International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp.3356–3365

Marathi Text Summarization for News Articles using Sequence To Sequence with Attention Mechanism

Kavya Nair[1], Kotian Snita[2], Aarati Lomte[3], Praharsha Gottapu[4] and Rushali Deshmukh[5]

[1234]UG Student, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune

[5]Professor, Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune

[1]kavyanair1598@gmail.com, [2]kotiansnita@gmail.com,
[3]aaratilomte1997@gmail.com [4]prahu24@gmail.com
[5]radeshmukh_comp@jspmrscoe.edu.in

Abstract

Text summarization is a common problem in machine learning as it involves developing algorithms and statistical models to create a coherent summary to convey the intended message in fewer words, keeping in mind the user's time-constrained environment. We have built an automatic text summarization model to produce compact and concise summaries while retaining the vital information from the news article using RNN's encoder-decoder using the Sequence-to-Sequence model with an attention mechanism. This model ensures to summarize news articles in the Marathi language on which there have been fewer previous contributions in the field. We have created our dataset, which has proved to be extremely beneficial to validate our model.

Keywords: *Abstractive, Bahdanau, attention model, decoder, encoder, summarization.*

1. Introduction

Machine learning uses algorithms to produce rational, Human-Like results in scenarios that are impracticable for a human to perform in a time-constrained environment. When we look up a particular topic on the internet, there is an unimaginably large quantity of suggestions popping up, which may or may not be useful for a particular user. The task of extracting precise and accurate data as per the user requirement from this huge bulk is a challenge. Manual extraction is an unachievable task when the data is large, so we use the machine learning application of automatic text summarization. Automatic text summarization is a method to generate a compact and accurate summary from the source data. Deep learning is a complex mechanism that involves computing an enormous dataset generating precise results depending on the mechanism used. It can be used for diverse types of datasets that are both labeled and unlabeled. This learning approach has proved to be very useful in image processing as the features to be selected are picked by the model itself after training it; hence guaranteeing less human intervention. In text summarization, usage of deep learning has opened a wide number of techniques that have better results and accuracy. Being a type of machine learning, inspired by the structure of the human brain, this approach gives a human touch to the working of the model and ensures results which are similar to the Human-Generated result. Also, deep learning makes the system powerful enough to process complex tasks and improve its accuracy with the increase in its dataset, which is very beneficial in text summarization of news articles.

Text summarization is deducing the original text by reducing the word-range to increase its readability thereby and reduce the amount of time taken to analyze it. Similarly, when we consider a news article, going through the entire article may not

ISSN: 2233-7857 IJFGCN

always be possible, which may lead the reader to skip the significant and prime aspects of it. The primary purpose of automatic text summarization is to create a fluent summary containing the significant points of the document, formatted in a way that the user understands the entire document by reading the same. For example: The INSHORTS app on Google play store where every news is limited to a 60-word summary, updated to capture daily news. The other fields where deep learning is thriving are text sentiment analysis, medical care by detection of cancer cells or analyzing MRI images for better detection of diseases, customer support by the use of chat bots, voice recognition, etc. These are just a few glimpses of this approach's vast applications.

Text summarization can be further classified into extractive summarization and abstractive summarization. Extractive summarization, as the name suggests is a technique to extract words or sentences from the original text that have higher importance. It is calculated by the frequency count of the most recurring word to generate a summary. The importance is given based on statistical and linguistic features. Its main aim is to retain the salient information that is present in summary by eliminating stop words. Stop words are a set of words, frequently used in a language that does not hold any significance on its own. For example, the prepositions in the English language. An approach to the extractive method is by using TF-IDF, where 'tf' calculates the number of times a word occurs in the particular document, and 'idf' is the number of different documents the word appears in. The ratio of the frequencies of words is compared and a score is created for every word also known as thematic word. On similar lines, these thematic words are calculated per sentence, generating priorities for the same to be a part of the summary. This is just one of the numerous techniques used to implement this approach. The extractive summarization approach has proved to be less efficient, as the generated summary is sometimes grammatically inaccurate. The alternative is to use abstractive text summarization. This approach trains the model to generate self-summaries that implicate Human-Like behavior to fulfill the purpose of using deep learning and overcoming the shortcomings of the extractive approach. It is also capable of creating new summaries by using the words not present in the original text. It involves training the models using different algorithms to teach the system, hence it is considered to be more complex to implement. There is a lot of work done in extractive text summarization to improve its efficiency by using optimization techniques. Still, we have found the abstractive method more useful based on the results achieved for each model, described in section 3.

