Comparative Study of Sentence Embedding Techniques on News Summarization

K. Vishnu Sainadh, K. Satwik, V. Ashrith, Deepa Gupta
Department of Computer Science,
Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.

bl.en.u4aie19028@bl.students.amrita.edu, bl.en.u4aie190034@bl.students.amrita.edu, bl.en.u4aie19066@bl.students.amrita.edu,
g_deepa@blr.amrita.edu

***Abstract*— Text summarization has arisen as an important topic of debate in the academic world in recent years as a direct result of the advancements that have been made in natural language processing. Abstractive summarizing and extractive summarization are the two sorts of methodologies that are available for doing summarization. The extractive summarization using LexRank is going to be the primary focus of this project. These algorithms have been put to the test in a variety of situations to see which ones perform the best, and they all outperform each other on key criteria. LexRank is being performed using five different sentence transformers. The purpose of this article is to assess whether or not an algorithm performs better than extractive summaries that were generated by people using the BBC news dataset. ROUGE Score was utilized to assign a value to each summary after all of the methods had been applied to the same dataset.**

Keywords—Extractive Summarization, LexRank, Sentence transformers, ROUGE.

I. INTRODUCTION

Because people's lives are so hectic, they don't have the time to read lengthy texts very often. This causes them annoyance due to the fact that they are regularly obliged to read lengthy documents. As a consequence of this, in order to save time, a shortened version of these documents is necessary. Getting your hands on a condensed version of these resources would be time efficient due to the fact that shorter texts are considerably simpler to read and grasp. As many sectors are faced with an increasing volume of textual data to process, summarization remains a subject of study. Because creating a summary by hand is an expensive and time-consuming operation, automated methods to generate them are required. The purpose of text summary is to create a concise and intelligible version of a big text, such as an article or a review. The two approaches of summarizing a document are known as the abstractive and extractive summarizations, respectively.

The generation of new textual pieces is the goal of abstractive summarization, which is done in order to summarize the source material. The process of summarization may be seen as a matching sequences issue. The objective of extractive summarizing is to generate a summary by using a section of the text that is extracted from the original text. The two most common approaches are an activity called sequence labelling, the objective of which is to select the phrases that have been flagged as being a part of the summary, and a rating task, in which the sentences deemed to be of the utmost significance are ranked first. Both of these approaches are described in more detail below. Because most summaries produced by humans are abstract, it's difficult to locate datasets for these jobs.

LexRank is the name of the strategy for text summarization that we built and analyzed for this particular piece of research. We used five Sentence representations as input to LexRank. We evaluated these two methods on the BBC news dataset.

II. LITERATURE SURVEY

In this part of the article, we discussed some previous research that was conducted in the field of text summarization. In this study, we compared the Universal Sentence Encoder (USE), Sentence BERT, MPNet, MiniLM, and LaBSE sentence embedding's, all of which are used on the Lex Rank Algorithm.

In a recently published research with the title "Automatic Summary Generation Using TextRank based Extractive Text Summarization Technique," They discussed the implementation of the TextRank algorithm in order to efficiently summarize texts in a variety of formats. The benefits of using the TextRank algorithm were also mentioned in the aforementioned study. The query-time cost of the TextRank algorithm is significantly lower when contrasted with the costs of other extractive summarization techniques. This is only one of the many advantages that come with using the approach. TextRank is thought to be more practical than other algorithms due to the fact that it does not train the entire model and instead focuses intently on the data that is now accessible. In addition to this, it is considered that the TextRank algorithm is more accurate than other algorithms.

ROUGE is an example of the kind of standard assessment measure that may be found in the text summarization field. Recall Oriented Understudy for Gisting Evaluation is what the abbreviation ROUGE stands for. This measure was first presented in a piece of research written by the authors of "ROGUE: A Package for Automatic Evaluation of Summaries". Lin presented ROGUE, an algorithm for the automated assessment of summaries, in the paper that was ultimately published. Comparing the newly produced summary to the ideal summary that was established in advance is how it assesses the quality of the summary.

There are four distinct ROUGE measures, which are denoted by the notations ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S respectively. In this particular research, we make use of two of these metrics.

III. CONCEPTS

Summarization:

The act of generating a shortened version of one or more documents, called a summary that preserves the most of the meaning of the original text is known as summarization.
Types of Summarization –

- Extractive summaries, often known as "extracts," are created by stringing together a number of phrases that are taken from the source material precisely as they are presented there.
- In order to communicate the most important information from the input, abstractive summaries (also known as abstracts) are generated. These summaries may repeat phrases or sentences from the original text, but they are, for the most part, conveyed in the author's own words.

