***Text summarization of Covid-19 dataset***

***Course Name:*** Undergraduate Research Opportunity Program (UROP)

***Course Code:*** CSE340

Submitted by

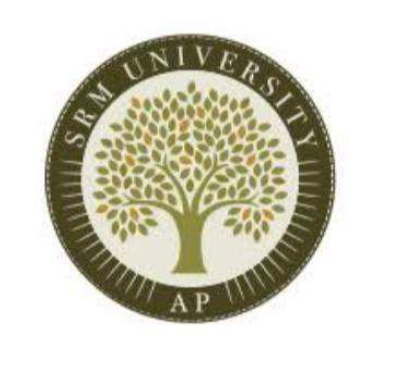
Addepalli Srilekha – AP19110010013

Sri Lahari Sivanvitha Nori– AP19110010076

K.Sriharsha – AP19110010252

Research Supervisor

**Dr. Priyanka Singh**



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SRM UNIVERSITY AP, ANDHRA PRADESH

GUNTUR– 522502, INDIA

APRIL 2022

**Table of contents**

| **Heading** | Pages |
| --- | --- |
| **Abstract** | 3 |
| **Introduction** | 4 |
| **Related work** | 6 |
| **Dataset description** | 8 |
| **Procedure** | 8 |
| **Results** | 11 |
| **Conclusion** | 18 |
| **Reference** | 19 |

**Abstract**

The amount of text data available from various sources has increased dramatically in recent years. This large volume of literature has a wealth of information and knowledge that must be adequately summarised in order to be useful. The main approaches to automatic text summarization are presented in this review. We examine the various summarising approaches and discuss their efficacy and drawbacks. With growing data and information all around and the daily improvement in technology, no one has the time and need to read all the documents present and manually summarize them. A solution to this is text summarization. It has had a huge impact on the current industrial field. Implementing this, we can generate a summary that is easy to understand. For this project, after the preprocessing of data like removing null values, and removing the stop words required to summarize the data, we are implementing Extractive Text summarization and then Abstractive text summarization to the summarized data taken from the extractive summarization. To check the performance implemented the Rouge score and ranked the resultant output.

**Introduction**

Given the massive amounts of information available today, both on the Web and elsewhere, methodologies for effective automatic text summarization are critical for improving access to that data. In natural language comprehension, text summarization is a well-known task. In general, summarization refers to the problem of presenting information in a compact manner that focuses on the most relevant aspects of the data while maintaining the meaning. The basic goal of summarization is to select a subset of data that contains all of the data's "information." Data generation and consumption are booming at an exponential rate in today's environment. As a result, text summary has become a requirement for many applications, including search engines, business analysis, and market research. Text summarization is diminishing the word count without losing the essence of the text. This will help us in saving time as it has the crucial part of the text and will eliminate the noise in the sentence. Text summarization can be done in 2 ways.

In-human text summarization is assigning a person to read and understand the text and writing the truly crucial part of the text. Human beings have the capacity and skills to differentiate between what is important and what is not. But their knowledge is limited to some words and they cant go through the whole text to put their best. Each person will do the text summarization according to his knowledge and will get different outputs. In humans text, summarization is an expensive than automatic text summarization.

Automatic text summarization is using a tool that has the knowledge to combine some words. These automatic text summarization tools are made using Machine learning algorithms. Text summarization is one of those Natural Language Processing (NLP) technologies that will undoubtedly have a significant impact on our lives. It is cheaper than In-human text summarization. Even in automatic text summarization, there are two types of summarization.

Abstractive text summarization works on generating the new sentences from the scratch. It interprets the document and provides the output with the same essence as the original document. It takes a lot of time to run as it has to generate new sentences. Extractive text summarization extracts important sentences from the document. It either keeps the whole sentence or removes the whole sentence. Abstractive text summarizations are faster when compared to extractive text summarizations.

Abstractive summarization produces the output in a new way after interpreting and examining the text using the natural language algorithms for a better and shorter text which consists of the crucial information from the original document whereas extractive summarization produces the output by extracting the important sentences from the original document which are interpreted essensing to the document. Abstractive text summarization is a more advanced and more human-like interpretation when compared to extractive text summarisation whereas extractive text summarization is faster and easier to understand when compared to abstractive.

TextRank Algorithm is a graph-based text processing ranking model that may be used to determine the most relevant phrases in a text as well as keywords.

To arrive at a final output, it does numerous iterations on the pages. TextRank is unsupervised and after splitting the text into sentences, the computer creates a graph with sentences as nodes and overlapped words as links and provides the output.

Sequencetosequence (Seq2seq) takes a set of words (phrases or sentences) as input and outputs a string of words. This is accomplished by employing a recurrent neural network (RNN). Although the simple RNN is rarely utilized. It consists of two components.The encoder uses deep neural network layers to transform input words into hidden vectors. The current word and its context are represented by each vector. The decoder uses the encoder’s hidden vector, its own hidden states, and the current word as input to construct the next hidden vector and forecast the next word.

The GloVe(global vectors for word representation) technique is an unsupervised learning method for obtaining word vector representations. The resulting representations highlight intriguing linear substructures of the word vector space, which are trained using aggregated global word-word co-occurrence information from a corpus. Each word is a point in vector space that is learned and moved around the target word while semantic links are preserved. The vector space representation of words creates a projection in which words with similar content are grouped together.

Transformers is an architecture that runs on the approach of self-attention to make important representations of the dataset. Transformer models use a developing collection of mathematical techniques known as attention or self-attention to detect subtle ways that even far-flung data pieces in a series impact and rely on one another.

The model sshleifer distilbart CNN 12-6 is a Natural Language Processing (NLP) Model implemented in the Transformer library, generally using the Python programming language. The model is based on the model of the DistilBART model. . For instance, the model was constructed with the model name "Portrait of the Distilbart model" and the model's speedup function, MSSPL speedup, from MS Speedup. The Hugging Face transformers package is a widely used Python library that provides pre-trained models that can be used for a range of natural language processing (NLP) purposes. It formerly exclusively supported PyTorch, however as of late 2019, it also supports TensorFlow 2. Text categorization is one of the most common and practical use cases for the library, which may be utilized for a variety of tasks ranging from Natural Language Inference (NLI) to Question-Answering. The Hugging Face transformers package is a widely used Python library that provides pre-trained models that can be used for a range of natural language processing (NLP) purposes. It formerly exclusively supported PyTorch, however as of late 2019, it also supports TensorFlow 2. Text categorization is one of the most common and practical use cases for the library, which may be utilized for a variety of tasks ranging from Natural Language Inference (NLI) to Question-Answering.

