IJNRD.ORG ISSN : 2456-4184

International Research Journal IJNRD Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Big-BirdPegasus based Abstractive Multi-Document Summarization

P Spandana Valli [1], N.Usha Satwika , S.Vijay Swaroop , V.V.L.Prasad , B.Venkata Lakshmi

[1]Assistant Professor, Dept of Computer Science and Engineering, Anil Neerukonda Institutes of Technology and Sciences, Visakhapatnam, India.

Abstract:

We are introducing a new method of multi-document summarization[2] that utilizes our Big Bird Pegasus[3] transformers (BBPT) in conjunction with Maximal Marginal Relevance (MMR)[2][3][18]. Our approach generates candidate breviloquent sentences and then uses MMR[2][3] to select the most relevant and diverse sentences for inclusion in the arbitrary. In this research, we propose a method for multi-document summarization[2] using Big Bird Pegasus[3] along with MMR[2][3][18]. We import the dataset directly from the ski-learn library and split it into 70% for training and 30% for testing the model. Our proposed model achieved an accuracy of 80% when tested on the same dataset by extracting 30% of the information.

Keywords: Maximal Marginal Relevance, Big Bird Pegasus Transformers, Summary.

1. Introduction:

[1]The job of creating a concise synopsis from several papers on the same subject is known as multi-document summarization[2]. It entails selecting the most crucial details from each document and then fusing them into a single synopsis that sums up the main ideas and points. This job is crucial for a variety of uses, including business intelligence, scientific book reviews, and news arbitrary. conglomerate-document summarization can help users who need to rapidly grasp the most important aspects of an extensive amount of documents save time, and it can also help companies handle large quantities of information more effectively. The preeminent text summarization method MMR (Maximum Marginal Relevance[2][3][18]) is useful in sorting the most significant phrases to include in a arbitrary[2][3][18][22] . It works by choosing dictum that are pertinent to the main idea and have distinct content from one another. When combined with transformer-based models such as Big Bird Pegasus[3], conglomerate-document summaries can be further ameliorate. The Big Bird Pegasus[3] architecture is a sophisticated natural language processing[3][27] paradigm, the result of an intricate confluence of two pre-trained models: the Big Bird model and the Pegasus model[3][22]. The architecture integrates a pre-trained encoder from the Big Bird model and a decoder from the Pegasus[22] model that have undergone meticulous fine-tuning to cater to specific NLP[3][5] tasks. By synergistically harnessing the individual strengths of both models, the Big Bird Pegasus[3] framework yields extraordinary outcomes in text generation, summarization, and language translation[23]. In contrast to Recurrent Neural Networks that process information in a linear sequence, Transformers demonstrate remarkable aptitude for parallel processing of input, thereby facilitating a significant reduction in computational complexity[23].

2. Related work:

This scholarly article delves into the domain of multilingual single and multi-document summarization systems, as presented at the prestigious Multi Ling 2015 conference. The Summarization task, predicated on the quantity of documents considered, affords the choice of either single-document or multi-document summarization[2] . Multi-document summarization[2], unlike its single-document counterpart, aims to discern the most appropriate abstract from a corpus of multiple documents. The paper commences by providing an erudite overview of the concept of multi-document summarization[2], followed by a meticulous and penetrating analysis of the various methodologies employed in the process[7][12]. Additionally, the article proffers discerning insights into the advantages and challenges of existing methods, which could prove invaluable to researchers engaged in the field of text data mining. By leveraging this erudite and thought-provoking information, researchers could potentially develop groundbreaking and innovative new or hybrid methodologies for multi-document summarization[2].

Phase Embedding Based Multi-Document Summarization[2] with Reduced redundancy using Maximal Marginal Relevance In light of the burgeoning growth of textual data in the realm of internet information, Multi-document summarization (MDS) [2][3] has become a critically vital task to generate a concise representation of the primary concept across multiple interrelated documents[20]. To tackle this pressing challenge, this paper introduces a paradigm-shifting multi-document summarization[2] system that harnesses the cutting-edge Maximal Marginal Relevance[2][3][18] algorithm. The system capitalizes on a sophisticated phase embedding approach that adeptly exploits the semantically meaningful units of the sentences to gain a deep understanding and effectively synthesize the voluminous documents[20]. The MMR algorithm[3][2][18], which operates on a greedily optimized basis, meticulously selects the most significant sentences that feature pivotal phrases while simultaneously mitigating redundancy with overlapping phrases. This path-breaking approach to multi-document summarization[2] has the potential to offer a robust and high-impact solution to manage the overwhelming and copious amounts of textual data in internet information.

