HINDI DOCUMENT CLASSIFICATION USING HYBRID SYSTEM

Pratyaksha Pokhriyal[1], Piyush Pratap Singh[2]

M.Tech Student in Computer Science and Engineering[1]

Assosiate Professor, SCSS, JNU[2]

School of Computer and System Sciences, Jawaharlal Nehru University, Delhi, India

pratyakshapokhriyal@gmail.com[1], piyushpsingh@mail.jnu.ac.in[2]

***Abstract*:** Due to the advancement in the field of information and communication technologies, there is a rapid growth in text documents generated online. There is a need for an automatic text classification system to maintain this text data for easy retrieval of information as it is available in unstructured form. Today documents are available not only in English but in any regional language. There is a lot of content generated in digital form in many Indian languages. The purpose of this work is to build a hybrid classification system that can classify documents in one of the many Indian languages, the Hindi language. The system that has been created is not domain-specific. Some Linguistic rules have been used to extract the important words from the text documents and the classification of documents is done by using a machine learning model.

Keywords: Natural language processing, Hindi language, Karaka Theory, Linguistic rules, Document classification

1. INTRODUCTION

Today enormous numbers of documents and textual information are being introduced every day on the Internet, from various sources. With time, the textual data available online is increasing continuously. A considerable amount of this data is in unstructured form. Therefore, it is very important to organize it into a structured form for easy retrieval of information.

Text classification[1] is the process in which the documents are separated into predefined classes or categories. In classification, the main topics that the document covers are identified, and relationships are identified by looking for major-minor terms, synonyms, and related terms based on which classes are assigned to documents. Consider a set of labeled documents from a source D = [d1,d2,….dn] belonging to a set of classes C = [c1,c2,…,cp]. The text classification task is to train the classifier using these documents and assign classes to new documents.

India is home to many languages, due to its cultural and geographical diversity. Through the introduction of Unicode standards, there is a growth in Web content of Indian languages. The Hindi language is one of the official languages of India. It is the most widely spoken Indian language with over 600 million speakers around the world[2]. In the last few years, there has been extensive growth in Hindi content online. Several documents like blogs, news articles, etc are produced online in Hindi. Therefore, in this paper, we propose a hybrid system that can classify documents in the Hindi language based on their content for easy retrieval of data. The model analyses the data and extracts the relevant information required using Linguistic rules and classifies documents into different categories using the machine learning model.

2. LITERATURE SURVEY

Many kinds of research have been done on the classification of Hindi text documents and other related systems.

In the paper[3], Navneet Garg et al presented a Part of Speech tagger which uses a rule-based technique to tag the words in sentences. The system mainly works in two steps-firstly the input words are searched in the database; if present then they are tagged and if not present then appropriate rules are applied to tag the words. In the paper[4], Deepa Modi and Neeta Nain proposed a designed POS tagging system for the Hindi language. This system is developed in which corpus matching is applied while tagging known words and for unknown words, grammatical rules (based on prefixes and suffixes) and regular expression-based rules are used. In this paper[5], Shalini Puri and Satya Prakash Singh proposed a Hindi Text Classification model, which accepts Hindi documents, preprocesses them at the document, sentence and word levels, extracts features, and performs classification using an SVM classifier. In the paper[6], V. B. Parthiv Dupakuntla et al used the Naïve Bayes classifier for the classification of documents in the Hindi language. They also used Laplacian smoothing for each word in classes. Laplacian Smoothing is used to avoid the problem of zero probability. In paper [7], the authors have used the Naïve Bayes classifier for the classification of documents in the Hindi language and also suggest that the usage of ensemble methods like a combination of Naïve Bayes with neural networks or Support Vector Machines(SVM) can improve the accuracy of the classifier further. In this paper[8], Prafulla B. Bafna and Jatinderkumar R. Saini performed the classification of Hindi poems with different machine-learning algorithms(SVM, Naive Bayes, Decision Tree) and neural networks, and evaluated the performance of the algorithms used. In the absence of any other technique, which achieves prediction on Hindi corpus, misclassification error is used and compared to prove the betterment of the technique. Support vector machine performs best amongst all. Aspect-Based Sentiment Analysis (ABSA) identifies the aspects within the given sentence and the sentiment that was expressed for each aspect. In paper[9]. Abhilash Pathak et al performed classification on datasets belonging to different domains in the Hindi language. Using different methods, they construct an auxiliary sentence from this aspect and convert the Aspect-Based Sentiment Analysis (ABSA) problem into a sentence-pair classification task. They used two ensemble models based on Multilingual BERT, namely mBERT-E-MV and mBERT-E-AS, and then fine-tuned different pre-trained BERT models and ensemble them for a final prediction based on the proposed model.

