Journal of Philanthropy and Marketing
Vol 3 No 1 (2023)
E-ISSN: 2691-1361

A MULTILEVEL APPROACH FOR KANNADA STOPWORDS GENERATION USING SENTENCE FEATURES

Sowmya M S
Research Scholar, Dept of CSE, Jain Deemed to be University,Bengaluru Rural

Dr. Panduranga Rao M V
Professor, Dept of CSE-IOT, Jain Deemed to be University, Bengaluru Rural

Dr. Ashok Kumar P S
Professor, Dept. of CSE, ACSCE, Mysore Road, Bengaluru

Abstract—
The generation of kannada stopwords with multiple iterations from the e-newspaper is a logical task. It proposes work that makes use of POS tagging and stemming approach for Kannada text preprocessing with NLP algorithms. Text summary is the process of condensing the original text's content into a shorter version that nevertheless gives the consumer essential information. The extractive summaries of Kannada text documents are generated by the summarizer that is provided in this research. The significant sentences in the paper are determined by the proposed summarizer system based on five features.Sentence length, sentence location, keywords feature, term frequency, and term frequency-inverse sentence frequency are the features that are used. Each feature's value is calculated, and the average of all the feature score values is used to get the score for each sentence in the document. The extracted summary contains the sentences that have received the highest marks.After applying TF and IDF, ML algorithms normalize the e-newspaper to choose the appropriate stopwords. Experiments conducted on a specially constructed dataset including fifty Kannada text documents demonstrate notably superior performance in extractive summarization as compared to human summaries. After evaluation, the improved stopword set for additional Kannada NLP study was discovered.
Keywords — Kannada, Stopwords, NLP, TF, TFIDF, Summarization, Sentence

I. INTRODUCTION

Kannada is the oldest languages spoken today, which accounts for the many fascinating stories that are connected to it. The term Kanarese or Canarese is well-known. It is a Dravidian language that is primarily used in Karnataka among Kannada speakers. The language, which is listed in Schedule Eight of the Indian Constitution, has an impact on many other Indian languages.

It is one of India's ancient languages. According to the earliest literary works in the Epigraphs, the old Kannada language flourished in the 6th century AD during the Ganga dynasty and the 9th century AD during the Rashtrakuta dynasty. The Mudbidri palm leaf manuscript of Dhavala, known as Jain Bhandar, is the earliest Kannada manuscript still in existence. It has 1478 antique Kannada leaves from the ninth century AD.

Techniques for summarizing texts can be broadly categorized into two categories: extractive and abstractive. The document's summary is produced via extractive techniques, which use statistical and linguistic aspects based on word and sentence content to find key sentences. Using linguistic analysis of the text, a summary is produced using abstractive approaches by either choosing or rearranging the words in the original document or by adding new terms that were absent before. Depending on the number of documents supplied as input, text summarization systems can also be divided into single-document and multi-document categories.
Summarization systems are categorized as generic, domain-specific, or query-based based on their intended use.Based on the number of documents sent in as input, text summarization systems can also be divided into single-document and multi-document categories via classification. Summarization systems fall into three categories: query-based, domain-specific, and generic, depending on their intended use. The Indian state of Karnataka is where most Kannada speakers reside. Using Sentence characteristics, the approach presented in this research creates an extracted summary of a Kannada text content.

It can be surprising that every word in Kannada ends with one among the 10 vowels that make up the language. The most distinctive feature of the language is this in and of itself. Conferring to the Literary Giants, Sanskrit had a significant influence on this language. Prakrit and Pali are two more languages that have had an impact on this Dravidian language.

In Kannada, there isn't silent letter. The Kannada language lacks the silent letters that are so well-known in the English language and the term structure that is dependent on them. It is mentioned to as the most scientific language on Earth for this reason as well.

The only Indian language for which a dictionary was written by a foreigner is Kannada. Most people know Reverend Ferdinand Kittel for his research on the Kannada language. He is renowned for creating a Kannada-English dictionary in 1894 that contains over 70,000 terms.

The world's most rational and scientific language is kannada. Sanskrit is the foundation of and a major influence on Kannada. It is simple to understand and learn, yet it has a really unusual grammar. The Kannada script, which developed from the Kadamba script of the fifth century, is used to write the Kannada language. Eight Jnanpith prizes have been given to Kannada literature.

Hindi and English are not older than Kannada literature. The oldest language is Kannada, followed by Prakrit, Sanskrit, and Tamil. Kannada is thought to have split out from the proto-Tamil South Dravidian group even before the advent of Christianity, according to linguists. It was therefore spoken much before Hindi and English.
Numerous historical facts about the Kannada language show that even in the 4th century BCE, the great Greek dramatists were familiar with both the Kannada people and the language. The dialogue

of some of the characters in Euripides' and Aristophanes' comedies has been found to borrow Kannada terms and phrases in various works of local literature.