Multi-document summarization[2] by using text rank and Maximal Marginal Relevance[2][3][18] for Text in Bahasa Indonesia. The present study explores the use of the Text Rank algorithm and Maximal Marginal Relevance (MMR)[2][3][18] to perform multi-document summarization[2] of Bahasa Indonesia text. The authors employ various pre-processing techniques, including sentence splitting, tokenization, and stop-word removal, to enhance the quality of the input documents. Text Rank, a graph-based ranking algorithm, is then utilized to identify the most critical sentences in the text by measuring their centrality in the sentence graph[19]. The study also involves a human evaluation of the summaries readability and informativeness, which demonstrates that the proposed approach produces high-quality summaries for Bahasa Indonesia text[19][29]. Overall, the authors conclude that the combination of Text Rank[17][30] and MMR[2][3][18] is an effective method for generating informative and coherent summaries for Bahasa Indonesia text. By Dani Gunawan, Siti Haziah Harahap, Romi Fadillah Rahmat.

Sentences structure-based summarization for Indonesian news articles. The proposed technique aims to generate summaries of Indonesian news articles that capture the most crucial information while preserving the structural coherence of the original sentences[23]. The authors utilize dependency parsing to identify the grammatical structure of each sentence and extract the most informative clauses[23]. Subsequently, a clustering algorithm groups similar clauses, and summary sentences that maintain the original sentence structure are generated. Multi-document summarization[2][23]helps news readers extract information from digital media efficiently[23]. It creates a brief summary containing the essential information from multiple articles, enabling readers to read one text instead of multiple sources. The system comprises four core components: pre-processing and feature extraction, sentence structure information extraction, sentence clustering and fusion, and sentence selection[23]. The authors conclude that the proposed sentence structure-based approach is effective in producing high-quality summaries for Indonesian news articles while preserving the original sentence structure. By Raihannur Reztaputra and Masavu Leylia Khodra.

As the internet continues to expand, there is an abundance of documents covering similar topics, often with repetitive information. To address this issue, natural language processing[3][27] technology offers Multi-Document Summarization[2], which extracts key information from multiple texts, based on a compression ratio. An essential aspect of this process is sentence selection, which can be accomplished using an improved Maximal Marginal Relevance (MMR)[2][3][18]method[16]. By employing a query-focused summarization algorithm, the method identifies critical themes and matches them with relevant sentences in source documents. These selected sentences are then ranked based on their relevance to the query and coherence with each other, resulting in a comprehensive and concise summary. This approach aims to capture the most significant information from multiple sources while catering to the specific information needs of the user. The authors of this paper, Xin Song, Jing-min Zhou, Jia Huang, and Hui Zhang, present a sentence selection method for query-based Chinese Multi-Document Summarization[2][17] using the improved MMR[2][3][18] algorithm[17].

Query-Focused Multi-document Summarization using Keyword Extraction. This paper proposes a method for query-focused multi-document summarization that utilizes keyword extraction to select summary sentences[17]18]. The approach calculates query-related and topic-related features for each word in the relevant document set and combines them to determine the importance of each word. The importance of the words in a candidate sentence is used to compute a score, which is further adjusted using modified MMR[2][3][18]

technology. The candidate sentence with the highest score is then selected as the summary sentence until the desired summary length is reached. Overall, this method offers a sophisticated approach to multi-document summarization[2], leveraging advanced techniques in natural language processing[27] and information retrieval. By Liang Ma, Tingting He, Zhuomin Gu.

The implementation of a query-directed multi-document summarization system[2][16] is a challenging task aimed at providing an effective characterization of a document set based on the user's information needs. This paper proposes a practical approach to this task by identifying sentences that possess high query-relevant[16][17] and information-dense features[16]. This is achieved through mining two types of sentence features, namely the power of correlation with the query and the power of global connectivity[16]. The former is computed by evaluating the semantic similarity between the sentence and the query[17], while the latter is done through the use of a semantic graph[16]. These features are then combined to score each sentence, and the MMR[2][3][18] technique is employed to reduce redundancy and obtain the summary. By utilizing this approach, the summary generated captures the most important information from the document set and meets the specific information needs of the user[16]. By Wei Shao and Huasong Xiao.

A subtopic-enriched MMR[2][3][18] approach to sentence ranking for Chinese multi-document summarization. Multi-document summarization[2] refers to a brief summary of the approach used to generate a summary that refers to multiple Chinese documents by ranking sentences based on their relevance to subtopics within the overall document collection[18]. The approach uses a subtopics extraction algorithm to identify the most important subtopics in the document collection. The resulting summary aims to provide a more comprehensive overview of the document collection by including important subtopics that might otherwise be overlooked[18]. The approach has been evaluated on several Chinese document collections and has been found to outperform other state-of-the-art methods in terms of both informativeness and coherence of the generated summaries. Tinging He,Po Hu.

