IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Telugu News Classification

S Rohit Jaiswal[1], Geeda Dilip Reddy[2], Kodhati Sidharth Rao[3], Durbhakula Haranath[4], Dr. Subhranginee Das[5]

Department of Computer Science and Engineering, Koneru Lakshmaiah University,Hyderabad, India

Abstract: The rapid progress in the field of IT has resulted in the accumulation of a vast volume of textual files. While researchers have predominantly focused on English text files in the past, there is now a substantial amount of content available in various Indian regional languages. Presently, there is a continuous generation of digitized Telugu text documents from government websites, news portals, and both public and private sectors on a daily basis. It is crucial to efficiently categorize these documents into predefined classes for easy retrieval. Text classification, a branch of Text Mining, offers a viable solution to this challenge. Text classification involves the automated categorization of these documents, which is otherwise a labor-intensive and time-consuming task when done manually. Automating this classification process can greatly enhance document management and retrieval. In this research paper, we present the results obtained through the application of methodologies such as Heap's law and n-gram models.

Keywords: Text classification, Telugu text classification, Heap’s law, tokenization, stop words, n-grammodel, punctuation.

1. Introduction

The surge in the availability of digital text material in recent years has given rise to new perspectives and research opportunities. The analysis and categorization of this data have become crucial due to its exponential growth, highlighting the significance of text classification. The World Wide Web's evolution has led to a substantial increase in web-based data. While much of this data is in textual form, it can be overwhelming for humans to sift through and identify what truly matters. Initially, research efforts primarily centered on English text documents. However, the landscape has evolved, with lakhs of files now available in various languages that are spoken in the interior regions of India. Consequently, manually classifying these files has become both expensive and time-consuming [1].

Given the abundance of electronic files from diverse origins, which consists of all forms of data, text mining has gained immense importance. Text mining aims to enable users to extract valuable information from textual resources and address tasks like classification, retrieval, and summarization.

By combining data mining, NLP, and machine learning techniques, text mining automates the categorization and pattern identification in various document types. Text classification, a fundamental aspect of text mining, involves categorizing a set of files into predefined and labelled classes. This is typically achieved through automatic text classification, with machine learning methods playing a central role. These techniques fall into two categories: supervised learning, where classification algorithms assign specific class labels to testing documents, and unsupervised learning, which groups documents based on strategies like clustering. Text classification addresses the challenge of categorizing input documents into one or more classes. It is a valuable tool for organizing knowledge, structuring large text collections into topical categories, combating spam emails, and enhancing internet search engines. The text categorization process is a pivotal component of this approach [2].

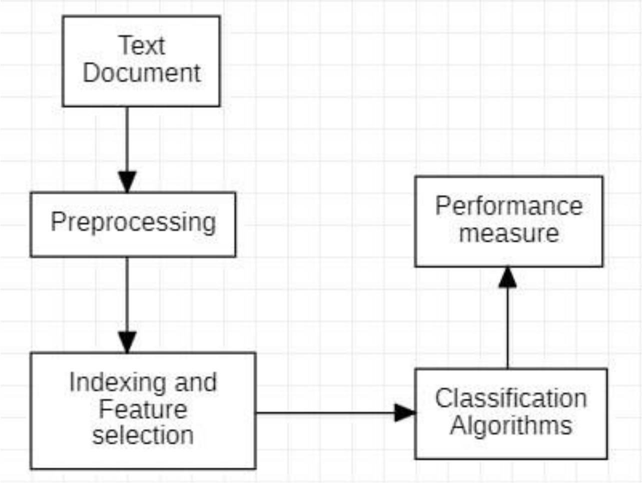


Fig. 1. Text categorization work flow

Machine learning models have garnered notable interest in recent times. Traditional models typically adhere to a conventional two-step methodology. In this approach, the initial phase involves the extraction of particular features from textual documents, which are subsequently input into a classifier during the second phase for prediction purposes. In the realm of NLP, the advent of deep learning models has revolutionized language modeling, offering a broader context. These models, designed to learn feature representations and carry out the processes like classification in a complete fashion, not only uncover underlying data patterns but also offer greater versatility when transitioning between different projects. This shift has made deep learning models the new standard paradigm for various text classification tasks. Several natural language challenges have been successfully addressed through text classification, underscoring the significance of these models, which have recently become the norm for such tasks.

