International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

An Adroit Approach for Extractive Text Summarization

Tulasi Prasad Sariki, G. Bharadwaja Kumar, Utkarsh Shukla, Ayush Mishra

Abstract: ***Over recent years, there has been growing amount of textual data on the World Wide Web. Hence, there is an increasing need for condensing the humungous text information while retaining its content and complete meaning. Text Summarization is the process of shortening the source text into a more concise form without losing the essence of the original text. Out of the two fundamental approaches i.e. abstractive and extractive, extractive summarization is the predominant approach in literature which fetches the significant sentences by using statistical and linguistic characteristics. In this paper, a judicious framework for extractive text summarization has been presented. The proposed approach contains three different concurrent pipelines to improve the effectiveness of the Summarization process. The proposed framework combines Statistical, NER-based and CUE-phrase methods in an effectual way to extract the summary. The novelty in our approach is to use semantic distance between the sentences to remove the redundant sentences in the final phase. The experimental results show that the proposed framework surpassed the ROUGE-L scores given by state-of-art summarization techniques.***

Index Terms**: *Text Summarization, Extractive Summary Sentence Scoring, Statistical Analysis, Semantic Analysis.***

I. INTRODUCTION

Text summarization is a subfield under the domain of Natural Language Processing which aims at distilling the most important information from a source (or sources) to produce an abridged version for a particular user(s) and task(s). Text summarization is being presently used in various search engines to match the contextual similarity between the search query and the information present on the pages. It also plays vital role in question-answering systems. Recently, there has been increasing interest in developing summarization systems for various domains such as summarizing finance articles, biomedical documents, weather news, terrorist events and many more. This process reduces the problem of information overload because only a summary needs to be read instead of reading the entire document. Hence, the summarization techniques used for these purposes must have a high reliability such that the important aspects of the given textual data should be included but redundant information must be excluded from the summary. Since computers lack world knowledge and language learning capability, it makes automatic text summarization a very difficult and non-trivial task. There are many issues like redundancy, temporal dimension, co-reference, sentence ordering, etc., that make summarization task more complex [1]. Also, the evaluation of the summarization systems is potentially challenging because of the subjectivity of the human created summaries as well as choosing the appropriate metrics for evaluation [2, 3]. In general, there are two different approaches for automatic summarization: extraction and abstraction. Both the techniques are used for summarizing text either for single document or for multi-documents. Extractive methods aim at selecting salient phrases, words or sentences and sometimes passages from documents which perfectly summarizes the document and presents them as a summary. Instead, abstractive summarization techniques aim to concisely paraphrase the information contained in the documents which usually needs information fusion, sentence compression and reformulation. Because of the fact that abstractive summarization methods cope with problems such as semantic representation, inference and natural language generation, data-driven approaches like extractive summarization methods have been extensively studied in the literature [4].

This paper aims at providing a novel approach for robust, fast and effective means of summarization of the text. The approach has different phases. At first, it extracts the candidate sentences from the original text using three independent approaches i.e. Statistical, NER and CUE phrase. Then, we use the majority voting method to extract the summary. Finally, we use Word Mover's Distance (WMD) semantic distance to eliminate redundant sentences while maintaining disparity amongst each other so that they cover the whole document. It is also worth noting that the time complexity of the proposed method is on par with the requirements for the real time implementation in different fields.The paper is organized as follows. Section-2 gives an overview of the state-of-art works done in this field. The proposed model is presented in Section-3. Section-4 explicates the experimental set up and datasets used in experiments. Section-5 discusses the empirical results and inferences. Finally, Section-6 presents the conclusions of this work.

II. RELATED WORK

Even though exhaustive research has been done in the field of text summarization, the new models, algorithms and architectures are being constantly proposed. In this section, we discuss about the papers that have set up the benchmark in extractive summarization on standard datasets.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Tulasi Prasad Sariki, SCSE, Vellore Institute of Technology, Chennai.
G. Bharadwaja Kumar, SCSE, Vellore Institute of Technology, Chennai.
Utkarsh Shukla, SCSE, Vellore Institute of Technology, Chennai.
Ayush Mishra, SCSE, Vellore Institute of Technology, Chennai.