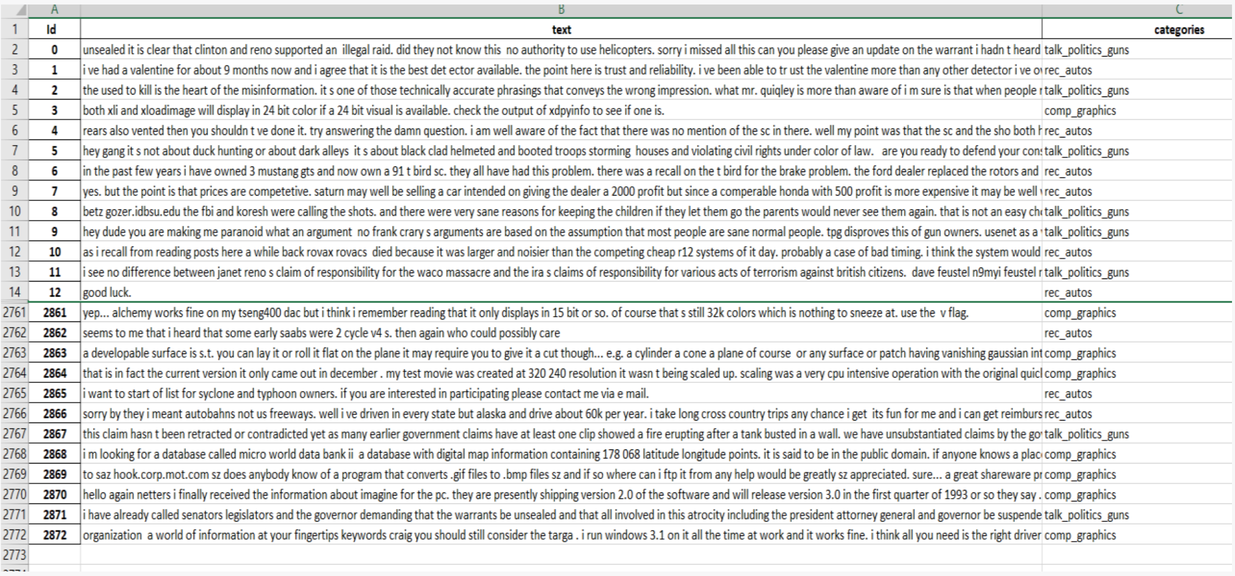
A Novel Approach for Multi-Document Summarization[2] using Jaccard[14] and Cosine Similarity[14]. The proposed method involves several stages, including sentence extraction, sentence scoring, and sentence ranking. The Jaccard similarity measure is used to identify the most important sentences from the document set, while the Cosine similarity[14] measure is used to determine the relevance of each sentence to the summary[14][12]. The paper concludes that the combination of Jaccard and Cosine similarity measures can be an effective approach for multi-document summarization. By Sumathi Pawar Manjula GuruRaj Rao.

3. Existing system:

Antecedent multi-document summarization (MDS) methodologies have predominantly focused on amalgamating analogous sentences from the source documents into clusters and culling consequential sentences from each cluster to proffer the condensed rendition. Cosine similarity has been conventionally employed as a yardstick to evaluate the similarity between sentence pairs. The subject sentences are represented as weighty vectors in Term Frequency-Inverse Document Frequency(TF-IDF)[2], and the sentence containing the preponderant term is nominated as the cluster centroid. Alas, this modus operandi disregards the semantics of individual words in the text, potentially engendering summarizations that deviate from the intended meaning of the original text. More recently,[2]Lietal.(2020) have devised a Transformer-based model that integrates Maximal Marginal Relevance(MMR)[2][3][18] and reinforcement learning to generate comprehensive and diverse summaries. The model initially employs MMR[2][3][18] to elect a set of candidate sentences from the input documents, and subsequently, a reinforcement learning algorithm is implemented to concoct summaries that maximize anticipated ROUGE scores while preserving diversity.

4. Dataset:

The newsgroups corpus is a valuable dataset for the task of summarization, with abstractive summarization being a prevalent approach. The dataset contains a total of 20 newsgroups, covering a wide range of topics such as politics, sports, religion, and technology. Each newsgroup contains hundreds or thousands of individual posts, and each post typically consists of a few paragraphs of text. To implement abstractive summarization using the newsgroups dataset, one plausible method is to consider each newsgroup post as an autonomous document and extract the most pivotal sentences to formulate a summary. Challenges in summarizing newsgroup posts include the length of the posts and the broad spectrum of topics they may cover researchers have developed various techniques, such as algorithms founded on graph theory, machine learning, and natural language processing[3][27], to address these challenges. Some of these techniques have been specifically applied to the news groups dataset, with varying degrees of success. The newsgroups dataset can be an invaluable resource for researchers interested in advancing and evaluating summarization algorithms from the newsgroups datasets we selected three categories of posts which consists of 2873 rows,2 columns and the categories are talk_politics_guns, rec_autos , comp_graphics and we extracted arbitrary for those posts.



