Document Summarization for Telugu News Articles

Naga Sudha D[1], Y Madhavee Latha [2]
1Research Scholar JNTUH, Hyderabad
2 MRECW Hyderabad

***Abstract* :** Automatic text summarization is a technique which compresses large text into a shorter text which includes the important information. Telugu is the most popular language used in south India and there is a lack of proper summarization system for telugu text. Hence,in this paper, we present an approach to the design an automatic text summarizer for telugu text that generates a summary by extracting sentences. It deals with a single document summarization based on feature weighting approach. Each sentence in the document is represented by a set of various features namely- key phrase ,sentence paragraph position, sentence overall position, numeric data, presence of inverted commas, sentence length and keywords in sentences. The sentence score is calculated and based on their score is used based on the feature vector. Then based on the required compression ratio, sentences are included in the final summary. The experiment was performed on news articles of different category such as education,Technology, movies, business and sports. The performance of the technique is then compared with the human generated summaries.

***Index terms*: Telugu text Summarization, keyword, Tokenization, Sentence Extraction, Summary Generation.**

I.Introduction : In recent decades, the information contents such as text, video and audio are easily generated by everyone, due to the development of internet techniques and digital capturing system [1]. So, the search for the required data in the large stored data, increases the difficulty and consumes more time. Recently, automatic text summarization is an emerging research topic and also showed great interest among the researchers [2], [3]. The automatic text summarization is used to develop a condensed version of original document. Manual text summarization needs a considerable number of qualified unbiased professionals, high considerable time and budget. To overcome these concerns, automatic text summarization is a growing and interesting field in natural language processing [4]. A lot of text summarization systems are developed for summarizing the documents in various languages such as English, Hindi, Telugu, Malayalam, etc. The summarization process could be done in two different ways such as abstractive and extractive. The extractive approach identifies the more used words and then score the sentences from different perspective [5]. In other words, the abstractive summarization clarify the contents and then improve the coherence among sentences by eliminating redundancies [6].

Currently, there are various approaches available for automatic summarization of multiple documents like feature extraction approaches, clustering approaches, graph based approaches, optimization approaches, etc. [7], [8]. A major problem with these conventional approaches is, time to summarize the multiple documents is raised incredibly, and it makes most of the systems unable to deal with the increase of data sizes [9], [10]. To overcome this problem, an automatic effective text summarization methodology is developed. In this paper, Telugu language was considered for text summarization, because it was the second most prevalent language in India just after Hindi. Also, Telugu ranks fifteenth in the Ethnologue list of most-spoken languages worldwide. Usually, the raw Telugu language dataset consists of more noises in terms of stop-words, low frequency words, etc., which were significantly reduced or pre-processed using stemming procedure.

The output of summary can be of two types: Extractive summaries and abstractive summaries. Extractive summaries[11] are produced by extracting the whole sentences from the source text.

The importance of sentences is determined based on statistical and linguistic features of

sentences. Abstractive summaries are produced by reformulating sentences of the source text. An Abstractive summarizers [12] [13] understands the main concepts in a document and then convey those

concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and terms to best describe it by generating new shorter text that conveys the most significant information from the original text document [14].

Although the various approaches are proposed for summarization of major languages like English, Hindi, Swedish, etc. some challenging problems are still open for other languages of the world. India has around 120 major languages and 1500 other languages. Since recent years many researchers are working on Indian and other languages but less work has done on Indian languages. Though Telugu is the second most popular language used in India (Telangana and Andhra Pradesh) just after Hindi, and there is a lack of proper summarization system for Telugu text. Hence, the technique for Telugu text summarization has been proposed in this paper.

The rest of the paper is structured as follows: Section 2 describes the related work in text summarization, and Section 3 describes the proposed summarization technique. Section 4 presents the experimental work and its results and Section 5 concludes the paper by summarizing the study and gives some future work.