Jun-Ping Ng et al. [5] proposed a heuristic approach for extractive text summarization in which scores of each sentence are given on the basis of features extracted from the sentences such as title of the document, length of the sentences (ideally taken as 20), sentence position in the whole document and keyword frequency.

Rada Mihalcea et al. [6] proposed text Rank method in which they followed four steps to extract the summary of the given document. Initially, they tagged the words present in the document and then applied Lemmatization. Then, the key words are extracted with their normalized frequencies. Thirdly, the scores are calculated using Jaccard Distance between the keywords and sentences. Finally, the summarization of the document is done using most significant keywords and sentences. In the whole method, they have assumed keywords as the nodes of a graph and the weights of the edges are defined using the Jaccard distance between two nodes i.e. key-phrases.

SummaRuNNer introduced by Ramesh Nallapati et al. [7] puts forward a deep learning model. They proposed a Recurrent Neural Network (RNN) based sequence model for extractive summarization of documents. The advantage of the approach is its easy interpretability; because it gives a provision for visualization of its learning which are the results of the abstract features namely information content, salience and novelty.

Y Wu and B Hu designed a reinforcement learning method named Reinforced Neural Extractive Summarization (RNES) model for Extractive Coherent Summary using Deep Reinforcement Learning framework [8]. A Jadhav and V Rajan presented a novel neural sequence-to-sequence model for Extractive Summarization called SWAP-NET (Sentences and Words from Alternating Pointer Networks) [9]. K Al-Sabahi, Z Zuping and M Nadher considered the summarization task as classification problem and used hierarchical structured self-attention mechanism to create the sentence and document embeddings [10].

Chenliang Li et al. [11] proposed a guiding generation model that combines the Extractive method and the Abstractive method by introducing a Key Information Guide Network (KIGN), which encodes the keywords to the key information representation. YC Chen and M Bansal Introduced fast summarization model that first selects salient sentences and then rewrites them abstractively [12].

III. PROPOSED METHOD

The proposed approach combines the simplicity of the statistical method coupled with the use of semantic information present in a document. Our work aims at creating a balance between the two aspects and includes the ingenuity held by the two approaches. The statistical aspect involved in Key Phrase Extraction method uses Term Frequency for ranking the sentences is based on the idea that if some words or phrases occur more number of times within a document they might hold more importance within the document. The Semantic approach is based on the fact that every document conveys a central idea and every phrase or sentence closely represents or depicts the idea of the document in which they occur. Our proposed method first scores each sentence using the basic statistical approaches.

The proposed work consists of following modules.

A. KEYWORD/KEYPHRASE EXTRACTION & SCORING

In this step, the whole document is consumed as a raw data and outputs keywords and key-phrases. These keywords and key-phrases are extracted on the basis of stopwords in the document. A keyphrase is considered as the sequence of words occurring in between two stopwords having length greater than one otherwise is called as keyword. The extracted keywords and keyphrases in a sentence or document play an important role in predicting the generalized meaning for a sentence or document.

The extracted keyphrases are the candidate features for statistical pipeline process. The next task in this pipeline is scoring the sentences with the help of candidate features. For scoring the sentences, we have adopted LexRank based approach which is based on the concept of eigenvector centrality in a graph representation of sentences. The scores are calculated on the basis of frequency (termed as Freq) and degree (termed as Deg) of a particular word and finally summing scores of all the words present in a keyphrase to get a score for a keyphrase. Degree of a word ***Deg(w)*** is defined as the frequency of a word ***w*** co-occurring with the extracted keyphrase while the frequency of a word ***Freq(w)*** is defined as the frequency of a particular word w occurring in the document. After calculating the degree and frequency of words, finally the score of each word is calculated by dividing degree of the word by the frequency of that word.

After calculating the scores of each word, the scores for keyphrases are calculated on the basis of scores of the words in a keyphrase. Individual scores of the words present in a specific keyphrase are summed up to get the final score of the keyphrases. On the basis of calculated scores of keyphrases, scores of sentences in a document is calculated by summing up all the scores of keyphrases present in a sentence.

