Summa: A Text Summarizer using LSTM based Encoder - Decoder Architecture with Attention Mechanism

Digvijay S. Patil
UG Student
Department of Information Technology
NBNSSOE, Pune

Rahul M. Samant
Head of Department
Department of Information Technology
NBNSSOE, Pune

Abhin A. Shetty
UG Student
Department of Information Technology
NBNSSOE, Pune

Danish S. Shaikh
UG Student
Department of Information Technology
NBNSSOE, Pune

Shrinath G. Gutte
UG Student
Department of Information Technology
NBNSSOE, Pune

ABSTRACT
Due to the exponential growth of online textual data, and rapid expansion of internet usage, tasks such as document management, text classification, and information retrieval have been significantly posing a challenge due to its huge size. A key tool for addressing this issue is Automatic Text Summarization (ATS). It is one of the most challenging tasks since there is no precise computational method for evaluating accuracy of a summary.[1] The main function of ATS is to automatically create a concise and apprehensive summary by extracting the main ideas from the source text. Models for abstractive summarization based on deep learning (DL) have recently been created to better balance and improve these two elements. The field of DL-based text summarization currently lacks a thorough literature review, nevertheless. This work offers researchers a thorough analysis of DL-based text summarization in order to close this gap. In this paper, many prevalent frameworks for text summarization are listed. We also propose Summa - an Abstractive text summarizer which uses Long Short Term Memory(LSTM) based Sequence to Sequence Model. Accuracy matrix or F1 score are not recommended, traditional evaluation measures for text summarisation are not that much relevant so, a novel measure, Bi-Lingual Evaluation Understudy(BLEU)[2] measure is used. The study reported the BLEU score of 0.71, with attention mechanism and 0.3, without Attention Mechanism, which is considered a good score.

General Terms
NLP, Text summarizer, Deep Learning.

Keywords
Extractive, Abstractive, Long Short Term Memory(Stacked LSTM), Attention, BLEU.

1. INTRODUCTION
Cloud resources, including websites, blogs, news, user communications, and social networking sites, have accumulated enormous amounts of textual data in the digital age and are continuing to grow rapidly every day. Rich textual data is also present in a variety of publications, including books, novels, legal records, scientific papers, biomedical documents, and other archives. Information overload is thus getting more and more problematic and a good opportunity for developing better insights about the information. Text summaries are a method of choosing key ideas from an article or document that a computer program can condense. As the issue of data overload got worse, more people became interested in text-capture systems. Since it takes a lot of time and labor to manually summarize a lengthy document.

In general, there are two ways to summarize a text document utilizing extractive and abstractive approaches.[3] Extractive summaries are focused on taking the key ideas from the original text and constricting them into a manageable form. The significance of sentences is determined, on the basis of the sentence's semantic properties.

Text mining requires removal of stop words, high frequency words from the text corpus, in contrast, abstractive summaries are focused on the fact that the summary is not produced by choosing specific sentences from the source text passage. Instead, using a vocabulary set different from that of the original text, they generate a paraphrase of the essential points of the provided text. To sum up, this is quite comparable to what we humans do.

3. RELATED WORK
Du et. al (2020) **[4]** presented an innovative automatic summarizing model for news articles that were based on fuzzy logic, multi-features, and genetic algorithms(GA) . This model used word features. Each word was given a score, and the extracted words with scores over the predetermined threshold are referred to as keywords. Spatial and temporal features can sometimes also be extracted directly as keywords as the text corpus belongs to News contents. The second component was sentence features, a linear combination of these features which demonstrates the importance of each sentence. The Genetic Algorithm (GA) then assigns a specific weight to each feature individually. The fuzzy logic system is used to determine the final score and accomplish automatic text summarization. The proposed model's results were compared to those of existing techniques, such as MsWord, System19, System21, System 31, SDS-NNGA, GCD, SOM, and Ranking SVM, using the ROUGE assessment method and the DUC2002 dataset. The experimental results demonstrate that the new technique performs better than the traditional approaches.

Shahabi et. al (2022) [5] In this study, The author has covered the approaches that have been put forth so far to address automatic summarization, which involves summarizing both single-text and multi-text, with a focus on experimental approaches and text extraction methods. The Author described an extraction method based on various techniques for detecting sentence similarity and a meta-heuristic optimization algorithm tweaked and refined for quicker convergence. Improvements are made to this technique based on density detection using the probability distribution function in order to avoid being placed in local optimization and to try to search the response space more thoroughly. The experimental findings from the use of the algorithm demonstrate that the effectiveness of the suggested method is successfully improved in terms of efficiency on criteria like ROUGE and accuracy.