2 Related Work:

Automatic text summarization approaches [15] are also classified as:

i) Vector based approach - The summary generated for each document will consist of sentences that are extracted from it using the Vector Space Model (VSM). After the preprocessing step each text element, a sentence in the case of text summarization, is considered as N-dimensional vector [16]. The sentences are then ranked using the VSM according to their similarity within the document.

ii) Fuzzy based approach - All the rules needed for summarization, are included in the knowledge base of the fuzzy system [17]. Different characteristic of a text such as sentence length, location in paragraph, similarity to key word etc, is given as input to fuzzy system. A value from zero to one is obtained as an output for each sentence based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary.

iii)Genetic algorithm based approach - In Genetic Algorithm, the solutions are called individuals or chromosomes. After the initial population is generated randomly, selection and variation function are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into the next
generation. The quality of an individual is measured by a fitness function. Fattah and Ren [18] proposed an automatic text summarizer using several feature score functions like sentence position, positive and negative keyword, sentence centrality etc. to train genetic algorithm and mathematical regression models to obtain a suitable combination of feature weights.

iv) Neural Network based approach - A neural network is trained on a corpus of documents. The neural network is then modified, through feature fusion, to produce a summary of highly ranked sentences in the document. Through feature fusion, the network discovers the importance (and unimportance) of various features used to determine the summary-worthiness of each sentence [19]. The input to the neural network can be either real or binary vectors.

3. Proposed Method:

The goal of automatic text summarization is to select the most important sentences of the Hindi text document. The proposed method consists of text preprocessing, features extraction, sentence evaluation and selection stage .In the preprocessing stage the document is prepared and converted in a structured form which is given to the following stages. A set of statistical features computed for each sentence to reflect its importance and used in sentence evaluation. Based on sentence score summary is extracted and compared with human generated summary.

3.1Text Preprocessing:This is the first stage of summary generation. Its main purpose is to prepare the input text document for processing for other stages. It transforms the input document into structured format.Text preprocessing includes Normalization, tokenization, stops word removal and stemming.

i)Normalization: The first phase in pre processing is the normalization and segmentation. This phase outputs the article as a set of words by removing the unnecessary symbols like semicolons, colons, exclamation marks, hyphens, bullets, parenthesis, numbers etc. The full-stops are not eliminated to indicate end of a sentence which is required to identify topic sentences from the article for summary generation.
ii)Tokenization:Tokenization process is splitting the input documents into their units with different levels to facilitate accessing all parts of the input document.These units are paragraphs,sentences,tokens or any other appropriate unit.

iii)Stemming : The third phase in pre processing is the transformation of the inflected words into their root form. A stemming is a process of linguistic normalization. In stemming variant forms of a word are transformed to a common form termed as root stem.Combination of Linguistic approaches with machine learning approaches results a new development in technology. Parts-of-speech taggers, constraint-based parsers, statistical parsers, chunkers, tree banks and semantic analyzers are developed for Indian languages. IIIT, Hyderabad and Central university of Hyderabad together developed a software tool, Telugu morphological analyser (TMA) to get the stem forms of the inflected words.

iv)POS Tagging: In this phase each of the words are assigned with the part-of-speech (POS) information. This is done using Natural Language tool kit (NLTK) called Telugu tagger. The Telugu tagger is trained on a tagger named as telugu.pos from the Indian corpus that comes with NLTK. The accuracy is almost 98%.

v)Stop Word removal: Stop words are noisy words and need to be removed from the text. They are the functional words act as determinants, prepositions, conjunctions. Their presence is to form a grammatical sentence and not as a meaningful word in the sentence. They usually occur very frequently and tend to be small in size. Stop word lists are not well established in Indian languages. In this , a set of stop words are identified with POS tagging and document frequency. The words except nouns, verbs, adjectives and adverbs are considered as stop words. 3256 stop words are identified from the data corpus and are removed from the articles.

3.2 Feature extraction: An extractive based text summary involves selecting sentences of high relevance or importance is based on set of features to generate coherent summaries that state the main idea of the given document. Therefore selecting and designing these features will greatly affect the quality of the generated summaries. The feature are classified into four levels including word based level, sentence based level, paragraph based level and graph based features. Since the quality of the generated extractive summary is highly affected by the selected features along with their design, these features that identify the most important sentences in addition to maximizing content converage and diversity between sentences with in the summary.

