ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 11, Issue 2, 2023

Leveraging NLP and Machine Learning for Effective Text Summarization

Mr. K.Appa Rao
Assistant Professor, Department of CSE,
Malla Reddy College of Engineering for Women.,
Maisammaguda., Medchal., TS, India

Abstract—

With so much information at our fingertips, the ability to distil the most relevant details from massive texts has become more important. Numerous publications include in-depth information about Web pages like news sites, weblogs, and consumer review forums. Several methods are presented in this review study for synthesizing brief summaries of lengthy texts. Methods utilized so far for text summarizing have been researched and analyzed in a number of published studies. Abstractive summaries (ABS) and extractive summaries (EXT) are the most common outputs of the techniques discussed in this work. Methods for summarizing information depending on a user's queries are also examined. Most of the text is devoted to a discussion of structured and semantic bases. The summaries generated by these models were put to the test on several datasets such the CNN corpus, DUC2000, single and multiple text documents, etc. We have investigated these strategies together with current and potential developments, successes, and applications in text summarization and other areas.

I. INTRODUCTION

These days, everything revolves on data and technology. We might think of data as our intangible thoughts and creative ideas. At the same time that we create data, we also use it. Everything is included we constantly generate or consume information in the course of our daily routines. When we drive, for instance, there are a variety of metrics at play, including vehicle speed, gas mileage, travelled distance, and so on. Data has been important to us since the 20th century, but we now draw more conclusions from it than ever before. They are kept in electronic and wireless storage systems, which we use to retrieve them as needed.

The proliferation of online resources has resulted in a deluge of information. It's safe to say that you can find just about anything on the Internet. The internet is a great resource for learning about almost any topic imaginable: current events, popular culture, science, history, politics, medicine, the environment, geography, and even meteorology and climatology. This information might be textual, quantitative, mathematical, or graphical. The greater number of characters in text data makes it more challenging to comprehend. This massive quantity of data necessitates a method for extracting just the most relevant aspects of the information we need. One approach is to summarize the text. Research and development into the art of summarizing texts dates back decades. To do this, several alternative models have been developed and evaluated on a variety of datasets. Various comparison scores are used to evaluate them. It is possible to do EXT or ABS summary on a text, as well as single document or multidocument, query-based or general summarizing. The EXT text summarizing method use the source texts very own phrases to create summaries. The ABS approach is broader and emphasizes the document's major ideas. Similar to how single-document summarizing methods summarize a single document's text, multi-document summarization methods do the same for a set of related documents. Moreover, query-based text summarization is becoming important. Generic summaries are often ABS that concentrates on the overall region of the text input, while query-focused summarization models provide summaries of the text depending on a particular area as indicated by the user's question. Use of text summary has spread widely across disciplines like science, medicine, law, engineering, etc. One area of study that has proven valuable to patients is the creation of summaries of prescriptions written by doctors. In a similar vein, lengthy news stories have been condensed so that readers may absorb a great deal of knowledge quickly and easily [1].

This study reviews the literature on text summarizing techniques over the last five years. Machine Learning was shown to be the most popular technique. Fuzzy logic, sequence to sequence modelling, sequence to sequence learning, neural networks, and reinforcement learning. Similar optimization techniques have been used to the suggested goal function to further the cause of text summarization. Here we see that several approaches were evaluated on the same dataset, with varying degrees of success. Some studies have shown that when researchers utilize a combination of methodologies, the resulting summaries are even more reliable than when just one is used. We notice that python libraries like sickest learn, notch, spicy, and festal have been employed when NLP processing has been used to summarize text content.