It is primarily spoken in the Indian state of Karnataka, even if a sizable population speaks Kannada in the further states as well. Additionally, it is articulated in the close-by states of Tamil Nadu, Maharashtra, and Andhra Pradesh. It is estimated that up to 44 million individuals worldwide express it, including those who speak it as a second language. Since the BCs, Kannada has undergone changes as a language. PurvaHalegannada, Halegannada, Nadugannada, and Hosagannada are the four categories that it can be categorized.

II. Literature review

Pre-processing of the gathered information is the most crucial activity for the recovery of data from various sources and formats. One such pre-processing step is the Stop-Word removal method. The list of stopwords, stop stems, and stop lemmas for the Indian language of Malayalam is presented in this publication for the first time. A corpus of Malayalam languages was first produced. More than 21 million words in all, of which 0.33 million were unique words, were found in the corpus. This was processed to produce 153 Stop-words in total. Only 25 words could be lemmatized, and only 20 words could be stemmed. There are 123 Stop-words in the final, polished stopword list. The majority of people in India and many other countries around the world speak Malayalam. Any NLP activity for this language will undoubtedly utilize the results offered here [1].

Stopword s arerecurrent words in a document that place unrealistic demands on the classifier in terms of complexity in both time and space. Information retrieval in English has been the subject of a great deal of research, but it is a relatively new idea in Kannada. The discovery and eradication of stopword s in the Kannada dialectal might be a significant undertaking because doing so would lower the feature space and, in turn, the complexity of time and space. It must be noted that the Kannada language does not have a standard stopword list. This justifies our undertaking the challenge of creating an algorithm to rejectstopword s that are structurally identical. Conclusions show that while the stopword removal does lower the amount of feature space, it may not actually increase the performance of the classifiers [2].

The expansion of the internet has increased the demand for better information retrieval (IR) methods that facilitate finding pertinent information more quickly. One such technique that looks for to produce a brief and simple summary of the text is text summarization. Recently, key word-based summaries have attracted the interest of several researchers in the field of natural language processing. By combining GSS (Galavotti, Sebastiani, Simi) coefficients, IDF (Inverse Document Frequency) methods, and TF (Term Frequency) methods, the algorithm we created identifies key words from Kannada text documents and uses them for summarization.

The main goal of our work is to give each word in a sentence a weight; the weight of a sentence is calculated based on the scores of the sentences, and we select the top'm' sentences. Our specially designed database is used to choose a document from a specific category. Kannada Webdunia is where the files are obtained. A Kannada portal called Kannada Webdunia provides political news, movie news, sports news, shopping news, and jokes. A summary is produced centered on the user's specified amount of sentences. Finally, we compare the machine-generated summary to the one created by a human. The exclusion of stopword s from feature extraction is yet another goal of this effort. We have introduced a novel method for deleting stopword s that looks for structurally related terms in article [3].

Here, we take into account the Kannada language of southern India and suggest a suffix stripping algorithm. This algorithm is for the online, unicode-formatted Kannada text. It is a rule-based methodology that eliminates suffixes (pratyaya in Kannada) and some subclasses from fourteen different primary classifications. It also encompasses stopword s, articles, adjectives, and suffixes connected to nouns, verbs, and articles. The ranking of Kannada documents (which are represented by a bag of words) using this method will be very helpful for web searches. It can also be used in Kannada for text extraction, NLP tools, and speech recognition engines, among other things. We put this suffix stripping stemming technique into practice and tested it on Kannada documents from the online magazine "Kendasampige" both with and without our stemming algorithm. For the evaluation, we made use of many metrics. Conclusions show that stemming greatly improves the recall factor. This algorithm can be used for the aforementioned applications, according to these encouraging preliminary results [4].

Stopword s arerepeated words in a document that place unrealistic demands on the classifier in terms of complexity in both time and space. Information retrieval in English has been the subject of a great deal of research, but it is a relatively new idea in Kannada. The discovery and eradication of stopword s in the Kannada language might be a significant undertaking because doing so would lower the feature space and, in turn, the complexity of time and space.It should be noted that the Kannada language does not have a standard stopword list. This justifies our undertaking the challenge of creating an algorithm to eliminate stopword s that are structurally identical. Conclusions show that while the stopword removal does lower the amount of feature space, it may not actually increase the performance of the classifiers [5].

A corpus is one of the first things needed for activities involving natural language processing (NLP). Corpus, which is Latin for "body," refers to a group of texts in linguistics and NLP. The knowledge foundation of corpus linguistics is corpora. This task uses natural language processing and is primarily focused on building a corpus of Kannada words. Texts gathered from the Internet make up the Kannada Corpus, a linguistic corpus. The primary goal of the task is to extract various Kannada terms from newspapers, pre-process them in various stages, and create a big collection

of Kannada words in a text file. This text file will then be made accessible for use in any future Kannada-related projects, such as annotation algorithms [6].