Telugu, a Dravidian language, is the native and official language of the Indian states of Andhra Pradesh and Telangana. It is the most widely spoken Dravidian language and one of India's twenty-two scheduled languages, spoken by approximately 96 million people as of 2022. Due to its morphological richness, classifying Telugu text documents presents a challenging task. Telugu often referred to as the "Italian of the East" because of its vowel-ending words, has a rich tradition in lyrical writing. The Telugu language boasts 16 vowels and 40 consonants and is the second most spoken language in India after Hindi. In this research paper, our focus is on electronic documents written in the Telugu language.

The structure of this paper comprises five components. Component 1 provides a brief understanding of the entire concept, while Component 2 presents a brief survey conducted to understand the various methods that can be used for classification of text. Section 3 outlines the proposed work for classification of text in Telugu files. Finally, the last two components talk about the results obtained, and the conclusion of the paper.

2. Literature Survey

This literature survey explores various techniques for text classification and offers valuable insights into the field:

The increased accessibility of digital documents has led to the growing need for automated text categorization. This survey talks about the major methods to carry out the process of text categorization, covering three key aspects: document representation, classifier construction, and classifier evaluation [3].

Text classification analysis in Indian languages is addressed in [4]. The study highlights the challenges posed by Indian-written text in the field of NLP. It suggests that supervised learning algorithms have proven effective for text classification in this context.

The authors in [5] delve into advanced learning models for text analysis and classification, with a specific focus on addressing intricate challenges in natural language processing (NLP). Their paper offers a concise overview of both traditional machine learning and cutting-edge deep learning techniques applied in text categorization. It underscores the significance of comprehending diverse subtasks within the field and the pertinent scholarly works.

D Naga Sudha [6] explores semi-supervised learning methods in the context of Telugu text classification. The study evaluates the influence of n-gram feature selection on categorization of articles obtained from newspapers. The proposed model uses TF-IDF N-gram features and count vectorizers, demonstrating the effectiveness of Support Vector Machine with a classification rate of 93.64%.

A novel ontology-based text classification method for Telugu documents is introduced by D. Naga Sudha [7]. This approach effectively captures the concepts in text, distinguishing between relevant and irrelevant terms based on sentence semantics. The results indicate improved efficiency compared to conventional methods.

Kavi Narayana Murthy's research [8] focuses on automating Telugu news article classification into politics, sports, business, and cinema categories. The Naïve Bayes (NB) classifier is employed, and its performance is assessed using precision, recall, and F-measure metrics. This approach aims to effectively categorize Telugu news content for information retrieval and organization.

Ashish Kumar, et al. [9], explored various supervised machine learning methods for classifying web files that are in Bangla language. Through his study, he has observed that SVM outperforms the others, particularly in high-dimensional data scenarios.

In [10], the authors investigated the influence of n-gram feature selection on the classification of text obtained from the news articles, emphasizing the use of semi-supervised learning methods due to a largely unlabeled dataset. The study validates the results with respect to classification rate, precision, recall, and F-score.

In [11], the author compares the performance of various classification approaches and term weighting techniques for Telugu text classification. Support Vector Machine (SVM) is found to be more effective than Naive Bayes (NB) and k Nearest Neighbour (kNN) classifiers in most categories.

In [12], the authors attempt to classify text from online Telugu newspapers using a generative model. They introduce language-dependent features such as contextual words, part of speech (POS), post-position features, clue word features, and gazetteer features to enhance the model's performance. The average F1-Scores for Person, Location, and Organization for the entire model were 88.87%, 87.32%, and 72.69%, respectively.

The literature review concludes by noting the limited work of research done on text classification on various languages spoken in the interior regions of India such as Bangla, Telugu, Punjabi, and Urdu. Supervised learning methods have been applied to these languages. This comprehensive survey provides a valuable overview of text classification methods and their application in various linguistic contexts.

