Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm

by

Mofiz Mojib Haider
16301038
Md. Arman Hossin
17301214
Hasibur Rashid Mahi
16301035

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
April 2020

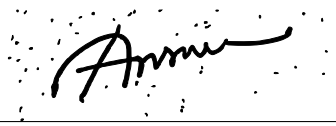
Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Mofiz Mojib Haider
16301038



Md. Arman Hossin
17301214

Hasibur Rashid Mahi
16301035

Approval

The thesis/project titled "Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm" submitted by

1. Mofiz Mojib Haider (16301038)
2. Md. Arman Hossin (17301214)
3. Hasibur Rashid Mahi (16301035)

Of Spring, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 07,2020.

Examining Committee:

Supervisor:

Hossain Arif

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:

Rabiul

Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:

Prof. Mahbub Majumdar
Chairperson
Dept. of Computer Science & Engineering
Brac University

Mahbub Alam Majumdar
Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

The significance of text summarization in the Natural Language Processing (NLP) community has now expanded because of the staggering increase in virtual textual materials. Text summary is the process created from one or multiple texts which convey important insight in a little form of the main text. Multiple text summarization technique assists to pick indispensable points of the original texts reducing time and effort require reading the whole document. The question was approached from a different point of view, in a different domain by using different concepts. Extractive and abstractive are the two main methods of summing up text. Though extractive summary is primarily concerned with what summary content the frequency of words, phrases, and sentences from the original document should be used. This research proposes a sentence based clustering algorithm (K-Means) for a single document. For feature extraction, we have used Gensim word2vec which is intended to automatically extract semantic topics from documents in the most efficient way possible.

Keywords: Text summarization, Extractive, Single Document, NLP, Gensim, Word2Vec, K-Means.

Dedication

We dedicate this thesis to our loving parents.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption.

Secondly, we want to thank our advisor Hossain Arif sir to guide us and support us in a very well manner which helps us to finish the thesis at a particular time.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer, we are now on the verge of our graduation

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$BLEU$ Bilingual Evaluation Understudy

FCM Fuzzy C Means

LDA Latent Dirichlet Allocation

LSA Latent Semantic Analysis

NLP Natural Language Processing

$NLTK$ Natural Language Toolkit

Chapter 1

Introduction

1.1 Introduction

The subfield of text summarization has increased over the half-century past. DR Radev [7] a text generated from one or even more texts conveying vital information in the actual document, not more than half of the actual document and generally less than that. Redundant texts exist in these daily generated documents and the size of the documents is enlarging bit by bit.
It's convenient for people to summarize the document and bringing out the implicated meaning of that particular document whereas the machine can't resolve the same problem as efficiently as expected that is why various methods of summarization have been tested to bring out the best possible outcome. However, a universal strategy of the summarization is not available yet.
The summarized document reflects the important aspects of the large text. Different text summarization technique has been implied over time. An extractive approach of summarizing is to pick relevant sentences, paragraphs, etc. from the actual document and concatenate them towards a simplified form. The meaning of sentences is determined based on the numerical and linguistic characteristics of sentences [13]. On the other hand, abstract text summarization systems create new sentences, likely rephrasing by using terms, not in the original document [18]. There are some other unconventional approaches like sequential document, sentence compression summary has been applied which could be useful in future research of the text summarization.
The purpose of that research is to summarize the single document through sentence based model using clusters, whereas using Gensim word2vec for features extraction for the sentence-based model evaluates for figuring out the main ideas through all the sentences in the text.

1.2 Thesis Orientation

The rest of the report is as follows. Chapter 2 contains the Literature Review which says about the previous work done on text summarization and the algorithms which are used to do that. Then, Chapter 3 contains the Proposed Model where the system has been described properly which apprehends System Overflow Dataset, Word2Vec, Gensim Word2Vec, K-Means CLUSTERING, Elbow method to find K, Summary Extraction. System Overflow represents the model which shows graphically how the system works. Versatile datasets have been brought to implement with various algorithms. Word2Vec is one kind of embedding system in neural techniques. Gensim Word2Vec is used to make a word vector from the tokenizer list. K-Means CLUSTERING makes clusters from the clustered sentence. It is an iterative process. At last, the summary is extracted from the whole text by selecting the most valuable sentence of the higher frequency cluster. Furthermore, the algorithms are described properly which has been used. Chapter 4 represents the Result Evaluation based on the several kinds of datasets that are processed by the algorithms. Moreover, the tables of the result part show the different scores for different kinds of articles. Furthermore, it shows the comparison between the results of the versatile datasets. Finally, chapter 5 concludes the whole report with a summary of the work and future works.