Stopwords are frequently eliminated from text data in order to lower the size of the dataset and enhance the performance of machine learning models. As a result, scientists have worked to create methods for creating potent stopword sets. Precedingapproaches have ranged from qualitative ones that rely on linguists to statistical ones that use correlations or metrics that are frequency dependent and generated on a corpus to extract word importance. Iterative and recursive feature deletion algorithms are used in study to determine which words can be eliminated from a pre-trained transformer's vocabulary with the least impact on the performance of the transformer, specifically for the task of sentiment analysis. Empirically, stopword lists created using this method significantly reduce dataset size while barely affecting model performance. For instance, in one case, the corpus was reduced by 28:4% while the trained logistic regression model's accuracy increased by 0:25%. Another time, the accuracy dropped by 2:8% while the corpus was reduced by 63:7%. These encouraging conclusions show that our method can provide very effective stopword sets for particular NLP tasks [7].

Stopwords are words that add little to the semantic information contained in article and are used in info retrieval. Identification of stopwords is regarded as a benefit for many retrieval tasks since the regularization of the dataset is increased and the volume and computational complexity are decreased when these words are uninvolved from the text collection. Both manually creating stop-word lists by examining each term individually and sorting-based procedures that use global term features as a proxy are traditional approaches for detecting stopwords. Transfer and application of these conventional approaches to new languages or topics allows for standardized and comprehensible outcomes due to the context-dependent and imprecise definition of what actually qualifies as a stopword. We provide a feature-based supervised machine learning technique for the automatic detection of stopwords in order to allay this worry. We conducted comprehensive trials to verify the efficacy of the suggested method, and we compared the outcomes with a list of common English stopwords. Additionally, we tested the suggested method using formal written language as well as text from social media. The conclusions show that the suggested method achieves positive outcomes and is able to take into account dialect variations [8].

Effective indexing in contemporary information retrieval systems can be accomplished by eliminating stopword s. For the English language, numerous stopword lists have been produced to date. For the Chinese language, no standard stopword list has yet been created. Exploring Chinese stopword lists is essential given the rapid development of info retrieval in the Chinese language. In this research, we offer anspontaneouscombined approach based on arithmetical and info models for extraction of a stopword list in Chinese language in order to save time and relieve the load of manual stopword selection. Outcomesstudy reveals that our stop list is substantially more general than existing Chinese stop lists and comparable to a general English stopword list. Our stopword

extraction algorithm is a promising method that develops a standard while saving the time required for manual production. In the future, it might be used with other languages [9].

Information retrieval calls for the removal of stopword s. To save on memory storage, it might eliminate terms that are frequently used and generic. Every word that exactly matches a term on the stopword list is disregarded by the algorithm. However, creating the list could take some time. The words in a particular language and domain need to be gathered and verified by experts. The goal of this reading is to create a brand-new K-means Clustering stopword list generation technique. The suggested method classifies words depending on how frequently they occur. A genuine stopword list created by a Javanese linguist is used by the confusion matrix to calculate the discrepancy between the results. The proposed technique has a 78.28% (K=7) accuracy rate. The outcome demonstrates the dependability of creating Javanese stopword lists using a clustering algorithm [10].

III. Methodology

The Dravidian linguistic family includes the Kannada language, similarly famous as Kanarese or Kannana, which serves as the state of Karnataka's official tongue. The states that border Karnataka also speak Kannada. Kannada was estimated to be the 38 million people's first language, according to the census statistics from the early twenty-first century; another 9 to 10 million people were estimated it as a second language to speak. The Indian government recognized Kannada as a classical language in 2008[11].

3.1 Kannada usage variety

Of the four main Dravidian languages, Kannada has the second-oldest literary history. The earliest known Kannada inscription, which dates to around 450 CE, was found in the little village of Halmidi. The Ashokan Brahmi script's southern variants gave rise to the Kannada script. The Telugu script and the Kannada script both descended from the Old Kannarese script. Old Kannada, Middle Kannada and Modern Kannada since 1700–present are the three historical phases that are recognized.

Subject, object, and verb are the order of the words, just like in other Dravidian languages. Person, number, and gender indicators are present in verbs. The case-marking pattern is nominative-accusative, with the dative inflection used by experience in subjects. The majority of inflection is produced by affixation, particularly of suffixes. The language uses a variety of voiced and voiceless aspirates that are derived from the Indo-Aryan language family, as well as characteristic Dravidian retroflex consonants like sounds uttered with the tip of the tongue curled back against the roof of the mouth.

There are three distinct regional variations of Kannada. Mysore and Bangalore are connected with the southern variety, Hubli-Dharwad with the northern, and Mangalore with the coastal. The Mysore-Bangalore diversity serves as the foundation for the prestige varieties. There are currently

atleast three separate social languages: Brahman, non-Brahman, and Dalit, which are all branded by tutoring and class or caste. There is also a dichotomy or diglossia between spoken and formal literary forms.

Beginning in the ninth century CE with Nripatunga'sKavirajamarga, Kannada literature continued with Pampa's Bharata in the ninth century CE. The grammar of Keshiraja, which is still revered, is the first existing grammar and was written by Nagavarma in the early 12th century. The Virasiva and Lingayat movements had an impact on Kannada writings. The Haridasa movement of popular devotional song reached its pinnacle in the 16th century with the influences of Purandaradasa and Kanakadasa, the former of whom is regarded as the founder of Karnatak music, the conventional style of southern India.

