Automatic Text Summarization System

Dr. R. Deepa[1], J. Konshi[2], A. Haritha[3], K. Shobini[4]

[1]Associate Professor, [2,3,4] Final year, B.Tech Information Technology
Department of Information Technology,
Loyola- ICAM College of Engineering and Technology, Chennai.

Abstract - Text mining or text analytics is the process of deriving high-quality information from text. It is challenging for users to go through the entire content available on the internet. Text summarization is part of text mining. Text summarization methods are greatly needed to address the ever-growing amount of text data available online to discover better relevant information. The text summary is extracted using natural language processing and text mining techniques. The natural language processing involves various steps like normalization, tokenizing and word embedding. Then text mining technique is applied to extract the summary of the text automatically. The existing systems use bag-of-words, TF-IDF and Jaccard Similarity techniques for generating summaries, which leads to redundancy. The proposed system resolves the redundancy of the summary by using glove word embedding and pairwise cosine similarity sentence ranking. Glove word embedding is used because it takes into account the order of words in the sentences, unlike bag-of-words and TF-IDF. Sentence ranking is done and higher ranked sentences are extracted which forms the summary of the text. The summary of the text is then sentimentally analyzed for polarity and subjectivity parameters. The summarized text is also subjected to speech conversion.

Keywords: Text mining, text summarization, natural language processing, sentimental analysis, text-to-speech.

1. INTRODUCTION

Text mining identifies and extracts facts from the massive amount of textual data and then converts into structured data, for analysis, visualization, integration with structured data [1]. An automatic text summarizing is the part of text mining. Automatic Text Summarizer determines most informative sentences from the entire document and then creates a representative summary.

Summarization is of two types, extractive and abstractive [2]. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express [3]. This involves using natural language processing (NLP) and text mining techniques to parse unstructured input data into more structured forms and deriving patterns, insights from the data that would be helpful for the end user [4]**.** The processing involves various steps like tokenizing, stopwords, normalization and word embedding. Tokenization refers to splitting up a larger body of text into smaller lines, words or even creating words for a non-English language. Stop words are English words that do not add much meaning to a sentence and can be safely ignored. Normalization reduces derivationally correlated forms of a word to a common root word. Word embedding is a process of mapping collective name for a set of language modelling and feature learning techniques in NLP, where words or phrases from the vocabulary are mapped to vectors of real numbers. The existing systems use bag-of-words [5], TF-IDF (term frequency-inverse document frequency) [6] and Jaccard similarity [7], and lead to redundancy problem, when there is a duplication of words in the same sentence [3].

In this paper, the proposed system solves this problem using Glove and Cosine similarity. Glove model is an unsupervised learning algorithm for obtaining vector representation of words [8]. It takes into account the order of words in sentences, unlike Word2Vec [9] and fastText [10] models. The cosine similarity [11] determines the similarity between the vectors of words. Then the summarized text is obtained by computing cosine similarity and extracting higher ranked sentence. Then the summarized text is also subjected to speech conversion using Google Text-to-Speech engine [12][13].

Finally, the sentimental analysis is carried out to determine the attitude or the emotion of the writer using polarity and subjectivity parameters. Polarity identifies positive and negative statements. Subjective of then sentences generally refer to personal opinion, emotion or judgment [14].

2. RELATED WORK

Namita Mittal et al [1] proposed a text summarization approach based on the removal of redundant sentences. The summarization takes places in two stages where the input of a stage is the output of the previous stage and after each stage, the output of the summary is less redundant than the previous one. In this approach, more than 60% of the generated sentences match with the original input text.

Santosh Kumar Bharti et al [15] proposed a hybrid approach to extract keyword automatically for multi-document text summarization in e-newspaper articles. This approach has extracted around 96.23% of semantically common sentences among the articles.

Roshna Chettri [4] proposed a text summarization approach using natural language processing and various extractive summary approaches like statistical based, topic-based, graph-based and machine learning based. The features with better results of extractive summarization can be combined together to make better summarization of the text.

Shahana Bano et al [16] proposed a method for document summarization using Agglomerative hierarchical clustering, K means clustering, DBSCAN clustering and text analytic techniques to reduce the data redundancy.

Arvind Singh Raghuwanshi and Satish Kumar Pawar [14] finds out the polarity of twitter data using sentimental analysis. It evaluates two classifiers, one is linear and other is probabilistic for sentiment polarity categorization.