(f1) **Existence of numerical data** : The existence of numerical data rather than enumerations or bullets such as numbers,dates and time can affect the importance of a sentence. The normalization score of this feature is computed

$$\text{Numbers Score} = \frac{\#\text{Occurences of numbers in the sentence}}{\#\text{Occurences of numbers in the document}}$$

(f2) **Sentence location feature**: The importance of a sentence is dependent on its location. Generallly the first paragraph and last paragraph are important since they provide a summary about the whole document. In each paragraph the first and last sentence are very important and strong candidates to be included in the summary.Sentences that are away from the beginning of the document are less important.

$$\text{Sentence Length Score} = \frac{\#\text{Total number of words in the sentence}}{\#\text{Total number of words in the longest sentence}}$$

(f3) **Sentence centrality feature**: This feature is defined as the similarity or the overlap between a sentence and other sentences in the document. Thus a sentence might be central in the document and many sentences might explain it. Thus a sentence is given a high score when its words occur in a greater number

of other sentences in the document.The centrality score is by computing the similarity matrix using cosine similarity measure similar where each item in the matrix represents the similarity between the corresponding sentences.

$$\text{Centrality} = \frac{\text{Similarity degree of Si}}{\text{Maximum similarity degree in the document}}$$

(f4) **Sentence length feature** : The length of a sentence might affect its importance. Too long or too short sentences are excluded from the summary.Too long sentence will increase its information content and short sentences tend to include less information compared to other sentences and thus they are less important. Thus very short and very long sentences are given a score equal to 0.For other sentences their score are calculated.

$$\text{Length Score} = \frac{\text{\# Words in the sentence}}{\text{\#Words in the longest sentence}}$$

(f5) **Key-Phrases feature**:It is a short list of important and topical keywords that provide a condensed summary of the main topic in the document.It may be a single word or a composite of multiple words.The importance of sentence is conditioned by containing a key Phrase[20] .The score of the key phrase feature depends on many factors including the frequency of the candidate phrase,number of words in each phrase , frequency of the most recurring single word in a candidate phrase , location of the phrase with in the document, location of the candidate phrase within its sentence, relative phrase length to its containing sentence and assessment of the phrase sentence verb content .

Key phrase feature is computed based on three of the most important factors

Key Phrase frequency: It indicates how many times the key phrase appeared in the sentence and it is calculated by KPF= #SKPi/KPd where KPF is the key Phrase feequency. #SKPi is the number of sentences that contain the key phrase (KPi) and KPd is the total number of key phrases in the document.

Key Phrase Length: It is the number of words that the key phrase has. The length of the Key Phrase plays a role in its importance and consequently the sentence importance.The value of this feature is calculated as √KPL where KPL represents the length of KPi. The aim of using square root is to smoothly increase the score if the length is more than one term.

Proper Name:The importance of a key phrase increase if it is a proper name which is a noun

corresponding to a particular person, place or thing. Telugu POS tagger is used.If the key phrase is a proper name then the value of this feature is set to 2 Otherwise it is set to 1.

Key Phrase feature score is calculated as

$$\text{Key phrase score} = \sum\_KPi\ [(KPFi * \sqrt{KPF}] * PNV)$$

Where KPFi is the key phrase frequency of KPi,KPL is the length of KPi and PNV is the proper name value of KPi.The above equation will give a higher score if the length of the key phrase is more than one or if it is a proper noun and will give more score if both factors are found.

(f6) **Semantic Score** :The semantic Score is a numeric value that reflects the degree of semantic closeness between two sentences.The cosine similarity is used to measure the correlation between corresponding sentences.To calculate semantic score and build a similarity matrix between sentences ,each sentence in the document is represented as TF-IDF vector

$$\text{TF} = \frac{\text{Frequrncy of term}}{\text{No of document}}$$

$$\text{IDF} = \log \frac{\text{Frequency of term}}{\text{Number of docs containing term}}$$

TF-IDF= TF*IDF

3.3 Sentence extraction and summary generation:In extractive text summarization, important text segments of the original document usually are identified based on a set of important feature extracted from different levels. In the score based method important sentences are extracted based on the total scores that are assigned to them. The score of each sentence is given by

$$\text{Sentence Score} = \sum_{i=1} Wi * Si$$

Where Si and Wi represent the weight and score of the feature I defined previously.nWi set to one because we take into consideration the importance and contribution of each feature during the processing. For example all features except key phrase feature,and sentence location have a value between 0 and 1.The other features are more important and have a value as following sentence location have valued either 3 or 2 or 1 or less than 1.Key phrase have a value greater than or equal 0.These values reflect the contribution of the feature in the total score and for this assign weigth to be 1.After computing the total score, sentences are ranked in a descending orderbased on their total score.

The top ranked sentences will be selected to included in the output summary based on the required summary ratio. Indeed the sentences that have the highest scores will be representing the most important content of the document and thus will be selected to be included in the final summary. Finally the extracted sentences will be reorder based on their original position on the document to preserve text coherency in the generated summary.