Afzal et. al (2020) [6] This research introduces a novel framework called Biomed-Summarizer that offers quality-aware Patient/Problem, Intervention, Comparison and Outcome (PICO)-based clever and context-enabled summarizing of biomedical material. The Biomed-Summarizer combines a clinical context-aware model and a prognosis quality recognition algorithm to locate text sequences in a biomedical article's body for usage in the final summary. First, in order to identify high-quality studies and exclude inferior ones, the researchers created a deep neural network binary classifier. Researchers created a bidirectional long-short-term memory recurrent neural network as a second phase. Finally, the researchers calculated the similarity between query and pico text sequences. This clinical context–aware multiclass classifier outperformed traditional machine-learning algorithms, including support vector machine, gradient boosted tree, etc.

Li et al. (2020) [7] In this research, a Double Attention Pointer (DAPT) Network-based encoder-decoder model is proposed. In DAPT, the self-attention mechanism collects important information from the encoder, the pointer network and soft attention give more coherent core content, and the two together provide precise and coherent summaries. The improved coverage approach results in summaries that are of higher quality and also handle the repetition problem. Scheduled sampling and reinforcement learning (RL) is used at the same time to create novel training strategies for the model. The CNN/Daily Mail Dataset and the LCSTS dataset have both been used to test the suggested model, and the results reveal that it performs as well as many state-of-the-art models. The outcomes of the provided model show that the recommended strategy enhances summary performance while reducing repetition. The major concepts of the book are therefore included in the summary created by this method. Information selection is controlled by an integrated gate mechanism. Truncation parameters depending on the present coverage strategy are recommended to keep this method from interfering with the production of additional targets.

Cheng et. al (2021) [8] The generative summarization technique is examined in this study, and a semantic similarity model is suggested. The model's coding and decoding section use an LSTM network. By using the encoder and decoder, respectively, the text semantic vector and abstract semantic vector are obtained. In order to decrease the loss function as much as feasible and increase the semantic similarity, the semantic similarity is calculated using the Jaccard similarity measure. The model based on semantic similarity has superior quality and readability when compared to the experimental findings.

Abdeljaber et. al (2020) [9] The XAI-RL model has two stages. The first stage is creating the summary text. The encoder (module A) uses the AL-BERT component to obtain the encoded representation Ea of the original text set. The coherence measurement module (module B) obtains auxiliary information H and extracts the set of critical sentences Input. Next, the decoder (module C) decodes Input and H, and afterword search, the initial output for the content of the critical sentences is generated abstract text. The second stage is sentence coherence reinforcement.Multiple sets of experiments have shown that the evaluation accuracy of the XAI-RL model including coherence measure and coherence augmentation is superior to the other available approaches. Future work on this paper will be to enhance the efficacy of self-attention weights on coherence measurement. To enhance sentence coherence for the next-generation models, new measuring techniques can be built, taking into account the coherence aspects from many angles, such as semantic linkage, grammatical regularity, and coreference disambiguation.

Kasimahanthi et. al(2020) [10] The sequence-to-sequence model is the model used for text summarization in this paper to summarize abstractive texts. The idea behind this is to use two LSTMs to predict the following state sequence from the previous sequence that will use a special token. Algorithm used in this model is the LSTM, a form of recurrent neural network model. It is generally used when a model's input and output can have different lengths. This model remembers past data easily by resolving the Vanishing Gradient problem. In this text summarization model, researchers used ROUGE as the validation metric. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. Machine translation and automatic text summarization are both evaluated using this method. The end result is a quick and effective text summarizing model.

Tang et. al (2021) [11] In this paper, a new training framework 'GOLD-FACTUAL' is proposed for text summarization that addresses the hallucination problem in abstractive summarization. The framework is built upon GOLD, an existing offline reinforcement learning training framework that is originally designed to alleviate long-standing problems in conventional maximum likelihood estimation(MLE) training. Also, the GOLD-FACTUAL framework can improve factual consistency in generated summaries and even outperform human reference summaries.. Rather than the standard training procedure that uses (d, y) pairs from (D, Y), authors leverage a trained summarization model and keep fine-tuning it using pairs of document d and SOTA-model generated summaries y'. In addition, researchers provide token-level factual consistency labels l for each token in y' to encourage the model to not generate factually inconsistent words. In this experiment, the pre-trained Facebook models are fine-tuned for 6 epochs on the proposed training set by the GOLD-FACTUAL with non-factual penalty λ =0.1, 0.5,1,2,5. Then the models are evaluated on the test set, and the ROUGE scores and factual consistency accuracy. In the remaining part, then analyzed the factual consistency of the text summarizations of both humans and models, and investigated the trade-off between factual consistency and informativeness by sweeping the non-factual penalty λ.