**Related work**

In this particular section, we discuss the recent well-known works in the domain of text summarization. A number of methodologies have been proposed in this area.

Jianpeng Cheng presents a data-driven strategy based on continuous sentence characteristics and neural networks. They create a system for single-document summarization that includes a hierarchical document encoder and an attention-based extractor. They can use this architecture to create various types of summarization models that can extract sentences or words. Proposed and compared two unique ways for finding keywords is done by Marina Litvak should be used in extractive text summarization of documents: supervised and unsupervised. Both of these methods are expansions of the graph-based syntactic representation of text and online documents, which improves on the classic vector-space model by accounting for some structural document characteristics. Many ways for summarising English text have been developed. However, just a few attempts at Bengali text summary have been done. Kamal Sarkar describes a method for summarising Bengali text that takes relevant sentences from a Bengali document and generates a summary. To merge diverse sentence elements, Kam-Fai Wong suggests a learning-based technique. Surface, content, relevance, and event features are the four categories. Extrinsic parts of a sentence are linked to surface features. Content characteristics are words that communicate information in a sentence. The event features represent the events that were contained in the sentences. Finally, relevance features assess a sentence based on its similarity to other sentences. Using a deep auto-encoder to construct a feature space from the term-frequency input, Mahmood Yousefi-Azar and Len Hamey offer approaches for extractive query-oriented single-document summarization. Konstantin Lopyrev explains the use of an encoder-decoder recurrent neural network with LSTM units and attention to generate headlines from news article text in a recent paper.

Zhong et al. (2015) used a deep architecture that is akin to an AE for extractive summarization. They used the information acquired as a result to filter away extraneous terms from a site from the first layer and to discover keywords in the second layer. During the first stage of their training approach, sample queries are required. The goal of the concept space is to uncover potential summary sentences. However, both the phrase and conceptual space in our approach are formed around a document phrase. In the ranking function, we employ the learned representations to reflect the semantic similarity of sentences. Denil et al. (2014) suggested a technique based on a CNN to extract candidate sentences to be included in the summary in supervised models. Because of the large number of records available on the internet, as well as communications and complex libraries, archive grouping is becoming an important and necessary task. It is usually done after completing highlight selection, which entails selecting appropriate highlights to improve grouping precision. The production of a term-frequency contradicting record frequency including vector is used in the majority of element evaluations to bring altogether content order processes, which is not always successful. Similarly, many archive order studies focus on the English language. Garg and Saini proposed an Arabic Text Classification technique, which has gotten little attention due to the language's complex structure. Another component selection technique based on firefly calculations is proposed. A lot of computational challenges have been associated with this calculation. In any case, it hasn't been tied to the Arabic Text Classification management highlight choice idea. One of three assessment methods used to validate this method is the Support Vector Machine classifier: exactness, review, and F-measure. In addition, the OSAC authentic dataset is probed, as well as an examination Text Summarization: A brief review of 7 new solutions is carried out. The proposed method achieves a precision of 0.994. The findings support the proposed highlight choice technique's effectiveness in boosting Arabic Text Classification precision.

**Experiment and results**

**Dataset description**

The dataset which we have used for our Undergraduate Research Opportunity Program is the Covid-19 Open research dataset(CORD-19) which we were able to get in the kaggle. This dataset is prepared by the white house and an alliance of top research groups. This dataset is so much popular that it is used in more than 10 lakh scholarly articles as coronavirus literature is booming in this era. This dataset has increased machine learning jobs by a high margin. Medical institutes also believed that using this dataset will help the team to better understand the covid-19 virus. 27706 new samples were updated to the dataset and removed 69 datasets. In the pdf, there was a total of 334572 JSON files out of which 14648 files were added to the dataset on 2022-01-03 and removed 429 JSON files from the dataset. The 12GB zip dataset was downloaded from the kaggle website and after extraction, the size of the dataset was 64 GB. Total metadata columns in the dataset are 902600 and rows in the dataset are 19. The rows for the dataset are ‘ cord\_uid ’, ‘ sha ‘, ‘ source\_x ‘, ‘ title ‘, ‘ doi ‘, ‘ pmcid ‘, ‘ pubmed\_id ‘, ‘ license ’, ‘ abstract ’, ‘ publish\_time ’, ‘ authors ’, ‘ journal ’, ‘ mag\_id ’, ‘ who\_covidence\_id ’, ‘ arxiv\_id ’, ‘ pdf\_json\_files’, ‘ pmc\_json ‘, ‘ url ’ and ‘ s2\_id ’.

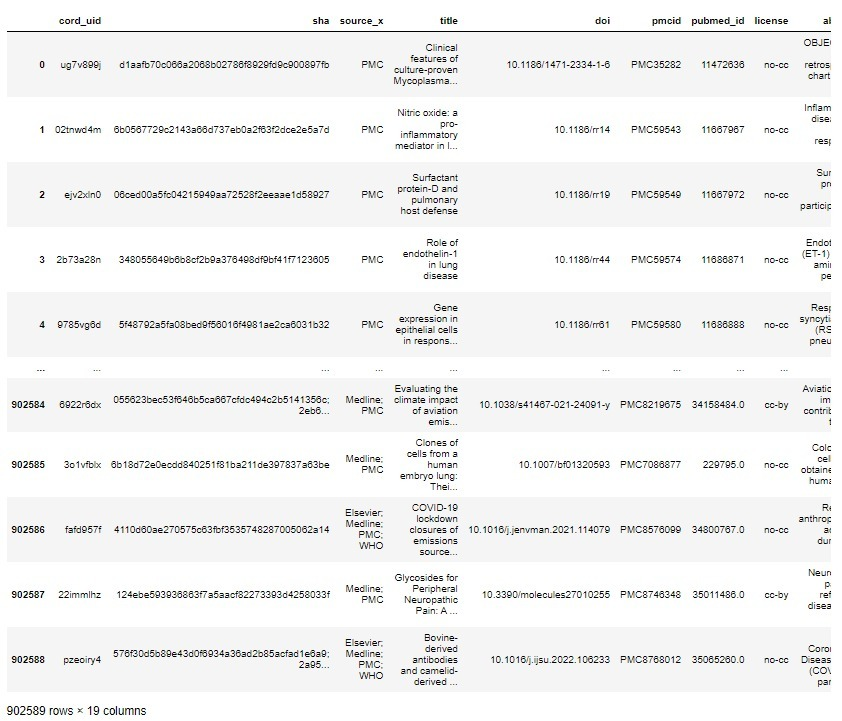
**Procedure**

We will first describe the research project process in detail, followed by a brief introduction to the experimental datasets and evaluation methods. Finally, we will conduct a thorough analysis of the results obtained and identify areas of our model that can be improved in the future.

