ है कि लोग इस क़ानून का ज़िक्र तक करने से कतराते हैं.इस क़ानून के आलोचकों का
कहना है कि इसका इस्तेमाल देश के अल्पसंख्यकों के ख़िलाफ़ किया गया है और कई बार तो व्यक्तिगत दुश्मनी के मामलों में भी इसका दुरुपयोग
होता है.ह्यमन राइट्स वॉच की वरिष्ठ अधिकारी बेडी शेपर्ड का कहना है, ""समीउल्लाह के ख़िलाफ एक स्कूल अधिकारी के मामले की शुरुआत किया
जाना ही चिंता का विषय है, लेकिन फिर पुलिस और न्यायालय के एक किशोर को जेल भेज देने की घटना आश्चर्यचकित करती है.""पुलिस का कहना है कि
मोहम्मद समीउल्लाह के ख़िलाफ़ स्कूल बोर्ड के अधिकारी की शिकायत के बाद केस दर्ज किया गया था.(बीबीसी हिंदी के टर पर भी फ़ॉलो कर सकते
हैं.)
```

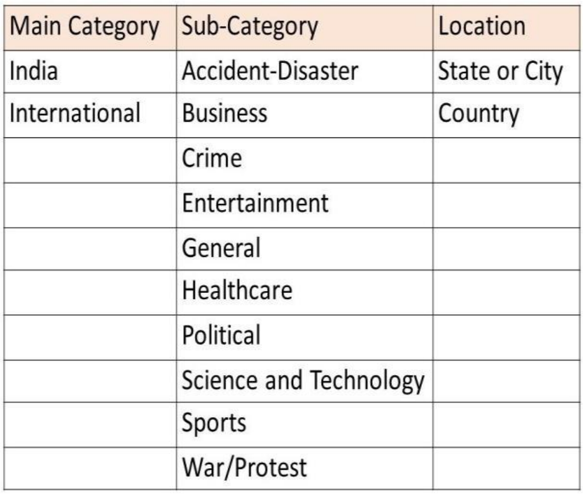
Figure 2: Dataset Before removing blank spaces and unwanted lines



```
india   "पिछले हफ़्ते रिलीज़ हुई उनकी फ़िल्म 'आर.राजकुमार' को बॉक्स ऑफ़िस पर दर्शकों का अच्छा समर्थन मिला.(कहां घिर गए शाहिद कपूर)
फ़िल्म व्यापार विशेषज्ञों के मुताबिक़ फ़िल्म ने पहले सप्ताहांत में क़रीब 31 करोड़ रुपए का कारोबार किया. उम्मीद की जा रही है कि
फ़िल्म पहले सप्ताह में ही अपनी लागत वसूल कर लेगी.शाहिद की इस फ़िल्म को भी समीक्षकों ने बकवास करार दिया था लेकिन दर्शकों ने फ़िल्म
को पसंद किया. शाहिद के लिए ये राहत की ख़बर है क्योंकि कुछ दिनों पहले रिलीज़ हुई उनकी फ़िल्म 'फटा पोस्टर निकला हीरो' फ़्लॉप हो गई थी.
बॉलीवुड के 'दबंग' और 'मास्टर ब्लास्टर' सचिन तेंदुलकर एक साथ नज़र आने वाले हैं. किसी फ़िल्म में नहीं बल्कि सेलेब्रिटी क्रिकेट लीग
यानी सीसीएल के चौथे संस्करण के लॉन्च के मौक़े पर.इसमें सलमान के छोटे भाई सोहैल की टीम भी शामिल है. ये कार्यक्रम 20 दिसंबर को मुंबई
के एक फ़ाइव स्टार होटल में आयोजित होगा. सीसीएल-4 की शुरुआत 25 जनवरी से होगी. इस टूर्नामेंट में विभिन्न फ़िल्मी कलाकारों की टीमें
शामिल होंगी.हाल ही में सुप्रीम कोर्ट ने भारत में समलैंगिकता को अपराध घोषित किया है. इस सिलसिले में बॉलीवुड ने समलैंगिकों के पक्ष
में आवाज़ उठाई है. सुपरस्टार आमिर ख़ान ने कहा, "मैं बहुत ही निराश हूं. ये फ़ैसला मानवाधिकारों का उल्लंघन है. ये बेहद शर्मनाक बात है.