4 Data set: As our main focus is on Telugu news domain, we have collected data set from various telugu e- news papers like Eenadu, Andhra jyothi,Saskhi. etc. We have collected total 6000 telugu news articles from different categories of news namely Business,Education,Technology, Movies and sports.

No of articles collected	6000 telugu News articles
Domain Categories	Technology,Sports,Education,Business,Movies
Average Sentences	30-50
Average Words	100-200

4.1 Evaluation It is required to check accuracy, relevance and usefulness of the summary created by an automatic summarizer. For evaluating results of summarization, accuracy has been used as evaluation parameters. The generated summaries are evaluated against human generated summaries. The experts were asked to create summaries of our collected dataset.

Accuracy is the percentage from which the sentences are correctly classified for the inclusion in the summary .

$$\text{Accuracy}() = \frac{\text{Number of sentences correctly classified}}{\text{Total number of sentences in summary}} \text{X } 100$$

4.2 Experimental Setup with Results: In this experiment, we have used various combinations of features to identify the importance of each sentence and tested on news articles. Total 6000 documents are evaluated with combination of feature sets and compared with human summaries. It can be observed from the results of Table6 that as the compression ratio decreases the accuracy also decreases. Another interesting observation is as the size of input document increases,the accuracy at 50% compression improves as compared for the small sized input.

Input Document: పెను ప్రమాదంలో ఆర్థిక వ్యవస్థ..!_ - ఇష్టం వచ్చినట్టు వ్యవహరిస్తోన్న మోడీ సర్కారు! - నిబంధనల్ని పక్కనబెట్టి అనూహ్య నిర్ణయాలు.. - ప్రశ్నిస్తున్న అధికారులపై కక్ష సాధింపు ధోరణి - ప్రమాదాన్ని గుర్తించి తప్పుకుంటున్న 'బాస్ 'లు - సుబ్రమణియన్ , పనగరియా నమస్కారం!