The dataset we used for our Undergraduate Research Opportunity Program is the Covid-19 Open research dataset (CORD-19), which we obtained from Kaggle. Initially, we loaded the metadata file into the notebook. There we observed numerous columns regarding the abstracts. Every abstract is stored in the form of a JSON file whose path is also specified in a column. Using file open methods, we were able to link the JSON file to the metadata table. Since our aim is to summarize the body\_text of the abstract, we dropped all the irrelevant columns and our final dataset contains the title, abstract, body\_text, and date of publication. So the first step is to pre-process the data. We have checked for null values and we found almost six lakh null value rows. Since all these are textual data we cannot use any fill method. So we dropped all the null rows. As the textual data is imported from a JSON file, all the data must be formatted and irrelevant spaces, lines must be deleted. The model which we are proposing can only summarize abstracts of the English language. So, we have used langdetect model to detect the language of each row entry. To detect this language, we took the first fifty characters of the row for detection. We have created a new column, named language to enter the detected language of each row entry. We have filtered out all the English abstracts. Since the dataset is huge we have converted this dataset into a new CSV file.

For the next steps, we have imported the new CSV file. We need PyTorch. So we went to the official website of PyTorch and selected install, according to the PyTorch build, the OS, package, language, and compute platform we will get a command and by running it directly in the pip, the required version of PyTorch is installed. Then with the help of the command PIP install Transformers we installed Transformers. Then we imported basic libraries like Numpy, Pandas, nltk regex, auto tokenizer, an auto model for sequence-to-sequence, cosine similarity, xlsx writer, etc. Then we initialized an auto tokenizer using sshleifer/distilbart-cnn-12-6 and we imported A sequence-to-sequence model which is a pre-trained model from sshleifer/distilbart-cnn-12-6.From nltk Corpus, we imported stopwords and list out all the stop words from the language English. We defined a function to remove stopwords and pass every sentence into the function and the output will be a sentence that is free from stopwords. We loaded the CSV file that we have generated in the above step. We have used a glove text file which is of UTF 8 encoding and create all the word embeddings. So the abstracts which you are trying to filter out must be related to covid-19, so we have taken a few keywords such as coronavirus, corona, virus, PPE, social distancing, vaccine, Quarantine, etc and filtered all the abstracts that contain any of these words and created a new data set. The summary which we are trying to generate is stored in an excel file for that we have used xlsx writer which is the module used to insert data into an excel sheet with the help of python. We ran a for loop for all the abstracts that were filtered out earlier and then passed the body text into an extractive summarizer then passed the result of the above step into an abstractive summarizer and then passed this summarised data into the Excel sheet where the results will be stored. For the Extractive Text summarisation, we have used a text rank algorithm which is used to rank the text and find out the importance with the help of which will be able to sort out the important sentences for our final summary. We have defined a function for extractive text summarisation in that the first step is to remove the stop words, the second step is to flatten the two-dimensional list that will be created, and in the next step, we removed all the unnecessary punctuations. Then we created a matrix with initial zeros and then assigned the values from the word embeddings and inserted them into the matrix. With the help of cosine similarity, we have reshaped the entire data into one cross hundred and using NX graph which is the module from networkx we converted everything into a NumPy array and then applied the page rank algorithm directly from the module. So the result obtained will be the sentences that are ranked in the form of a list inside which contains a set that has the rank of the sentence in its zeroth index and the sentence itself in the first index. Then we sorted the list based on the the text rank and then took out the 10 most important text for our next step. The texts from the above step will be returned as a result of the extractive text summarisation which will be the return value to the extractive text summarizer. The abstractive text summarizer inputs a sentence and tokenizes the sentences where the tokenizer is initialized at the beginning of a program and it generates new inputs which will be further batch decoded and the output obtained will be the result of the abstractive text summarizer which will be returned by the function. This summary will be appended to the Excel sheet. For the accuracy metric, we have used the ROGUE score which compares the summary generated from the extracted summarizer with the summary generated from the abstractive summarisation and then generates different scores. This will be the final result of the project

**Results**

****

**Fig - 1**

**Initial Dataset**

**Original text**

Introduction

Anaemia is a common condition in critically ill patients, and RBC transfusions are often used in the treatment and management of this patient population. In fact, one study [1] reported that 25% of all critically ill patients received RBC transfusions. Many laboratory studies [2] [3] [4] [5] [6] [7] [8] have examined the physiological responses (ie compensatory mechanisms) of the body to anaemia, which include the following [9] : increased cardiac output, decreased blood viscosity, capillary changes, increased oxygen extraction, and other tissue adaptations to meet oxygen requirements. Although critically ill patients are affected by a number of factors that predispose them to the adverse consequences of anaemia, persistence of this condition is of particular concern because it may cause the compensatory mechanisms in these patients to become impaired, risking oxygen deprivation in vital organs [9] . However, critically ill patients may also be at increased risk from the adverse effects of RBC transfusions, such as pulmonary oedema from volume overload, immune suppression resulting in increased risk of infection, and microcirculatory injury from poorly deformable RBCs.It is the aim of the present commentary to justify the statement 'Transfusing to normal haemoglobin concentration will not improve outcome.' If we define normal haemoglobin as being greater than 115 g/l for women and greater than 125 g/l for men, then there is no evidence in the literature to justify maintaining such high concentrations by the use of RBC transfusions in any anaemic patient. There may, however, be some debate about adopting a transfusion threshold of 100 g/l, which is well below 'normal'. transfusion threshold would, obviously, reduce the number of allogeneic RBCs transfused. It is our goal to impress upon the reader that transfusing to a level equal to or greater than 100 g/l does not improve mortality and other clinically important outcomes in a critical care setting. We first explore the reasons why a reduction in the total number of allogeneic blood transfusions would be beneficial. Second, we examine the current evidence for using a lower transfusion strategy, specifically that employed in the Transfusion Requirements In Critical Care (TRICC) trial.