3. Dataset

The dataset used in this project has been taken from Kaggle. This is a labelled dataset which comprises of various news articles obtained from different newspapers. There are 5 news categories in this dataset, which are,
Entertainment news, Sports news, Business news, National news and editorial news.

4. Methodology

The proposed approach consists of following phases:

1. Pre-processing
2. Implementation of N-gram model on the entire corpus.
3. Classification of Telugu text based on topic

The initial source of our text data comprises old Telugu newspapers. Our process begins with the crucial step of text preprocessing, also known as text cleaning. This step is fundamental in news classification as it serves to reduce data storage requirements and enhance classification accuracy. Typically, datasets from such sources are unstructured and contain a mixture of valuable and irrelevant information. Elements such as stop words, special characters, and incomplete sentences do not contribute to the classifier's predictive capabilities. Therefore, it is imperative to perform a cleaning process on the raw dataset before considering the removal of any specific elements. This cleaning step aims to mitigate distortions that may adversely affect the model's performance.

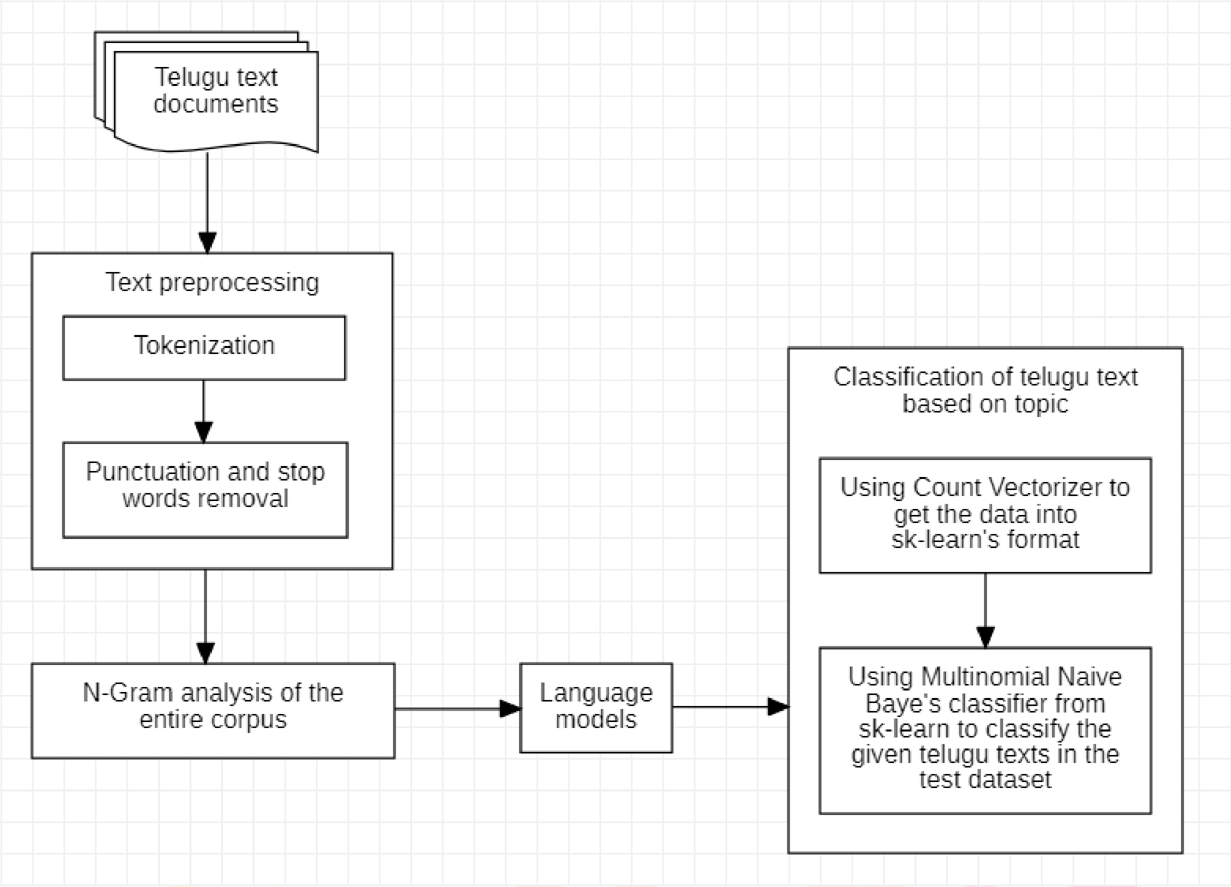

Telugu text documents
Text preprocessing
Tokenization
Punctuation and stop words removal
N-Gram analysis of the entire corpus
Language models
Classification of telugu text based on topic
Using Count Vectorizer to get the data into sk-learn's format
Using Multinomial Naive Baye's classifier from sk-learn to classify the given telugu texts in the test dataset

Fig. 2. Proposed work

4.1 Pre-processing

4.1.1. Tokenization:

Tokenization is a crucial process in text analysis where a large text is divided into smaller units called tokens. This step is vital for building a vocabulary, which consists of a unique set of tokens within a given corpus. There are multiple methods available to carry out this process

In our project, we have employed word tokenization, which is the most commonly used tokenization algorithm. Word tokenization breaks down the text into tokens based on certain delimiters, typically spaces or punctuation marks. In the context of Indian languages, we have utilized the "indicnlp" library, specifically the "sentence_tokenize" module, to facilitate this processing. This library allows us to work effectively with Indian languages and perform word tokenization, enabling us to create a vocabulary from the text corpus.

4.1.2. Punctuation and Stop word removal:

The process of removing punctuation is essential to insure that each textbook is treated constantly. In English, when removing punctuation, we must exercise caution as some words may lose their meaning. For illustration, the word" don't" can come" do nt," which has no meaningful interpretation. Still, in the Telugu language, similar issues don't arise, making it doable to directly remove punctuation from the textbook. To further reuse the textbook, CountVectorizers are used to prize features from words in the dataset, followed by the junking of stop words. Stop words are common words in a language that are constantly used but convey limited meaningful information. exemplifications in English include" a,"" the,"" is,"" are," etc. Removing stop words is a common practice in textbook mining and natural language processing (NLP) to exclude overused words that don't contribute significantly to the analysis. After textbook preprocessing, an n- gram model is employed for farther analysis.
This algorithm takes a textbook input and a parameter' n,' which specifies the number of commemoratives to be considered as an n-gram. It utilizes the n- gram language model constructed from the corpus to determine the liability that the input textbook belongs to each predefined class. The program also assigns the prognosticated class for the input textbook grounded on the loftiest probability.

4.2 Implementation of N-gram Model on the entire corpus

An N- gram model is a probabilistic model that helps in predicting the word that would most likely follow a sequence of N- 1 words. It's trained on a textbook corpus and finds operations in colorful NLP tasks similar as speech recognition, machine restatement, and prophetic textbook input. N-gram models are created by calculating chances and counting the frequency of word sequences in a corpus. To address limitations of simple N- gram models, ways like smoothing, interpolation, and back- out are frequently employed to enhance their performance. Specifically, in this environment, we explore bi-gram, tri-gram, 4- gram, and 5-gram models. The top 30 most constantly being n- grams are uprooted for each n- gram model. The process involves forming a vocabulary, tokenizing words, and generating a list (tuple) of the asked n-grams.

4.3 Classification of Telugu text based on topic

The textbook categorization algorithm mentioned utilizes n- gram language modeling and comprises functions for creating an n-gram vocabulary from a textbook corpus, calculating the liability of a given textbook belonging to a class, and grading input textbook into one of the classes grounded on liability. Laplace smoothing is applied when an n- gram isn't set up in the vocabulary. Tokenization and judgment segmentation are performed using the IndicNLP library. Log chances are preferred over regular probability addition to avoid dwindling vaticination chances as the length of the input judgment increases. Multinomial Naïve Bayes Classifiers are employed for this purpose.