- మాట్లాడినందుకు ఆర్ బీఐ డైరెక్టర్ నచికేత్ వేటు... - తాజాగా ఆర్ బీఐ గవర్నర్ ఉర్జిత్ పటేల్ పైనా గురి! దేశ ఆర్థిక వ్యవస్థ పెను ప్రమాదంలోకి జారుకుంటోంది.. వ్యవస్థలను, ప్రజా ప్రయోజనాలను పట్టించుకోకుండా ప్రభుత్వం కార్పొరేట్ సంస్థలకు మేలు చేసేలా నిర్ణయాలు తీసుకుంటోంది. సర్కారు నిర్ణయాలు సరైనవి కాదని వాటివల్ల దేశ ప్రయోజనాలకు విఘాతం కలుగుతుందని ఆర్థికవేత్తలు, అధికారులు చెబుతున్నా వినకుండా ప్రభుత్వం ముందుకు సాగుతోంది. ఎదురొచ్చిన అధికారులపై విరుచుకుపడుతోంది. 'ఉంటే ఉండండి లేదంటే సర్దుకోండి..' అన్న విధంగా తన విధానాన్ని స్పష్టం చేస్తోంది. అయినా దేశంపై ప్రేమకలిగిన అధికారులు కొందరు సర్కారుకు.. వాస్తవాలను వెల్లడించి దారిలో పెట్టే ప్రయత్నం చేస్తున్నప్పటికీ.. ప్రభుత్వంలో వినిపించుకొనే ధోరణిలో కనిపించడం లేదు. పెద్దనోట్ల రద్దు విషయంలో మోడీ సర్కారు అనాలోచిత ధోరణి దేశ ప్రజలకు పూర్తిగా స్పష్టమైంది. జీఎస్టీ విషయంలో ఇది మరింత తేటతెల్లమైంది. మోడీ ప్రభుత్వం మొండి వైఖరి గమనించి అధికారులు చేసేది లేక.. ఆర్థిక వ్యవస్థ పూర్తిగా కుప్పకూలక ముందే తమ బాధ్యతాయుత పదవుల నుంచి ఒక్కొక్కరు క్రమంగా తప్పుకొంటూ వస్తున్నారు. కేంద్ర ప్రభుత్వ ముఖ్య ఆర్థిక సలహాదారు అరవింద్ సుబ్రమణియన్, నిటి ఆయోగ్ వైస్ ఛైర్మన్ పనగరియాలూ... మోడీ ప్రభుత్వానికి నమస్కారం పెట్టేసి వెళ్లిపోయారు. ప్రభుత్వంలో మమేకం అయినవారిగా పేరున్న ఈ ఆర్థిక వేత్తలుగా నిలిచిన వీరు ఎక్కువకాలం తమతమ హోదాల్లో పని చేయలేకపోయారు. సుదీర్ఘకాలం సేవలు అందిస్తారనుకుంటే, తక్కువ సమయంలోనే తప్పుకొని వెళ్లిపోయారు. ఇంకా మూడేండ్ల పాటు ఆర్ బీఐ బోర్డులో డైరెక్టర్ కొనసాగాల్సిన నచికేత్ మోర్ ను ప్రభుత్వం అర్ధంతరంగా ఇంటికి పంపేసింది. తాజాగా ప్రస్తుత ఆర్ బీఐ గవర్నర్ ఉర్జిత్ పటేల్ విషయంలోనూ ఇదే జరగబోతోందనే ప్రచారం జరుగుతోంది.నవతెలంగాణ, వాణిజ్య విభాగం: దేశంలో అత్యున్నత బ్యాంక్ ఉంటూ వస్తోన్న భారతీయ రిజర్వు బ్యాంక్ ను (ఆర్ బీఐ) తమ ఇంటి సంస్థగా మార్చుకొని ఆర్థిక వ్యవస్థలో అనూహ్య సంస్కరణలను తీసుకురావాలని మోడీ సర్కారు భావిస్తోంది. విశాలమైన దేశ ఆర్థిక ప్రయోజనాల దృష్ట్యా ఆర్ బీఐ సర్కారు నిర్ణయాలకు తలొగ్గకుండా స్వతంత్రంగా వ్యవహరిస్తూపోతోంది. ఈ పరిణామాలు మింగుడు పడని సర్కారు దేశ చరిత్రలోనే తొలిసారిగా ఆర్ బీఐ చట్టంలోని సెక్షన్ -7ను తెరపైకి తెస్తూ ఆర్ బీఐని బెదిరించే ప్రయత్నం చేసింది. 80 ఏండ్ల ఆర్ బీఐ చరిత్రలో ఎప్పుడూ సెక్షన్ 7ను ఉపయోగించలేదు. నయానోభయానో ఆర్ బీఐ ద్వారా తమకు కావాల్సిన విధంగా నిర్ణయాలు వెలువడేలా ప్రభుత్వాలు చేసుకున్నాయే తప్ప ఇంత రాక్షసంగా వ్యవహరించిన దాఖలాలు లేవు. కొన్ని సార్లు ఆర్ బీఐ గవర్నర్లు తమ స్వయంప్రతిపత్తిని ఉపయోగిస్తూ, ప్రభుత్వ సూచనలను పక్కన పెట్టేవారు. ఇప్పుడు మోడీ సర్కారు ఈ సెక్షన్ నే ఆయుధంగా ఆర్ బీఐని తమ చెప్పు చేతల్లోకి తెచ్చుకోవాలని భావిస్తోంది. రూ.3.6 లక్షల కోట్లపై ప్రభుత్వం కన్ను.. ప్రజా ప్రయోజనాల కోసం ఆర్ బీఐకి ఆదేశాలు జారీ చేసే అధికారాన్ని ఈ ఆర్ బీఐ చట్టం-1934 సెక్షన్ 7 ప్రభుత్వానికి కట్టబెట్టింది. ఇప్పుడు మోడీ ప్రభుత్వం చట్టంలోని సెక్షన్ 7(1) ఆర్ బీఐకి ప్రభుత్వం కనీసం మూడు లేఖలను పంపినట్టుగా సమాచారం. తద్వారా ప్రజాప్రయోజనాల విషయంలో ఆర్ బీఐ గవర్నర్ కు ప్రభుత్వం ఎటువంటి ఆదేశాన్నైనా ఇవ్వగలం అనే సందేశాన్ని పంపింది. ఈ లేఖలేమిటంటే.. తొలి లేఖలో సత్వర దిద్దుబాటు చర్యల (పీసీఏ) నుంచి విద్యుత్ కంపెనీలకు ఆర్ బీఐ మినహాయింపునివ్వాలని కోరింది ఎందుకంటే ఇందులో అత్యధికులు తమ పార్టీకి కార్పొరేట్ దిగ్గజాలే ఉండడం విశేషం. రెండో లేఖలో ఆర్ బీఐ ఉన్న రూ.3.6 లక్షల కోట్ల రిజర్వ్ నిధులను మార్కెట్లో ద్రవ్యలభ్యత పెంచేందుకు, ద్రవ్యలోటు పూడ్చుకునేందుకు ఉపయోగించాలని కోరింది. ఎన్ బీఐఎఫ్ సీ సంస్థలో ద్రవ్య లభ్యత కొరత ఉంది. దీనిని ఆసరగా చేసుకొని ఆర్ బీఐ రిజర్వును కొల్లగొట్టాలన్నది సర్కారు లక్ష్యంగా కనిపిస్తోంది. ఈ సొమ్మును బ్యాంకు రుణాల రూపంలో తిరిగి కార్పొరేట్ సంస్థలకు దోచిపెట్టాలన్న దురాలోచన కనిపిస్తోంది. ఇదే జరిగితే దేశ ఆర్థిక వ్యవస్థ పెను ప్రమాదంలోకి జారుకోవడం తథ్యం. మూడో లేఖలో చిన్న, స్థాయి కంపెనీ (ఎస్ ఎమ్ ఈ)లకు ఇచ్చిన రుణాల విషయంలో బ్యాంకులు నిబంధనలను సడలించాలని సూచించింది. ఎన్ పీఏల వర్గీకరణ, విద్యుత్ కంపెనీలకు నిబంధనల సరళీకరణలోనూ ఆర్ బీఐ కాస్త దిగిరావలని సూచించింది. లేదంటే ఇదే సెక్షన్ కింద ఉన్న అధికారాన్ని ఉపయోగించి ప్రభుత్వం అటు గవర్నర్ ను కానీ.. డిప్యూటీ గవర్నర్ ను కానీ.. ఇతర డైరెక్టర్ నైనా తొలగించడానికి అధికారం ఉందని పరోక్షంగా హెచ్చరించింది. . ఆర్ బీఐ గవర్నర్ , నలు గురు డిప్యూటీ గవర్నర్ లను ప్రధాన మంత్రి ఆధ్వర్యంలోని నియామకాల మంత్రివర్గ సంఘం (ఏసీసీ) నియమిస్తుందన్న సంగతి తెలిసిందే. ఆర్ బీఐ కేంద్ర బోర్డు లోని స్వతంత్ర డైరెక్టర్లను సైతం ఏసీసీ నియమిస్తుంది. తాజా పరిణామాలు చూస్తుంటే అధికారులపై ప్రభుత్వం కక్ష సాధింపులకు దిగుతోందా.. తమకు ఎదురు తిరిగిన వారు ఉన్నత హోదాల్లో ఉండొద్దన్న ధోరణి సర్కారు వర్గాల్లో కనిపిస్తోంది.