Fitrianah et. al (2022) [12] This research compares the performance of LSTM and GRU and focuses mostly on extractive text summarization using both of these techniques. The practice of gathering datasets of scientific journal articles from online free libraries constitutes the initial step. The information about the abstract section, title, author, and publication year will then be extracted from the data. Following the pre-processing stage, the dataset will move on to the feature engineering stage in order to extract semantic information from the text. The final 75 features, all of which are binary values, will be created when the features have been extracted. The researcher employed the latent semantic analysis method to identify the key phrase in the text while labeling the data. The researchers compared two algorithms-LSTM and GRU during the model training phase. The findings demonstrate that the novel method we developed, which combines one-hot encoding for each sentence with feature engineering to extract semantic meaning, can provide extractive summaries that are equally as good as those produced by the LSA algorithm. Unlike LSA, which uses a machine learning approach, our method may be altered by altering the way the data is labeled using a different strategy, such as clustering or human methods. Due to its more intricate Gate Architecture, LSTM performed a little bit better in terms of accuracy, although GRU excelled in training speed due to each cell having only two gates.

Koh et. al. (2021) **[13]** In order to determine how significant a sentence is in a discussion, the author of this research devised a bug report summarization approach that incorporates two different values, text ranking score, and credibility score. The suggested strategy surpasses BugSum, which just takes sentence credibility scores into account, according to experimental results. Future research on this topic may take into account several variables in addition to text ranking and plausibility to assess sentence relevance. One can also look into the best deep-learning methods for maximizing summary quality.

Ghodratnama et. al. (2020) **[14]** The authors of this work have put forth ExDoS, a unique intelligent strategy that simultaneously reaps the rewards of supervised and unsupervised algorithms. ExDoS employs dynamic local feature weighting to repeatedly reduce the classifier's error rate for each cluster. Additionally, this method describes the contribution of features to each class's ability to be distinguished, which is a difficult problem in the process of summarization. ExDoS may therefore measure the significance of each character in the summarization process in addition to summarizing text. ExDoS is a technique that can assess the contribution of features to the discrimination of each class separately through the summarizing process by creating distinct feature spaces. It can quantify the relative value of various aspects in addition to summarizing the documents with the aid of local feature weighting.The goal of this feature-weighting method is to close the gap between same-label and different-label samples while advancing the former. The simplest methods for clustering (k-means) and classifying (KNN) are employed to support efficiency because the algorithm operates in an iterative manner using gradient descent. They have assessed their model using the benchmark datasets, the DUC2002 and CNN/DailyMail, both automatically (ROUGE factor) and empirically (human analysis). The algorithm outperformed the majority of cutting-edge techniques in terms of performance and efficiency.

Jang et. al. (2021) **[15]** The purpose of this work is to improve the quality of the summary statement by proposing an RL-based incentive mechanism for text summarization. ROUGE-SIM and ROUGE-WMD are two different modified versions of the The authors proposed the ROUGE function. ROUGE-SIM supports semantically linked words as opposed to ROUGE-L. The Word Mover's Distance (WMD) method was used to determine how semantically related articles and summary text were. Their algorithm is based on a simple sequence-to-sequence LSTM model with attention, a pointer mechanism for handling words that are out of vocabulary (OOV), and intra-decoder attention for handling repeated words. Their model beat models based on sequence-to-sequence learning, Transformer-based pre-learning models, and other models based on reinforcement learning. The results of the grammatical and abstract assessments were also improved. Compared to human-written summaries, there were less new representations in Abstractness Analysis, but overall, pointer-based models provided the most recent representations.