Potential consequences of allogeneic red blood cell transfusion

RBC transfusions have inherent risks that may be categorized as follows [11] [12] [13] [14] [15] : transfusion-transmitted infections; immune-related reactions (acute or delayed haemolytic reactions, febrile, allergic, anaphylactic reactions and graft-versus-host disease); and nonimmunerelated reactions (fluid overload, hypothermia, electrolyte toxicity and iron overload).Major improvements in donor screening procedures and laboratory testing have dramatically improved the safety of the blood supply [16] . Currently, the risk of transmitting an infectious agent through blood transfusion ranges from 1:100,000 for hepatitis B virus to 1:1,000,000 for HIV (Canadian Blood Services, personal communication, 2000). The most important threats to the blood supply remain new and unknown pathogens. More recently, concern has focused on the potential transmission of prions through RBCs. Also, infectious agents with long latency periods pose particular risks to young individuals who require RBCs, such as multiple trauma victims. The risk : benefit ratio for a 24-year-old trauma victim with a 50-year life expectancy differs markedly from that for a person aged 80 years who is undergoing coronary artery bypass surgery. In summary, because there is a risk of transmitting diseases through the blood supply, we should always strive to use RBCs according to the best available evidence in order to ensure that we do more good than harm to our patients.It is a long-standing observation [17] [18] [19] [20] [21] that blood transfusions are associated with immune suppression. This clinical phenomenon was first observed in renal transplant patients who had received blood transfusions while on dialysis before the transplant [22] , and has been observed repeatedly in transplant centres around the world [23, 24] . Recently, Opelz et al [25] reported a multicentre clinical trial in which all renal allograft recipients received modern immunosuppressive regimens. Those patients who were allocated to receive three allogeneic RBC units before renal transplant had a 1-year graft survival rate of 90%, as compared with 82% for patients who were not transfused (P = 0.02). These data suggest that there are long-term immunosuppressive effects following transfusion of nonleukocyte-reduced allogeneic RBCs.A large number of studies [26] [27] [28] [29] [30] [31] [32] [33] [34] have also suggested that allogeneic transfusions accelerate cancer growth, perhaps due to altered immune surveillance. These altered immune responses after allogeneic RBC transfusions may also predispose critically ill transfusion recipients to nosocomial infections [35] [36] [37] [38] [39] [40] and increased rates of multiplesystem organ failure [41] , which may ultimately result in higher mortality rates. However, most studies that examined the association between cancer recurrence and postoperative infection after transfusion [42, 43] only provided weak causal inferences because of poor study design and the lack of independence between allogeneic RBC transfusions and the potential complication.A recent meta-analysis [44] combined the results from seven RCTs, and was unable to detect clinically important decreases in mortality and postoperative infections. We added the results of a new RCT by van de Watering et al [45] to the above meta-analysis. The relative risk for allcause mortality was 1.05 (95% confidence interval 0.88-1.25), and was 1.10 (95% confidence interval 0.85-1.43) for postoperative infections. However, this meta-analysis excluded two positive RCTs [40, 46] because of the significant statistical heterogeneity introduced by these two studies. If all available RCTs are combined, ignoring heterogeneity, then the relative risk difference for postoperative infections across all studies is 1.60 (95% confidence interval 1.00-2.56; P = 0.05). Thus, the available evidence suggests that universal prestorage leukoreduction could have clinical effects that range from none to decreasing rates of infection by as much as 50% in high-risk patients. In summary, despite convincing laboratory evidence and some supportive clinical studies, the clinical significance of the immunosuppressive effects of allogeneic RBC transfusions have not been clearly established [47] . More importantly, the impact of a leukoreduction programme has not been studied in a large population of patients who are expected to have significant exposure to allogeneic RBCs.The majority of complications from allogeneic RBC transfusion, however, are no more frequent in the intensive care setting than in other patient populations, with the possible exception of pulmonary oedema, hypothermia and electrolyte disturbance. Hypothermia and electrolyte disturbances occur most frequently with massive transfusions. In critically ill patients, the optimal effective circulatory volume may be difficult to determine, and as a consequence pulmonary oedema may be a much more frequent complication of RBC transfusion than in other patient populations. This may explain the significantly higher rate of pulmonary oedema in patients transfused using a threshold of 100 g/l (5.3% in the restrictive transfusion group versus 10.7% in the liberal transfusion group; P < 0.01), as reported in the TRICC trial [10] . As an alternative explanation, the more frequent use of RBCs might have resulted in more frequent episodes of transfusion-related acute lung injury in the liberal strategy group (7.7% in the restrictive strategy versus 11.4% in the liberal strategy; P = 0.06), as reported in the TRICC trial.Clinical evidence is also insufficient to definitively establish a correlation between the age of RBCs being transfused and patient mortality; however, laboratory evidence has shown many storage-related changes that may result in impairment of blood flow and oxygen delivery at the microcirculatory level. Marik et al [48] demonstrated an association between a fall in gastric intramucosal pH and transfusion of RBCs stored for longer than 15 days. In addition, there is ample laboratory evidence that prolonged RBC storage adversely affects RBCs, potentially results in the generation of cytokines, and alters host immune function. In another study, Fitzgerald et al [49] , using an animal model of transfusion, consistently observed a lack of efficacy of transfused, stored rat blood to improve tissue oxygen consumption as compared with fresh cells or other blood substitutes.Three retrospective clinical studies tested the association between the age of transfused blood and duration of stay in the intensive care unit (ICU) [50] and mortality [51, 52] . Martin et al [50] observed a statistically significant association between the transfusion of aged blood (>14 days old) and increased duration of ICU stay (P = 0.003) in 698 critically ill patients. In patients who received a transfusion, aged RBCs was the only predictor of duration of stay (P < 0.0001). In survivors, only median age of blood was predictive of duration of stay (P < 0.0001). Purdy et al [51] demonstrated a negative correlation (r = -0.73) between the proportion of RBC units of a given age transfused to survivors and increasing age of RBCs in patients admitted to the ICU with a diagnosis of severe sepsis (n = 31). Those investigators also noted that these latter units were more likely to be older. A recently reported study by Vamvakas and Carven [52] evaluated the effect of duration of RBC storage on postoperative pneumonia in 416 consecutive patients undergoing coronary artery bypass grafting. Those investigators noted an adjusted increase of 1% in the risk of postoperative pneumonia per day of average increase in the duration of RBC storage (P < 0.005) in transfused patients. Each of these three studies also noted that patients who received a large number of RBC units had a higher mortality. Although these risks are relatively small when viewed collectively, they become significant when one considers that 25% of all critically ill patients in Canada are transfused during their ICU stay [1] .