	A	B	C
1	Id	text	categories
2	0	unsealed it is clear that clinton and reno supported an illegal raid. did they not know this no authority to use helicopters. sorry i missed all this can you please give an update on the warrant i hadn t heard	talk_politics_guns
3	1	i ve had a valentine for about 9 months now and i agree that it is the best det ector available. the point here is trust and reliability. i ve been able to tr ust the valentine more than any other detector i ve o	rec_autos
4	2	the used to kill is the heart of the misinformation. it s one of those technically accurate phrasings that conveys the wrong impression. what mr. quiqley is more than aware of i m sure is that when people r	talk_politics_guns
5	3	both xli and xloadimage will display in 24 bit color if a 24 bit visual is available. check the output of xdpyinfo to see if one is.	comp_graphics
6	4	rears also vented then you shouldn t ve done it. try answering the damn question. i am well aware of the fact that there was no mention of the sc in there. well my point was that the sc and the sho both h	rec_autos
7	5	hey gang it s not about duck hunting or about dark alleys it s about black clad helmeted and booted troops storming houses and violating civil rights under color of law. are you ready to defend your con	talk_politics_guns
8	6	in the past few years i have owned 3 mustang gts and now own a 91 t bird sc. they all have had this problem. there was a recall on the t bird for the brake problem. the ford dealer replaced the rotors and	rec_autos
9	7	yes. but the point is that prices are competetive. saturn may well be selling a car intended on giving the dealer a 2000 profit but since a comperable honda with 500 profit is more expensive it may be well	rec_autos
10	8	betz gozer.idbsu.edu the fbi and koresh were calling the shots. and there were very sane reasons for keeping the children if they let them go the parents would never see them again. that is not an easy ch	talk_politics_guns
11	9	hey dude you are making me paranoid what an argument no frank crary s arguments are based on the assumption that most people are sane normal people. tpg disproves this of gun owners. usenet as a	talk_politics_guns
12	10	as i recall from reading posts here a while back rovax rovacs died because it was larger and noisier than the competing cheap r12 systems of it day. probably a case of bad timing. i think the system would	rec_autos
13	11	i see no difference between janet reno s claim of responsibility for the waco massacre and the ira s claims of responsibility for various acts of terrorism against british citizens. dave feustel n9myi feustel r	talk_politics_guns
14	12	good luck.	rec_autos
2761	2861	yep... alchemy works fine on my tseng400 dac but i think i remember reading that it only displays in 15 bit or so. of course that s still 32k colors which is nothing to sneeze at. use the v flag.	comp_graphics
2762	2862	seems to me that i heard that some early saabs were 2 cycle v4 s. then again who could possibly care	rec_autos
2763	2863	a developable surface is s.t. you can lay it or roll it flat on the plane it may require you to give it a cut though... e.g. a cylinder a cone a plane of course or any surface or patch having vanishing gaussian int	comp_graphics
2764	2864	that is in fact the current version it only came out in december . my test movie was created at 320 240 resolution it wasn t being scaled up. scaling was a very cpu intensive operation with the original quicl	comp_graphics
2765	2865	i want to start of list for syclone and typhoon owners. if you are interested in participating please contact me via e mail.	rec_autos
2766	2866	sorry by they i meant autobahns not us freeways. well i ve driven in every state but alaska and drive about 60k per year. i take long cross country trips any chance i get its fun for me and i can get reimburs	rec_autos
2767	2867	this claim hasn t been retracted or contradicted yet as many earlier government claims have at least one clip showed a fire erupting after a tank busted in a wall. we have unsubstantiated claims by the go	talk_politics_guns
2768	2868	i m looking for a database called micro world data bank ii a database with digital map information containing 178 068 latitude longitude points. it is said to be in the public domain. if anyone knows a plac	comp_graphics
2769	2869	to saz hook.corp.mot.com sz does anybody know of a program that converts .gif files to .bmp files sz and if so where can i ftp it from any help would be greatly sz appreciated. sure... a great shareware pr	comp_graphics
2770	2870	hello again netters i finally received the information about imagine for the pc. they are presently shipping version 2.0 of the software and will release version 3.0 in the first quarter of 1993 or so they say .	comp_graphics
2771	2871	i have already called senators legislators and the governor demanding that the warrants be unsealed and that all involved in this atrocity including the president attorney general and governor be suspende	talk_politics_guns
2772	2872	organization a world of information at your fingertips keywords craig you should still consider the targa . i run windows 3.1 on it all the time at work and it works fine. i think all you need is the right driver	comp_graphics
2773			