Beginning in the ninth century CE with Nripatunga'sKavirajamarga, Kannada literature continued with Pampa's Bharata in the ninth century CE. The syntax of Keshiraja (1260 CE), which is still revered, is the first existing grammar and was written by Nagavarma in the early 12th century. The Virasiva and Lingayat movements had an effect on Kannada literature. The Haridasa movement of popular devotional song reached its pinnacle in the 16th century with the influences of Purandaradasa and Kanakadasa, the former of whom is regarded as the founder of Karnatak music, the traditional style of southern India.

The Halegannada dialect of Kannada is spoken by the one million Komarpants in and around Goa. They have established themselves all across the state of Goa, as well as in the Karnataka districts of Belagavi and Uttara Kannada's Khanapur Taluk [12][13][14]. Halakki Kannada, also known as Achchagannada, is the dialect used by the HalakkiVokkaliga people of Karnataka's Uttara Kannada and Shimoga districts. About 75,000 people are estimated to live there in that moment [15][16][17].

The language employs 49 phonemic characters, which are broken down into three groups: yogavaahakagalu (neither vowel nor consonant - two letters: anusvara and visarga), swaragalu (vowels - thirteen letters), and vyanjanagalu (consonants - thirty-four). The character set resembles other Indian languages almost exactly. Almost all of the Kannada alphabet is phonetic, with the exemption of the sound of a "half n" (which becomes a half m). There are many more written symbols than there are in the alphabet's forty-nine letters since distinct characters can be merged to create compound characters (ottakshara). The Kannada script is syllabic, meaning that each printed symbol resembles to one syllable rather than one phoneme as it does in languages like English.

Along with early Kannada grammar works, there was a Kannada-Kannada dictionary. The poet 'Ranna' called 'Ranna Kanda' was created the earliest known Kannada dictionary in 996 AD. 'AbhidhanaVastukosha' by Nagavarma (1045 AD), 'AmarakoshadaTeeku' by Vittala (1300),

'Abhinavaabhidaana' byAbhinavaMangaraja (1398 AD), and many others are examples of other dictionaries [18]. Ferdinand Kittel created a Kannada-English dictionary with about 70,000 terms in it [19].

The Kannada SahityaParishat released an eight-volume, 9,000-page series of the first contemporary Kannada-Kannada dictionary, which was edited by G. Venkatasubbaiah. He also created a glossary of challenging terms called klitapadaka and a Kannada-English dictionary [20][21].

In order to extract the most significant information from a text document, our work proposed the Kannada document summarizer, an application of natural language processing (NLP). Text extraction and text abstraction are the two primary methods used in automatic summarization. The extraction method takes the key words, phrases, or sentences from the input content and uses them to create a summary. By inserting a few additional terms that aren't in the input document, an abstraction approach generates the summary. Pre-processing, feature extraction, and summary synthesis are the three primary basic processes in the summary production process [9].
Fig. 1 shows the proposed model to generation of Kannada stopwords using Kannada e-newspaper.
.

A. Pre-processing

In the process of Kannada document summarization, some pre-processing operations are carried out before the sentence scoring algorithm is executed. The pre-processing function prepares the document for ranking of sentences and the generation of summary.
The following pre-processing actions were taken on the documents:

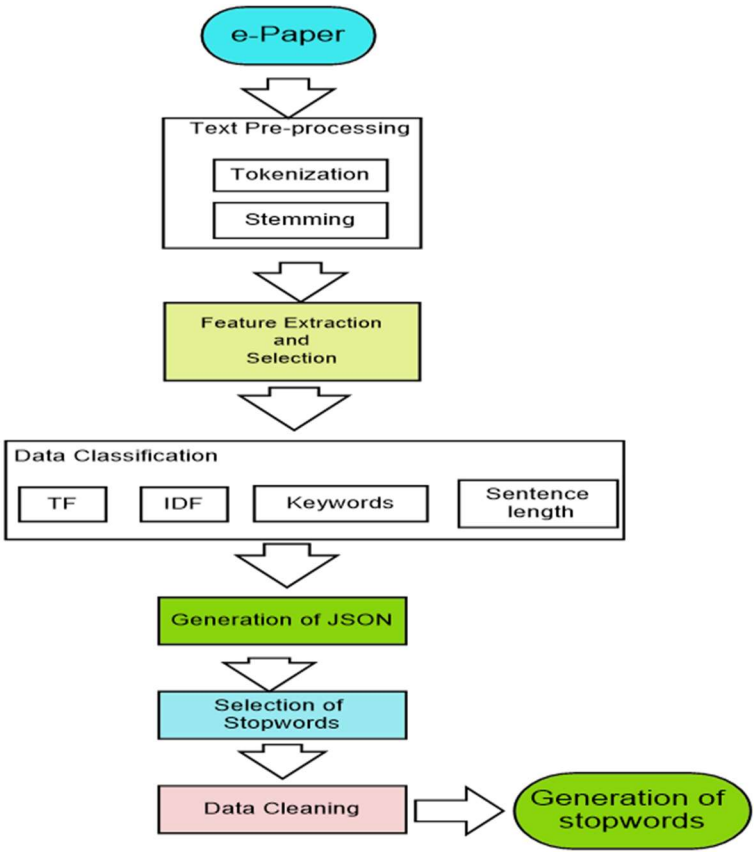

Fig 1: System Architecture

a) Tokenization – Sentences are arranged in a document, and each sentence is made up of a collection of words. Every word is handled as a token. Tokenization divides the text into phrases, which are then broken down into individual words.

b) Stemming – Many words in the document may have the same root but various spellings. For convenience, all such nouns are reduced to their basic form. The stemming process is used to transform words into their canonical forms. For instance, the terms 'ಗಳು', 'ಗ' brought to be reverted to their initial form. In this work, all inflected words are stemmed using a specified suffix list to return them to their root form.

Feature Extraction & Sentence Scoring

Following pre-processing, an input document's phrases are scored according to five key characteristics: Sentence length, position in the document, keywords, term frequency, and term frequency-inverse sentence frequency.

a) Term Frequency (TF) – A measure of a term's frequency of occurrence in the document is called term frequency. The calculation of term frequency is as follows:

$$\mathrm{TF(t)} = \frac{Nt}{TW} \text{ ---------- (1)}$$

Where,
TW is the total number of terms included in the document,
Nt is the number of times a term appears in the document.
TF is the total term frequency(TF) values for every word that appears in a sentence is used to calculate the phrase's score.

b) Term Frequency-Inverse Sentence Frequency (TF-IDF) – The significance of a word in the document's phrase is determined by this numerical statistic. The frequency of a word in the document balances out the TF-IDF value, which rises in direct proportion to the number of times a word appears in a sentence. This measure aids in managing the phenomenon of some words being used more frequently than others. The formula for the inverse sentence frequency is (2).

$$IDF\ (t,d) = \log\ (N/Nt) \text{ ---------- (2)}$$

Where,
N represents the total number of sentences in the document d, and nt is the total number of sentences that contain the word w.As shown in (3), the term frequency and inverse sentence frequency are combined to provide the term frequency-inverse sentence frequency (TF-IDF) feature score.

$$TF - IDF(t) = TF(t) * IDF\ (t,d) \text{ ---------- (3)}$$

Each sentence's value is determined by adding the tf-IDF scores of all the words that make up the sentence.

c) Keywords – Ten percent of all words that are frequently occurring in the document are classified as keywords. The ratio of the total number of times a keyword appeared in a sentence to the total number of times a keyword appeared in the document is how this feature is scored (as shown in (4)). Keywords are words that appear frequently in a document and are likely related to the subject of the document.

$$\textit{Keywords score} = Ks / Kd \text{ ----------(4)}$$

The total number of times a keyword appeared in the sentence is denoted by Ks, while the total number of times a keyword appeared in the document is represented by Kd. Each sentence's score is determined by adding the scores of all the keywords that are included in the sentence.

d) Sentence length –The document's brief sentences are filtered out using the sentence length tool. The ratio of a sentence's length to the longest sentence in the document is used to calculate the sentence length feature score,

Sentence length = Ls / Lls ----------(5)

Where, Ls denotes the sentence's length and Lls is the longest sentence in the manuscript. One can determine the length of a sentence by counting the total number of words in the sentence [10].

In order to prepare unstructured text data for analysis, text preprocessing—a crucial stage in natural language processing, it involves cleansing and modifying the data. Tokenization, stemming, lemmatization, stop-word elimination, and part-of-speech labeling are all included in this process.The technique of giving each word in a sentence its own part of speech tag is known as parts of speech (POS) tagging. The work's unique Kannada language native features are used in the POS tagger. Sandhi splitting, which divides a compound word into two or more meaningful constituent words, is one of the unique features.NLP lexical or morphological analysis involves word structure analysis and recognition. The technique of dissecting a text file into words, phrases, and paragraphs is known as lexical analysis. In this step, the source code is read as a stream of characters and transformed into comprehensible lexemes. Words, sentences, and paragraphs make up the whole book.

The proposed architecture shown in Fig 1, it considers the e-newspaper as the input for the system. It applies the NLP native language text preprocessing process by making use of POS tagging and Stemming. Feature is extracted using lexical and morphological analysis. Applying the ML language algorithms,the. json output is generated. The corpus of terms is generating by applying TF and IDF methods. POS based classification selects the required stopword terms from the list generated from TF and IDF. The set generated is normalized so that to get the stopword list that has no irrelevant or redundant terms by iterating the process, here we kept the iteration for 10 times.