Our model is designed to generate summaries on news articles in the Marathi language by abstractive text summarization using RNN sequence to sequence model. The usage of encoder-decoder LSTM has proved to be beneficial in this context. The attention mechanism has improved the quality of the generated summaries. It was originally developed for problems that involve human-computer interaction in the area of text translation along with image captioning, movement classification, etc. The paper is classified into two phases; the first phase includes the implementation of various models with their drawbacks. The second phase is the detailed working of our model.

2. Related Work

Numerous researches are going on in automatic text summarization by using machine learning algorithms. But more work is done using the Extractive approach. Chi Zhang et al. (2004) [1] used the technique of sentence selection with semantic representation. It has helped to develop the best summary from the source document, overcoming the shortcomings of the earlier models, and validating their results by processing it using two diverse datasets. Pratibha Devihosur and Naseer R (2019) [2] uses lesk function for Word Sense Disambiguation (WSD) on a dataset from word-net. It helps to summarize text from blogs, web search results, and many more. Sushma R. Vispute and M. A. Potey (2013) [3] built a model for automatic categorization of Marathi documents to show personalized documents to a user

ISSN: 2233-7857 IJFGCN

based on previous searchers using a Lingo clustering algorithm based on Vector space model (VSM). It has many categories ranging from health to tourism. Nallapati et al. (2016) [4] used the sequence to sequence RNN for summarization on the Gigaword corpus dataset and increasing the performance using Large Vocabulary Trick (LVT). This paper concludes that using the sequence to sequence model produces promising outcomes in text summarization. Nallapati et al. (2016) [5] applied an attentional recurrent neural network to achieve reliable results by using two distinct corpora. Dalal et al. (2018) [6] applied a data clustering approach for summarizing Hindi text by Particle Swarm Optimization (PSO) algorithm. Unlike the traditional languages used for such purposes, they have used the Java platform and have achieved reasonably acceptable results. Bhosale, S. Et al. (2018) [7] This paper proposes a summary generation method by fetching the top-ranked words and extracting lines comprising of those keywords which are restricted to a certain pre-specified length. Hao Xu et al. (2018) [8] This paper proposes a Generative Adversarial Network (GAN) for summarization. It consists of two models: A generator and a triple-RNN discriminator to generate summaries and evaluate them respectively. Shah Chintan and Anjali Jivani (2018) [9] This paper proposes usage of an adversarial network for text summarization. It has a combination of supervised and unsupervised approaches for text summarization, as Self-Organizing Maps (SOM) to learn features, map them and re-train which is unsupervised and Artificial Neural Network (ANN) for summarization which is supervised. Nihar Ranjan et al. (2016) [10] This paper proposes a method for automatic text document summarization using the extractive approach for multiple text documents of the same domain. Cosine similarity is used to calculate the similarity measure and sort them into unique clusters generating summaries that will be presented to the client by a web service. Jobson Elliott and Abiel Gutiérrez [11] This paper is an enhancement of certain attributes from [5] considering it as their baseline. They have implemented different models with variations of word embeddings, the complexity of encoder-decoder and attention mechanism. Shimpikar Sheetal and Sharvari Govilkar (2017) [12] contributed by comparing different approaches for text summarization in different regional languages along with their types. Also, the paper concluded that TF-IDF, Graph-based are recognized as the most effective text summarization methods for most of the Indian Regional Languages among Bengali, Malayalam, Hindi, Odia, Gujarati, Tamil, Telugu, Gujarati and Punjabi. As described above, deep learning has proven to be reliable in text summarization by using various techniques. Also, much of the work is done on other Indian regional languages except the Marathi language. Even though there exist some models for Marathi text summarization, our model can be used as the base for future research in this field as it generates reliable results.

3. Model phases

3.1. Model 1

The first model we implemented is based on the extractive text summarization, and the method used here was the text ranking system. It provided a score for each sentence in a text, and the top-n sentences were chosen and sorted as they appeared in the text to build an automatic summary. A dataset consisting of news articles about tennis was used for training and testing purposes. Each of the articles consists of a minimum of 200 words. The generated summary as a result of the assigned values was inaccurate due to its grammatical inconsistency.