Transfusion strategies

Until recently, physicians have depended on clinical judgement when deciding at what haemoglobin level to transfuse a critically ill patient. As a result, significant variation has been shown to exist in transfusion practice among Canadian critical care physicians [53] , which is due largely to a lack of published data on the subject. An arbitrary haemoglobin level of 100 g/l has historically been used as a threshold to transfuse critically ill patients.Six observational studies investigated the importance of anaemia on transfusion practices in various settings. Of these, three large cohort studies, which were performed in different patient populations (intensive care [1] , coronary artery bypass surgery [54] and hip fractures [55]), reached different conclusions. RBC transfusions in particular improved outcome in critically ill patients with cardiovascular disease, but increased the risk of myocardial infarction in coronary artery bypass surgery patients. Transfusion had no impact on short-term or long-term mortality in hip-fracture patients. Three smaller studies [56] [57] [58] evaluated the relationship between anaemia and adverse outcomes in vascular disease patients. Although increased numbers of ischaemic events were observed in anaemic patients, the validity of these studies is uncertain, given that the decision to transfuse a patient was often correlated with illness burden of the patient. It is also possible that comorbidity was not adequately adjusted for in those studies.Transfusion thresholds were compared in 10 randomized clinical trials [10, [59] [60] [61] [62] [63] [64] [65] [66] [67] . Although the clinical settings varied, each trial randomized patients to be transfused on the basis of a 'conservative' or a 'liberal' strategy. The definitions of conservative and liberal strategies varied, and actually overlapped between studies. Of these 10 trials, only three included more than 100 patients and only one trial evaluated the impact of transfusion on symptoms. In the first trial of patients undergoing elective coronary artery bypass surgery [65] , the differences between perioperative haemoglobin levels were small, event rates were very low, and there were no differences in any outcome. In the second trial [67] , patients undergoing knee arthroplasty were randomly assigned to receive autologous blood transfusion immediately after surgery or to receive autologous blood if haemoglobin level fell below 9 g/dl [67] . Again, no differences in outcome were observed. The third trial of hip fracture patients undergoing surgical repair [64] found no differences in outcomes; however, five deaths were recorded at 60 days after surgery in the symptomatic group, and two deaths occurred in the 10 g/dl group. The numbers of patients in these trials were too small to evaluate the effect of lower transfusion triggers on clinically important outcomes such as mortality, morbidity and functional status.In 1999, Hebert et al [10] reported the results of the TRICC trial. Patients (n = 838) were randomized either to a restrictive strategy (haemoglobin concentration maintained between 70 and 90 g/l, with a trigger set at 70 g/l) or to a liberal strategy (haemoglobin concentration maintained between 100 and 120 g/l, with a trigger at 100 g/l). To date, the TRICC trial is the only large study that has investigated these parameters. The groups were comparable at baseline. The average daily haemoglobin concentration ranged from 85 Â± 7.2 g/l in the restrictive group to 107 Â± 7.3 g/l in the liberal group (P < 0.01). The average number of transfusions was reduced by 52% in the restrictive group (2.6 Â± 4.1 versus 5.6 Â± 5.3 RBCs/patient; P < 0.01). Cardiac events, primarily pulmonary oedema and myocardial infarction, were more frequent in the liberal strategy (P < 0.01; Table 1 ). On examination of composite outcomes, the number of patients with multiorgan failure was found to be substantially increased in both groups, with the results being marginally better in the restrictive strategy group (20.6% versus 26.0%; P = 0.07; Table 2 ). Overall, the restrictive transfusion group showed a lower 30-day mortality (18.7% versus 23.3%; P = 0.11; Fig. 1 ). Kaplan-Meier survival curves, however, were significantly different in the subgroup of patients with an Acute Physiology and Chronic Health Evaluation II score of 20 or less (P = 0.02; Fig. 2 ). In addition, although 60-day mortality (22.8% versus 26.5%; P = 0.23) and ICU mortality (13.9% versus 16.2%; P = 0.29) were not statistically significant, they did show a consistent trend in terms of absolute values that favoured the restrictive strategy. The key observation from the TRICC trial is not that the restrictive strategy is better, but rather that it is at worst equivalent to the liberal strategy and at best superior to the liberal strategy.

Subgroups that are at increased risk from anaemia

At this juncture, preclinical and clinical evidence support the adoption of a more restrictive transfusion strategy in most critically ill patients. However, there remain divergent views regarding the risks and benefits of treating anaemia in patients with cardiovascular disease. Laboratory-based studies [68, 69] suggest that patients with cardiovascular disease may require higher haemoglobin concentrations to maintain oxygen delivery in partially occluded or diseased coronary arteries. Studies to demonstrate how anaemia affects contractile function of the left ventricle have rarely shown important effects above haemoglobin concentrations of 70 g/l. Indeed, it is more important to address the underlying pathophysiological causes of the acute coronary syndrome with proven therapy such as aspirin and Î²-blockers, rather than treating mild-to-moderate anaemia as an initial step. If the effects of RBC transfusion were either limited or increased then there would be no debate; however, the use of allogeneic RBCs has been shown to be associated with immunomodulation [12, 47] and/or alteration in the delivery of oxygen in the microcirculation [70, 71] , resulting in increased rates of infections and organ failure.Few clinical studies have attempted to elucidate the risk : benefit ratio of anaemia and transfusion in cardiac patients. Two small RCTs [62, 72] examined transfusion practice in patients undergoing coronary artery bypass grafting, and concluded that a conservative approach to the administration of RBCs may be safe. However, two recent cohort studies suggested that anaemia may increase the risk of mortality in critical illness [73] and following surgery in patients with cardiovascular disease [74] . There were 418 and 420 patients in the restrictive and liberal transfusion groups, respectively. \*Difference calculated by subtracting mean values of restrictive group from those of liberal group. â Three patients were lost to 60-day mortality rate; therefore n = 835. â¡ Nonsurvivors are considered to have all organs failing on date of death. Â§ Changes in MOD score from baseline, while also incorporating adjustment for death. Data from HÃ©bert et al [10] .In a study of Jehovah's Witnesses (a group that refuses RBC transfusion on religious grounds) undergoing surgical procedures [74] , it was noted that mortality was significantly increased in patients with cardiac disease after a decrease in haemoglobin levels from 100-110 g/l to 60-69 g/l. In that study, patients with no cardiac disease and similar changes in haemoglobin levels showed no increase in mortality, which is in accordance with the results of the TRICC trial [10] . In the study by HÃ©bert et al. [73] of 4470 critically ill patients, a correlation between Critical Care Vol 5 No 2 Alvarez et al [10] .

Figure 1

Kaplan-Meier estimates of survival in the 30 days after admission to the ICU in the restrictive and liberal transfusion strategy groups (all patients). Data from HÃ©bert et al [10] .