ISSN 2321-2152
www.ijmece.com
Vol 11, Issue 2, 2023

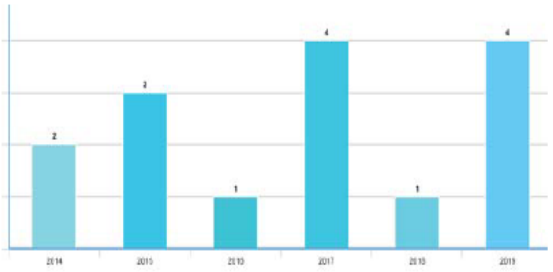

Fig. 1. Distribution of papers studied over the years

II. RELATED WORK

EXT summaries were created using a sentence extraction technology developed by Massimo Mauro and colleagues. Relevance is determined and evaluated for each phrase in this approach. Accordingly. Next, similar sentences were grouped together to find the most illuminating sentences, and these were chosen using the sentence scores [2]. In order to generate EXT summaries, Sarda A.T. et al. suggest using NNs and Rhetorical Structure Theory (RST). The summary statement is constructed by comparing and combining features. First, the NN must be trained to learn how to choose which phrases to include in the summary [3]. After that, sentences are picked out of the paper and reviewed to see whether they match the summary. Once these phrases are located, they are input into the rhetorical framework, where language linkages are differentiated to provide a more accurate summary. Experiments to produce EXT summaries of the papers in CNNCorpus have been conducted by Gabriel Silva et al. After pictures, movies, and other non-textual elements have been tables were given feature vectors to be used in the scoring process. By using WEKA's CFS Subset Evaluator, Information Gain Evaluator, and SVM Attribute, the dimensionality of the feature vectors was decreased. Among the five classifiers evaluated on WEKA's platform, Naive Bayes was shown to be the most effective [4].

Taeho Jo has described a method for summarizing text that makes use of the KNN algorithm and takes into account the similarity between characteristics. As input, a paragraph is provided, and it is broken down into sentences. We use vectors to represent words. The inclusion of a sentence in the summary is then determined by its similarity score to a human-generated summary, which is based on the classification of each phrase as either summary or remaining [5]. In this work, we focus on features as a means to summarize text passages. Multiple fields, including medicine, law, and engineering, may benefit from this method [5]. Paragraphs are summarized using the KNN technique, which takes into consideration similarities between just a subset of attributes and generates a similarity score. A query-based method for EXT text summarization was described by Mash Afsharizadeh et al. It takes just the most relevant sentences from the source material and incorporates them into the summary. Eleven in the article, relevant phrases were identified using a combination of query-dependent features [6]. Each phrase is given a score based on the linear function of its feature values, and this score is then used to locate the valuable sentences in the text.

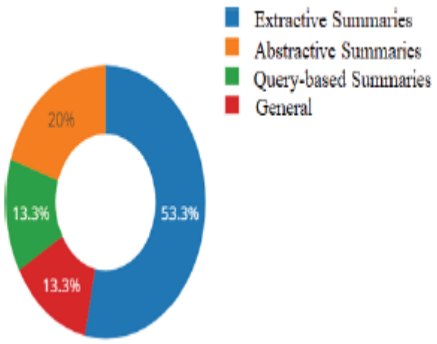

Fig 2. Types of Summaries Studied

The DUC 2007 corpus was utilized for both training and assessment. Average accuracy, average recall, and average F-score were all improved above prior approaches by using this one. Contrasted in the article [7].

TABLE I. SUMMARY OF THE PAPERS ANALYZED

Name of Author	Year of Publication	Methods Used	Dataset used	Remarks
Kamalanathan Kandasamy et. Al[9]	2014	URL analysis, NLP, and Supervised ML Techniques.	Tweets from users.	Combining three models gives more accuracy than single method used.
Mohamed Abdel Fattah [10]	2014	k—means clustering, differential evolutionary algorithm	data sets from DUC2001 and DUC2002	Results were compared with ROGUE 1 and ROGUE 2 scores. It showed better results than other text summarizers. Graph based algorithms for clustering could improve the result.
Mr. Sarda A.T .et al [3]	2015	NN, ML, RST	Input text documents	Due to inclusion of numerical data features, the output of this summarization method is better than other online text summarizers.
Gabriel Silva et al[4]	2015	STOM, CSF, Information Gain Evaluator and SVM attribute	news texts from CNN corpus	The results confirmed ML techniques improves the text summarization results.
K. Vimal Kumar, et al[15]	2015	Sentence ranking, Sentential semantic analysis and Sentence extraction	Hindi-text Single documents	This method can be applied for multi documents.
Rasim Alguliyev et al[12]	2016	Human Learning Optimization Algorithm	Takes a document as input.	This model tries to find balance between the topic coverage and sentence repetition in a summary, for generating concise summary.
Taeho Jo[5]	2017	k-nearest neighbor algorithm	Input paragraph	Similarity score is taken into account. Can be used in medicine, science and law.