"अभिनेता-निर्देशक फ़रहान अख़्तर ने कहा कि सुप्रीम कोर्ट का फ़ैसला ग़लत है. वहीं अभिनेत्री श्रुति हासन ने ट्वीट किया, "ये बात सोच के
ही कितनी डरावनी लगती है कि कोई और ये फ़ैसला करे कि हमें किससे प्यार करना चाहिए. यानी अपना साथी चुनने की आज़ादी ही ग़ैरकानूनी घोषित
कर दी गई है."करण जोहर और ओनीर ने भी सुप्रीम कोर्ट के फ़ैसले पर निराशा जताई. अभिनेत्री अनुष्का शर्मा ने भी इस फ़ैसले को आज़ादी पर
हमला बताया."
international   "सत्रह साल के मोहम्मद समीउल्लाह दक्षिण के शहर कराची में क़ैद हैं.उन पर आरोप है कि उसने एक इम्तिहान के दौरान
पैग़म्बर मोहम्मद पर अभद्र टिप्पणी की.ह्यूमन राइट्स वॉच ने इस पूरे मामले को 'स्तब्ध करनेवाला' बताया है.पिछले साल नवंबर में एक ईसाई
महिला आसिया बीबी को हुई सज़ा के बाद ईश निंदा क़ानून चर्चा में रहा है.हालांकि आसिया बीबी पैग़म्बर मोहम्मद के शान में किसी गुस्ताख़ी
की बात से इनकार करती हैं.इस साल जनवरी में ही पंजाब के गवर्नर सलमान तासीर की हत्या कर दी गई थी.पुलिस के अनुसार हत्या करनेवाले सलमान
तासीर के अंगरक्षक ने कहा कि उसने ऐसा इसीलिए किया क्योंकि तासीर ईश निंदा क़ानून का विरोध कर रहे थे.भय का माहोलसंवाददाताओं का कहना है
कि इस घटना के बाद से पाकिस्तान में भय का ऐसा माहोल पैदा हुआ है कि लोग इस क़ानून का ज़िक्र तक करने से कतराते हैं.इस क़ानून के
आलोचकों का कहना है कि इसका इस्तेमाल देश के अल्पसंख्यकों के ख़िलाफ़ किया गया है और कई बार तो व्यक्तिगत दुश्मनी के मामलों में भी इसका
दुरुपयोग होता है.ह्यमन राइट्स वॉच की वरिष्ठ अधिकारी बेडी शेपर्ड का कहना है, "समीउल्लाह के ख़िलाफ एक स्कूल अधिकारी के मामले की
शुरुआत किया जाना ही चिंता का विषय है, लेकिन फिर पुलिस और न्यायालय के एक किशोर को जेल भेज देने की घटना आश्चर्यचकित करती है."पुलिस
कहना है कि मोहम्मद समीउल्लाह के ख़िलाफ़ स्कूल बोर्ड के अधिकारी की शिकायत के बाद केस दर्ज किया गया था."
international   "अमरीकी सीनेट कमेटी ने बैंक के बहुत से अधिकारियों से 11 घंटों से भी अधिक समय तक पूछताछ की है.लॉयड ब्लैंकफ़ेन और
```

Figure 3: Dataset Prepared for Binary Classification

ii) Dataset Classification: Table 1 shows the categories and subcategories created in the dataset for the classification process. Figure 8 shows the categories and subcategories converted to Hindi Language. Then, the articles are manually tagged in each data row according to the type of categories and sub-categories. Each article unit is categorized into three columns, as shown below.