ISSN: 2233-7857 IJFGCN

3.2. Model 2

The second model is based on an abstractive sequence-to-sequence model, which consists of encoder and decoder. Training of the model was executed using the Amazon reviews dataset (2016) consisting of five lakh records, and each of these reviews was given ratings as 1 for a positive response and 0 for a negative response. The dataset consisted of reviews for various items like books, accessories, etc. in the English language. By using variants of Recurrent Neural Networks, which are preferred as the encoder and decoder, helps in overcoming the problem of vanishing gradient as they are capable of retaining long term sequences. "The input sequence is converted into a fixed-length vector by the encoder, and thereby the decoder predicts the output sequence". The problem here is, it only works for short sequences as it becomes difficult for the encoder to memorize long sequences.

3.3. Model 3

Our third model is based on an improvement over the second model. We have introduced an attention mechanism here. It aims to predict a word by only looking at a specific part of the sequence. We have used a local attention mechanism that focuses on only a few source positions. The attended context vector is derived by considering some hidden states of the encoder. The results generated may not match with the actual summary in terms of words, but they both convey similar meanings. The same dataset as used for model 2 is used here for training and testing purposes on which it gives an output of accuracy 86%. The model is also trained and tested on an updated Amazon reviews dataset (2019), which consists of over 5 lakh records on which it gives an output of accuracy 88%.

3.4. Model 4

Our fourth model is based on an abstraction text summarization of Marathi news articles, which is created after comparing and evaluating various metrics of the above-mentioned models. The dataset used here is a self-generated dataset of source news articles fetched from various Marathi e-newspapers and thereby creating its summaries. This model shows the implementation of a sequence to sequence model using RNN with an attention mechanism. The model generates results with an overall accuracy of 60-70%. Due to the lack of a pre-existing Marathi inbuilt library, it is very difficult to interpret the appropriate Marathi words for the summary.

4. Proposed Model

The baseline of our model is the encoder-decoder's sequence to sequence mechanism. "A sequence to sequence model is designed to map a fixed-length input with a fixed-length output, where the input and output length can vary". The model is trained with a sequential input, and it generates an output likewise. The news article is fetched from an online source that dynamically updates. The basic flow of the system includes an encoder to process the input sequence and a decoder to generate the output sequence shown in Figure 1.

ISSN: 2233-7857 IJFGCN

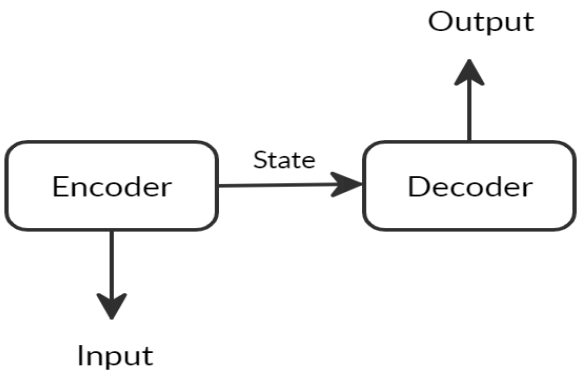

Figure 1. Encoder decoder structure

As our domain deals with the prediction of words that must be covered in the summary, we use a stacked LSTM(long short term memory) to overcome the diminishing gradience problem along with the sequence to sequence prediction problem due to the varying size of input and output. The sequence to sequence prediction problem requires the input and output to be in a sequential format with varying lengths.