Total_Score(s) = [TF(t)+ TF-IDF(t) + Keywords score+ Sentence length+ Sentence Position] / Nf ---- (7)

The final score for each sentence in the document. It is calculated by averaging all the feature score values (Term frequency, Term frequency-Inverse sentence frequency, Keywords, Sentence length, and Sentence position in the document) for that specific sentence.
Where Nf represents all the qualities that were used to score the sentence.

Summarization Algorithm

The Kannada text document is fed into the sentence characteristics based summarizing algorithm, which outputs a summary with the amount of sentences the user specifies.
Input: the number of sentences (n) required for the summary in the Kannada text document
Output: An extractive summary with the right amount of phrases
Start

Step 1: Take the input text document.
Step 2: Divide the text into phrases
Step 3: For each statement
(a) Tokenize each statement into words.
(b) Join the significant words that remain into the sentence.
Step 4: Determine the characteristics' values - Sentence length for every sentence in the doc as
Term frequency: TF(t) = Nt / TW; -
Keywords; - Term frequency-Inverse sentence frequency;
TF-IDF(t) = TF(t)*IDF(t,d)
b) Term frequency-Inverse sentence frequency TF-IDF(t) = TF(t)*IDF(t,d)
c) Keywords Score for keywords = Ks / Kd
d) Sentence Position: Sentence Position = (Ts – SP) / Ts
e) Sentence Length: Sentence length = Ls / Lls
Step 5: Determine each sentence's overall score. The average of all feature scores for every sentence is used to calculate it.
Total_Score(s) =(TF(t)+ TF-IDF(t) + Keywords score+ Sentence length+ Sentence Position) / Nf
Step 6: Choose the top n sentences that scored the highest.
Step 7: Arrange the sentences such that the selected sentences in the summary remain in the same sequence as they were in the original document.
Step 8: To create a summary, take a few sentences out of the original document.
Stop

V Experimental details

NPTEL course document that is translated into Kannada language, is proposed, where the input is a kannada stopword literature. Moreover,input is also a document that is generated by the audio input of three styles of Kannada speaking society, namely Dharwad Kannada, Malnad Kannada, and Mysore Kannada.

Fifty documents from various categories on the Kannada Webdunia website are gathered to produce the dataset. The papers are stored in text files in the Unicode standard UTF-8 format. Movies, Business, Sports, and Politics are the five categories that were selected. Table 1 displays the dataset statistics that take into account four papers in each category for evaluation.

Table 1: kannada Dataset Statistics

Category	Total Number of Sentences	Total Number of Words
Politics (A)	76	731
Business (B)	53	804
Sports (C)	82	722
Movies (D)	60	943

To evaluate the proposed system, documents are taken from four different dataset groups. There is just one human summary considered for evaluation each document. The generated system summary is evaluated by comparing it with the human summary using the NLTK toolbox.

Using the NLTK toolkit, the generated system summaries are assessed in terms of recall, precision, and F1-score. The results for the five documents in each category are displayed in Tables II, III, and IV, in that order.

The ratio of common sentences found in both the system and model summaries to the total number of sentences in the system summary is known as precision.

The ratio of common sentences found in both the system and model summaries to the total amount of sentences in the model summary is known as recall.

Recall and precision are combined into a composite statistic known as the F1-score. The harmonic average of recall and precision is used to calculate it. The document numbers are indicated in the table by the notations D1, D2,... D5.

Table 2: Categorical value of Precision, recall and F1-score

Article	Categories of Precision				Categories of Recall				Categories of F1-score			
	A	B	C	D	A	B	C	D	A	B	C	D
D1	0.8	0.6	0.9	0.56	1.1	0.6	0.53	0.46	0.87	0.6	0.7	0.5
D2	0.85	0.52	0.78	0.13	0.95	1	1	0.08	0.9	0.72	0.78	0.03
D3	0.51	0.73	0.4	0.46	0.61	0.53	0.48	0.46	0.56	0.53	0.48	0.43
D4	0.65	0.54	0.61	0.8	0.6	0.55	0.66	1	0.65	0.54	0.51	0.87
D5	0.41	0.51	0.61	0.7	1.01	0.51	0.91	0.65	0.61	0.51	0.61	0.73
Average	0.64	0.58	0.66	0.53	0.85	0.64	0.72	0.53	0.72	0.58	0.62	0.51