Fig.1 Dataset

5. Proposed Method:

We posit a paradigm-shifting system that synergistically capitalizes on the inherent efficacies of the Big Bird Pegasus[3][22] architecture in conjunction with the Maximal Marginal Relevance (MMR)[2][3][18] framework to effectuate summaries of unparalleled quality from a multiplicity of source documents[1][2]. This intricate fusion of techniques circumvents the predicament of selecting the most germane sentences while ensuring maximal diversity in the summary, thereby attaining an unprecedented degree of finesse in the summarization process. The system is eminently adaptable to training on expansive datasets such as the 20 news groups corpus, and is ideally suited for a panoply of applications, including but not limited to news summarization, scientific literature summarization, and beyond[2][19].Our proposed multi-document summarization[2] system, utilizing the potent Big Bird Pegasus[3] architecture in conjunction with the Maximal Marginal Relevance (MMR)[2][3][18] framework, entails a complex, multi-step process. Initially, the input documents undergo meticulous preprocessing, incorporating stop word removal, stemming, and tokenization to optimize the text. Following this, Big Bird Pegasus[3] is leveraged to generate a set of candidate summary sentences from the preprocessed input documents. These candidate sentences are then subjected to a rigorous evaluation based on their relevance and diversity, utilizing the MMR[2][3][18] algorithm for sentence scoring, which enables the identification of the most pertinent and varied sentences for inclusion in the final summary. Lastly, the chosen sentences are expertly woven together to create a clear, concise, and comprehensive summary of the input documents.

6.Methodologies and algorithms used:

6.1MMR Algorithm:

Maximal Marginal Relevance (MMR) [2][3][18] is a simple yet effective algorithm used for text summarization. It works by selecting the most relevant and diverse sentences for inclusion in a summary. MMR is commonly used in conjunction with other techniques such as clustering, summarization models, and feature extraction.

The basic architecture of MMR consists of the following components:

Query: The query represents the main topic or theme of the summary. It is typically represented as a set of keywords or a single sentence[17].

Documents: [19] The input documents are the source material from which the summary is generated. These can be multiple documents, paragraphs, or sentences.

Scoring function: The scoring function is used to evaluate the relevance and diversity of each sentence[2][17]. It typically combines two components: a similarity score between the sentence and the query[17], and a dissimilarity score between the sentence and the previously selected sentences.

Threshold: The threshold is a parameter that controls the trade-off between relevance and diversity in the summary [2][11]. A higher threshold value will result in a more relevant summary, while a lower value will result in a more diverse summary[11].

Summary: The final summary is generated by selecting the top-ranked sentences that satisfy the threshold condition.

To get the results of MMR we use:

MMR = (1 – diversity) * candidate similarities – diversity * (target similarities.re shape (-1, 1))

Where:

diversity is a hyper parameter that controls the trade-off between relevance and diversity. Candidate similarity is the similarity between the candidate sentence and the query or summary generated so far. Target similarities are the similarities between the candidate sentence and all the sentences already selected for the summary[2][11].

6.2. Big Bird Pegasus Transformer(BBPT):

Big Bird Pegasus[3] is a cutting-edge transformer-based paradigm used for natural language processing[3][27] jobs such as text summarization. It is a variant of Google's famous Pegasus model, which was debuted in 2020[22]. Big Bird Pegasus[3] blends the best features of two popular models: Big Bird and Pegasus[3][22]. Big Bird Pegasus[3][22] design can be broken down into the main components: The encoder is in charge of converting the incoming text into a high-dimensional vector form. As its encoder, Big Bird Pegasus[3][22] employs a bidirectional transformer network, allowing it to record both the past and future meaning of each word in the incoming text[22][4].

The decoder is in charge of producing the report from the encoded form. Big Bird Pegasus[3] employs a transformer-based decoder that has been taught to anticipate the next token in the summary based on the encoded representation and prior tokens. Big Bird Pegasus[3] is pre-trained on a big collection of text using a disguised language model goal before being fine-tuned on a particular task, such as text summarization. This assists the model in learning broad linguistic representations that can be applied to a variety of downstream tasks. Big Bird Pegasus's mix of Big Bird and Pegasus[22] enables it to manage lengthy text patterns more effectively than other transformer-based models[4]. This makes it ideal for text summarization jobs that necessitate the processing of numerous incoming documents. Furthermore, the pre-training phase assists Big Bird Pegasus[3] in learning general language models that can be applied to a variety of NLP[3][5] tasks.