1.3 Motivation

A lot of writing is produced every day. Both papers can be read or reviewed less often. People tend not to read a huge amount of content. Therefore, the whole essay takes a lot of time, which is why people move to a condensed version of the original post. The vast size of texts is generated every day. There are redundant texts created daily.

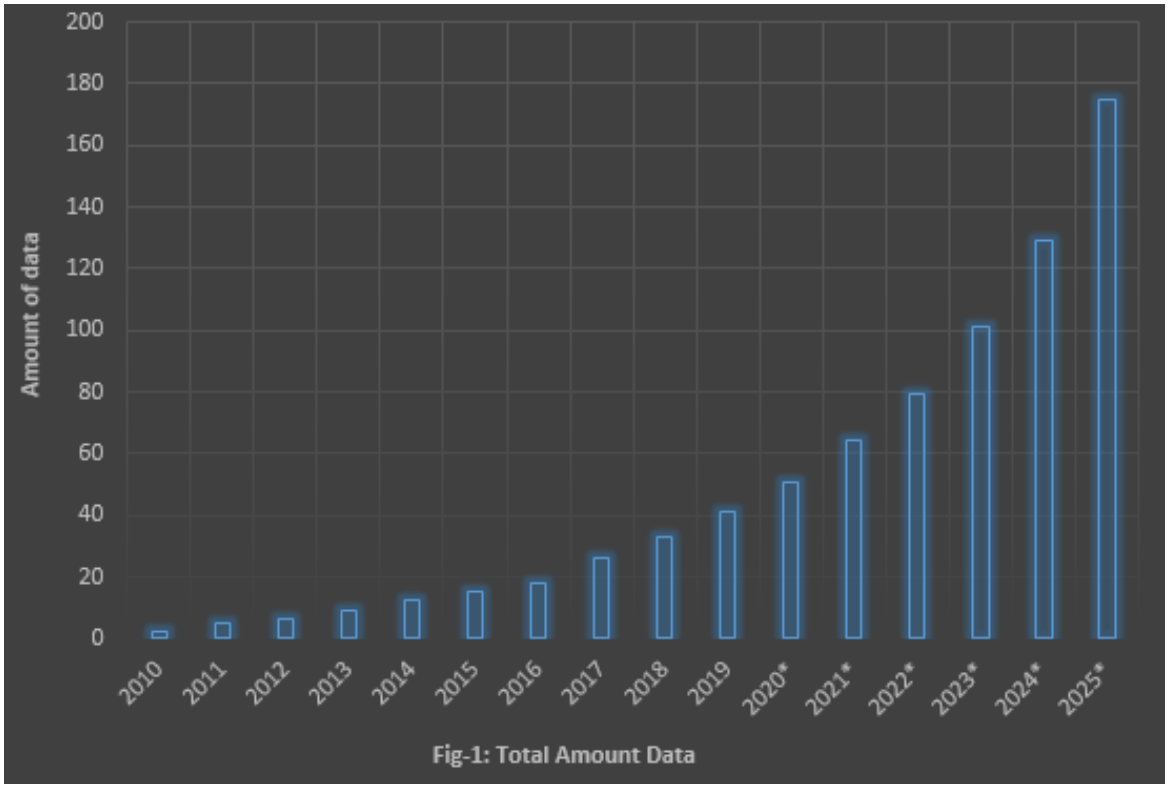


Figure 1.1: Information created worldwide from 2010 to 2025

The total amount of data is increasing gradually. In 2019, the total amount of data is 41 zettabytes [3]. According to statistics, this amount will become 175 zettabytes. Every year data generation will be increased more than the previous year. This amount of data will need to be stored and will have to be retrieved every time which is fluttering and time lessening. So the data needs to be transformed into a small one. Summarization is the way to make the writing in a short version. Since a large amount of data is in a form of articles, newspapers, research papers, journals so if these writings convert into shorter ones then it is easy to read and there will be space for upcoming data.