Figure 2

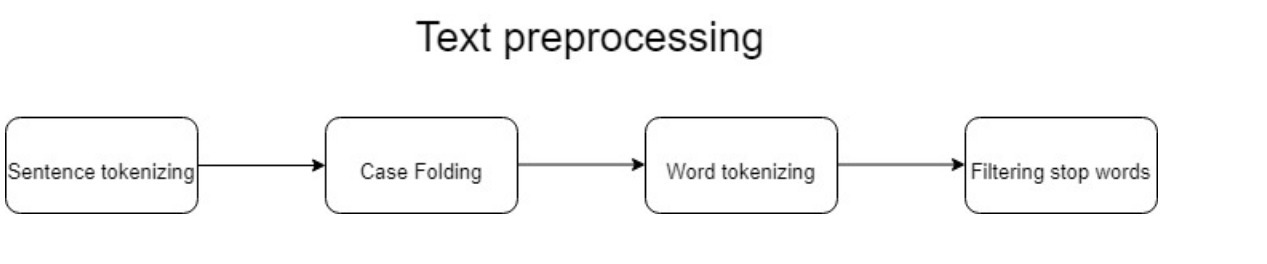
Kaplan-Meier estimates of survival in the 30 days after admission to the ICU in the restrictive and liberal transfusion strategy groups (patients with APACHE II score â¤20). Data from HÃ©bert et al [10] .anaemia and mortality rates was observed. Those investigators also found that the risk of anaemia appeared to decrease with RBC transfusion in patients with cardiac disease. In patients with cardiac disease, mortality increased when haemoglobin concentrations were below 95 g/l, as compared with anaemic patients with other diagnoses (55% versus 42%; P = 0.09). In the subgroup of patients with cardiac disease, increasing haemoglobin values in anaemic patients was associated with improved survival (odds ratio 0.80 for each 10 g/l increase; P = 0.012). One possible explanation for the discrepancy between the TRICC trial and this observational study may be that the attending physicians who recruited patients into the study did not enter those patients who were considered to have severe cardiac disease.HÃ©bert et al. [73] sought to examine further whether a restrictive transfusion strategy was at least as effective as a liberal strategy in critically ill patients with cardiac disease. In the subgroup of patients with cardiovascular disease from the TRICC trial, those investigators suggested that most haemodynamically stable critically ill patients with cardiovascular disease may be transfused when haemoglobin concentrations fall below 70 g/l, and that the hemoglobin concentration should be maintained between 70 and 90 g/l. Average daily haemoglobin concentrations were 85 Â± 6.2 g/l in the restrictive transfusion group and 103 Â± 6.7 g/l in the liberal transfusion group (P < 0.01). In the 357 patients with cardiovascular disease, the 30-day mortality rate was 23% in the restrictive transfusion group versus 23% in the liberal group (95% confidence interval of the difference -8.4% to 9.1%; P = 1.00). Other mortality rates, including 60-day (26% versus 27%; P = 0.90), ICU (19% versus 16%; P = 0.49) and hospital mortality (27% versus 28%; P = 0.81), were not significantly different between groups. Kaplan-Meier survival curves comparing time to death demonstrated similar trends in the two groups ( Fig. 3 ; P = 0.98). The multiple organ dysfunction (MOD) scores, during the entire study period, were also not significantly different between groups (8.6 Â± 4.9 versus 9.0 Â± 4.4; P = 0.40), but the change in MOD score from baseline values was significantly lower in the restrictive group than in the liberal group (0.2 Â± 4.2 versus 1.3 Â± 4.4; P = 0.02).Combined measures of morbidity and mortality, or composite outcomes, were also examined. When all patients who died were given a score of 24, the total MOD score between groups was not different (P = 0.39), or were the changes in MOD scores significantly different from baseline (2.7 Â± 6.9 versus 4.0 Â± 7.3; P = 0.08). Among the specific subset of cardiac patients with ischaemic heart disease (n = 257), there were no discernible differences in 30-day and 60-day as well as ICU mortality rates. However, a nonsignificant (P = 0.3) decrease in overall survival rate in the restrictive group was noted in those patients with confirmed ischaemic heart disease, severe peripheral vascular disease or severe comorbid cardiac disease (Fig. 4) .In conclusion, a restrictive RBC transfusion strategy generally appears to be safe in most critically ill patients with cardiovascular disease, with the possible exception of patients experiencing acute myocardial infarction or unstable angina.

Figure 4

Survival over 30 days in patients with ischemic heart disease in the restrictive and liberal allogeneic RBC transfusion strategy groups. This graph illustrates Kaplan-Meier survival curves for all patients with ischemic heart disease in both study groups. There is no difference in mortality in patients in the restrictive group (dashed line) as compared to the liberal group (solid line) (P = 0.30).

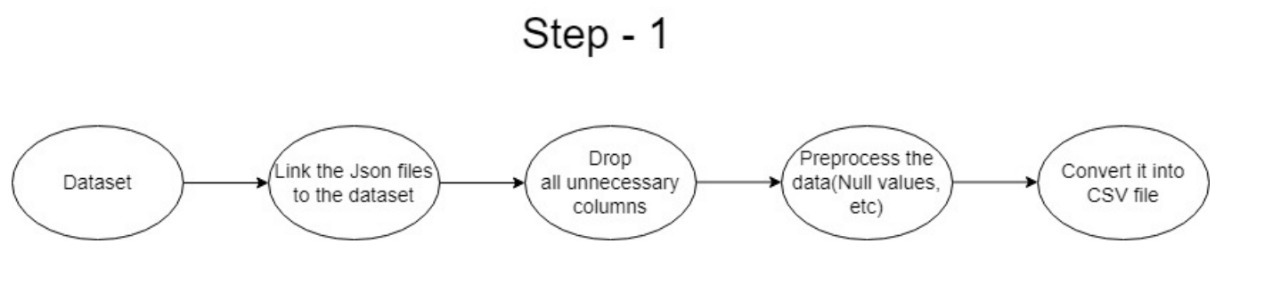
Conclusion

The need to reduce the amount of allogeneic blood transfusions in order to reduce the associated risks has been firmly established. RBCs are associated with clinically important immune suppression, and stored RBCs have been shown to cause adverse microcirculatory effects that result in increased organ failure.The question for some time has been whether critically ill patients are able to tolerate lower levels of haemoglobin without deleterious effects, thus reducing the amount of exposure to allogeneic transfusions. In the only large RCT, HÃ©bert et al [10] established that there was no difference in mortality rates between restrictive and liberal transfusion strategies in noncardiac, critically ill patients. Although those investigators were able to show convincing trends that the liberal strategy may in fact be deleterious in terms of absolute values, statistical significance was not achieved. However, the fact that no difference between the two strategies was achieved is of great importance, because this means that the total number of transfusions can be reduced by approximately half without any impact on mortality. In addition, these findings are easily put into clinical practice. Although many questions remain, the TRICC trial and many laboratory and clinical studies have established that transfusing to normal haemoglobin concentrations does not improve organ failure and mortality in the critically ill patient. As such, a restrictive transfusion strategy will reduce exposure to allogeneic transfusions, result in more efficient use of RBCs, save blood overall, and decrease health care costs.