Finally, we calculate the scores based on both Key Phrase Method and also the basic Term Frequency method and then the scored sentences are arranged in decreasing order so that the highest scored sentence is ranked first. On the basis of compression ratio given by the user, set of sentences are extracted from the document.

B. NER

One important task that has been largely neglected in the literature for automatic text Summarization is Named Entity Recognition. Named Entities are usually considered as important gestures to the topic or theme of a text. They contribute largely in defining the domain of the text. Therefore, Named Entity Recognition should greatly enhance the identification of important text segments in text summarization process. SpaCy is an efficient natural language processing toolkit for named entity recognition, which can identify the labels of the contiguous tokens. It can label a broad spectrum of named or numerical entities, which includes Persons, Events, Locations, Organizations, Product-Names, Dates, Times etc. In the proposed framework, we have used spaCy to label the named entities and all the found entities are given equal weightage.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

C. CUE Phrase

Cue phrases are words or phrases that depicts the structure of a discourse. They are also known as discourse markers, discourse connectives, and discourse particles in natural language processing. Cue phrases, such as "significantly", "conclusion" or "in particular" are often followed by important information. Thus, sentences that contain one or more of these cue phrases are considered more important than sentences without cue phrases. Hence, we have collected the cue phrase list from various sources and used in our experiments.

D. FILTERING SENTENCES

Initial Filter: On the basis of compression ratio given by the user, set of sentences are extracted from the document. Depending upon the user preference ***top N*** sentences are taken where N is the compression ratio.

Final Filter: After selecting the first N percent of sentences from the document, selected sentences are then compared with each other so that similar sentences can be removed from the selected sentences. Similarity between the sentences is calculated using WMD described in paper [15] in which a pre-trained model is used to find how similar two sentences are with each other.

WMD is a method that allows us to assess the "distance" between two sentences or documents in a meaningful way, even when they have no words in common. It uses word2vec vector embeddings for calculating the semantic distance. This Word Movers Distance can be seen as a special case of Earth Movers Distance (EMD), or Wasserstein distance, which is better than bag-of-words (BOW) model in a way that the word vectors capture the semantic similarities between words. The threshold value of the semantic similarity was chosen after considerable number of experiments and the algorithm performs best when the value is about 0.78 which means 78% similarity. If the similarity between the sentences is greater than 0.78 then the sentence having lower score is removed from the selected set of sentences. The proposed algorithm is centered over the fact that we should not just include the sentences that only depict the central idea of the document, that doesn't depict the complete summary. The summary after removal of redundant set of sentences from the selected set of sentences, final set of sentences are given as the summary of the given document.

IV. PROCEDURE

In order to extract the summary, we pass our documents into three independent pipelines and then combine the results.

The first pipeline takes in the raw data and performs preprocessing such as stop-word removal, lemmatization of the words in the sentences. The next task in the pipeline is to calculate candidate keywords and phrases using similarity graphs to compute degree of centrality. Based on the degree and frequency calculated, respective scores are assigned to keyword and keyphrases. Each sentence in the document is then scored based on the scores of the different keywords they possess. The sentence scores are normalized within the range of [0-1]. Second pipeline takes the raw data without any pre-processing and is used to find the NER scores for each sentence wherein we count the number of named identities that occur in a sentence and score the sentence by dividing the NER count of each sentence by the number of words. The NER are identified using the spacy library implemented in python. After the sentences are scored based on NER the scores are normalized between the ranges of [0-1] based on the NER scores.

The third pipeline also uses the raw data without any pre-processing and the sentences are scored based on the number of cue phrases that occur in the sentences. The cue phrases are identified from a set of predetermined cue phrase. The scores of the sentences are then normalized in between [0-1] only based on cue phrase.