Summarization is really tough because it is difficult to find the main theme of the topic. Summarization is possible in many ways by focusing on various topics of the content. It is not easy to fetch out the gist of the contents. Extractive text summarization just finds out the important word and makes a summary of the text where it does not add extra words. That is why extractive text summarization is average. On the other hand, by understanding a topic and write the summary with new words is abstractive summarization. Automatic generated abstractive summarization is not up to the mark because it does not summarize the text according to the main topic rather it follows an algorithm to summarize which is implemented. Therefore, the

abstract summary of the text does not yield better results. This creates a problem in comparing the summary created by the computer to the one generated by humans. Another problem is that the main theme may not be understood from the automatic text summarization. The machine finds important words through the frequency of the words. If the frequency of any word is high then the machine considers it as an important word and uses it at the time of summarization. But maybe the word is not needed for summarization. So, automatic text summarization needs to be upgraded.

1.4 Challenges

Several works had been performed on text summarization of the English language. However, the first challenge was that the summarization was not performed well with the title of the article. It was not given an exact result. So, the title of the article had to be removed.
Another challenge that the system could not recognize special characters like ($,*,,@), etc. These types of special characters generated complications in text summarization. It could not identify these special characters and could not use them to make a summary of the text. Moreover, repetition of a word in a single sentence was a challenge too. As a summary is a condensed version of any article when summing up the article it created a problem when every word came in a sentence multiple times.
Making the K-Means cluster was also a big challenge. K-Means cluster iterative algorithm so every time it made different clusters before making a summary. So it was difficult to make a cluster each time before preparing a summary for the article. Each time the clusters of the word had been changed and it gave different scores for the summary.

Chapter 2

Literature Review

Back in 1958 the very first text summary technique was implemented. Rafael Ferreira et al. [11] addressed the method of extractive text summarization using various sentence scoring method. The proposed model is based on tokenizing the words and scoring the words to identify the importance of the document. They have taken CNN, Blog Summarization and SUMMAC these three types of datasets to test the algorithm. They have followed three approaches word based scoring, sentence scoring, and system of graph scoring and compare these methods to identify which method is the best for text summarization for particular datasets.

Kupiec et al. [5] constructed a trainable summarization program based on statistical classification. They build a classification method that calculates a given sentence's likelihood using Bayes's rule. They used Frequency-Keyword, Tittle-Keyword, and Location as a heuristic.

Anam et al. [19] suggested a model based on sentences using Fuzzy C-Means Clustering Algorithm. FCM uses fuzzy sets and fuzzy subset matrix to predominate the relation among various cluster elements.

Das, D. and Martins, A. F. [10] showed a single document summarization and multi-document summarization using ex-tractive and abstractive text summarization approach. Where there are various algorithms has been applied like Naıve-Bayes method, Rich features and Decision trees, Hidden Markov methods and Long Linear models and manifest the performance depending on data set was implied.

Romain Paulus et al. [18] proposed a deeply strengthened model for summarizing abstract texts. Neural network model with a novel intra-attention has been used over input. Basic word prediction blends with reinforcement learning of training in global sequence prediction to make the description more legible.

Vishal Gupta et al. [13] submitted a survey on extractive text description techniques. Features of extractive text summa-rization like title word feature, sentence location feature, cue-phrase feature along with cluster-based, LSA, neural network methods are also explained in detail.

Bofang li et al. [21] they heuristically develop a Word2Vec variant to ensure that each pair of terms comprise a non-based word and a universally sampled descriptive term. They "freeze" the batch context and only adjust the insignificant part to resolve conflicts. This change also explicitly monitors training iterations by determining the number of samples and manages the high frequency and low-frequency conditions. Detailed experiments are carried out in a variety of NLP activities. The results show that your model is 7.5 times precise on 16 GPUs.