ISSN 2321-2152
www.ijmece.com
Vol 11, Issue 2, 2023

Massimo Mauro et al.[2]	2017	sentence extraction, sentence scoring and sentence ordering using Python libraries such as scikit learn, nltk.	DUC2001 dataset	It was mostly useful for multi-text summaries when compared with ROUGE score
Aditya Jain, Divij Bhatia et. Al[14]	2017	Word vector embedding, NN	100 news articles from CNN news with their ABS summaries.	the proposed model outperformed other online text summarizers such as SPLITBRAIN, AutoSummarizer when the ROUGE scores were compared.
Nicholas Giamblanco et al[11]	2017	Newtonian method	SemEval from the University of Waikato	The method outperformed RAKE and TF-IDF scores.
Mahsa Afsharizadeh, et al[7]	2018	Query based EXT summarization using TF-IDF, fuzzy logic and LSA.	DUC 2007 corpus	A fluent 250-word summary is generated and compared to ROGUE score.
J.N. Madhuri et al[8]	2019	Sentence ranking method using Python 3.6 and NLTK[8]	Takes text documents as input	Those sentences whose rank is greater than 8 are included in the summary. The summarized text is then converted to audio form.
Milad Moradi et al[13]	2019	BERT model	Test on preliminary experiments of the development corpus	The size of summary would be good if it is 30% of the original text. This method worked best on biomedical texts.
Begum Mutluet al[6]	2019	ML, fuzzy network	Input text documents and their features.	The resulting rule sets provide more coverage and generalization

III. OVERVIEW OF MACHINE LEARNING METHODS FOR TEXT SUMMARIZATION

A number of academics have recently presented their machine learning (ML) algorithms for summarizing text. Text summaries in EXT format have been generated using a wide range of supervised ML models, including Naive Bayes, Random Forest, and SVD. The work of Kamalanathan Kandasamy et al. classification of Twitter spam using ML techniques. Nave Bayes was determined to be more accurate when testing the data sets, while SVM was also used. One hundred participants were assessed for this study. Only sixty were genuine, while forty were spam. We were able to accurately classify 98 users [9]. Many studies have shown that deep learning methods perform better than traditional methods for both EXT and ABS text summarization, and this trend is only growing. This kind of ML is called "deep learning." Various NN approaches have been used. Similarly, reinforcement learning, Convolution NN (CNN), RNN have also been employed to create text summaries [10]. There's also a research on sequence-to-sequence models for text summarization these days. These approaches are extension of ML. We discuss some of the articles that employ the above described methodologies to produce summaries of text. Nicholas Giamblanco ET colleagues developed a Newtonian method \sot produce key words. The author has utilized four processes for keyword and key phrase extraction. First the stop words are removed which is known as noise filtering, next each words is assigned a mass, then the relations between words is calculated which is also known as word attraction, and lastly a key phrase is generated [11]. To assess the correctness of this algorithm, it was tested against RAKE and TF-IDF scores. The data used \swaps from Seminal from the University of Waikato [12]. The author has also showed how Key-LUG has outperformed RAKE and TF-IDF. The criteria employed for this aim included recall, \precision and F1 score [13].

Aditya Jain ET. Al study introduced a binary strategy for EXT text summarization. The material is divided into sentences and assessed whether it's relevant or unrelated to the primary topic of the paper. A NN is then employed and checked whether the sentence is to be included or not. 100 news stories from CNN \snows along with their related ABS summaries have been utilized as data sets for this model. Sentences in the news \article and summary are examined and those with higher similarity score between them are picked for EXT summary [14]. The author has utilized vector embedding to represent each word in a text as 100 dimensional vectors. For assessing the performance of the proposed model, first 284 documents of the DUC 2002 dataset was employed. It was found that \she suggested model was more accurate other online text summarizers [15].

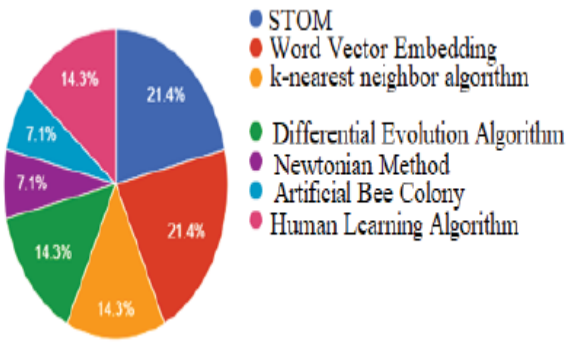

Fig 3. Different algorithms used

IV. CONCLUSION AND FUTURE WORK

As we've seen, text summarizing plays a crucial role in helping users save time and effort in a world where there's an abundance of data. Indeed, summary of texts is a crucial today's practical instrument. Many distinct algorithms and