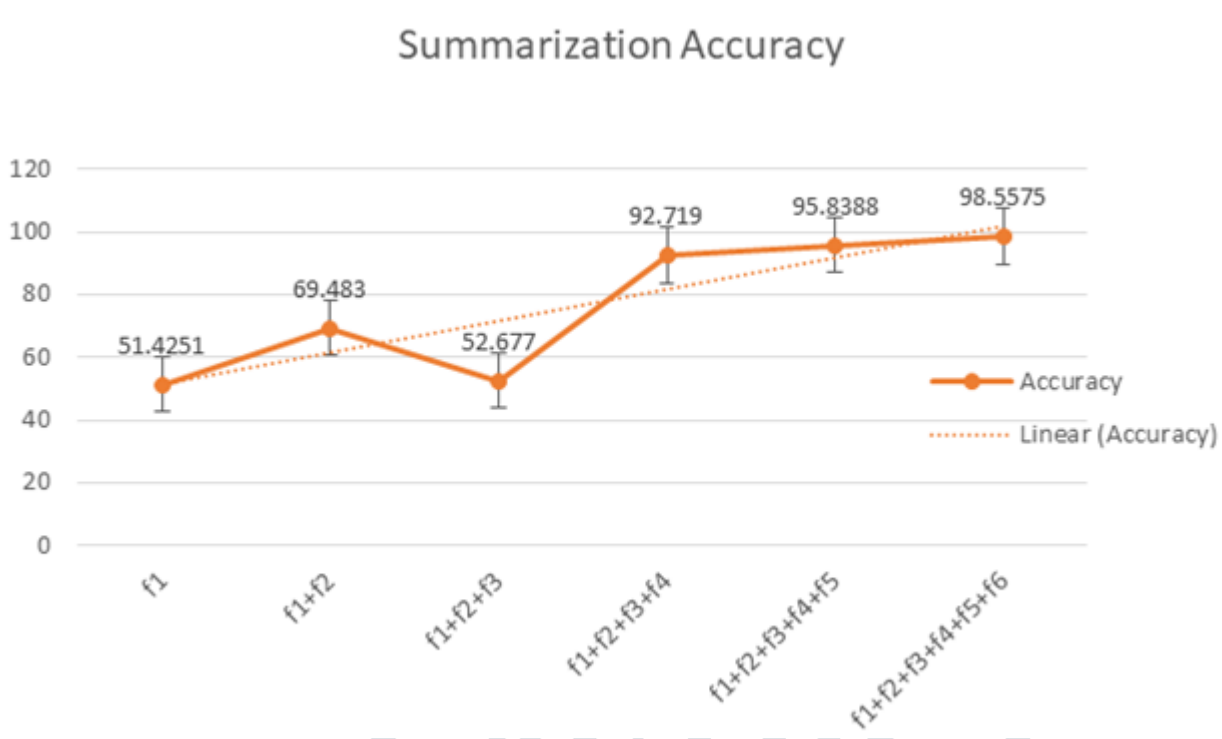
Model Generated Summary :-పెను ప్రమాదంలో ఆర్థిక వ్యవస్థ.!_ - ఇష్టం వచ్చినట్టు వ్యవహరిస్తోన్న మోడీ సర్కారు! - నిబంధనల్ని పక్కనబెట్టి అనూహ్య నిర్ణయాలు.! - నిబంధనల్ని పక్కనబెట్టి అనూహ్య నిర్ణయాలు.పెను ప్రమాదంలో ఆర్థిక వ్యవస్థ. - ప్రశ్నిస్తున్న అధికారులపై కక్ష సాధింపు ధోరణి - ప్రమాదాన్ని గుర్తించి తప్పుకుంటున్న 'బాస్' లు - సుబ్రమణియన్ , పనగరియా నమస్కారం! - మాట్లాడినందుకు ఆర్ బీఐ డైరెక్టర్ నచికేత్ వేటు.! తాజాగా ఆర్ బీఐ గవర్నర్ ఉర్జిత్ పటేల్ పైనా గురి! దేశ ఆర్థిక వ్యవస్థ పెను ప్రమాదంలోకి జారుకుంటోంది.!_