Kaladharan [12] proposed a system to convert the international language English text into speech sign. Text handling and speech generation are two main mechanisms of text to speech system. Although many texts to speech system are available in the text-to-speech field, the .NET framework system contributes satisfactory results.

3. PROPOSED SYSTEM

Reading the large content of text available online is challenging for the users as it consumes a large amount of time. The proposed system, Automatic Text Summarization is much more practical and

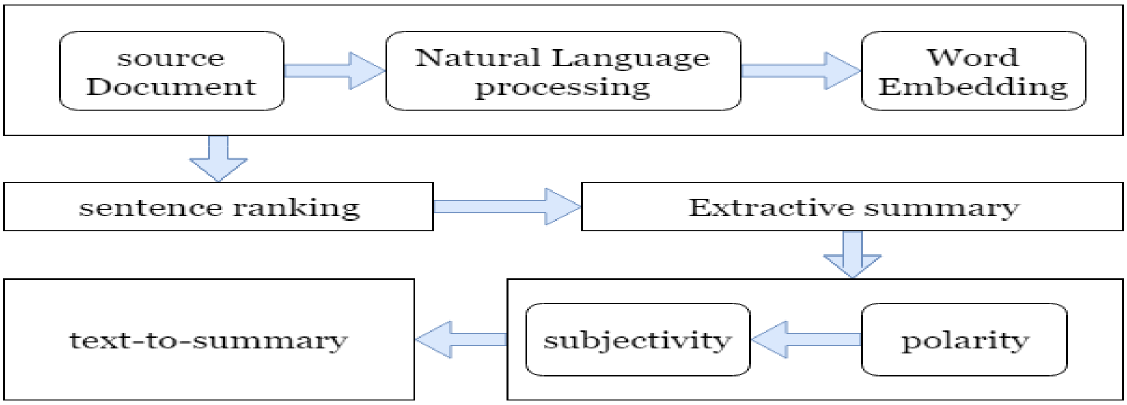


Fig.1: Architecture of Automatic text summarization

applicable in real time. Figure 1 shows the architecture of the proposed system. The input text is processed using natural language processing and processed input is converted into vector form using word embedding. Word embedding is the collective name for a set of language modelling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers. Sentence ranking is done between sentences to extract higher ranked sentence, which forms the extractive summary of the input. The summarized text is then analyzed using polarity and subjectivity parameters. The summarized text is also subjected to speech conversion.

3.1 SUMMARIZATION

The input text is tokenized and then normalization is done by removing the stop words. Stop words are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. The processed input is converted into vector form using Glove word embedding model. Glove is an unsupervised learning algorithm for obtaining vector representation of words. The vector input is subjected to sentence ranking using cosine similarity and higher ranked sentences are extracted to form the extractive summary of the input. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary.

3.2 SENTIMENTAL ANALYSIS

Sentimental analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Sentimental analysis is done on summarized data using polarity and subjectivity analysis. Polarity analysis takes into account the amount of positive or negative terms that appear in a given sentence. Subjectivity analysis generally refer to personal opinion, emotion or judgment.

3.3 TEXT-TO-SPEECH CONVERSION

The summarized data is also subjected to speech conversion using screen reader application. There are several APIs available to convert text to speech in python. One of such APIs is the commonly known as the gTTS (Google Text-to-Speech) API. gTTS is a very easy to use tool which converts the text entered, into audio which can be saved as a mp3 file.

4. PERFORMANCE EVALUATION

The proposed system uses Cosine similarity and its performance is compared with Jaccard Similarity. Jaccard similarity coefficient measures the similarities between sets. It is defined as the size of the intersection divided by the size of the union of the two sets. Cosine similarity is the cosine of the angle between two *n*-dimensional vectors in an *n*-dimensional space. It is the dot product of the two vectors divided by the product of the two vectors' lengths (or magnitudes).

Example sentences for calculating similarity
Sentence 1: AI is our friend and it has been friendly
Sentence 2: AI and humans have always been friendly

Figure 2 shows the Venn diagram of these sentences.