Lex Rank:

LexRank is an unsupervised approach for summarizing text that is based on ranking the degree to which phrases are key to graphs. The primary concept here is that individual phrases might "suggest" more examples of their kind to the reader. As a consequence of this, a sentence is likely to be of utmost significance if it has a great number of characteristics with a number of other sentences. The significance of the sentences that "recommend" it is the source of the relevance that this phrase possesses. As a consequence of this, a sentence needs to be comparable to a significant number of other sentences in order for it to get a high ranking and be included in a summary.

Sentence Embedding:
Sentence embedding refers to a group of natural language processing (NLP) approaches in which phrases are translated to real-number vectors. Methods for embedding sentences employ the usage of vectors to encode whole sentences together with the semantic content of those phrases. This makes it easier for the computer to understand the context of the text, as well as its intent and any other complexities.

Five different sentence transformers are being used:

- Universal Sentence Encoder
- Sentence BERT
- MPNet
- MiniLM
- LaBSE

The Universal Sentence Encoder is a sentence transformer that converts text into high-dimensional vectors. This sentence transformer assists with text classification, semantic similarity, clustering, and other natural language processing issues.

Sentence BERT: The Sentence-BERT (SBERT) network is developed from the BERT network utilizing Siamese and triplet networks. Because of this, it is possible to generate semantically relevant sentence embedding's.

MPNet: New pre-training approach, MPNet, inherits the benefits of BERT and XLNet, and avoids their drawbacks. With the use of permuted language modelling and the use of auxiliary position information as input, MPNet is able to make the model perceive the entire sentence, hence lowering the position discrepancy. This is accomplished by capitalizing on the interdependence that exists among the predicted tokens.

MiniLM: all-MiniLM-L6-v2. An illustration of a paradigm for transforming sentences is as follows: As a result of the fact that it maps phrases and paragraphs to a 384-dimensional dense vector space, it may be utilized for activities such as clustering and semantic search.

LaBSE: LaBSE is a BERT embedding model for several languages. Sentence embedding models for more than 100 languages may be generated in one model. Similar embedding's are generated by training the model on pairs of translated multilingual sentences.

IV. MODEL ARCHITECTURE

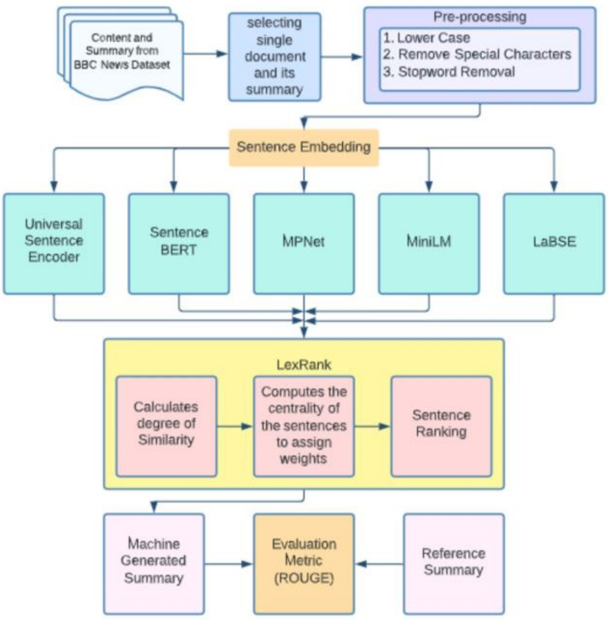

Fig – 1: Methodology

Initially, we extracted the content and summary from the BBC News dataset and then we preprocessed by changing them into lower case and by removing stop words & special characters / whitespaces. Then embedding's are drawn from the content which are given as input to the modals to generate summaries. The next step that we took was to apply the Universal Sentence Encoder (USE), Sentence BERT, MPNet, MiniLM, and LaBSE sentence embedding's to the LexRank algorithm. After that, we once more compared the results of our generated summary with the results that a human had generated using the ROUGE metric.

Dataset:

This dataset was created for the purpose of extractive text summarization, and the News Stories folder contains BBC news articles from 2004 to 2005. There is a folder labelled "Summaries" that contains a synopsis or summary of each individual article. The articles' titles appear in the very first line of the body of their respective texts.

	Number of files	Avg Number of Sentence in articles	Avg Number of Sentence in Summary
business	510	16	2
entertainment	386	16	2
politics	417	21	2
sport	511	17	3
tech	401	24	2
Total	2225	19	2

Fig – 2: Dataset Analysis

There are 2225 articles and their respective extractive summaries. In fig 3 we can see that average number of sentences in articles are 19 and average number of sentences in summary are 2.