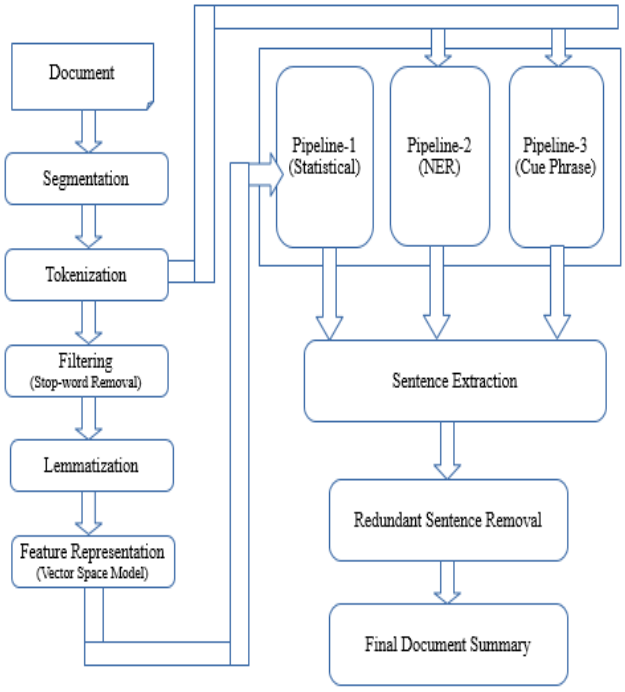

Figure 1: Overall Architecture of Text Summarization System

All these pipeline outputs are considered together and final score is generated using majority voting based on the normalized scores. Finally, they are sorted in decreasing order of the sentences scores.

The top N sentences are then taken based on the compression ratio. We then compare the sentences with each other based on their similarity score using word mover's distance, if two sentences are similar above a particular threshold the sentence with the lower score is removed. This ensures that every sentence in the summary is diverse and covers most of the document. For our case we have taken 40 percent that is the document would be reduced to 40 percent of its original size. The normalization of the scores at each pipeline is necessary so that we can give equal weight to each of the above method when considering scoring of the sentences.

V. RESULTS AND COMPARISON

Since the proposed methodology does not require any training data, we have used only the test set from Document Understanding Conference (DUC-2001) dataset.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

This dataset contains 60 reference sets, out of which 30 articles for training and 30 for testing. Every set comprises of documents, per-document summaries, and multi-document summaries, with sets defined by different types of criteria such as event sets, opinion sets, etc. Data has been provided on request from NIST. For evaluation purposes, we have used the ROUGE metrics [14]. Recall Oriented Understudy of Gisting Evaluation (ROUGE) is a set of metrics which gives a score based on the similarity in the sequences of words between a human-written model summary and the machine generated summary. Thus, it helps us to automatically evaluate the summary. ROUGE includes five measures like ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. ROUGE-1 refers to overlap of unigrams between the system summary and reference summary. ROUGE-2 refers to the overlap of bigrams between the system and reference summaries. ROUGE-L – measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length. From the literature, it can be understood that Rouge-L is very effective in automatic text Summarization. In the present work, we compared ROUGE-1, ROUGE-2 and ROUGE-L scores of different summarizers [14].

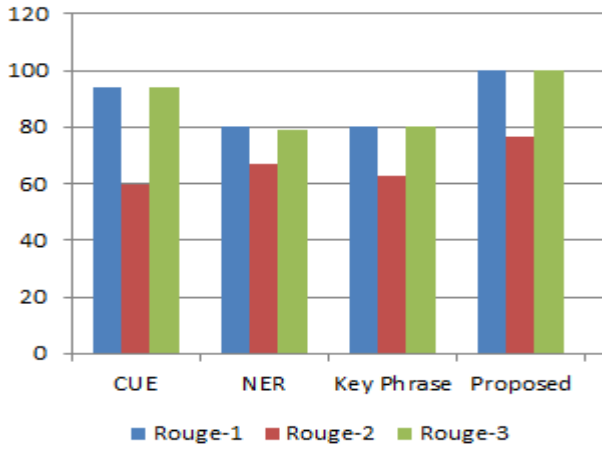

Figure 2: Comparison of the Rouge scores with all the individual methods

Table-1: Comparison of the proposed Algorithm with other state-of-art Text Summarization Algorithms by means of average Rouge score

	Rouge-1	Rouge-2	Rouge-L
Statistical Method	31.2	4.21	26.5
Rake Summary	**53.15**	11.29	**42.96**
Word Mover Distance	48.6	11.2	39.8
SummaRuNNer	46.0	23.1	41.1
Proposed Algorithm	**49.3**	**11.2**	**43.01**

From the Table-1 and Figure-2, clearly one can observe that the proposed model surpasses ROUGE-L scores of the state-of-art summarizers.