Human Generated Summary Setence::పెను ప్రమాదంలో ఆర్థిక వ్యవస్థ..!_ - ఇష్టం వచ్చినట్టు వ్యవహరిస్తోన్న మోడీ సర్కారు! - తాజాగా ఆర్ బీఐ గవర్నర్ ఉర్జిత్ పటేల్ పైనా గురి! ఇంకా మూడేండ్ల పాటు ఆర్ బీఐ బోర్డులో డైరెక్టర్ కొనసాగాల్సిన నచికేత్ మోర్ ను ప్రభుత్వం అర్ధంతరంగా ఇంటికి పంపేసింది. తాజాగా ప్రస్తుత ఆర్ బీఐ గవర్నర్ ఉర్జిత్ పటేల్ విషయంలోనూ ఇదే జరగబోతోందనే ప్రచారం జరుగుతోంది.నవతెలంగాణ, వాణిజ్య విభాగం: దేశంలో అత్యున్నత బ్యాంక్ ఉంటూ వస్తోన్న భారతీయ రిజర్వు బ్యాంక్ ను (ఆర్ బీఐ) తమ ఇంటి సంస్థగా మార్చుకొని ఆర్థిక వ్యవస్థలో అనూహ్య సంస్కరణలను తీసుకురావాలని మోడీ సర్కారు భావిస్తోంది. విశాలమైన దేశ ఆర్థిక ప్రయోజనాల దృష్ట్యా ఆర్ బీఐ సర్కారు నిర్ణయాలకు తలొగ్గకుండా స్వతంత్రంగా వ్యవహరిస్తూపోతోంది

Example of Telugu Article. Model Generated vs. Human Generated

Features used	Sentence Score	Accuracy
f1	22.30902	51.4251
f1+f2	26.64930	69.483
f1+f2+f3	26.649	52.677
f1+f2+f3+f4	23.68	82.719
f1+f2+f3+f4+f5	25.275	92.8388
f1+f2+f3+f4+f5+f6	26.95	95.5575

Sentence Summarization Accuracy



4. 3. Conclusion :This paper discusses single document automatic text summarization for telugu text using feature weighting technique. As expected it has become clear from the experimental results, the performance in each of the subtasks directly affects the ability to generate high quality summaries. Also it is noteworthy that the summarization is more difficult if we need more compression.