The attained confusion matrix after enforcing these processes is presented below (Confusion Matrix Data Then)

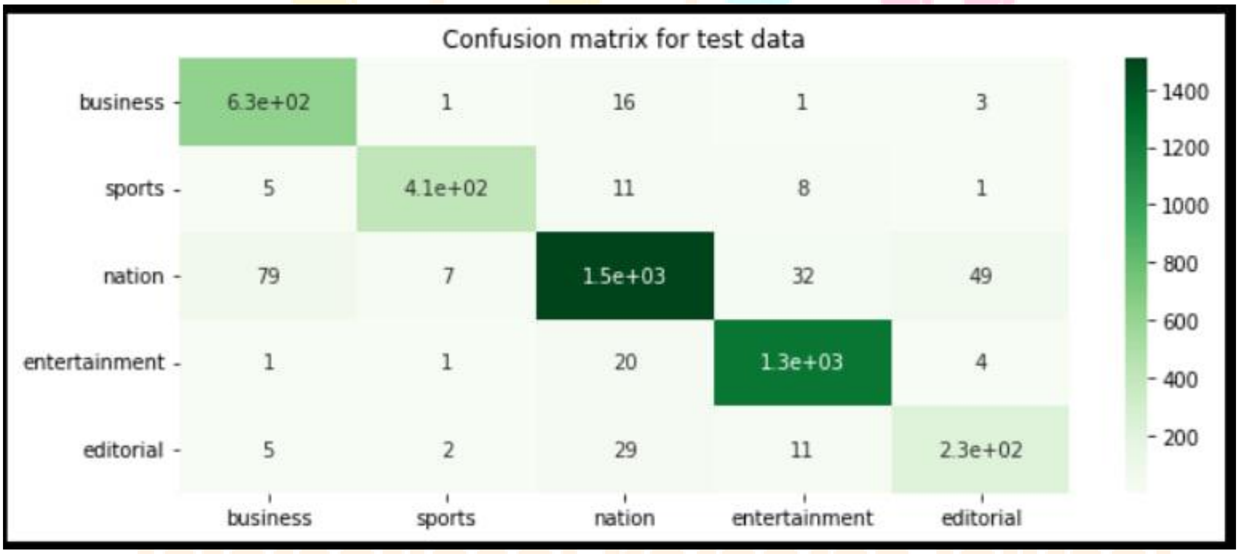
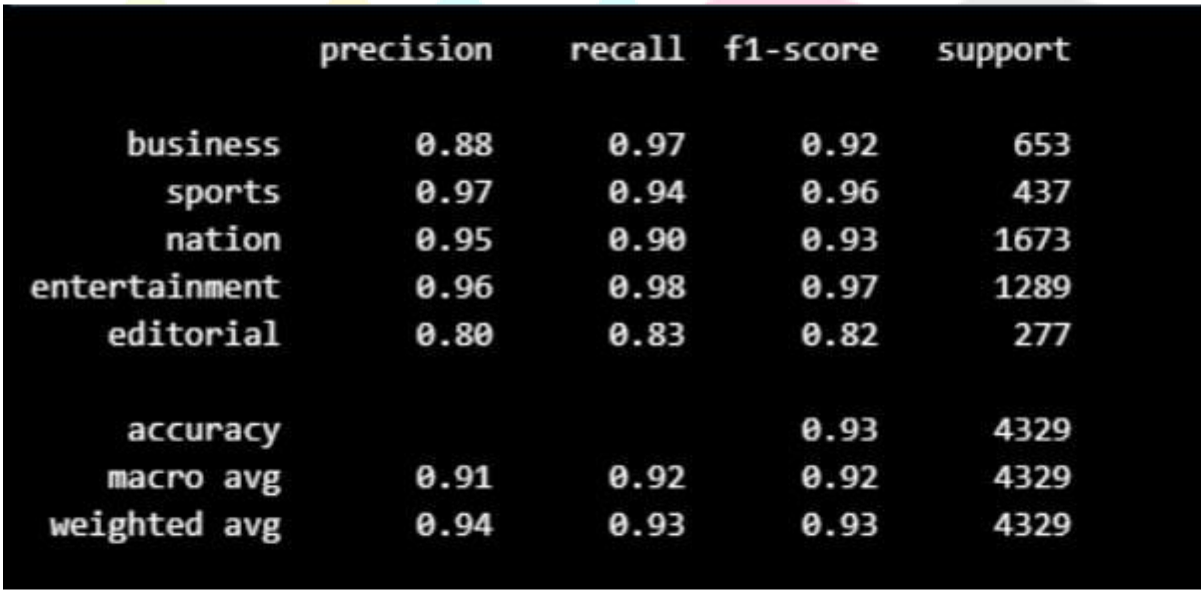