approaches have been used for this same function. Different kinds of summaries may be obtained using one of these approaches alone or in combination. In order to determine which summaries are most effective, one may evaluate their accuracy scores. More often than not, the ROGUE score has been employed for this function. Similarly, TF IDF scores have also been utilized sometimes. In general, the summaries produced by these methods are not of the highest quality all the time. The initial version of the paper may not always need this information. Because of this, the debate over this topic refuses to go down. Have conducted a great deal of research on it already. There is no hard evidence that any summarization approach is preferable. Therefore, the models going ahead will some modifications to the principles we've discussed so far might lead to more accurate descriptions. For Generative ad hoc network applications and transfer learning are feasible examples. Therefore, they may provide a way to expand upon existing textual notions and enhance summary.

REFERENCES

[1] C. Ordonez, Y. Zhang, and S. L. Johnson, "Scalable machine learning computing a data summarization matrix with a parallel array DBMS," Diatribe. Parallel Databases, vol. 37, no. 3, pp. 329–350, 2019, doe: 10.1007/s10619-018-7229-1.

[2] M. Mauro, L. Canini, S. Benini, N. Adami, A. Signoroni, and R. Leonardi, "A freeWeb API for single and multi-document summarization," ACM Int. Conf. Proceeding Ser., vol. Part F1301, 2017, doi: 10.1145/3095713.3095738.

[3] A. T. Sarda and M. Kulkarni, "Text Summarization using Neural Networks and Rhetorical Structure Theory," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 49–52, 2015, doi: 10.17148/IJARCCE.2015.4612.

[4] G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Riss, and H. O. Cabral, "Automatic text document summarization based on machine learning," Docent 2015 - Proc. 2015 ACM Symp. Doc. Eng., pp. 191–194, 2015, doi: 10.1145/2682571.2797099.

[5] T. Jo, "K nearest neighbour for text summarization using feature similarity," Proc. - 2017 Int. Conf. Common. Control. Compute. Electron. Eng. ICCCCEE 2017, pp. 1–5, 2017, doi: 10.1109/ICCCCEE.2017.7866705.

[6] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Multi-document extractive text summarization: A comparative assessment on features," Knowledge-Based Syst., vol. 183, p. 104848, 2019, doi: 10.1016/j.knosys.2019.07.019.

[7] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th Int. Conf. Web Res. ICWR 2018, pp. 128–132,2018, doi: 10.1109/ICWR.2018.8387248.

[8] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf. Data Sci.Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040.

[9] K. Kandasamy and P. Koroth, "An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques," 2014 IEEE Students' Conf. Electr. Electron. Comput. Sci. SCEECS 2014, pp. 1–5, 2014, doi: 10.1109/SCEECS.2014.6804508.

[10] M. A. Fattah, "A hybrid machine learning model for multidocument summarization," Appl. Intell., vol. 40, no. 4, pp. 592–600, 2014, doi: 10.1007/s10489-013-0490-0.

[11] N. Giamblanco and P. Siddavaatam, "Keyword and Keyphrase Extraction using Newton's Law of Universal Gravitation," Can. Conf. Electr. Comput. Eng., pp. 1–4, 2017, doi: 10.1109/CCECE.2017.7946724.

[12] R. Alguliyev, R. Aliguliyev, and N. Isazade, "A sentence selection model and HLO algorithm for extractive text summarization," Appl. Inf. Commun. Technol. AICT 2016 - Conf. Proc., pp. 1–4, 2017, doi: 10.1109/ICAICT.2016.7991686.

[13] M. Moradi, G. Dorffner, and M. Samwald, "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," Comput. Methods Programs Biomed., vol. 184, p. 105117, 2020, doi: 10.1016/j.cmpb.2019.105117.

[14] A. Jain, D. Bhatia, and M. K. Thakur, "Extractive Text Summarization Using Word Vector Embedding," Proc. - 2017 Int. Conf. Mach. Learn. Data Sci. MLDS 2017, vol. 2018-Janua, pp. 51– 55, 2018, doi: 10.1109/MLDS.2017.12.

[15] J. K. Mandal, S. C. Satapathy, M. K. Sanyal, P. P. Sarkar, and A. Mukhopadhyay, "Information systems design and intelligent applications: Proceedings of second international conference India 2015, volume 1," Adv. Intell. Syst. Comput., vol. 339, pp. 301–310, 2015, doi: 10.1007/978-81-322-2250-7.