3. ISSUES WITH HINDI LANGUAGE

Processing text in the Hindi language has several issues related to its linguistic features.

- No capitalization: Unlike English and most Western languages, Hindi doesn't have scripts with graphical cues like capitalization, which plays a vital role in identifying and could act as indicators for Name entity recognition.
- The word order in Hindi is somewhat flexible. Most of the sentences have the word order as <subject> <object> <verb>, but it changes depending on the other information and context.
- There is a lack of standardization in Hindi. The variations in the spellings are one major issue. People from different regions have different ways of pronouncing and writing words. This increases the number of tokens to train the machine with.
- The main issue with Hindi, like other Indian languages, is that it lacks resources and tools like annotated corpora, name dictionaries, suitable morphological analyzers, POS taggers, etc. required for the development of quality systems.

- Ambiguity in words: There is a high overlap between words used. E.g. A lot of Indian people's names can be used as common nouns. Also, the Hindi language contains masculine and feminine words, and the structure of sentences changes according to gender. All this makes recognition of important words a very difficult task.

4. *Karaka* THEORY

The concept of *karaka* relations is central to Paninian Grammar. Panini introduced six basic semantic notions that capture several aspects of action through its participants[10]. It is a semantic relation between the verb and a noun, and other related constituents in a sentence. Each participant in an activity denoted by a verbal root is assigned a distinct karaka.
The six different types of karaka relations in the Paninian grammar are listed below:

1. *Karta*: person or thing that does the action.
2. *Karma*: the thing/person that the action is done to.
3. *Karana*: an instrument that indicates association or mutual dealing.
4. S*ampradana*: beneficiary/recipient of the action.
5. *Apadana*: a participant which remains stationary in an act of separation or keeping away from something.
6. *Adhikarana*: real or conceptual space or time.

Vikhakti is that suffix or signs in front of the word which shows what is the relation of that word to the verb.

E,g, मीरा ने प्रीती को पत्र लिखा।
In the above sentence, 'ने' and 'को' are the vibhaktis (cases). 'ने' tells that Meera is doing the action of writing a letter, therefore is the *Karta karaka* (nominative noun). 'को' in the sentence tells that the effect is on Preeti who is the *Karma karaka*(accusative case).

5. DATASET USED

For this project work, the dataset used is the BBC news articles classification dataset taken from the Github repository of Nirant K [12]. The dataset contains 14 unique categories and has a total of 4335 documents written in the Hindi language with tags. Each document has exactly one tag associated with it. Since the dataset is too large, only 4 classes out of 14 classes have been used in this project. The categories (or classes) used are news, entertainment, sports, and science. Each category contains around 200 text documents (for training). In total, 1219 documents are used (including test documents).

6. METHODOLOGY

This section briefly describes the steps of the proposed methodology of the hybrid classification system. The first step is pre-processing, the next is feature extraction where only relevant features(word) are extracted from the text using Linguistic rules, and then the classification step where a Multi-nomial Naive Bayes classifier is used to classify textual documents into predefined categories based on extracted features.