Rene Arnulfo Grcıa-Hernandez et al. [11] they suggest an automated text review solution using an unsupervised learning algorithm by phrase extraction, independent of the language and domain. Their theory is that an algorithm unchecked will help to bring these ideas (sentences) together. The most descriptive term is then selected from each cluster to assemble the list. Many studies in the DUC-2002 model set indicate that the method proposed produces better results than any other solution.

Joel Larocca Neto et al. [8] they offer a description method based on the use of a machine learning algorithm that uses a set of features taken directly from the original text. These characteristically features are two: mathematical-based upon the occurrence of certain text elements and textual-from the simplistic statements of the language. They also show some computer results obtained by using their summarizer on a few well-known text bases, and we compare these results with some basic summary procedures.

Jincheng et al. [22] extensive studies on different benchmarking text classification datasets were performed to provide clear explanations of how fastText renders score matrix predictions, and a comprehensive analysis of frequency matrix based methods is given.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, Bing Xiang [17]shows abstract text summarization utilizing Attentional Encoder Decoder Recurrent Neural Networks and demonstrates that two separate companies produce state-of - the-art outputs. Some novel models, such as modelling main terms, capturing the hierarchy of the sentence-to-word structure, emitting terms that are uncommon or unrevealed at the time of training are proposed in revived problems which are not properly modelled on simple architecture.Research shows that all of our projects are contributing to greater efficiency enhancement.

Wesley T. Chuang and Jihoon Yang [6] they propose an extractive text summarization by segmenting the sentence using machine learning algorithms. Their approach is mainly segmenting the sentence with special cue markers. These segments have a set of specified functions. On the base of these functions, the algorithm trains the summarizer and gets the actual word to summarize the text.

Ibrahim F. Moawad and Mostafa Aref [15] throughout this paper they propose a modern method, utilizing a rich semantical graph methodology, to construct an analytical description for a single text. The method describes the input text, producing for the initial text a rich semantical graph, minimizing the resultant map and instead generating the abstractive description. In comparison, a hypothetical

case analysis reveals how the initial texts were limited to 50%.

Anjali R. Deshpande and Lobo L. M. R. J. [16]they represent text summarization using clustering methods. This approach works well in the multi-document text. It clusters the similar document into a group. Then from the similar document cluster, it makes the sentence cluster of each group. After that in the final summary, the best sentence scores from sentence clusters are selected

Chapter 3

Proposed Model

The proposed model of this paper is mainly a sentence based clustering approach to summarization a news article it is demonstrated that sentence based models are more efficient than graph and word-based modes [4]. At the very primary stage, a news article has been selected from the dataset and undergoes numerous pre-processing procedures. During preprocessing, the model will perform sentence tokenize, remove special characters, word tokenize, duplicate word remove and finally lemmatization to get the root word.