Akhmetov et. al. (2021) **[16]** In this paper, a greedy extractive summarization technique is used to summarize scientific papers from the arXive and PubMed datasets. They combined this method with Variable Neighborhood Search (VNS) to determine the highest level of quality for Extractive Text Summarization based on ROUGE scores. The algorithm relies on a minimum document frequency parameter tuning for TFIDF vectorization coupled with initially choosing for the summary the sentences from the text containing the greatest number of words with the highest TFIDF values.In addition, they have employed a greedy algorithm for the same objective. They employ an easy, time-tested method called extractive summarization. Both strategies performed similarly, but the Greedy strategy takes much less time to finish.

Ding et. al. (2020) **[17]** The traditional sequence mapping and word feature representation, as well as the model's semantic comprehension and summary quality, are optimized in this article. The proposed technique is validated using the LCSTS and SOGOU datasets. According to the testing findings, the suggested strategy can boost the ROUGE evaluation system's performance by 10–13 percentage points in comparison to other stated algorithms. They noticed that the generating impact is better and that the semantic interpretation of the text summaries is more precise.An improved semantic model based on dual-encoder is suggested in order to address the issues of inadequate use of semantic information, insufficient summary precision, and semantics loss in the current generated text summary method. Dual-encoder can provide richer semantic information for sequence-to-sequence architecture. The Gain-Benefit gate and empirical distribution are created for decoding, and the augmented attention architecture with dual-channel semantics is optimized. Additionally, the word embedding method now combines position embedding and word embedding, and it now incorporates part of speech, key score, and TF-IDF.

Shirwadkar et. al. (2020) **[18]** proposed a method for extractive text summarizing is created and used in the proposed work to summarize a single document. It picks out key passages from the text using a combination of fuzzy logic and restricted Boltzmann machine while still maintaining the summary's relevance and losslessness. English-language text documents were utilized to create the summaries. To create meaningful phrases, a variety of sentence and word-level properties are applied. The final summary of the document is created by combining the two summaries from fuzzy logic and restricted Boltzmann machine and a series of processes. The findings demonstrate that the intended strategy successfully addresses the issue of text overload by producing a useful summary. Suggested technique produces short and precise summaries without any irrelevant text. The connection of the sentences has been increased by the use of features like Sentence-Centroid similarity and thematic words. The suggested strategy yields an average of 88% precision, 80% recall, and 84% F measure. In comparison to the already used RBM approach, the findings obtained utilizing the suggested method provide better evaluation parameters.

3. DATASET

Amazon Fine Food Reviews dataset consists of ~500,000 reviews of fine foods from amazon. The data span a period of more than 10 years, up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

It is worth noting that when it comes to recurrent neural networks, accuracy and loss are not always the preferred evaluation metrics, particularly for deep learning models that generate natural language Output. In such cases, the BLEU (Bilingual Evaluation Understudy) score is often considered a better metric for assessing the quality of model-generated Natural Language Output.

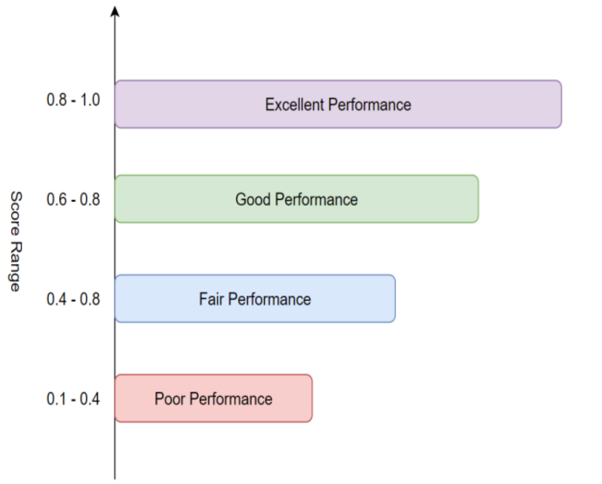


Fig.1. BLEU Metrics

4. PROPOSED SYSTEM

Sequence-to-sequence (seq2seq) models can help in building an automatic text summarization (ATS) model. When given an input, the encoder-decoder seq2seq model first generates an encoded representation of the model, which is then passed to the decoder to generate the desired output. In this case, the input and output vectors need not be fixed in size. Therefore, variable size input text can be passed to the model and it will give a summary of the text as output which is also dependent on the size of input text and hence It is also variable sized.