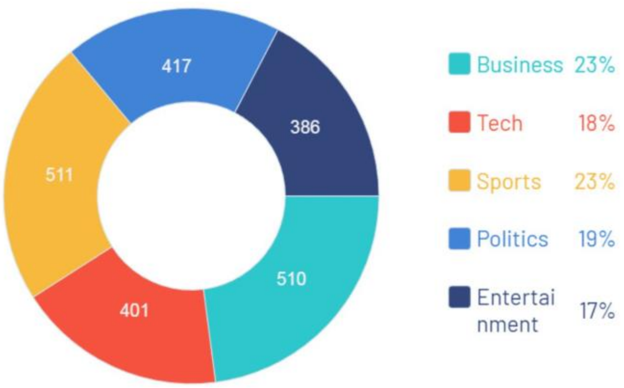

Fig – 3: Dataset Distribution

Evaluation Metrics:

The Evaluation metric used to evaluate our machine generated summary is ROUGE score.

ROUGE –
ROUGE is an abbreviation that refers for "recall-oriented understudy for gisting evaluation." It provides measures for automatically evaluating the quality of a summary by comparing it to other (ideal) summaries generated by people and it does this by comparing it to other summaries. Individuals were responsible for compiling these summaries. The summaries that were hand-written by humans serve as the standard against which the computer-generated summaries are evaluated. The metrics include word sequences, n-grams, and word pairs; they measure the total number of overlapping units.

ROUGE-N (N-gram Co-Occurrence Statistics):
The ROUGE-N algorithm, which is an n-gram recall between the two summaries, uses both a candidate summary and a collection of reference summaries. The following is how ROUGE-N is calculated:

$$ROUGE - N = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

ROUGE-L (Longest Common Subsequence):
When two sequences X and Y are analyzed together, the common subsequence that has the greatest length is referred to as the longest common subsequence (LCS) of the two sequences. During the process of building an N-best translation lexicon from parallel material, LCS was used to the task of detecting cognate candidates.

V. RESULTS

We performed a comparative study on six different models and the results are observed on 2225 articles and their respective extractive summaries. Rouge-1 Scores of five sentence embedding's are shown in below figures.

	Precision	Recall	F1-Score
business	0.842582	0.479898	0.567261
entertainment	0.789582	0.457575	0.540831
politics	0.799498	0.427306	0.512201
sport	0.777320	0.531011	0.594360
tech	0.831068	0.408156	0.502377
Avg	0.808010	0.460789	0.543406

Fig – 4: Rouge-1 Scores for Universal Sentence Encoder

In the above figure, Universal Sentence Encoder has performed well for sports domain and business domain. Rouge scores of all the five sentence embedding's are being calculated. Domain wise rouge scores are only shown for the top two sentence transformers which have performed well on the dataset.

	Precision	Recall	F1-Score
business	0.833765	0.475012	0.562407
entertainment	0.805610	0.473484	0.560209
politics	0.843496	0.457846	0.550918
sport	0.768916	0.507328	0.583112
tech	0.851369	0.422309	0.522471
Avg	0.820631	0.467196	0.555823

Fig – 5: Rouge-1 Scores for LaBSE

The above figures tells the Rouge-1 scores of two Sentence transformers which have performed well on five domains which are present in our dataset. In the same way Rouge-2 and Rouge-L are being calculated.

Rouge-1 Scores

	Precision	Recall	F1 - Score
USE	0.808010	0.460780	0.543400
MiniLM	0.802230	0.444060	0.531520
SBERT	0.744800	0.440190	0.521760
MPNET	0.785950	0.448090	0.532310
LaBSE	0.820630	0.467190	0.555820

Fig – 6: Rouge-1 scores of five Sentence Embedding's

Comparison of rouge-1 scores of five sentence embedding's are being shown in fig 9. LaBSE and Universal Sentence Encoder performed well and got a good average.

Rouge-2 Scores

	Precision	Recall	F1 - Score
USE	0.717512	0.384493	0.444098
MiniLM	0.710975	0.367289	0.430290
SBERT	0.636974	0.353037	0.408581
MPNET	0.690777	0.368193	0.428438
LaBSE	0.735935	0.391248	0.455330

Fig – 7: Rouge-2 scores of five Sentence Embedding's

Comparison of rouge-2 scores of five sentence embedding's are being shown in fig 7. LaBSE and Universal Sentence Encoder performed well and got a good average.