****

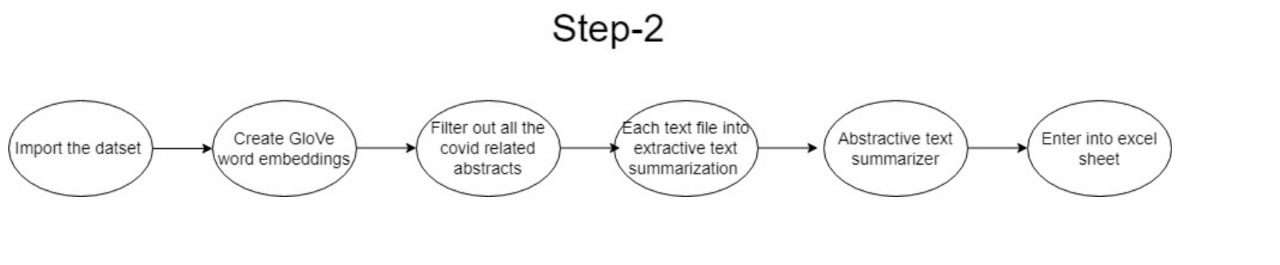
**Fig - 2**

**Textual data Pre-processing**

****

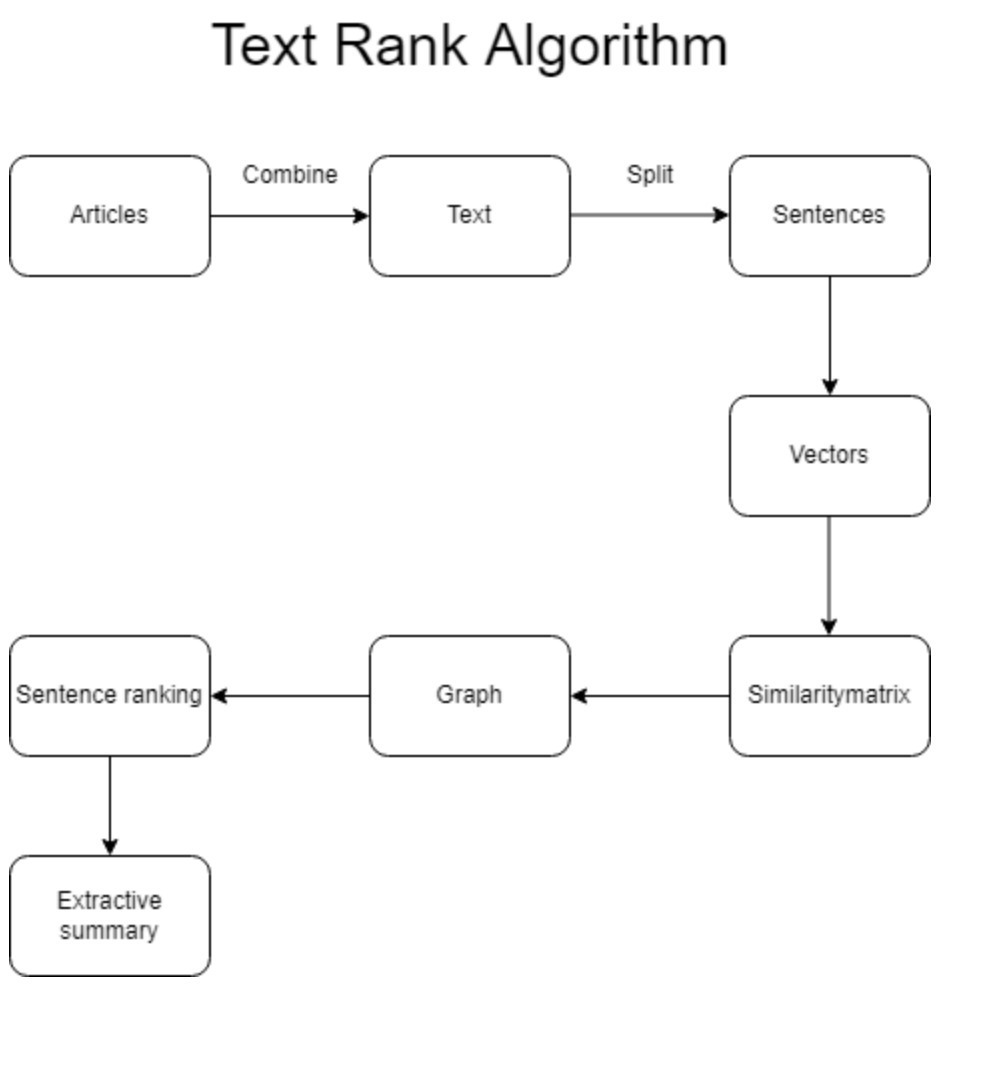
**Fig - 3**

**Pre-processing Step 1**

****

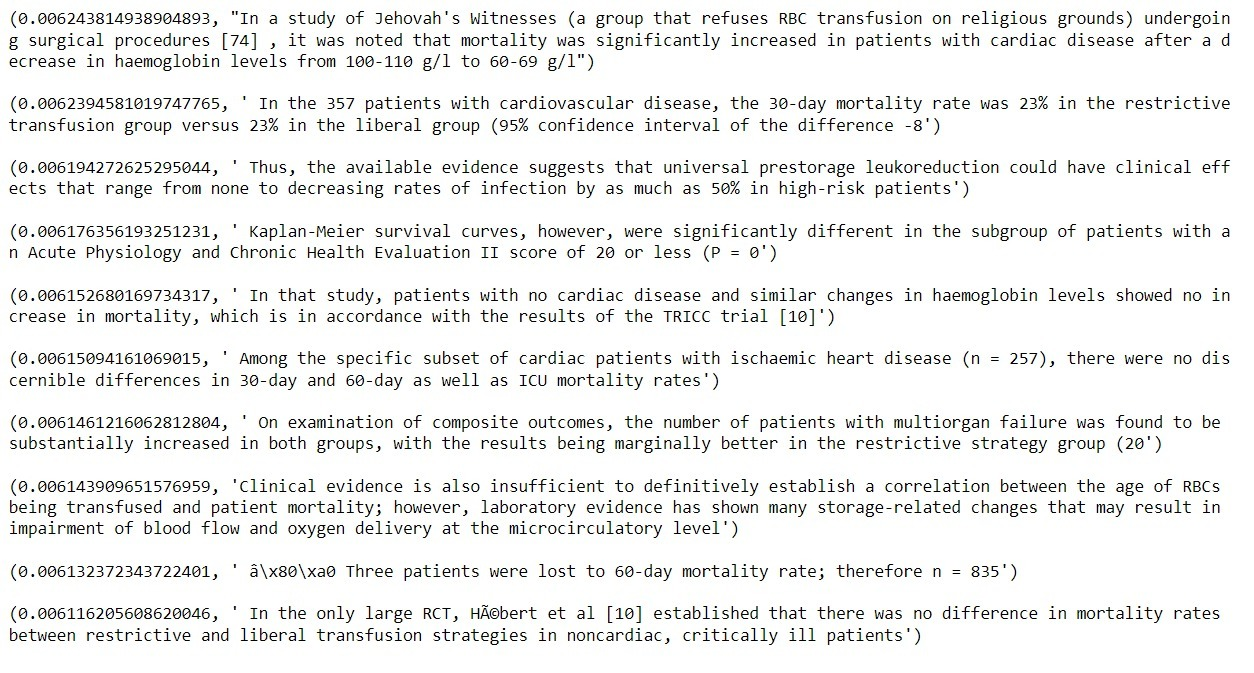
**Fig - 4**

**Pre- processing Step 2**

****

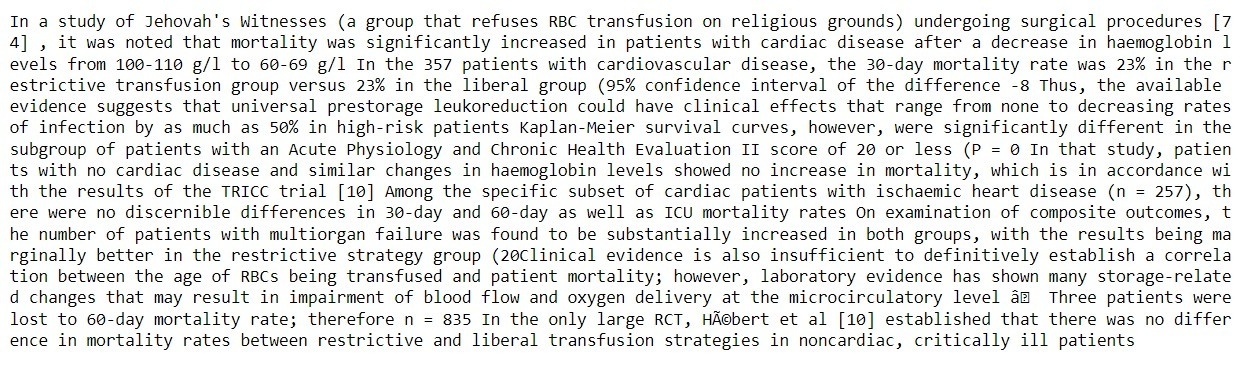
**Fig - 5**

**Text Rank Algorithm**

****

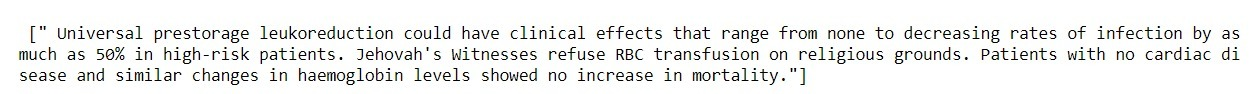
**Fig - 6**

**Ranked sentences**

****

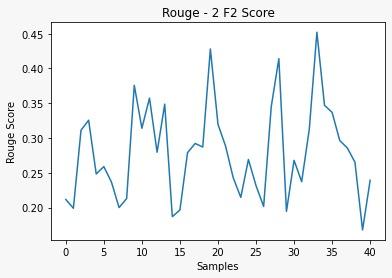
**Fig - 7**

**After extractive summarization**

****

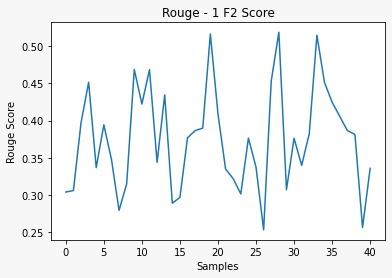
**Fig - 8**

**After abstractive summarization**



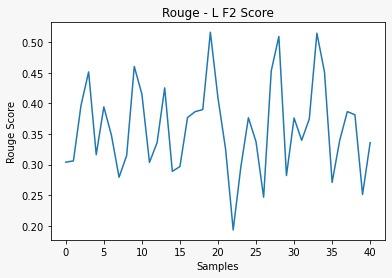
**Fig - 9**

**Rouge - 1 F1 score**



**Fig - 10**

**Rouge - 2 F1 score**



**Fig - 11**

**Rouge - L F1 score**

**Limitations**

The model we're using for abstractive text summarisation only supports summarising 1024 words at a time. Furthermore, because the data set is so large, it is difficult to summarise all 3,00,000 abstracts at once. To complete such complex tasks, we require a supercomputer.

**Conclusion**

Text summarization is in high demand due to the increase in data around the world. Because it is difficult for humans to summarise large amounts of data manually, we require a tool to do so. In this study, we used the Text rank algorithm and the Hugging face to implement both extractive and abstractive summarization techniques. With this model, we can also get the most important part of the text.

**References**

1. <https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715>
2. [https://metatext.io/models/sshleifer-distilbart-cnn-12-6](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
3. [https://www.aaai.org/Papers/AAAI/2005/ISD05-010.pdf](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
4. [https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
5. [https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
6. [https://towardsdatascience.com/text-classification-with-hugging-face-transformers-in-tensorflow-2-without-tears-ee50e4f3e7ed](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
7. [https://www.proquest.com/openview/e78c4bdbc451be6d8eebbf5ec822bee2/1?pq-origsite=gscholar&cbl=54626](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
8. [http://web.science.mq.edu.au/~len/preprint/yousefiazar-eswa-2016-preprint.pdf](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)
9. [https://analyticsindiamag.com/hands-on-guide-to-word-embeddings-using-glove/](https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715)