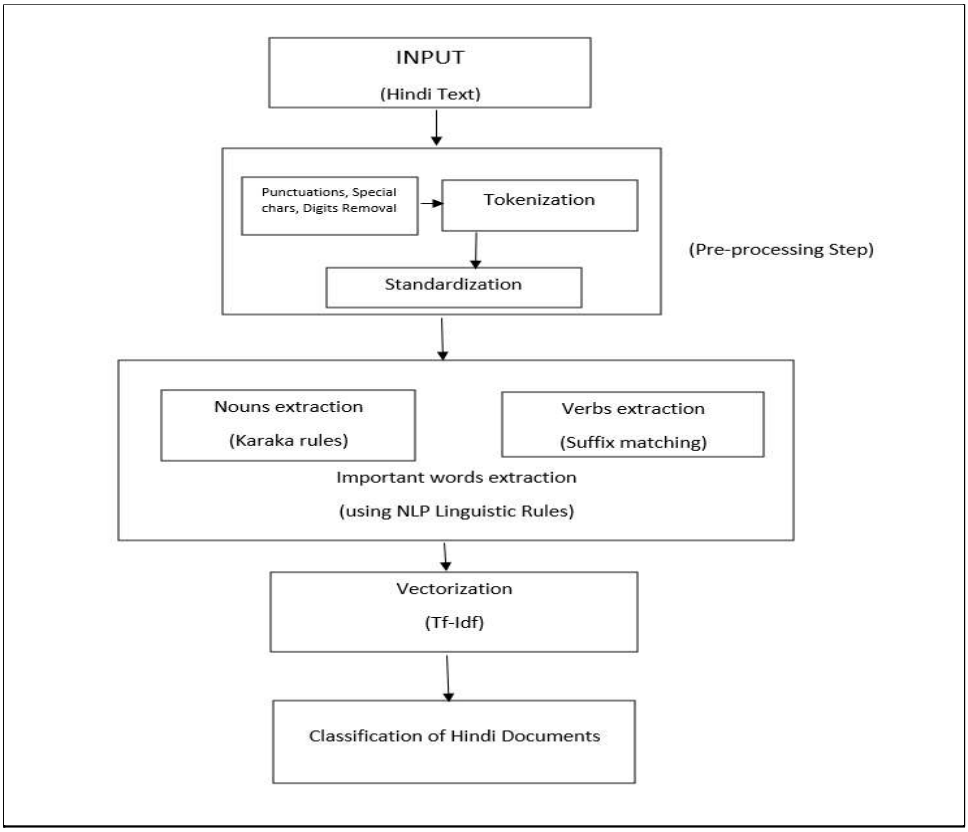

Figure 1. System Architecture

6.1. Pre-processing

Pre-processing is the primary step in classification. Pre-processing is carried out to clean, and remove unnecessary data and grammatical errors from documents.

- All the punctuations, Special characters, and digits have been removed.
- Tokenization: Paragraphs were broken down into individual words or tokens.
- Standardization: There are a lot of variations in spellings. Hence, the text has been standardized to avoid any error using the rules mentioned in the paper[11].

6.2. Important words Extraction

The different parts of a sentence impart a different amount of information on the actual topic of an article. For example, proper nouns or Named entities(name of the person or location) are likely to be most significant, verbs contain information about an action (taken) or state of being, and adjectives typically very little. Thus extracting these informative terms or parts of sentences helps recognize what the document is about.

To develop the vocabulary of the most discriminative features (words), the main goal is to identify all the terms in the given document or paragraph as much as possible that convey information pertinent to the document, those terms being the ones that are most likely to be shared with other documents belonging to the same category, and less likely to appear in documents of other categories. For this, we used some linguistic rules based on which the system

extracted the important words from the text document. Linguistic rules have been developed using karaka theory and Hindi suffixes.

6.2.1. Nouns Extraction:

For extraction of nouns or Named Entities, NLP rules are formed using Karaka Theory Principles where the Vibhaktis (Case), in Hindi, are taken as semantic markers for preceding and succeeding words. Vibhaktis do not have any meaning of their own but they give clues about whether a particular entity is a person or location, etc. ने/में/की/के/को/का, etc are considered as Vibhaktis in Hindi.

For example, if the current token or word is 'ने', then the previous token will be a proper noun.

If the above rule is applied to the following document


ब्रितानी सरकार ने यह घोषणा लंदन में परिवार नियोजन पर आयोजित एक सम्मलेन में की. इस सम्मलेन में ब्रितानी सरकार के साथ बिल और मेलिंडा गेट्स फाउंडेशन सह आयोजक थे. ऐसी अपेक्षा है कि बिल और मेलिंडा गेट्स फाउंडेशन भी गर्भ निरोधकों के प्रसार के लिए बड़ी रकम की घोषणा करेगा.मेलिंडा गेट्स ने इस मौके पर बोलते हुए कहा कि उनका संस्थान जो भी रकम घोषित करेगा वह कम से कम उतनी तो होगी ही जितनी वह मलेरिया, एड्स और टीबी जैसी बीमारियों से लड़ने के लिए खर्चा करता है. मेलिंडा गेट्स ने कहा कि उनका फाउंडेशन यह चाहते है कि विकाशील देशों में महिलाएं गर्भनिरोधकों तक पहुँच चाहती हैं, साथ ही वो इस तरह के गर्भनिरोधक चाहती हैं जो इंजेक्शन के ज़रिये लिए जा सकें. मेलिंडा गेट्स ने कहा " हम यह चाहते हैं कि साल 2020 तक 12 करोड़ महिलाओं को हम गर्भनिरोधक उपलब्ध करा पाएं. हमारे साथ हमारे सभी सहियोगी मौजूद हैं. हम वितरण की दिक्कतों से तो निपटना चाहते ही हैं लेकिन साथ ही हम इस दिशा में नए शोध भी करना चाहते हैं जो महिलाओं को नए तरह के ज़्यादा चलने वाले गर्भनिरोधक उपलब्ध करा पाएं." फाउंडेशन के परिवार नियोजन डिविज़न के के प्रमुख गैरी डार्मस्टेड ने विकाशील देशों की दिक्कतों के बारे बात करते हुए कहा " सबसे बड़ी समस्या है फंडिंग की, दूसरी समस्या है राजनीतिक इच्छाशक्ति की