Rouge-L Scores

	Precision	Recall	F1 - Score
USE	0.799363	0.406490	0.537674
MiniLM	0.792730	0.391956	0.525156
SBERT	0.731891	0.385843	0.512734
MPNET	0.775590	0.392966	0.525185
LaBSE	0.812303	0.413039	0.550080

Fig - 8: Rouge-L scores of five Sentence Embedding's

Among the five sentence embedding's LaBSE Sentence encoder has performed well and achieved an average F1-Score of 0.555823 in Rouge – 1 and an average F1-Score of 0.455330 in Rouge – 2 and an average F1-Score of 0.550080 in Rouge -L.

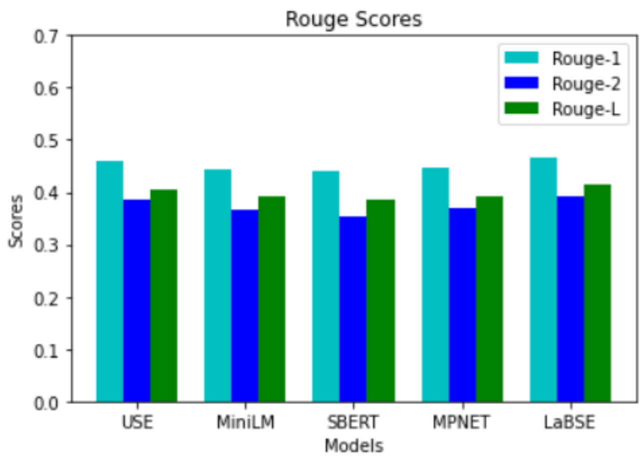

Fig 9: Graph Representation of all Rouge Scores for five Sentence Embedding's.

	Model Used	Precision	Recall	F1-Score
Paper-1 [11]	GloVe+TextRank	0.760000	0.570940	0.567600
Paper-2 [12]	KNN	0.605000	0.564000	0.575000
Proposed Model	**LaBSE+LexRank**	**0.820630**	**0.467190**	**0.555820**

Fig – 10: Comparative Study

In the above figure we have compared our model with other two model in two different papers Paper-1 [11] and Paper-2 [12].

VI. CONCLUSION

Within the scope of this study, we analyzed and evaluated five distinct sentence transformers for extractive summarization. We have evaluated the sentence transformers using Rouge – 1, Rouge – 2 and Rouge – L. As a result of our research, we have found that LaBSE performed better than other models when given the BBC News dataset, and the Lex Rank algorithm is the one that is used for ranking the sentences to get extractive summarization.

REFERENCES

[1] S. R. K. Harinatha, B. T. Tasara and N. N. Qomariyah, "Evaluating Extractive Summarization Techniques on News Articles," 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2021, pp. 88-94, doi: 10.1109/ISITIA52817.2021.9502230.

[2] H. Batra et al., "CoVShorts: News Summarization Application Based on Deep NLP Transformers for SARS-CoV-2," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596520.

[3] MPNet: Masked and Permuted Pre-training for Language Understanding, arXiv:2004.09297

[4] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv:1908.10084

[5] MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, arXiv:2002.10957

[6] Universal Sentence Encoder, arXiv:1803.11175

[7] Language-agnostic BERT Sentence Embedding, arXiv:2007.01852v2

[8] S M, Meena & M P, Ramkumar & R.E, Asmitha & Selvan, Emil. (2020). Text Summarization Using Text Frequency Ranking Sentence Prediction. 1-5. 10.1109/ICCCSP49186.2020.9315203

[9] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[10] D. S. J. A. J. A. Upansani A, Amin N, “Automatic summary generation using textrank based extractive tear summarization technique,” International Research Journal of Engineering and Technology, vol. 7, 2020

[11] Barman, Utpal, et al. "Graph Based Extractive News Articles Summarization Approach leveraging Static Word Embeddings." 2021 International Conference on Computational Performance Evaluation (ComPE). IEEE, 2021

[12] Krishnan, Devi, Preethi Bharathy, and Manju Venugopalan. "A supervised approach for extractive text summarization using minimal robust features." 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, 2019

[13] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of artificial intelligence research* 22 (2004): 457-479.

[14] Samuel, Avinash, and Dilip Kumar Sharma. "Modified lexrank for tweet summarization." *International Journal of Rough Sets and Data Analysis (IJRSDA)* 3.4 (2016): 79-90.

[15] Mallick, Chirantana, et al. "Graph-based text summarization using modified TextRank." *Soft computing in data analytics*. Springer, Singapore, 2019. 137-146.