Fig. 3. Confusion matrix



```
               precision    recall  f1-score   support

     business       0.88      0.97      0.92       653
       sports       0.97      0.94      0.96       437
       nation       0.95      0.90      0.93      1673
entertainment       0.96      0.98      0.97      1289
    editorial       0.80      0.83      0.82       277

     accuracy                           0.93      4329
    macro avg       0.91      0.92      0.92      4329
 weighted avg       0.94      0.93      0.93      4329
```

Fig. 4. Precision and recall values obtained on given dataset

5. Result

While predicting the class which the text belongs to, we have observed that the probability rates are constantly coming down. This is because of multiplication of probabilities. To eliminate this, we use Mulitinomial Naïve Bayes Classifier technique. With this technique, we can predict the exact class of the text and we can see that, using bigrams with the multinomial naive bayes classifier works absoutley well, even with text that is out of the given test corpus. This can be performed even on the text which is taken from outside the corpus.

6. Conclusion

Regional languages spoken in India, such as Telugu, have not received much attention. Therefore, the proposed system uses n-gram model and heap's law to provide text classification of documents in the Telugu language. With by using n-gram models and multi-nomial naïve bayes classifiers we could carry out the process of classifying the text based on the class which it belongs to. We break the corpus into tokens, and we convert them into n-grams and the model remembers which class it belongs and thus classifies the text. Future testing of the proposed Telugu classification system with a large corpus size and the addition of more domains are both possible.

References

1. Bolaj, Pooja, and Sharvari Govilkar. "Text classification for Marathi documents using supervised learning methods." International Journal of Computer Applications 155, no. 8 (2016): 6-10.
2. Dalwadi, Bijal, Vishal Polara, and Chintan Mahant. "A Review: Text Categorization for Indian Language."
3. Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34, no. 1 (2002): 1-47.
4. Kaur, Jasleen, and Jatinderkumar R. Saini. "A study of text classification natural language processing algorithms for Indian languages." VNSGU J Sci Technol 4, no. 1 (2015): 162-167.
5. Dogra, Varun, Sahil Verma, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. "A complete process of text classification system using state-of-the- art NLP models." Computational Intelligence and Neuroscience 2022 (2022).
6. Sudha, D. Naga. "Semi Supervised Multi Text Classifications for Telugu Documents." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 12 (2021): 644-648.
7. Durga, A. Kanaka, and A. Govardhan. "Ontology based text categorization-telugu document." International Journal of Scientific and Engineering Research 2, no. 9 (2011):1-4.
8. Murthy, Kavi Narayana. "Automatic categorization of Telugu news articles." Department of Computer and Information Sciences (2003).
9. Mandal, Ashis Kumar, and Rikta Sen. "Supervised learning methods for bangla web document categorization." arXiv preprint arXiv:1410.2045 (2014).
10. Sudha, D. Naga. "Semi Supervised Multi Text Classifications for Telugu Documents." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 12 (2021): 644-648.
11. Murthy, Vishnu G., B. Vishnu Vardhan, K. Sarangam, and P. Vijay Pal Reddy. "A comparative study on term weighting methods for automated telugu text categorization with effective classifiers." International Journal of Data Mining & Knowledge Management Process 3, no. 6 (2013): 95.
12. Gorla, SaiKiranmai, Sriharshitha Velivelli, NL Bhanu Murthy, and Aruna Malapati. "Named Entity Recognition for Telugu News Articles using Naïve Bayes Classifier." In NewsIR@ ECIR, pp. 33-38. 2018.