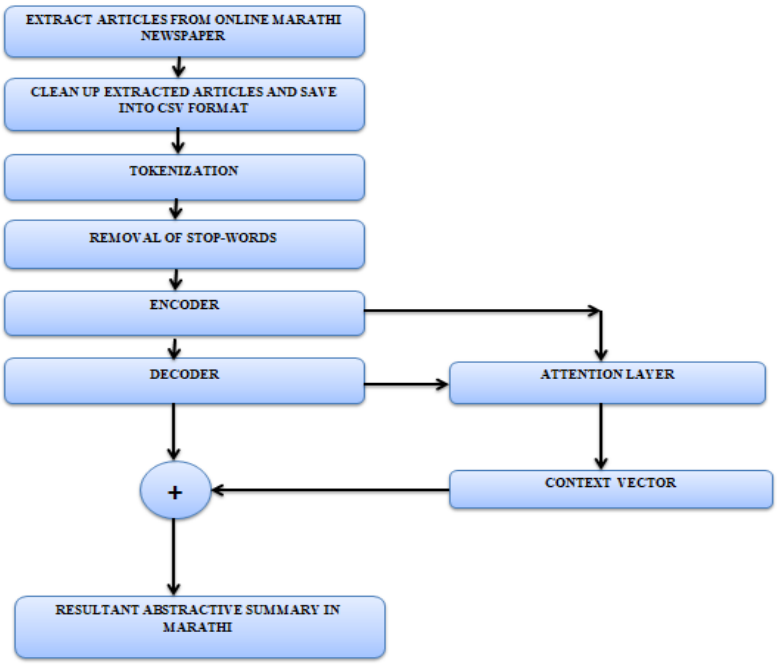

Figure2. Flow chart of the proposed model

4.1. Pre-processing

As shown in figure 2, pre-processing involves several steps starting from news extraction from a Marathi newspaper. All the articles are cleaned and saved into a .csv file. The process of cleaning the data mainly involves column titles conversion to match the format of News API, date format conversion to match the format of News API, removal of duplicate articles and many more secondary processes. The articles in the file are tokenized and stop-words are eliminated by the creation of a self-generated file.

ISSN: 2233-7857 IJFGCN

4.2. Encoder-Decoder

The encoder and decoder have two different sets of LSTM architecture. An encoder LSTM model reads the entire input sequence and it processes the information for each time step. At each time-step, one word is fed into the encoder. As decoder is trained to predict the next word in the sequence, given the previous word, the LSTM decoder model reads the entire target sequence word-by-word and processes it. Special tokens are added to the target sequence before feeding it to the decoder. In the first processing, the entire input sequence is encoded and the decoder is initialized using encoder's internal states and passed as an input to the decoder. For one time-step, the decoder runs with all the internal states. The probability of the next word is given as output and the word with the highest probability is selected. Then the selected word is given as input to the decoder for next time-step and the internal states of the current time-step are updated. This process repeats until repeated until the maximum length of the target sequence is reached or token is generated. Our model consists of three stacked LSTM encoders to generate a sequential output. As we have created the dataset ourselves, the entries used for training is pretty accurate in terms of its reliability. The next step involves the attention mechanism, which helps in subduing the shortcomings of the pre-existing system, explained in the next section.

5. Attention Mechanism

The neural network approach that we had implemented in section 3, model 2 had a potential issue as the encoder was not being able to capture all the data from the input sequence and compress it into a fixed-length vector as the entire input sequence is taken into consideration by the decoder for prediction purpose. As the length of the input increases, the capability to memorize it decreases, so we use this mechanism. The goal of the attention mechanism is to predict a word by focusing on the parts that are of prime importance during summary generation while excluding the other words of the source document. There are mainly two classes of attention mechanism namely, global attention and local attention. It focuses on all source positions and all the hidden states of the encoder into consideration and produces an attended context vector; whereas, the latter considers only a few hidden states of the encoder. We have implemented the local attention mechanism in our model. The following are the working steps of this mechanism.

For every time step 'j' in source sequence, the encoder outputs the hidden state (h_j). Similarly, for every time step 'i' in the target sequence, the decoder outputs the hidden states (s_i).

A score is computed, which is based on the alignment of the source word and the target word using score function, which is called an alignment score. The alignment score (e_{ij}) for target time step 'i' and source timestep 'j' is given as

$$e_{ij} = \text{score}(s_i, h_j) \quad (1)$$

We have used softmax function to normalize the alignment score and fetch the attention weights (A_{ij}):

$$Aij = e^{eij/\sum_{k=1}^{Tx} e^{eik}} \quad (2)$$

The attended context vector (C_i) is given as the linear sum of attention weights (A_{ij}) and hidden states (h_j) of the encoder.

ISSN: 2233-7857 IJFGCN

$$C_i = \sum_{j=1}^{Tx} A_{ij} h_j \quad (3)$$

Attended hidden vector (S_i) is produced by concatenating the hidden state (s_i) of the decoder and attended context vector (C_i) at the time step (i).

$$S_i = \text{concatenate}([S_i ; C_i]) \quad (4)$$

The attended hidden vector is provided into the dense layer to produce y_i

$$y_i = \text{dense}(S_i) \quad (5)$$

6. Dataset

We have created a dataset by fetching articles from Maharashtra times e-newspaper and devising summaries for the same. Due to the lack of a conventional Pre-Existing dataset, we have tried to train our model by creating it ourselves. The features taken under consideration are the original news article and its generated summary, which may or may not contain words from the original summary. The length of the article chosen is over 15 lines, and that of the summary is approximately 3 lines. The dataset consists of a variety of articles from sports, politics, business, technology, etc. The size of the dataset we used for training was 500 and above. We are still working on it and testing the accuracy after every training. Randomized shuffling of the dataset makes sure that only a specific criterion of data is not being provided for training as it might hamper the generation of new summaries. A ratio of training and testing is 90% - 10%.