4.1 Long-short term memory (LSTM)

Long-short term memory network is a special implementation of recurrent neural network. They are designed to be capable of learning long-term dependencies in the data and consider them while performing its task. Instead of neurons, LSTM networks have memory blocks (Cells) connected through layers. Each cell contains the information about its last state and output. A cell has three types of gates, which help it to decide how to use information from previous state and output as well as current input to update cell state and produce output.

i.Forget Gate: decides what information to throw away from the block.
ii.Input Gate: decides which values from the input to update the memory state.
iii.Output Gate: decides what to output based on input and the memory of the block.

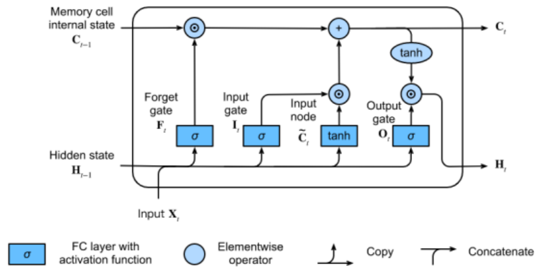


Fig.2. Architecture of LSTM [19]

In the figure, we can see the structure of a LSTM cell. C_t, X_t and h_t denote cell state, input and output at time step t respectively. LSTMs are trained using backpropagation through time algorithm, which is a standard technique for training recurrent neural networks. LSTMs are widely successful in time-series prediction, Natural language processing, and machine translation etc. tasks.

4.2 Model Architecture:

The idea behind the design of this model is to enable it to process input where we do not constrain the length. A stacked LSTM will be used as an encoder, and a single LSTM as a decoder. The output vector generated by the encoder and the input vector given to the decoder will posses a fixed size. However, they need not be equal. The output generated by the encoder can either be given as a whole chunk or can be connected to the hidden units of the decoder unit at every time step.

4.2.1 Encoder

The input length that the encoder accepts is equal to the maximum estimated text length. This is then given to an Embedding Layer of dimension (total number of words captured in the text vocabulary) x (number of nodes in an embedding layer). This is followed by three LSTM networks wherein each layer returns the LSTM output, as well as the hidden and cell states observed at the previous time steps.

4.2.2 Decoder

In the decoder, an embedding layer is defined followed by an LSTM network. The initial state of the LSTM network is the last hidden cell state taken from the encoder. The output of the LSTM is given to a Dense layer wrapped in a TimeDistributed layer with an attached softmax activation function.Altogether, the model accepts encoder (text) and decoder (summary) as input and it outputs the summary.

4.2.3 Shortcomings of LSTM

The encoder converts the entire input sequence into a fixed length vector and then the decoder predicts the output sequence. This works only for short sequences since the decoder is looking at the entire input sequence for the prediction.
Here comes the problem with long sequences. It is difficult for the encoder to memorize long sequences into a fixed length vector as document management, text classification, and information retrieval have been significantly hampered. A key tool for addressing this issue is Attention Mechanism.

4.2.4 Attention Mechanism

Attention is the idea of freeing the encoder-decoder architecture from the fixed-length internal representation. This is achieved by keeping the intermediate outputs from the encoder LSTM from each step of the input sequence and training the model to learn to pay selective attention to these inputs and relate them to items in the output sequence. Put another way, each item in the output sequence is conditional on selective items in the input sequence.

This increases the computational burden of the model, but results in a more targeted and better-performing model. In addition, the model is also able to show how attention is paid to the input sequence when predicting the output sequence. This can help in understanding and diagnosing exactly what the model is considering and to what degree for specific input-output pairs.