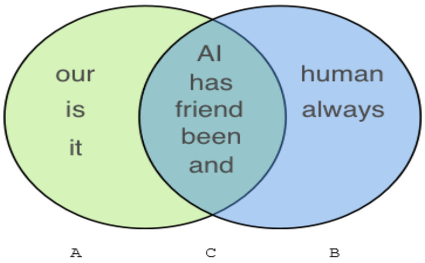


Fig. 2: Venn diagram of the two sentences

Here, A=(sentence1).difference(sentence2)
B=(sentence2).difference(sentence1)
C=(sentence1).intersection(sentence2)

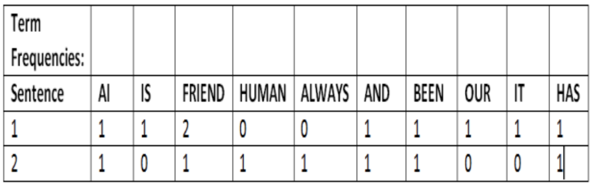
Jaccard Similarity for the above sentence is shown as calculated as:

(len(C)) / (len(A) + len(B)+ len(C) =5/(5+3+2) = 0.5

Where, len(A), len(B), len(C) are number of words in A, B and C respectively.

Cosine Similarity is calculated as follows:
(i) Tabulate the count of each word with respect to its occurrence as shown in table 1.

Table1: Frequency of each word

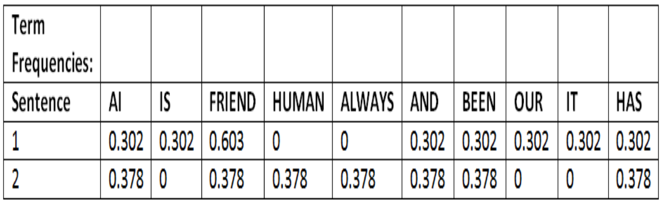


Term Frequencies:										
Sentence	AI	IS	FRIEND	HUMAN	ALWAYS	AND	BEEN	OUR	IT	HAS
1	1	1	2	0	0	1	1	1	1	1
2	1	0	1	1	1	1	1	0	0	1

ii) Then, summing up squares of each frequency and taking a square root, L2 normalization [17] of Sentence 1 is 3.3166 and Sentence 2 is 2.6458.

iii) Dividing above term frequencies with these norms, we get the values as shown in table 2.

Table 2: Values after applying norms



Term Frequencies:										
Sentence	AI	IS	FRIEND	HUMAN	ALWAYS	AND	BEEN	OUR	IT	HAS
1	0.302	0.302	0.603	0	0	0.302	0.302	0.302	0.302	0.302
2	0.378	0	0.378	0.378	0.378	0.378	0.378	0	0	0.378

Cosine Similarity = (0.302*0.378) + (0.603*0.378) + (0.302*0.378) + (0.302*0.378) + (0.302*0.378) = 0.684

Therefore, cosine similarity of the two sentences is 0.684, which is different from Jaccard Similarity of the exact same two sentences with value 0.5.

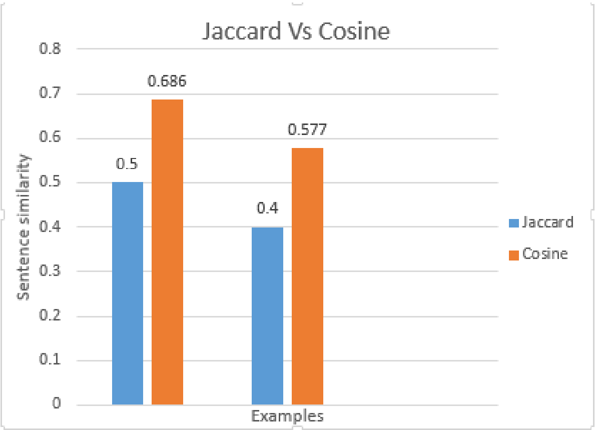


Fig. 3: Comparison of Jaccard and cosine similarity algorithm in sentence ranking.

Thus, the cosine similarity is good for cases where duplication matters while analyzing text similarity. Figure 3 represents the Jaccard and cosine similarity for different sentences.

Figure 4 represents the number of sentences in the given input text and the number of sentences generated as the output based on the cosine similarity with thresholds of 0.07 and above. The evaluation at the threshold of 0.07 is chosen as a good trade-off between quality and number of sentences over the threshold.