Main Category	Sub-Category	Location
India	Accident-Disaster	State or City
International	Business	Country
	Crime	
	Entertainment	
	General	
	Healthcare	
	Political	
	Science and Technology	
	Sports	
	War/Protest	

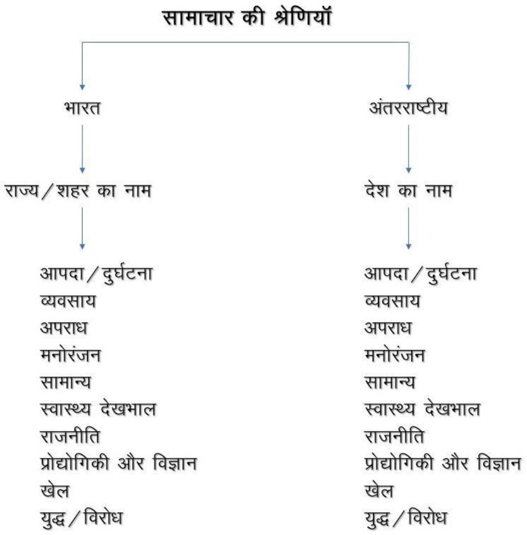


Figure 4: Classification to tag news articles in Hindi

iii) Preprocessing: Machine Learning and NLP techniques can efficiently organize Hindi Newswire articles into various sections. Tokenization, Embedding, stopword removable, Text filtering & cleaning, and vectorization are techniques used in binary classification.

3.2. Binary Classification of News: The system shown in Figure 4 used the binary classification strategy to divide the Hindi Newswire articles into India and International classes. Figure 5 shows the block diagram of the overall categorization system. Two modules comprise the entire system. Machine and deep learning algorithms will be combined with the preprocessing methods mentioned in this work to create a text classification model for news articles. The first module covers preparing and preprocessing the dataset, and the second covers creating a comprehensive machine-learning model. This work focuses on classifying Hindi texts because little research has been done.

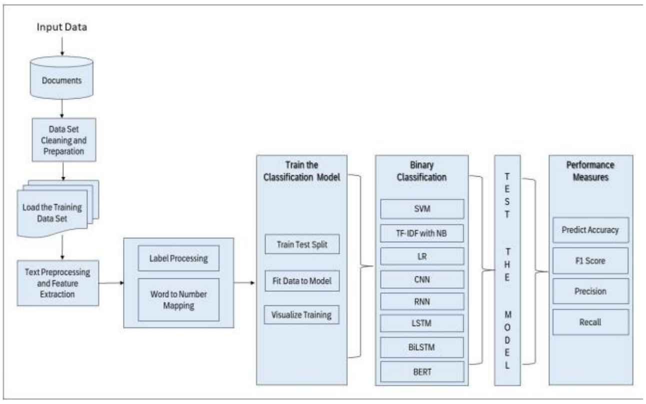


Figure 5: Block Diagram for Newswire Article Binary Classification

The proposed binary classification model was developed using several algorithms, including the RNN, LSTM, BiLSTM, BeRT technique, CNN, NB, LR, and SVM classifier. The SVM model is prepared with the Radial Basis Function kernel and hyperparameters to capture the linear relationships between the words and the news classes. It also helps make the classification very fast and practical. Stopwords and tokenizer methods are used in data preprocessing, and for feature extraction, TF-IDF vectorizer is implemented with specified parameters. The SVM model performs about 89% and above testing accuracy on the sample data predictions. The LR is used on vectorized training data. Two functions can be utilized in this statistical technique, which will work well for binary classification. The logistic function, the sigmoid function, and binary values. The LR performs 99.96% on the training dataset and 88.20% on testing the dataset. The multinomial NB classifier with a TF-IDF vectorizer with a specified alpha value of 0.7 is used. At first, Naive Bayes relies on the size of the dataset. Still, as we add more data, its performance eventually reaches a plateau, and more training data is needed to improve the Naive Bayes classifier's performance. However, we are getting improved accuracy if we increase the dataset. Our dataset's performance is recorded as 75.73% of testing accuracy.