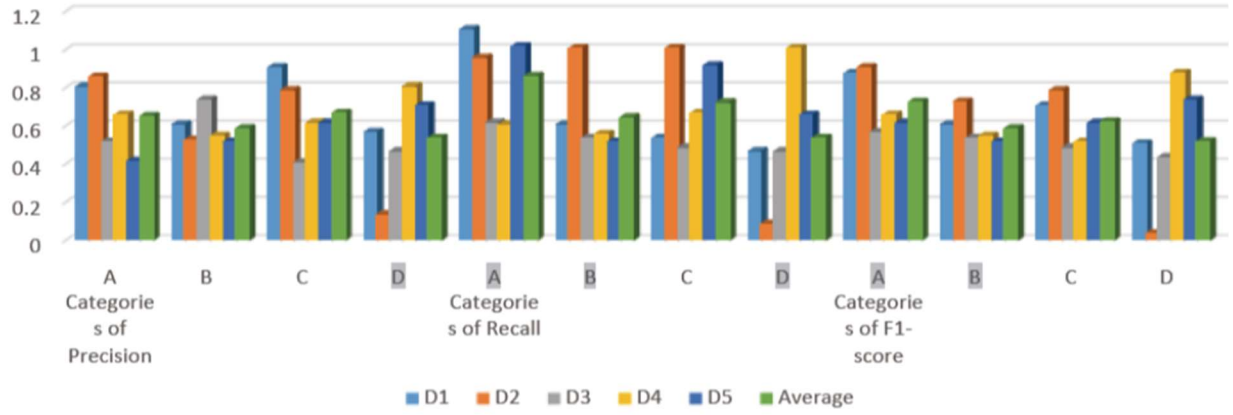

Fig2: Average value of Precision, recall and F1-score

A. Experiment1:

Articles with names, contents, and URLs more than 500 words were removed. All 155 articles were saved as supporting files; there were 155 in all.

The "prajavani.json" file was read using the textblob module, and blobs were made out of the "content" part of each JSON line. The words of each blob were read, and the regularity of every phrase was entered into a dictionary. The dictionary's list of terms ranked according to frequency printed the top 102 terms. Fig 3 shows the sample output. Fig 4 and Fig 5, represents the sample e-newspaper.

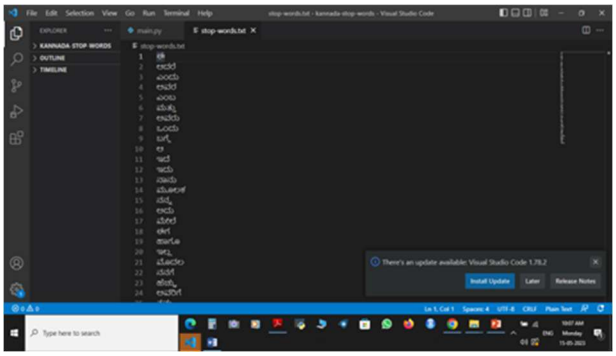

Fig 3: Output of Stopwords list generated from newspaper

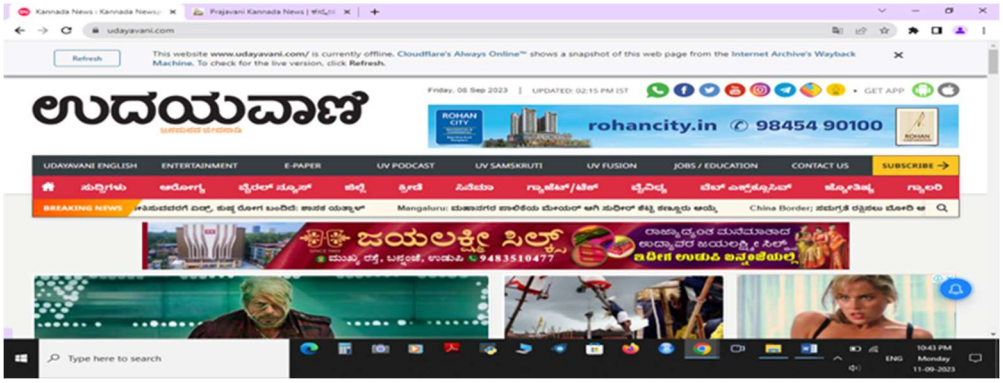

Fig 4: sample of e-newspaper Udayavani

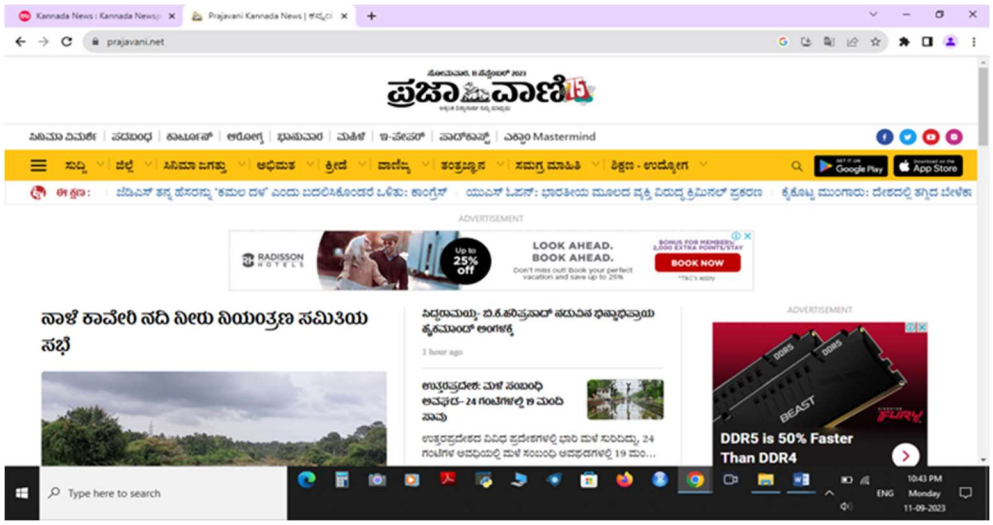

Fig 5: sample of e-newspaper Prajavani

VII. CONCLUSION

With several iterations, this inquiry has created a better number of Kannada stopwords for the e-newspaper. The suggested method uses NLP techniques for Kannada text preparation, POS tagging, and stemming. After applying TF and TFIDF, ML algorithms assist in choosing the appropriate stopwords through normalization.