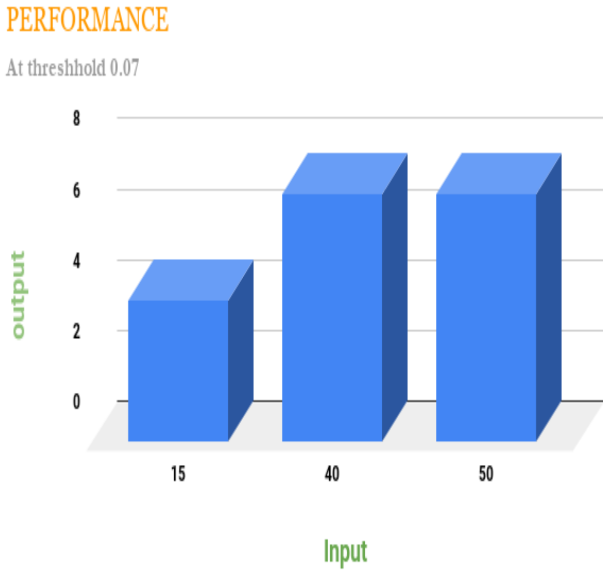


Fig. 4: Input and output sentences

5. CONCLUSION

This approach gives the brief account of the ideas and techniques used to produce reliable and meaningful summaries using cosine similarity sentence ranking algorithm, the summaries are analyzed sentimentally to identify the contextual emotions and the google-text-to-speech engine is used for text-to-speech conversions of the summaries. The main aim of an automatic summarization system is to produce a precise meaningful summary for a large volume of information available. But, it still requires a lot of improvement due to the huge amount of data available online in different formats. Hence, the work can be extended to scrap different formats of data available online to produce robust summaries and also could be used as a groundwork to probe the abstractive summary by generating novel sentences by either rephrasing or using the new words to be more expressible as humans rather than simply extracting important sentences.

6. REFERENCES

[1] Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal and Manoj Kumar Jain, “Extractive Text Summarization”, International Journal of Current Engineering and Technology, Vol.4, No.2, pp. 870-872, 2014

[2] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B.Gutierrez, Krys Kochut, “Text Summarization Techniques: A Brief Survey”, July 2017, Retrieved from https://arxiv.org/abs/1707.02268

[3] Ch. Nanda Krishna, S. S. R. K. K. G. S. B. S. “Automatic Text Summarization: Comparison of Various Techniques”, *International Journal of Engineering Technology and Computer Research*, Vol. 5, *Issue*. 2, pp. 8-14, 2017

[4] Roshna Chettri, Udit Kr. Chakraborty, “Automatic Text Summarization”, International Journal of Computer Applications, Vol. 161, Issue 1, pp. 5-7, 2017.

[5] Zhang, Y., Jin, R. & Zhou, ZH., “Understanding bag-of-words model: a statistical framework”, International Journal of Machine Learning and Cybernetics, Vol. 1, Issue 1–4, pp. 43–52, 2010.

[6] J. Ramos, “Using tf-idf to determine word relevance in document queries”, In ICML, 2003.

[7] S. Niwattanakul, J. Singthongchai, E. Naenudorn and S. Wanapu, “Using of Jaccard Coefficient for Keywords Similarity”, Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I, IMECS 2013, pp. 1-5, 2013.

[8] Pennington J, Socher R, Manning CD. “GloVe: Global vectors for word representation”, In EMNLP, pp. 1532-1543, 2014.

[9] Xin Rong, “word2vec Parameter Learning Explained” CoRR abs/1411.2738, 2014.

[10] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T., “Enriching word vectors with subword information”, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146, 2017.

[11] Ye, J., “Cosine similarity measures for intuitionistic fuzzy sets and their applications”, Mathematical and Computer Modelling, Vol. 53, pp. 91 – 97, 2007.

[12] Kaladharan N, “An English Text to Speech Conversion System”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 10, 2015.

[13] K. Samudravijaya and M. Barol, “Comparison of Public Domain Software Tools for Speech Recognition”, ISCA Archive, 2013.

[14] Arvind Singh Raghuwanshi, Satish Kumar Pawar, “Polarity Classification of Twitter Data using Sentiment Analysis”, International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 5, Issue. 6, pp. 434-439, 2017.

[15] Santosh Kumar Bharti, “Automatic Keyword Extraction for Text Summarization in Multi-document eNewspapers Articles”, European Journal of Advances in Engineering and Technology, Vol. 4, Issue 6, pp. 410-427, 2017.

[16] Shahana Bano, B Divyanjali, A K M L R V Virajitha, M Tejaswi, “Document Summarization Using Clustering and Text Analysis”, International Journal of Engineering & Technology, Vol. 7, Issue 2.32, pp. 456-458, 2018.

[17] Wang, X., Wang, L., Qiao, Y, “A comparative study of encoding, pooling and normalization methods for action recognition” ACCV, pp. 572–585, 2012.