Text classification issues involving binary and multi-class data are handled by CNNs[12]. Pre-trained Word embeddings can be used to prepare an embedding matrix. Model Accuracy of 62.18% is calculated. Despite its remarkable effectiveness in gathering features for relation extraction, CNN only gathers local features and ignores the long-distance dependency between the nominal pairings. RNN can be the solution to this problem. The model is prepared using binary cross-entropy loss, RMSprop optimizer, and accuracy and precision metrics. Tokenizer is also used in the initial preprocessing of the dataset. The testing results are 88.94%. Interpreting an RNN's output can be challenging, particularly with complex inputs like natural language. Because of this, it could be challenging to comprehend how the network generates its predictions. Ordinary RNN models perform poorly and cannot support text with complicated sequence memory. Among the several RNN variations, LSTM and BiLSTM are the most often utilized models. A sequential model is created, and an embedding layer for word embeddings is added. Dropout for overfitting and Dense layers for classification were used. model is used with binary cross entropy loss and Adam optimizer and gets 62% testing accuracy results. BERT uses transfer learning to understand the context of textual data fully. Attention masks are created, and inputs, masks, and labels are given to PyTorch tensors. For training and validation sets, batch size and data loaders are created, and then a pre-trained BERT model is used for sequence classification, and good results in testing accuracy of 73% are obtained.

The following sections describe the processes required to construct and analyze the output of these already existing models. Only Bert and CNN's structured procedures are shown here. The remaining classification model algorithms are also prepared and analyzed using the same hindi news dataset.

3.3 Steps involved in the process: Bert for Binary Classification

1. Import necessary libraries:
   - **pandas** for data manipulation.
   - **tensor flow** for GPU device availability check.
   - **torch** for PyTorch framework.
   - **train_test_split** from **sklearn.model_selection** for splitting data.
   - **Electra Tokenizer** from **transformers** for tokenization.
   - **pad_sequences** from **tensorflow. keras. preprocessing.sequence** for padding sequences.
   - **TensorDataset**, **DataLoader**, **RandomSampler**, **SequentialSampler** from **torch.utils.data** for creating data loaders.
   - **Bert For Sequence Classification**, **AdamW**, **BertConfig** from **transformers** for BERT model and optimizer.
   - **get_linear_schedule_with_warmup** from **transformers** for creating a scheduler.
   - **numpy** for numerical operations.
   - **time** and **datetime** for timing operations.
   - **plotly.express** for visualization.
   - Various metrics from **sklearn.metrics** for evaluation.
2. Read data from CSV files (**india_news.csv** and **international_news.csv**) using **pd.read_csv**. Make the length of both data frames equal by trimming the longer one.
3. Add a label column to each data frame (**india_news** and **international_news**) with values 1 and 0, respectively. Concatenate the two data frames (**india_news** and **international_news**) into a single data frame **news**.
4. Check for GPU availability. Split the data into training and testing sets using **train_test_split**.
5. Tokenize the sentences using ElectraTokenizer and encode them.
6. Pad the sequences to a maximum length.
7. Create attention masks.
8. Split the data into training and validation sets. Convert inputs, masks, and labels to PyTorch tensors.
9. Define batch size and create data loaders for training and validation sets. Load the pre-trained BERT model for sequence classification.
10. Set up the optimizer and scheduler. Define functions for calculating accuracy, formatting time, and clipping gradients.
11. Train the model for the specified number of epochs. Evaluate the model on the validation set after each epoch. Store the training loss values for plotting. After training, plot the training loss. Tokenize the test sentences and pad sequences, create attention masks, and convert them to tensors.
12. Create a data loader for the test set. Put the model in evaluation mode and make predictions on the test set. Calculate accuracy and various evaluation metrics (F1 score, precision, recall).

3.4 Steps involved in the process: CNN for Binary Classification

1. Import necessary libraries:
   - **punctuation** from **string** for string operations.