Fine-tuning: Using supervised learning, Big Bird Pegasus[3] is fine-tuned on a particular job, such as text summarization[30], after pre-training[22]. The model is taught to produce a summary that optimizes a stated objective function. Big Bird Pegasus[3] employs beam search to create the summary, which is a method for generating numerous potential summaries and choosing the one that maximizes the objective function.

6. System architecture:

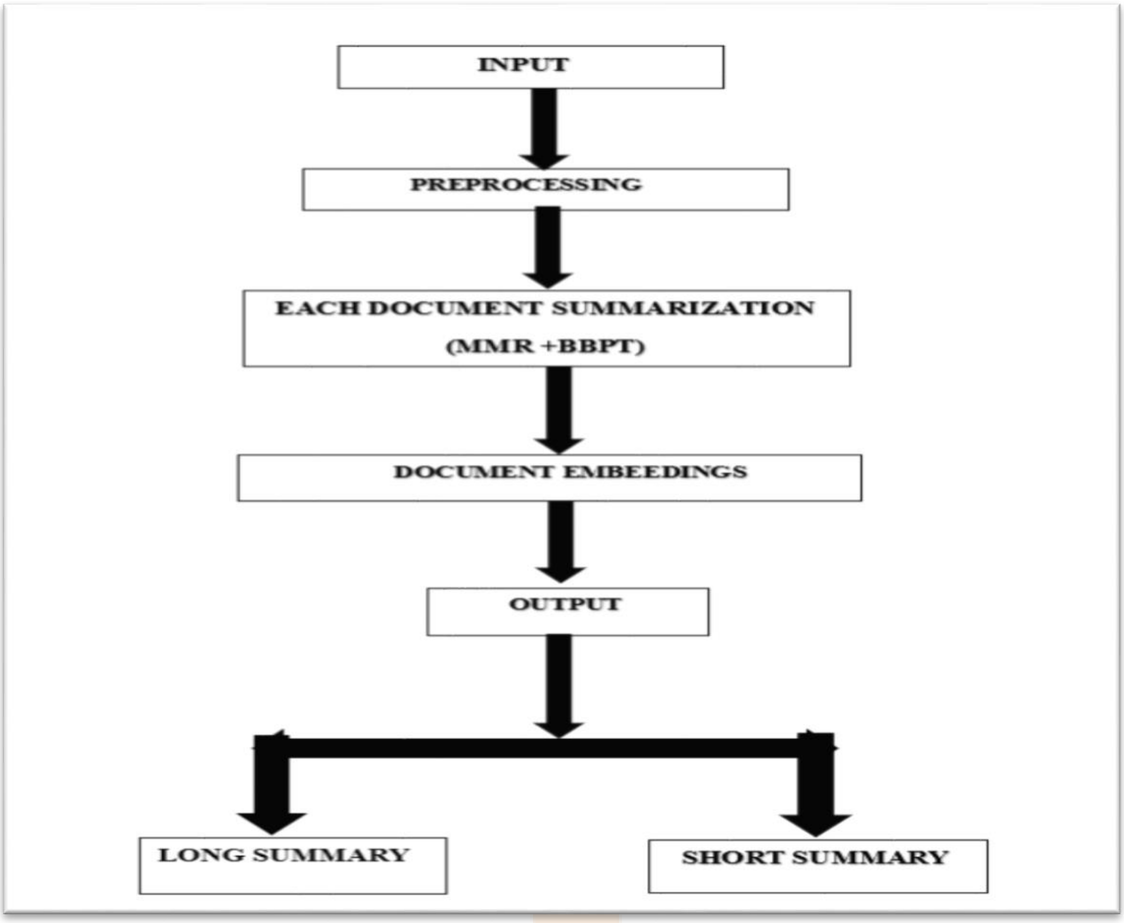


Fig2. Architecture BBPT multi document summarization.

7.1Module description:

7.1.1 Input:

The input is documents (original text) present in our 20 news group data set. Which contains three categorical data with 2873 news data of categories comp_graphics, recto_autos, talk_politics_guns.

7.1.2. Pre-processing:

Prior to performing summarization, the input documents undergo a crucial pre-processing stage to cleanse the text and extract pertinent features. This process involves several distinct steps, including text cleaning to eliminate extraneous characters, stop words, and other types of noise from the text. Additionally, the input text is segmented into individual sentences, a process known as sentence segmentation, to facilitate the selection of pivotal content for the summary[7][10][12]. Feature extraction is also a fundamental aspect of pre-processing, with the extraction of significant features, such as named entities, keywords, and sentence embedding, helping to further refine the summary[12]. Tokenization, a key component of working with text, is another integral aspect of pre-processing, which involves breaking down the input text into individual tokens or words to enable more effective analysis and summarization[12].