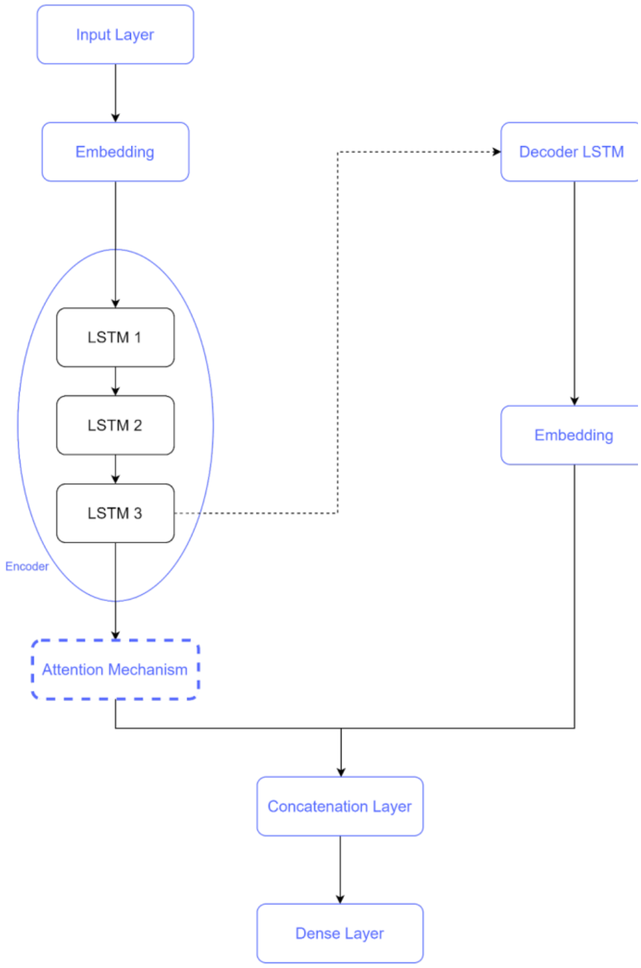


Fig.3. Stacked LSTM based text summarizer

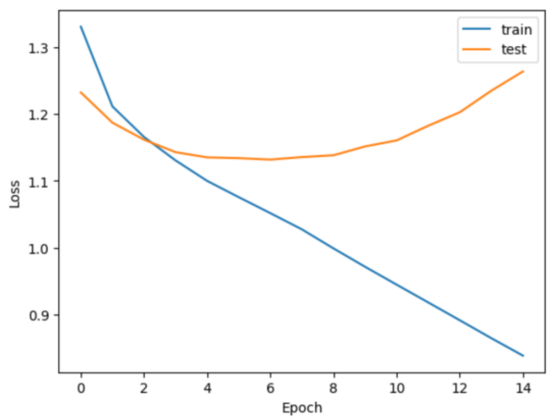
5. RESULTS AND DISCUSSION

In terms of hyperparameters, both models were trained with a learning rate of 0.001 and a batch size of 64, undergoing 15 epochs. The optimizer used for training was RMSprop, and the loss function employed was sparse categorical cross-entropy.

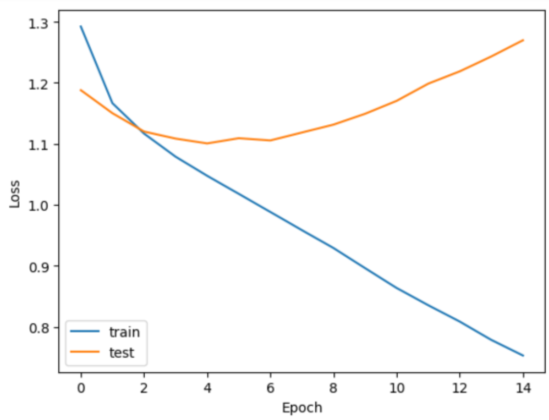
Two models, namely the Stacked LSTM and the Stacked LSTM with Attention Mechanism, were trained and evaluated for the given task. The Stacked LSTM model attained an accuracy of 85% on the test set, whereas the Stacked LSTM with Attention Mechanism achieved a slightly higher accuracy of 88%.

Both models utilized the RMSprop optimizer, which is commonly chosen for training recurrent neural networks like LSTMs due to its capability to adjust the learning rate for each weight based on the gradient history. The sparse categorical cross-entropy loss function was employed to measure the discrepancy between the predicted and actual labels. A lower cross-entropy loss implies better predictions made by the model. Although the Stacked LSTM with Attention Mechanism achieved superior performance in terms of both accuracy and loss.

When the BLEU score was calculated for both models, the Stacked LSTM achieved an average BLEU score of 0.3, indicating moderate performance. On the other hand, the Stacked LSTM with Attention Mechanism achieved a considerably higher BLEU score of 0.71, suggesting a significant improvement in the quality of the generated outputs.