   - **listdir** from **os** for listing directory contents.
   - Various modules from **numpy** for numerical operations.
   - **pd** for data manipulation using Pandas.
   - **keras** for building neural network models.
   - **Tokenizer** from **keras.preprocessing.text** for tokenization.
   - **pad_sequences** from **keras.preprocessing.sequence** for sequence padding.
   - **Sequential**, **Dense**, **Dropout**, **Flatten**, **Embedding**, **Conv1D**, and **MaxPooling1D** from **keras.layers** for building layers of the neural network.
   - **confusion_matrix**, **accuracy_score**, **precision_score**, **recall_score**, and **f1_score** from **sklearn.metrics** for evaluation metrics.
   - **matplotlib.pyplot** for plotting.
2. Read data from CSV files (**india_news.csv** and **international_news.csv**) using **pd.read_csv**.
3. Preprocess the data:
   - Combine the two data frames (**india_news** and **international_news**) into a single data frame **news**.
   - Encode labels using one-hot encoding.
   - Create train and test data sets using a given split ratio.
   - Tokenize documents and clean them by removing punctuation and filtering based on word frequency.
4. Load pre-trained word embeddings and create an embedding matrix.
5. Define and compile the Convolutional Neural Network (CNN) model using Keras.
6. Train the CNN model on the training data.
7. Predict the labels for test data and evaluate the model:
   - Make predictions on the test data.
   - Convert predicted probabilities to class labels.
   - Calculate precision, recall, and F1-score using sklearn metrics.

IV. THE EXPERIMENTAL RESULTS

This subsection contains the proper outcomes after training and testing the classifier models for binary classification of news articles.

4.1 Performance evaluation and analysis

Binary Classification system using Hindi News Text is profoundly complex, making its evaluation and analysis challenging. The content is intricate, requiring a deep understanding to capture the meaning in a sentence. Linking an incident with its associated argument set becomes even more challenging with increased distance. Various methods can be employed to evaluate and analyze the system's performance, each demanding a thorough understanding of the system's intricacies.

The research focused on classification using current techniques, such as RNN, LSTM, BiLSTM, BeRT, CNN, NB, LR, and SVM classifiers. This research needs a large Hindi dataset. With adequate training data, text classification is more straightforward and requires sufficient hand-labeled data to use supervised techniques[15]. Consequently, the best evaluation metric for classifying or defining the fundamental data categories would be an assessment made by a human assessor.

4.2 Experimental Setup and Results Obtained

Using the hand-labeled approach, labeled data preparation was done for different classification models, and it was discovered that the accuracy of the classifier is better if labeled data is increased to train the model. Instead of classifying the raw data into multiclass news categories, this research defined two primary categories देश, विदेश in the first phase and ten different news classes in secondary categorization. A total of 8 different classification models are used for binary classification, and Performance indicators are essential when assessing the effectiveness of machine learning algorithms and making judgments. After the model was successfully prepared and tested on the news dataset, the outputs were obtained and compared regarding F1 score, precision, and recall. These indicators are crucial for assessing algorithm performance, directing the optimization process, reporting outcomes, and discovering limitations. SVM, LR, and RNN models classify and give better results than the other models for our highly challenging news data. The study assessed the algorithms' effectiveness by evaluating their accuracy, F1 score, recall, and precision as performance metrics, and the comparison results are represented in Table 2.