A sample of the summaries generated by different summarizers including the proposed method for one DUC 2001 document and corresponding gold standard summary are shown below. Original Document consists of 17 sentences and gold standard summary consists of 5 sentences.

Gold standard Summary: "A strong earthquake that measured 5.2 on the Richter Scale rocked the northern Japanese Island of Kokkaido at about 7:45 local time today. Three hours earlier a quake of similar strength off the Philippine Island of Luzon was felt in Manila. Every increase of one integer on the scale represents a tenfold increase in strength, thus a quake measuring 5.2 is ten times stronger than one of 4.2. Today's quakes caused little damage because they occurred far below the earth's surface. The last quake to cause great destruction in Hokkaido occurred in 1968. It measure 7.9 and claimed 52 lives."

Statistical Summary: "Separate strong earthquakes rocked northern main Philippine island Luzon today swaying trees apparently causing serious injuries officials said. Second strong temblor hit northern Hokkaido less three days Central Meteorological said Philippine Institute Seismology said centered miles Manila Luzon Sea Earthquakes measuring Richter scale strong capable causing considerable damage areas However quakes located deep surface therefore caused little damage officials said earthquakes also occurred areas crustal plates thought rub together Meteorological Agency official said time since earthquake activity Hokkaido surprising caused pressure said speaking anonymity official said way knowing whether quake related one Philippines away Institute spokesman Ulpiano Trillanes said affected wide area Luzon island including apparently caused serious damage injuries.".

RAKE Summary: "Separate strong earthquakes rocked northern Japan and the main Philippine island of Luzon today swaying buildings and trees but apparently not causing any serious injuries or damage officials said. The Japanese quake registered a preliminary 5.2 on the Richter scale and was the second strong temblor to hit the northern island of Hokkaido in less than three days Central Meteorological Agency officials said. They said it struck at 7:44 (am Sunday) and was centered in the Hidaka mountain range 45 miles away from Saturday's earthquake which measured a preliminary 6 Local police said there were no immediate reports of damage or casualties. The pre-dawn Philippine earthquake struck at 3:45 am. The Japanese quake registered a preliminary 5.4 Richter reading the Royal Observatory of Hong Kong said.".

WMD Summary: "5 would be 10 times stronger than one of 5 The official said there was no way of knowing whether the Japanese quake was related to the one in the Philippines 2250 miles away However both quakes were located deep below the earth's surface and therefore caused little damage officials said In the past 35 years there have been two major earthquakes in Hokkaido The Richter scale measures the total magnitude of an earthquake. Separate strong earthquakes rocked northern Japan and the main Philippine island of Luzon today swaying buildings and trees but apparently not causing any serious injuries or damage officials said 2 on the Richter scale and was the second strong temblor to hit the northern island of Hokkaido in less than three days Central Meteorological Agency officials said They said it struck at 7:44 a Every increase of one digit in the scale represents a tenfold increase in magnitude thus a reading of 4.".

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

SummaRuNNer Summary: “Separate strong earthquakes rocked northern Japan and the main Philippine island of Luzon today, swaying buildings and trees but apparently not causing any serious injuries or damage, officials said. They said it struck at 7:44 a.m. (6:44 p.m. Sunday) and was centered in the Hidaka mountain range, 45 miles away from Saturday's earthquake, which measured a preliminary 6.4 on the scale. The Philippine Institute of Volcanology and Seismology said it was centered 110 miles northwest of Manila under the Luzon Sea. In the past 35 years, there have been two major earthquakes in Hokkaido. In 1968, the Tokachi-Oki earthquake measuring a 7.9 on the Richter scale killed 52 people. Every increase of one digit in the scale represents a tenfold increase in magnitude, thus a reading of 4.5 would be 10 times stronger than one of 5.5.”