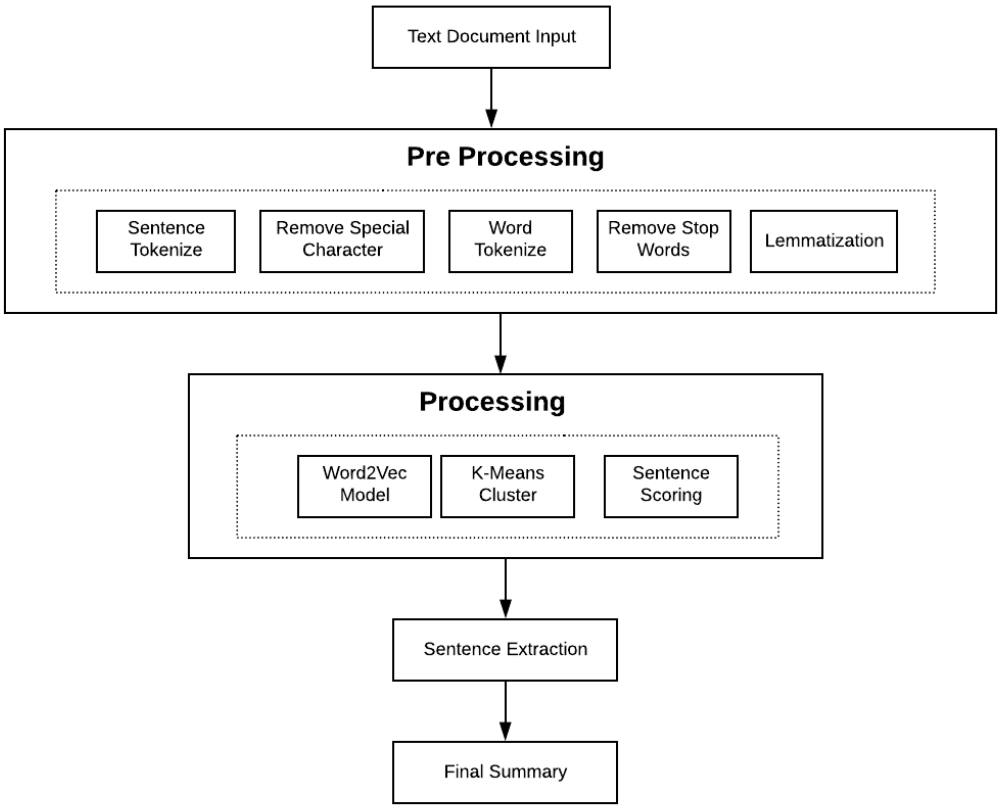


Figure 3.1: Flow diagram

After completion of the preprocessing, the model will perform the feature extraction process to score each sentence of the text. The model used Gensim Word2Vec to generate a vector representation of the text. Then, the model has distributed the vectorized sentence into k clusters based on the clustering algorithm K-Means where the number of clusters is k. To determine the perfect value of k, this model has used the Elbow method.
Finally, the model will generate a summary by picking up some important sentences from those clusters based on the score of each sentence given by our processing algorithm. The generated summary will be one-third of the given text.

3.1 Dataset

BBC news article dataset [9] contains 2225 news articles divided into 5 categories (business, entertainment, politics, sport, and tech). Randomly 10 news articles from each category total 50 news articles have been chosen for processing.

3.1.1 Preprocessing

Preprocessing is required to convert the data into a machine-readable form of the vector.

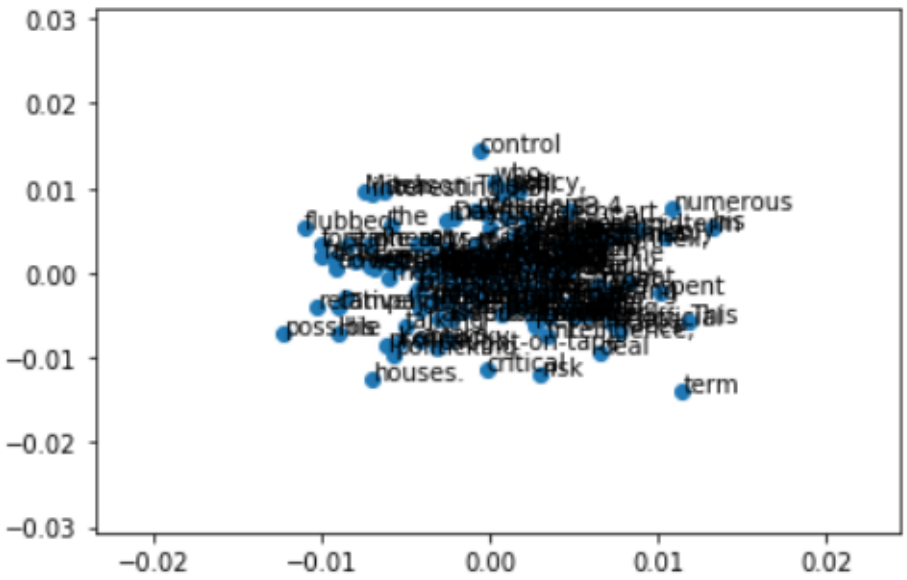


Figure 3.2: Unique words from the text

1. Sentence Tokenization: It is the process to split the text into sentences [19] [20]. Sentence tokenizer from NLTK library python was used to split the sentences.

2. Remove Special Character: It is possible that text may contain some unnecessary characters. All those unnecessary characters have been removed.

3. Word Tokenization: Each of the sentences of the article has been split into words by using word spaces [19] [20].

4. Removal of stop words: Stop words are those words that will be ignored while processing the text. All the words from the text which are considered as stop words have been removed [19] [20].