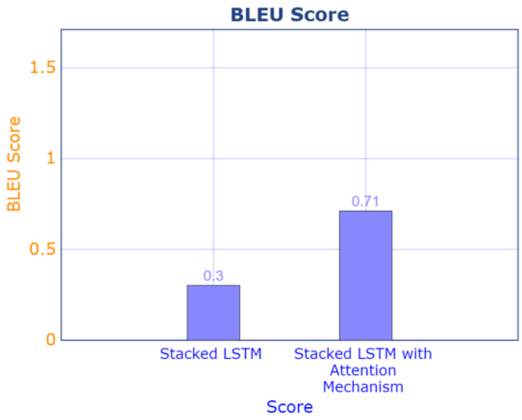


Stacked LSTM



Stacked LSTM with Attention Mechanism

Fig. 4. Network performance curves for proposed system



6. CONCLUSION

The technique for automated text summarization has steadily evolved from extractive to abstractive since it was first presented in the late 1950s. Abstractive summarization based on deep neural networks has advanced quickly in recent years as deep learning technology has grown in the NLP sector. In addition to being widely employed in the sectors of finance, journalism, and media, automatic text summarization is also important for content review, information retrieval, and public opinion analysis.

In this paper, a thorough analysis of the models for extractive and abstractive text summarization that are currently on the market was given. A few datasets and evaluation metrics that are frequently applied in the field of text summarization have also been discussed. In the last section, An LSTM based encoder and decoder model with attention mechanism to design an automatic text summarizer was proposed. The attention mechanism also emphasizes the important word of the sequence and copies the same in the output summary. Since The proposed method is based on sequence to sequence deep learning model it provides better results than several other approaches in the literature. When given a text article as input the model is producing a meaningful summary.

7. ACKNOWLEDGEMENT

We would like to thank Pratik S. Patil, Mtech, Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur for his valuable help in this study.

8. REFERENCES

[1] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation* (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318

[2] BLEU. In Wikipedia. https://en.wikipedia.org/wiki/BLEU

[3] From Extractive to Abstractive Summarization: A Journey, by Parth Mehta, Prasenjeet Mujumde

[4] Yan Du, Hua Huo* "News Text Summarization Based on Multi-Feature and Fuzzy Logic", 2020, IEEE

[5] Amir Shahab Shahabi, Mohammad Reza Kangavari, Amir Masoud Rahmani. " A Method for Multi-text Summarization Based on Multi-objective Optimization use Imperialist Competitive Algorithm" March 2022

[6] Muhammad Afzal, Fakhare Alam, Khalid Mahmood Malik, Ghaus M Malik, "Clinical Context–Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation" 2020 JMIR

[7] Zhixin Li, Zhi Peng, Suqin Tang, Canlong Zhang, Huifang Ma"Text Summarization Method Based on Double Attention Pointer Network" 2022

[8] Jingjing Cheng, Fucheng Yu. "Text summarization Generation Based On semantic similarity" 2021

[9] Hikmat A. M. Abdeljaber, Sultan Ahmad, Abdullah Alharbi, and Sudhir Kumar "XAI-Based Reinforcement Learning Approach for Text Summarization of Social IoT-Based Content" 2022

[10] Kasimahanthi Divya, Kambala Sneha, Bassetti Sowmya, G Sankara Rao, "Text Summarization using Deep Learning"2020, IRJET(ISSN: 2395-0056)

[11] Liyan Tang, Zifan Xu, "GOLD-FACTUAL: Learning to Generate Faithful Summaries from models Generations" 2021

[12] Devi Fitrianah, Raihan Nugroho Jauhari, "Extractive text summarization for scientific journal articles using long short-term memory and gated recurrent units" 2022 (ISSN: 2302-9285)

[13] Youngji Koh, Sungwon Kang, Seonah Lee. "Bug Report Summarization using Believability Score and Text Ranking" 2021

[14] SAMIRA GHODRATNAMA, AMIN BEHESHTI, MEHRDAD ZAKERSHAHRAK, AND FARIBORZ SOBHANMANESH "Extractive Document Summarization Based on Dynamic Feature Space Mapping" 2020

[15] HEEWON JANG AND WOOJU KIM "Reinforced Abstractive Text Summarization With Semantic Added Reward" 2021

[16] ISKANDER AKHMETOV, ALEXANDER GELBUKH, AND RUSTAM MUSSABAYEV "Greedy Optimization Method for Extractive Summarization of Scientific Articles" 2021

[17] JIANLI DING, YANG LI, HUIYU NI, AND ZHENGQUAN YANG "Generative Text Summary Based on Enhanced Semantic Attention and Gain-Benefit Gate" 2020

[18] Nikhil S. Shirwandkar, Samidha Kulkarni "Extractive Text Summarization using Deep Learning" 2020

[19] Architecture of LSTM [Image]. Retrieved from https://raw.githubusercontent.com/d2l-ai/d2l-en/master/img/ls