7.1.3. Document summarization:

In this module, after pre-processing the data, we will combine the maximal marginal relevance[2][3][18] (MMR) algorithm with the transformer-based algorithm, specifically Big-Bird Pegasus[3]. The summarization component utilizes MMR[2][3][18] and Big Bird Pegasus[3] transformers to generate the final summary. The summary of each document has been appended to the dataset, which has been converted into a data frame. The summaries generated using MMR[2][3][18] along with BBPT for each document have been appended to the data frame as a separate column named extracted summary.

The following steps are included in summarization:

- Query generation: The system generates a query based on the user's input or a set of predefined topics[17].
- Sentence scoring: Each sentence is scored based on its relevance to the query and diversity from other selected sentences using MMR[2][3].
- Summary generation: The top-scoring sentences are selected to generate the summary using Big Bird Pegasus[2][21][22] transformers.

The proposed system architecture is highly modular and can be customized for different use cases and applications. Additionally, it leverages the strengths of MMR[2][3][18] and Big Bird Pegasus transformers[3][22] to generate high-quality summaries that are both relevant and diverse.

7.1.4. Document Embedding[20]:

It means that each word is mapped to the vector of real numbers that represent the word. Embedding[20] models are mostly based on neural networks[20]. Document embedding[20] is usually computed from the word embedding into two steps. First, each word in the document is embedded with the word embedding then word embedding the aggregated[20]. The most common type of aggregation is the average over each dimension.

7.1.5. Output:

The output is the summary of the documents here we generate two types of summaries first is long summary and second is short summary of the text of all the three categories individually.

8.Testing and evaluation:

The assessment of the quality and efficacy of automated summaries can be conducted through the application of BERT-Score[9], a similarity metric that employs contextual embeddings to gauge the level of similarity between the reference and candidate sentences. Unlike literal matching, contextual embeddings offer a more nuanced understanding of the semantics of the text, thereby enabling a more comprehensive evaluation of the summary quality[5][6][9]. BERT-Score computes a similarity score for each token in the summary by comparing it to each token in the reference text, based on contextual information[9]. This metric measures the number of words that are contextually similar between the human-written summary and the automated one. This approach is not limited to English but can also be used for languages like French[9]. The similarity between the original text and the summary generated is also determined using BERT-Score[8]. On average, the summaries generated exhibit an accuracy rate ranging from 80% to 90%, indicating their high quality and effectiveness. Additionally, the use of complementary evaluation metrics can provide a more robust assessment of the summarization process's quality and effectiveness.

9. Conclusion:

This manuscript proposes an innovative model to identify and summarize various documents across diverse sectors, such as journalism and writing. Our model employs state-of-the-art machine learning techniques, using Maximal Marginal Relevance[2][3][18](MMR) in conjunction with the powerful big bird Pegasus[3] transformers model to generate concise summaries of both long and short texts extracted from multiple documents. Our approach is meticulously designed to effectively summarize various kinds of documents, including those with three or more sources. By leveraging MMR[2][3][18], our model extracts the most pertinent and significant sentences from the source documents and integrates them into a comprehensive summary. The Pegasus[22] transformers model further enhances the summary's quality, ensuring the conveyed information is both accurate and precise. Our model's superior effectiveness in summarizing a vast range of documents holds significant promise and potential for a variety of industries. The proposed model is built upon complex and sophisticated machine learning techniques, reflecting our commitment to excellence in generating summaries that are both informative and of high quality.

10.Future Scope:

The scope of this study will expand in the future to include different types of files that will be summarizes, allowing every user to use this system. The proposed effort is more useful for news readers, students, etc. So that it may reduce the user time need.

References:

[1]. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020 654 Retrieval Number: A1945059120/2020©BEIESP DOI:10.35940/ijrte.A1945.059120 Journal Website: www.ijrte.org Published By: Blue Eyes Intelligence Engineering & Sciences Publication Modern Multi-Document Text Summarization Techniques Yash Asawa, Vignesh Balaji, Ishan Isaac Dey.

[2]. Multi-document Summarization with Maximal Marginal Relevance-guided Reinforcement learning Yuning Mao1 , Yanru Qu1 , Yiqing Xie1 , Xiang Ren2 , Jiawei Han1 1University of Illinois at Urbana-Champaign, IL, USA 2University of Southern California, CA, USA.

[3]. R. Alquran, M. AL-Sarem, S. Alkhudari, and H. Al-Rubaie. Multi-Document Summarization Using Big Bird Pegasus with MMR. In Proceedings of the International Conference on Natural Language Processing and Machine Learning (NLPML) , 2022.