Table 2: Comparison Results of different machine and deep learning models

Model	Binary Classification					
	Epochs used	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Support Vector Machine	-	88.1	89	89	89	89
Logistic Regression	-	99.96	88.20	88	86	88.5
TF-IDF with Naïve Bayes	-	90.38	75.73	82	74	73.5
CNN	10	85.95	62.34	61.00	62.02	61.91
	20	93.52	61.06	62.03	61.14	64.63
	50	94.54	63.02	63.21	60.11	64.22
	100	95.28	58.12	63.14	62.44	59.64
RNN	10	94.94	76.81	97.21	75.5	78.46
	20	94.75	77.84	97.11	75.18	78.87
	50	95.42	79.03	97.44	77.65	80.55
	100	95.38	78.58	98.64	76.81	80.00
LSTM	10	98.9	62.6	67.00	61.5	62.6
	20	98.71	59.72	57.20	60.18	59.74
	50	98.66	59.61	59.13	59.65	59.61
	100	98.57	60.68	55.69	61.81	60.78
BiLSTM	10	98.81	59.84	65.63	65.16	62.41
	20	98.77	61.31	69.45	62.93	61.62
	50	99.02	61.41	67.09	63.62	61.54
	100	99.21	61.71	65.00	63.90	61.15
Bert	10	95.29	73.59	73.93	73.59	73.63
	20	98.12	78.74	78.76	78.74	78.74
	50	98.34	80.03	83.51	80,03	80,38
	100	98.74	88.74	88.31	88.74	88.35

The observations show the training and testing accuracy of various algorithms for binary classification. It illustrates the variation between training and validation data across different epochs, specifically 10, 20, 50, and 100. The experimental analysis includes evaluating the following performance parameters: training and testing accuracies, training and validation precision, training and validation recall, training and validation F1 score, training loss, and validation loss. The scatter plot in Figure 6 illustrates how the documents are distributed in the lower-dimensional space, with PCA capturing the most variance in the data using as few dimensions as possible. It highlights how the SVM model separates binary classes with a clear decision boundary, making the results more interpretable and impactful.

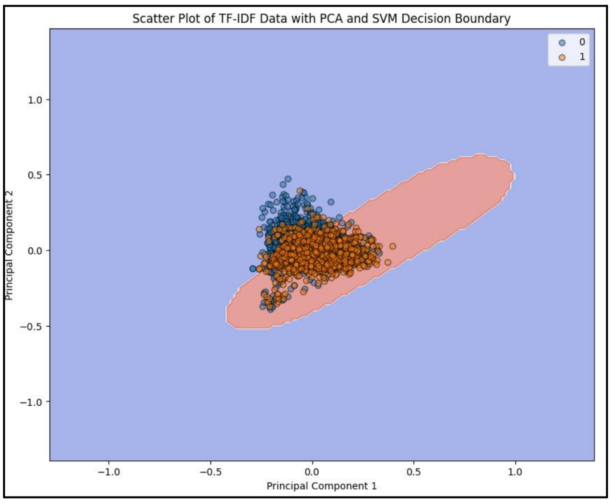


Figure 6: Scatter plot of TF-IDF with PCA, showing the SVM decision boundary

The confusion matrix for the LR model is represented using an 80-20 train-test split shown in Figure 7. The model excels in predicting the two classes. It performs moderately well with the India and International classes but struggles with some articles being incorrectly predicted as other classes, leading to confusion in predictions. Performance Assessment:

Class 0: The model performs reasonably well for Class 0, with 144 correct predictions and only 23 incorrect. This indicates good precision and recall for this class.

Class 1: However, the performance for Class 1 is concerning, as it has 71 false negatives, which is quite high. This suggests poor recall for Class 1, indicating that the model is not capturing the positive class effectively. It might need to consider improving the model by:

Addressing potential class imbalance through techniques like oversampling/undersampling.Tuning hyperparameters to improve recall for Class 1. Exploring alternative models or feature engineering to better capture the characteristics of both classes.