References:

[1] S. Akter, A.S. Asa, M.P. Uddin, M.D. Hossain, S.K. Roy, and M.I. Afjal, "An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm", In Imaging, Vision & Pattern Recognition (icIVPR), IEEE International Conference, pp.1-6, 2017.

[2] J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, and C.J. Pérez, "Extractive Multi-Document Text Summarization Using a Multi-Objective Artificial Bee Colony Optimization Approach", Knowledge-Based Systems, 2017.

[3] S.A. Babar and P.D. Patil, "Improving performance of text summarization", Procedia Computer Science, Vol.46, pp.354-363, 2015.

[4] M. Yousefi-Azar and L. Hamey, "Text summarization using effective deep learning", Expert Systems with Applications, Vol.68, pp.93-105, 2017.

[5] S. Abujar, M. Hasan, M.S.I. Shahin, and S.A. Hossain, "A Heuristic Approach of Text Summarization for Bengali Documentation", In 8th International Conference on Computing, Communication and Networking (8th ICCCNT), 8th International Conference on IEEE, 2017.

[6] T.B. Mirani and S. Sasi, "Two-level text summarization from online news sources with sentiment analysis", Networks & Advances in Computational Technologies (NetACT), International Conference on IEEE, 2017.

[7] A. Kumar, A. Sharma, S. Sharma, and S. Kashyap, "Performance analysis of keyword extraction algorithms assessing extractive text summarization", In Computer, Communications and Electronics (Comptelix), International Conference on IEEE, pp. 408-414, 2017.

[8] R.S. Baraka, and S.N. Al Breem, "Automatic Arabic Text Summarization for Large Scale Multiple Documents Using Genetic Algorithm and MapReduce", Information and Communication Technology (PICICT), Palestinian International Conference on IEEE, 2017.

[9] R.Z. Al-Abdallah and A.T. Al-Taani, "Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm", Procedia Computer Science, Vol.117, pp.30-37, 2017.

[10] S. Swamy, T. Shalini, S.P. Nagabhushan, S. Nawaz, and K.V. Ramakrishnan, "Text Dependent Speaker Identification and Speech Recognition Using Artificial Neural Network", In Global Trends in Computing and Communication Systems, Springer, Berlin, Heidelberg, pp.160-168, 2012.

[11]Kallimani, J. S., Srinivasa, K. G., Eswara Reddy, B., Information retrieval by text summarization for an Indian regional language. In 6th International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, IEEE NLP-KE 2010, 21–23 August 2010, IEEE Catalog Number: CFP10811-PRT,ISBN:978-1-4244-6897-3, pp. 596–599.

[12] . Sarkar, K., Bengali text summarization by sentence extraction. In Proceedings of International Conference on Business and Information Management, NIT, Durgapur, 2012, pp. 233–245.

[13]Genest, P.-E. and Lapalme, G., Text generation for abstractive summarization. In Proceedings of the Third Text Analysis Conference, National Institute of Standards and Technology, aryland,USA, 2010.

[14]Reddy, S. and Sharoff, S., Cross language POS taggers (and other tools) for Indian languages: an experiment with Kannada using Telugu resources. In Proceedings of IJCNLP Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies. Chiang Mai, Thailand, 2011.

[15] Das, D., Martins, A., 2007. A survey on automatic text summarization. Literature Survey for the Language and Statistics II.

[16] C Slamet, A R Atmadja , D S Maylawati , R S Lestari , W Darmalaksana and M A Ramdhani Automated Text Summarization for Indonesian Article Using Vector Space Model IOP Conference Series: Materials Science and Engineering(2018)

[17] Al-Radaideh, Q.A., Bataineh, D.Q., 2018. A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms. Cognitive Comput.

[18] Qazvinian, V., Hassanabadi, L., Beheshti, S., 2008. Summarising text with a genetic algorithm-based sentence extraction. Int. J. Knowl. Manage. Stud. (IJKMS) 4,426–444.

[19]k.KaikhahAutomatic text summarization with neural networks 2004 2nd International IEEE Conference on 'Intelligent Systems'. Proceedings (IEEE Cat. No.04EX791)

[20]A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences,2019