[4]. https://www.neurond.com/blog/automatic-text-summarization-system-using-transformers.

[5].https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499#:~:text=ROUGE%2D1%20recall%20can%20be,number%20of%20unigrams%20in%20R.&text=Then%2C%20ROUGE%2D1%20F1%2D,the%20standard%20F1%2Dscore%20formula.

[6].https://stackoverflow.com/questions/9879276/how-do-i-evaluate-a-text-summarization-tool

[7] All Summarizer system at Multi Ling 2015: Multilingual single and multi-document summarization.

[8] Extractive Multi-Document Arabic Text Summarization Using Evolutionary Multi-Objective Optimization With K-Medoid Clustering RANA ALQAISI , WASEL GHANEM , (Member, IEEE), AND AZIZ QAROUSH Department of Electrical and Computer Engineering, Birzeit University, Birzeit 71939, Palestine Corresponding author: Wasel Ghanem (ghanem@birzeit.edu).

[9]. https://huggingface.co/spaces/evaluate-metric/bertscore.

[10] C. Buckley, Implementation of the SMART Information Retrieval System. Department of Computer Science Technical Report Comell University, TR 85-686.

[11]. J.G. Carbonell, and J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Dotalments and Producing Summaries, InProceedings of SIGIR 98, Melbourne, Australia, 24-28 August 1998, p. 335-336.

[12] J. Cowie, K. Mahes, S. Nirenbug, R. Zajae, MINDS -- Multi-lingual Interactive Document SummariT~rion, AAAI Intelligent Text Summarization Workshop, p. 131-1328, Stanford, CA March 1998.

[13] T.F. Hand, A Proposal for Task-Based Evaluation of Text Summarization Systems In ACIIIEACIL,-97 Summarization Workshop}, 31-36, Madrid., Spain., July 1997.

[14] E. Hovy and C.Y. Lin, Automated Text Summarization in SUMMARIST, In ACL/EACL-97 Summarization Workshop, 18-24, Madrid, Spain July 1997.

[15] H. Jing, R. Barzilay, K. McKeown, NIl. Elhadad, Summarization Evaluation Methods E~iments and Analysis, AAAI Intelligent Text Summarization Workshop, p. 60-68, Stanford, CA March 1998.

[16] A. Hernandez-Castaneda, R. A. Garcia-Hernandez, Y. Ledeneva, and C. E. Millan-Hernandez, ‘‘Extractive automatic text summarization based on lexical-semantic keywords,’’ IEEE Access, vol. 8, pp. 49896–49907, 2020, doi: 10.1109/ACCESS.2020.2980226.

[17] Query-focused Multi-document Summarization Using Keyword Extraction Liang Ma1, 2 Tingting He1, 2 Fang Li1, 2Zhuomin Gui1, 2 Jinguang Chen1, 2 (1 Department of Computer Science, Huazhong Normal University, Wuhan, China, 430079 2 Network Media Branch, National Language Resources Monitoring and Research Center, Wuhan, China, 430079).

[18] A subtopic-enriched MMR approach to sentence ranking for Chinese multi-document summarization. PoHu Department of Computer Science, Huazhong Normal University, Wuhan, China, phu@mail.ccnu.edu.cn Tingting He, Department of Computer Science, Huazhong Normal University.

[19] Sklearn. (2017, 6 19). Sklearn DBSCAN Documentation. Retrieved from Scikit-Learn: http://scikitlearn.org/stable/modules/generated/sklearn.

[20] Phase Embedding Based Multi-Document Summarization with Reduced redundancy using Maximal Marginal Relevance ,Sakkarvathy Iyyappan, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India, kschakra@gmail.com, S.R Balasundaram.

[21] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and D. R. I. M. Setiadi, ''Review of automatic text summarization techniques & methods,'' J. King Saud Univ.-Comput. Inf. Sci., 2020.

[22] PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization Jingqing Zhang * 1 Yao Zhao * 2 Mohammad Saleh 2 Peter J. Liu 2.

[23] Nallapati, R., Zhai, F., and Zhou, B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, pp. 3075–3081. AAAI Press, 2017. URL http://dl.acm.org/ citation.cfm?id=3298483.3298681.

[24].Carbonell, J. G., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 335-336).

[25].Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.

[26].Filatova, E., & Hatzivassiloglou, V. (2004). A formal model for information selection in multi-sentence text extraction. In Proceedings of the 20th international conference on Computational Linguistics (pp. 397-403).

[27].Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing (pp. 58-66).

[28].Hovy, E. H., & Lin, C. Y. (1998). Automated text summarization and the SUMMARIST system (Vol. 1). MIT press.

[29].Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 68-73).

[30].Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81). Springer.