Figure 2. Hindi Text Document

Then the extracted words will be: ब्रिटेनी, सरकार, मेलिंडा, गेट्स, मेलिंडा, गेट्स, गैरी, डार्मस्टेड .

Here, the system extracts two previous words whenever the word 'ने' occurs as the length of the proper nouns in articles is generally 2.

Every time the vibhakti occurs, the previous word will be extracted and if the word is a Hindi stopword like 'इसमें', 'है', ' को' etc., then it is dropped. Similarly, rules are constructed to extract other important words from the text.

6.2.2. Verb Extraction:

To identify whether the given word is a verb or not, suffix matching has been used. First, the list of all suffixes, that describes the verbs, has been prepared. If a word ends with a postfix that has a match in the suffix list and is not a stopword then it is considered to be a verb and the system will extract the word from the document text.

A few suffixes used are given in the following table:

Table 1: Few Examples of Suffixes used to extract Verbs

S.No.	Verb Suffix	Words
1.	-ता	बढता, जाता, देता
2.	-एंगी	आजमाएंगी
3.	िए	कहिए, जानिए
4.	-या	बताया, टकराया

6.3. Classification

Multi-nomial classifier is used to classify data that cannot be represented numerically. The multinomial naïve Bayes is widely used for assigning documents to classes based on the statistical analysis of their contents. It is a probabilistic learning method based on the Bayes theorem that predicts the tag of a text. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. It considers each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature[13]. Hence, it is used for the classification of the extracted text.

7. RESULT AND DISCUSSION

Testing was performed on 244 documents of 4 different categories entertainment, news, science, and sport. The overall accuracy of the system is 79.9% over the test set. The classification accuracy of the system was also measured in terms of precision, recall, and F-measure. The precision and recall values of the classification are 82.9% and 77.5%. The F-Score is thus 78.8%.

The confusion matrix and classification report have been shown below that tell about the performance of the system class-wise.