7. Qualitative analysis of summary

The results given below are obtained after training different models containing a sample set of variations. Although we have displayed the discrete possibilities of summaries generated by the model, this is just a sample corresponding to the actual result set.

7.1. Model 1 result

Input Article: Federer, 37, first broke through on tour over two decades ago, and he has since gone on to enjoy a glittering career. Federer is hoping he can improve his service game as he hunts his ninth Swiss Indoors title this week.....
Result: Federer said earlier this month in Shanghai in that his chances of playing the Davis Cup were all but non-existent.

7.2. Model 2 result

Review: Excellent film...speaks to the penalty of interracial relationships at the time. I would surely recommend this film to others to watch.
Given_summary: Excellent film.
Result: Great film!

ISSN: 2233-7857 IJFGCN

7.3. Model 3 result

Review: I am very satisfied with the quality of the honey, the product is as advertised, I use honey on cereal, with raw vinegar, and as a general sweetener.
Given_summary: Quality Honey
Result: Great Honey

7.4. Model 4 result

Review: शस्त्र आयातीत भारत आता जगातील सर्वात मोठा देश राहिला नाही. सौदी अरेबिया शस्त्र आयातीत अव्वल स्थानी आहे तर भारत दुसऱ्या स्थानी. या यादीत मिस्त्र तिसऱ्या, ऑस्ट्रेलिया चौथ्या तर चीन पाचव्या स्थानावर आहे. हे अव्वल पाच देश जगातील एकूण शस्त्र आयातीच्या ३६ टक्के आयात करतात. शस्त्रांच्या आयात निर्यातीवर लक्ष ठेवणाऱ्या स्टॉकहोमच्या सिप्री संस्थेने २०१५ ते २०१९ दरम्यान शस्त्र खरेदी व विक्री व्यवसायाची आकडेवारी एका अहवालातून सादर केली आहे. जगात सध्याच्या घडीला अमेरिका सर्वाधिक शस्त्रनिर्यात करतो. त्यापाठोपाठ रशिया, फ्रान्स, जर्मनी व चीन यांचा क्रमांक लागतो. नव्या माहितीनुसार पश्चिम आशियाई देश सर्वाधिक शस्त्र खरेदी करीत आहेत. सध्या हैती बंडखोरांसोबत युद्धाचा सामना करीत असलेला सौदी अरेबिया सर्वाधिक शस्त्र आयात करीत आहे. शस्त्र निर्यातीच्या क्षेत्रात जागतिक स्तरावर अमेरिका व फ्रान्सने आघाडी घेतली आहे. अमेरिकेच्या एकूण निर्यातील २३ टक्क्यांनी वाढ नोंदविण्यात आली आहे. जगातील एकूण शस्त्रनिर्यातीत आता अमेरिकेचा वाटा ३६ टक्के येवढा आहे. २०१५-१९ दरम्यान अमेरिकेचा शस्त्रनिर्यात व्यवसाय दुसऱ्या क्रमांकावर असलेल्या फ्रान्सच्या तुलनेत ७६ टक्क्यांनी जास्त आहे. अमेरिका जगातील ९६ देशांना शस्त्रांचा पुरवठा करीत आहे. महत्त्वाचे म्हणजे अमेरिकेच्या एकूण निर्यातीमध्ये अर्धा वाटा पश्चिम आशियाई देशांचा आहे. त्यातही अर्धा वाटा एकट्या सौदी अरबने खरेदी केला आहे. सौदी अरबने एकूण शस्त्रास्त्र विक्रीच्या १२ टक्के शस्त्रांस्त्रांची खरेदी केली आहे. सौदी अरबला सर्वाधिक शस्त्रांची विक्री अमेरिकेने केली आहे. भारताची राफेल खरेदी व मिस्त्र व कतारने शस्त्रास्त्र खरेदी केल्याने फ्रान्सच्या शस्त्रास्त्र निर्यातीत मोठी वाढ झाली आहे.
Given_summary: सौदी अरेबिया शस्त्र आयातीत अव्वल स्थानी आहे तर भारत दुसऱ्या स्थानी. जगात सध्याच्या घडीला अमेरिका सर्वाधिक शस्त्रनिर्यात करतो. अमेरिका जगातील ९६ देशांना शस्त्रांचा पुरवठा करीत आहे. सौदी अरबला सर्वाधिक शस्त्रांची विक्री अमेरिकेने केली आहे. भारताने केलेल्या शस्त्रास्त्र खरेदीमुळे फ्रान्सच्या शस्त्रास्त्र निर्यातीत मोठी वाढ झाली आहे.