Proposed Method: “Separate strong earthquakes rocked northern Japan and the main Philippine island of Luzon today, swaying buildings and trees but apparently not causing any serious injuries or damage, officials said. The Japanese quake registered a preliminary 5.2 on the Richter scale and was the second strong temblor to hit the northern island of Hokkaido in less than three days, Central Meteorological Agency officials said. They said it struck at 7:44 a.m. (6:44 p.m. Sunday) and was centered in the Hidaka mountain range, 45 miles away from Saturday's earthquake, which measured a preliminary 6.4 on the scale. In the past 35 years, there have been two major earthquakes in Hokkaido. In 1968, the Tokachi-Oki earthquake measuring a 7.9 on the Richter scale killed 52 people. Every increase of one digit in the scale represents a tenfold increase in magnitude, thus a reading of 4.5 would be 10 times stronger than one of 5.5.”.

VI. CONCLUSION

In this paper, we proposed a framework for an effective text summarization which uses a judicious combination of the different phases. Our proposed framework surpasses the ROUGE-L score of state-of-art text summarizers in the literature. Since, NER is also included in the pipeline, it avoids most of the pronouns which may create ambiguity in the summary while co-referencing. The proposed approach has consistent quality for single document summarization on short news text in DUC 2001 documents. The proposed method can be utilized to generate the plot of novel or book to give book readers a brief idea about the book by saving time.

REFERENCES

1. J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, “multi-document summarization by sentence extraction”, in: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, Association for Computational Linguistics, 2000, pp. 40–48.
2. C.-Y. Lin, “looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?”, in: Proceedings of NTCIR Workshop-4, 2004, pp. 1–4.
3. E. Lloret, M. Palomar, Text summarisation in progress: a literature review, Artificial Intelligence Review 37 (1) (2012) 1–41.
4. M. Gambhir, V. Gupta, “recent automatic text summarization techniques: a survey”, Artificial Intelligence Review 47 (1) (2017) 1–66.
5. J.-P. Ng, P. Bysani, Z. Lin, M.-Y. Kan, C.-L. Tan, “exploiting category-specific information for multidocument summarization”, in: Proceedings of COLING 2012, 2012, pp. 2093–2108.
6. R. Mihalcea, P. Tarau, “textrank: Bringing order into text”, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 1–4.
7. R. Nallapati, F. Zhai, B. Zhou, “summarunner: A recurrent neural network-based sequence model for extractive summarization of documents”, in: Proceedings of AAAI, 2017, pp. 3075–3081.
8. Wu, Yuxiang, and Baotian Hu. "Learning to extract coherent summary via deep reinforcement learning." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
9. Jadhav, Aishwarya, and Vaibhav Rajan. "Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
10. Al-Sabahi, Kamal, Zhang Zuping, and Mohammed Nadher. "A hierarchical structured self-attentive model for extractive document summarization (HSSAS)." IEEE Access 6 (2018): 24205-24212.
11. Li, Chenliang, et al. "Guiding generation for abstractive text summarization based on key information guide network." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018.
12. Chen, Yen-Chun, and Mohit Bansal. "Fast abstractive summarization with reinforce-selected sentence rewriting." arXiv preprint arXiv:1805.11080 (2018).
13. Kryściński, Wojciech, et al. "Improving abstraction in text summarization." arXiv preprint arXiv:1808.07913 (2018).
14. C.-Y. Lin, “rouge: A package for automatic evaluation of summaries”, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.
15. Kusner, Matt, et al. "From word embeddings to document distances." International Conference on Machine Learning. 2015.

AUTHORS PROFILE

Tulasi Prasad Sariki, holds bachelor’s degree and master’s degree in computer science and engineering. He is currently pursuing his PhD degree in the field of computer science and engineering. His research interests include natural language processing, Machine Learning and Data Science.

G. Bharadwaja Kumar holds a PhD degree in computer science and his research interest include machine learning, data analytics, Internet of things, speech and natural language processing. He is very passionate about developing resources and applications for Indian Languages in the areas of Natural Language Processing and Speech.

Utkarsh Shukla, is currently bachelor’s degree in School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. His research interests include natural language processing, Machine Learning and Data Science.

Ayush Mishra, is currently bachelor’s degree in School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. His research interests include natural language processing, Machine Learning and Data Science.