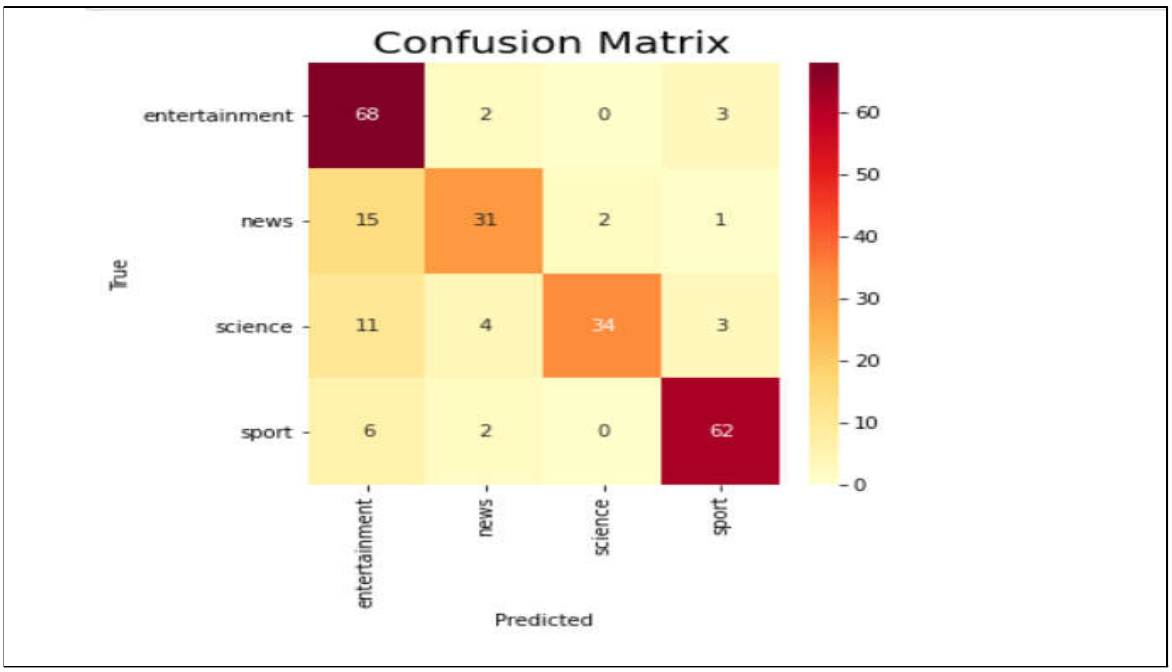

Figure 3. Confusion Matrix For Classification System

	precision	recall	f1-score	support
entertainment	0.68	0.93	0.79	73
news	0.79	0.63	0.70	49
science	0.94	0.65	0.77	52
sport	0.90	0.89	0.89	70
accuracy			0.80	244
macro avg	0.83	0.78	0.79	244
weighted avg	0.82	0.80	0.80	244

Figure 4. Classification Report

8. CONCLUSION

With the advent of technology and the growth of documents on digital platforms, we need classification systems to organize text documents for easy retrieval of information.

Very less work has been done on the classification of text in Hindi and other Indian languages. In this paper, we have presented a hybrid system for classifying Hindi text documents. The system uses Hindi Linguistic rules (Karaka rules and suffix matching) to extract important words from the text and with the help of the Multi-nomial Naive Bayes classifier classifies the text documents into different categories. The proposed system has an average accuracy of 79.9%.

9. FUTURE SCOPE

In the future, more linguistic rules can be incorporated to extract relevant features and rules to remove non-relevant features from the text. The proposed model accuracy can be further enhanced by performing the stemming of the verbs reduced to their root words.

10. REFERENCES

[1] Korde, Vandana. (2012). Text Classification and Classifiers: A Survey. International Journal of Artificial Intelligence & Applications. 3. 85-99. 10.5121/ijaia.2012.3208.

[2] https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2021.

[3] Garg, Navneet & Goyal, Vishal & Preet, Suman. (2012). Rule Based Hindi Part of Speech Tagger. 163-174.

[4] Modi, Deepa & Nain, Neeta. (2016). Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method. 10.1007/978-81-322-2638-3_28.

[5] Puri, Shalini & Singh, Satya. (2019). An Efficient Hindi Text Classification Model Using SVM. 10.1007/978-981-13-7150-9_24.

[6] V. B. PARTHIV DUPAKUNTLA, HEMISH VEERABOINA, & M. VAMSI KRISHNA REDDY. (2021). LEARNING BASED APPROACH FOR HINDI TEXT SENTIMENT ANALYSIS USING NAIVE BAYES CLASSIFIER. International Journal of Innovations in Engineering Research and Technology, 7(08), 40–47.

[7] Tamhankar, Ishaan & Chaturvedi, Ashysh. (2019). Classification of Spam Categorization on Hindi Documents using Bayesian Classifier. International Journal of Computer Trends and Technology. 66. 8-13. 10.14445/22312803/IJCTT-V66P102.

[8] Bafna, Prafulla & Saini, Jatinderkumar. (2020). On Exhaustive Evaluation of Eager Machine Learning Algorithms for Classification of Hindi Verses. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110224.

[9] Pathak, Abhilash & Kumar, Sudhanshu & Roy, Partha & Kim, Byung-Gyu. (2021). Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models. Electronics. 10. 2641. 10.3390/electronics10212641.

[10] Subhash C. Kak, The Paninian approach to natural language processing(1987), 10.1016/0888-613x(87)90007-7 http://www.sciencedirect.com/science/article/pii/0888613X87900077.

[11] V. Goyal and G. S. Lehal, "Automatic standardization of spelling variations of Hindi text," 2010 International Conference on Computer and Communication Technology (ICCCT), 2010, pp. 764-767, doi: 10.1109/ICCCT.2010.5640441.

[12] https://github.com/NirantK/hindi2vec/releases/tag/bbc-hindi-v0.1

[13]https://www.upgrad.com/blog/multinomial-naive-bayes-explained/#:~:text=The%20Multinomial%20Naive%20Bayes%20algorithm%20is%20a%20Bayesian%20learning%20approach,tag%20with%20the%20greatest%20chance.