5. Duplicate Word Removal: Words from each sentence that occur more than once have been removed except keeping it once.

6. Lemmatization: It's a method of finding the root of every word. The text's words have all been lemmatized [20].

3.1.2 Word2Vec

In neural network one of the most common word embedding techniques is Word2Vec. First of all, a vector representation for each word at a certain length where the vector would consist of zeros except for the element representing the words. The words that have similar meaning take a closer spatial position [2].

$$sin(X, Y) = cos(\theta) = \frac{(X \cdot Y)}{||X||||Y||} \tag{3.1}$$

It can be implemented in two ways, one is Skip Gram and another one is CBOW (Common Bag of Words).

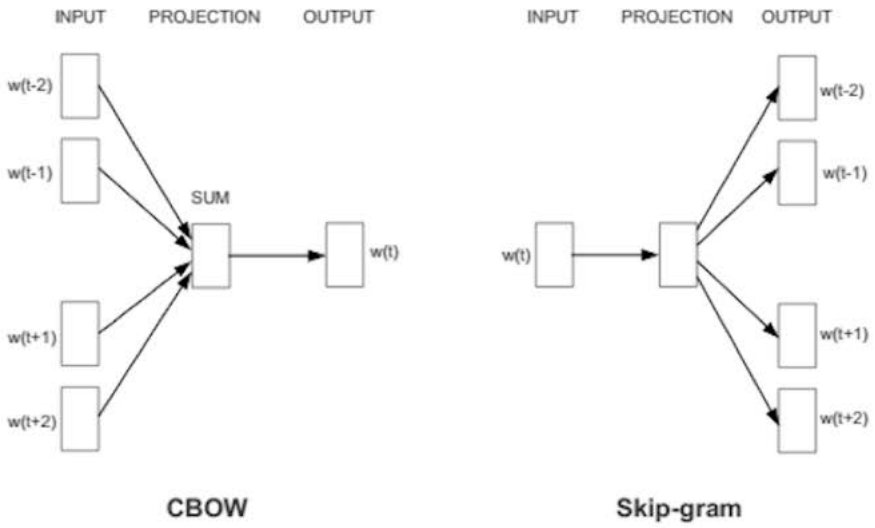


Figure 3.3: CBOW & Skip-Gram

In our research, Skip Gram techniques have used as it works better for the small amount data and for the words those are not that much common. As this process gives input the word, it gets output of probability distributions for each vector length for every single word where the backpropagation method is used to deal with it.

3.1.3 Gensim Word2Vec

Gensim is a very popular open-source library for unsupervised learning implemented in python[14]. Gensim implements Word2Vec based on Latent Dirichlet Allocation (LDA). Gensim Word2Vec has been used to generate a word vector from the tokenized word list. It is easy for a machine to some vectors instead of some words and it is also easy to implement mathematical operations on the vector.