This study discusses a Kannada text summarization technique for a single document that is extraction based. Sentence length in the document, keywords, term frequency, and term frequency-inverse sentence frequency are the several criteria that are used to grade sentences. The NLTK toolbox is utilized to assess the generated summaries through performance metrics, including recall, precision, and F1-score. In terms of average recall, average precision, and average F1-score values, the suggested approach performs well. Future improvements to the suggested system's effectiveness could come from giving sentence grading greater weight to linguistic and statistical characteristics.

REFERENCES

- Prasad Desai; Jatinderkumar R. Saini; Prafulla B. Bafna, POS-based Classification and Derivation of Kannada Stop-words using English Parallel Corpus, 2022 3rd International Conference for Emerging Technology (INCET), 27-29 May 2022, DOI: 10.1109/INCET54531.2022.9825429.
- Kallimani, J.S. and Srinivasa, K.G., 2010, August. Information retrieval by text summarization for an Indian regional language. In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010) (pp. 1-4). IEEE.
- Jayashree, R., Murthy, S.K. and Sunny, K., 2011. Keyword extraction based summarization of categorized Kannada text documents. International Journal on Soft Computing, 2(4), p.81.

- Jayashree, R., Murthy, S. and Anami, B.S., 2012, November. Categorized Text Document Summarization in the Kannada Language by Sentence Ranking. In 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA) (pp. 776-781). IEEE.
- ArpithaSwamy, Srinath S, Automated Kannada Text Summarization using Sentence Features, International Journal of Recent Technology and Engineering (IJRTE), Vol-8, Issue-2, 2019, PP – 470-474
- R. Jayashree, K. Srikanta Murthy, Basavaraj S. Anami, Effect of stopword removal on the performance of naïve Bayesian methods for text classification in the Kannada language, International Journal of Artificial Intelligence and Soft ComputingVolume 4Issue 2/3June 2014pp 264–282https://doi.org/10.1504/IJAISC.2014.062824.
- R. Jayashree; Murthy K. Srikanta; Basavaraj S. Anami, Categorized Text Document Summarization in the Kannada Language by sentence ranking, 2013, 12th International Conference on Intelligent Systems Design and Applications (ISDA), DOI: 10.1109/ISDA.2012.6416635
- YashaswiniHegde; Shubha Kadambe; PrashanthaNaduthota, Suffix stripping algorithm for Kannada information retrieval, 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), DOI: 10.1109/ICACCI.2013.6637227.
- Ranga Reddy. Jayashree, Srikanta Murthy K, BasavarajAnami, Effect of stopword removal on the performance of naïve Bayesian methods for text classification in the Kannada language, January 2014, International Journal of Artificial Intelligence and Soft Computing 4(2/3):264 – 282, DOI: 10.1504/IJAISC.2014.062824
- Kavyashree R. Bhat, Saritha Shetty, Kannada Shabhdakosha, International Journal of Research in Engineering, Science and Management, Volume-3, Issue-5, May-2020, ISSN (Online): 2581-5792.
- Daniel M. DiPietro, Quantitative Stopword Generation for Sentiment Analysis via Recursive and Iterative Deletion, arXiv:2209.01519v1 [cs.CL] 4 Sep 2022.
- Tayfun Kucukyilmaz, Tayfun Akin, A Feature Based Approach on Automatic Stopword Detection, Springer Nature 202, Research Square, DOI: https://doi.org/10.21203/rs.3.rs-1986294/v1.
- Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, Lu Sheng Wang, Automatic Construction of Chinese Stopword List, Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015).
- AjiPrasetyaWibawa, HidayahKariimaFithri, Ilham Ari ElbaithZaeni, Andrew Nafalski, Generating Javanese Stopword s List using K-means Clustering Algorithm, Knowledge Engineering and Data Science (KEDS) pISSN 2597-4602, Vol 3, No 2, December 2020, pp. 106–111.
- https://www.britannica.com/topic/Assamese-language, 19-Feb-2014
- Buchanan, Francis Hamilton (1807). A Journey from Madras through the Countries of Mysore, Canara, and Malabar. Volume 3. London: Cadell. ISBN 9781402146756.

- Naik, Vinayak K.; Naik, Yogesh (6 April 2007). "HISTORY OF KOMARPANTHS". hindu-kshatriya-komarpanth. Atom.