Result: शस्त्र आयातीत सौदी अरेबिया अव्वल स्थानी आहे तर भारत दुसऱ्या स्थानी तसेच अमेरिका सर्वाधिक शस्त्र निर्यात करत आहे.

8. Result

By the result from section 7.4, we can infer that the model we have built has generated the desired output. The given result is a sample generated a result that gives a glimpse of the actual result. The abstractive representation of the articles in a novel manner to preserves the user's time and efforts required in reading the entire article. Since the model is in its initial training phase, the reliability of the system can change according to the environment, input features, and size of the dataset. The dataset used was a self-generated dataset by fetching the source articles from Maharashtra times e-newspaper. The dataset set consisted of about 500 records from various domains like sports, politics, business, technology and many more, randomly shuffled to train the model with unique and diverse inputs. The length of the article was about 15 lines and that of the summary about 3 lines. The overall accuracy achieved was 76% by using a Bilingual Evaluation Understudy (BLEU) for estimating sentence quality. Taking into account the methods we have applied, there is a possibility of repetition of words in the final summary which can be improved with increased training.

ISSN: 2233-7857 IJFGCN

9. Conclusion

Thus we have used deep learning to implement an abstractive text summarization model for the Marathi language by using attentional Encoder-Decoder architecture with Sequence-to-Sequence mechanism. This model overcomes the problems of the previously implemented models and thus contributing to a promising outcome. The future scope is to create a robust model with better performance in terms of its prediction accuracy for various types of documents, develop interactive interfaces so that the user can easily explore news and other information of various domains in the Marathi language. This model can also be used to summarize the text in different languages in the future.

10. Acknowledgment

It gives us great pleasure in presenting the paper on “Marathi Text Summarization for News Articles using Sequence To Sequence with Attention Mechanism”. We would like to thank our project guide Prof. Rushali A. Deshmukh, for all the help and guidance provided.

References

[1] Zhang, C., Zhang, L., Wang, C. J., & Xie, J. Y. (**2014**, November). Text summarization based on sentence selection with semantic representation. In 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (pp. 584-590). IEEE.

[2] Pratibha Devihosur1, Naseer R2. "Automatic Text Summarization using Natural Language Processing." **(2017)**.

[3] Vispute, Sushma R., and M. A. Potey. "Automatic text categorization of Marathi documents using clustering technique." 2013 15th International Conference on Advanced Computing Technologies (ICACT). IEEE, **2013**.

[4] Nallapati, Ramesh, Bing Xiang, and Bowen Zhou. "Sequence-to-sequence rnns for text summarization." **(2016)**.

[5] Nallapati, Ramesh, Bowen Zhou, Caglar Gulcehre, and Bing Xiang. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 **(2016)**.

[6] Dalal, Vipul & Malik, Latesh. **(2018)**. Semantic Graph Based Automatic Text Summarization for Hindi Documents Using Particle Swarm Optimization. 284-289. 10.1007/978-3-319-63645-0_31.

[7] Bhosale, S., Joshi, D., Bhise, V., & Deshmukh, R. A. **(2018)**. Marathi e-Newspaper text summarization using automatic keyword extraction technique. International Journal of Advance Engineering and Research Development, 5(3), 789-792.

[8] Xu, H., Cao, Y., Jia, R., Liu, Y., & Tan, J. (**2018**, November). Sequence Generative Adversarial Network for Long Text Summarization. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 242-248). IEEE.

ISSN: 2233-7857 IJFGCN

[9] Shah, Chintan, and Anjali Jivani. "A hybrid approach of text summarization using latent semantic analysis and deep learning." 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, **2018**.

[10] Prof. Nihar Ranjan, Pranay N. Lonkar, Sanket M. Sathe, Nayan A. Shendre, and Sonali M. Shingade, "Automatic Text Document Summarization", IJRASET, **(2016)**, vol. IV, no.1, pp. 2321-9653.

[11] Jobson, Elliott, and Abiel Gutiérrez. "Abstractive Text Summarization using Attentive Sequence-to-Sequence RNNs."

[12] Shimpikar, Sheetal, and Sharvari Govilkar. "A Survey of Text Summarization Techniques for Different Regional Languages in India." International Journal of Computer Applications 165.11 **(2017)**.

ISSN: 2233-7857 IJFGCN