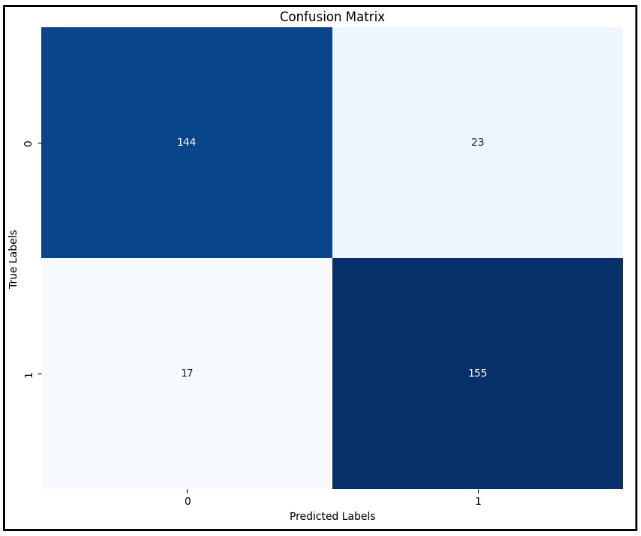


Figure 7: Confusion matrix for LR

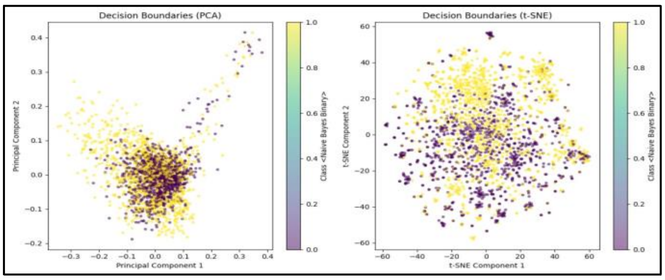


Figure 8: Decision Boundaries(PCA) and (t-SNE) for NB

Based on the analysis done, it can be observed that the BiLSTM model performs well only at 50 epochs. Among the machine learning models, Logistic Regression (LR) and Support Vector Machine (SVM) perform better on our dataset than other algorithms and some of the evaluation perimeters are already shown. The training accuracy of the BiLSTM and BERT models is higher than the other models. Additionally, the Recurrent Neural Network (RNN) and Bert models achieve higher testing accuracy than the other models.

V. Conclusion

In conclusion, the SVM, LR, and RNN outperformed all other algorithms regarding model performance measures in the binary text classification tasks examined. Text classification has significantly impacted fields, including language identification, sentiment and semantic analysis, news classification, and spam content detection. This work focuses on the classification of Hindi text because few studies have been done. In this research, eight separate classification techniques were applied, and the results were carefully examined. NLP techniques were used on the news articles and preprocessed data to examine the effects of data preparation. Text Data were normalized by Stopwords, tokenizer in the preprocessing stage, and vectorizer for feature extraction. The data were trained using the SVM, NB, LR, CNN, RNN, LSTM, BiLSTM, and Bert techniques. The performance of the various methods was analyzed after the classification. The SVM, LR, and RNN methods were the most effective in terms of accuracy, F1 score, precision, and recall. The results indicate that preprocessed datasets and coherence between the number of training sets and categories are important factors affecting how well the classifier works. Binary classification is applied in this work, having two categories only and, as of now, getting good results. The TF-IDF approach utilized in the feature extraction stage stopword and tokenizer used in data preprocessing significantly impact the results. Overall, 89% performance accuracy for SVM was attained in this study, significantly greater than the results of earlier text classification solutions suggested for the Hindi language. Currently, no available system classifies the Hindi text into predefined categories by considering newswire articles, processing them, and organizing random news articles. We plan to build a classification model using our own Hindi dataset, which uses NLP, Machine, and Deep Learning. This work can also be expanded further into the areas with the expansion of Hindi datasets and the development of a classification system for multiple classes with an innovative algorithm, providing better performance with outstanding results.

References

[1] Pallathadka, H., Pallathadka, L. K., & Devi, T. K. 2022. "Importance of Hindi Language and Its Significance in Nation-Building. Integrated Journal for Research in Arts and Humanities", 2(6), 92–98. https://doi.org/10.55544/ijrah.2.6.12

[2] Sahoo Kumar Sovan, Saha Saumajit, Ekbal Asif, Bhattacharyya Pushpak and Mathew Jimson 2019 "Event-Argument Linking in Hindi for Information Extraction in Disaster Domain" *CICLing 2019*.

[3] Ahmad Zishan, Sahoo Kumar Sovan, Ekbal Asif, Bhattacharyya Pushpak 2018 "A Deep Learning Model for Event Extraction and Classification in Hindi for Disaster Domain" *Proc. of ICON*-2018, Patiala, India. December 2018 c2018 NLPAI, pages 127–136.

[4] Shah, K., Patel, H., Sanghvi, D. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augment Hum Res 5, 12 (2020). https://doi.org/10.1007/s41133-020-00032-0.

[5] Alzoubi, Yehia Ibrahim, Ahmet E. Topcu, and Ahmed Enis Erkaya. 2023. "Machine Learning-Based Text Classification Comparison: Turkish Language Context" *Applied Sciences* 13, no. 16: 9428. https://doi.org/10.3390/app13169428.

[6] Mccallum A, Nigam K 2001 A comparison of event models for naive bayes text classification. Work LearnText Categ752:05

[7] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing. *ArXiv*. /abs/1702.01923.

[8] Umer Muhammad, Imtiaz Zainab, Ahmad Muhammad, Nappi Michele, Medaglia Carlo, Choi Gyu Sang Mehmood Arif 2022 " Impact of convolutional neural network and FastText embedding on text classification" Multimedia Tools and Applications" August 2022, volume 82, 10.1007/s11042-022-13459-x.

[9] Arora, M., Dhingra, B., Gupta, D., Singh, D. (2022). Performance Comparison of Different Machine

Learning Algorithms on Hindi News Classification. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1388. Springer, Singapore. https://doi.org/10.1007/978-981-16-2597-8_27.

[10] Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., Jagdale, J., Joshi, R. (2022). Experimental Evaluation of Deep Learning Models for Marathi Text Classification. In: Gunjan, V.K., Zurada, J.M. (eds) Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Lecture Notes in Networks and Systems, vol 237. Springer, Singapore. https://doi.org/10.1007/978-981-16-6407-6_53

[11] Khuntia, M., & Gupta, D. (2022). Indian News Headlines Classification using Word Embedding Techniques and LSTM Model. *Procedia Computer Science*, *218*, 899-907. https://doi.org/10.1016/j.procs.2023.01.070

[12] Soni, S., Chouhan, S.S. & Rathore, S.S. TextConvoNet: a convolutional neural network based architecture for text classification. Appl Intell 53, 14249–14268 (2023). https://doi.org/10.1007/s10489-022-04221-9

[13] Al-Qerem, A., Raja, M., Taqatqa, S., Sara, M.R.A. (2024). Utilizing Deep Learning Models (RNN, LSTM, CNN-LSTM, and Bi-LSTM) for Arabic Text Classification. In: Musleh Al-Sartawi, A.M.A., Al-Qudah, A.A., Shihadeh, F. (eds) Artificial Intelligence-Augmented Digital Twins. Studies in Systems, Decision and Control, vol 503. Springer, Cham. https://doi.org/10.1007/978-3-031-43490-7_22

[14] Wahdan, A., Hantoobi, S., Salloum, S. A., & Shaalan, K. (2020). A systematic review of text classification research based on deep learning models in Arabic language. *Int. J. Electr. Comput. Eng*, *10*(6), 6629-6643.

[15] S. Usmani and J. A. Shamsi, (2020) "News Headlines Categorization Scheme for Unlabelled Data," International Conference on Emerging Trends in Smart Technologies (ICETST), pp. 1-6.

[16] Mahato S., Thomas A. 2017 “Lexico-semantic analysis of essays in Hindi language” CEUR Workshop Proceedings 2017, 1819

[17] Thomas A., Kowar M. K., Sharma S. and Sharma H. R., "Exploring Text Semantics to Extract Key-Fragments for Model Answers," 2010 *International Conference on Advances in Recent Technologies in Communication and Computing*, Kottayam, India, 2010, pp. 255-257, doi: 10.1109/ARTCom.2010.110.