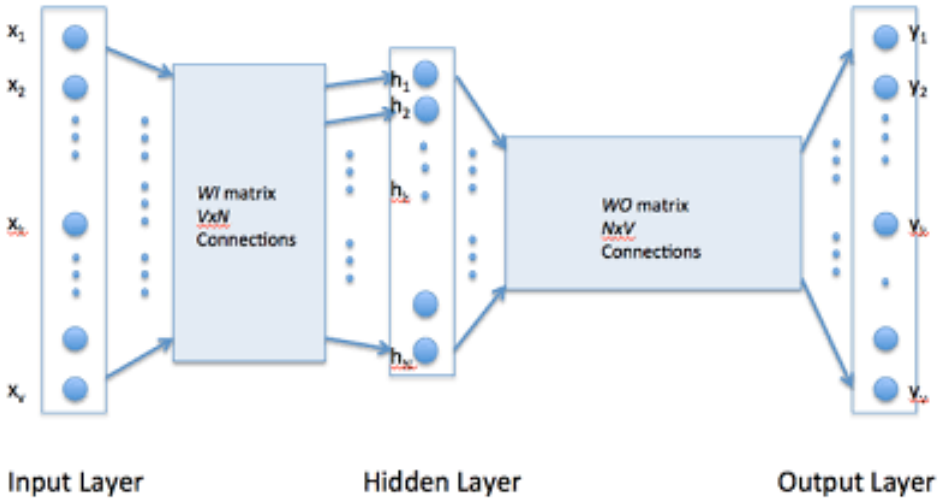


Figure 3.4: Gensim Word2Vec

```
[ 1.90746825e-04 -3.24315298e-03  8.65199661e-04 -4.65938076e-03
  1.03452876e-05  4.51342668e-03 -1.29584543e-04 -3.90595477e-03
 -7.67513935e-04  2.81555788e-03 -2.94687226e-03  1.23590464e-03
 -2.67795194e-03 -2.96570058e-03 -2.29436625e-03  1.38930464e-03
 -1.29379798e-03  3.23758274e-03 -2.09732633e-03 -3.03218770e-03
 -1.73784239e-04  2.48856866e-03 -8.82992579e-04 -1.41982362e-03
  4.87834634e-03  3.02469381e-03 -3.25092697e-05  8.97264836e-05
  1.64743897e-03 -4.67915274e-03  3.14930221e-04 -2.66643148e-03
 -1.16532785e-03  4.64198831e-03 -4.55797743e-03 -1.66505959e-03
 -1.41585351e-03  1.49267050e-03  1.49881328e-03  1.65657047e-03
  4.26740199e-03 -3.62392841e-03 -5.12270897e-04 -3.75866517e-03
 -4.05161688e-03 -3.63882608e-03  6.38694735e-04  1.93018839e-03
 -3.40933655e-03  2.91059585e-03  4.92762914e-03 -1.18519133e-03
  3.75425187e-03 -3.13054118e-03  4.59842989e-03  3.72342253e-03
 -2.81402655e-03  2.91750696e-03 -2.43111583e-03 -4.76850988e-03
 -4.92900331e-03 -3.98270739e-03 -4.53673908e-03 -1.19835015e-04
  5.81800879e-04  7.73843320e-04 -1.01654651e-03  3.29629355e-03
  4.66273027e-03 -3.35133821e-03 -3.20546725e-03 -1.32410962e-03
 -1.89637532e-03 -3.61338747e-03  4.93613118e-03  4.12057743e-05
  2.63816048e-03  2.18735402e-03 -2.65997672e-03 -4.67958534e-03
 -1.95616251e-03 -4.45109839e-03  2.14213855e-03  4.13563946e-04
 -1.48614869e-03  8.26021947e-04  3.97148263e-03 -3.00235988e-04
  2.31247753e-07 -2.61443271e-03  3.70247453e-03 -4.77531087e-03
 -1.07507408e-03  9.35588090e-04  4.75564227e-03  1.94002874e-03
  4.44599474e-03  2.99110147e-03  1.08190160e-03  2.52339896e-03]
```

Figure 3.5: Vector representation of the text

3.1.4 K-Means CLUSTERING

Clusters means a set of aggregated data points having certain similarities. K-Means is an iterative algorithm where it divides the dataset into distinct clusters keeping each data points in one group. K-Means aim is to reduce the square distance summation between data points and their respective cluster centers.
Algorithmic representation of K-Means[12]:
Let M = {m1,m2,m3,. . . . . ..,mn} be the Data points collection and V = {v1,v2,. . . . . .,vc} are the centers.

1. Select Cluster Centers 'c' by random selection.
2. The difference between individual data points and cluster centers is determined.
3. Allocate the cluster center data point with a minimum distance from the cluster center of all cluster centers.
4. Calculate the new center of clusters using:

$$v_i = \frac{(1)}{(c_i)} \sum_{j=1}^{c_i} m_i \tag{3.2}$$

5. Recalculate the distance between each data point and newly obtained cluster centers.
6. If there is no reassigned data point then, otherwise repeat step no 3.

3.1.5 Elbow method to find K

It is very much essential to determine the perfect value of k to get the best outcome from the K-Means algorithm. Elbow method is one of the most popular methods to determine the value of k which represents the number of clusters will be used in this model. In this model, the iterative range for k is 1 to 9.
The steps of Elbow method [20]:

- K starts from 1 to 9
- Increase k by 1
- Measure the distortion
- The point after which the distortion begin to decrease in a linear line.

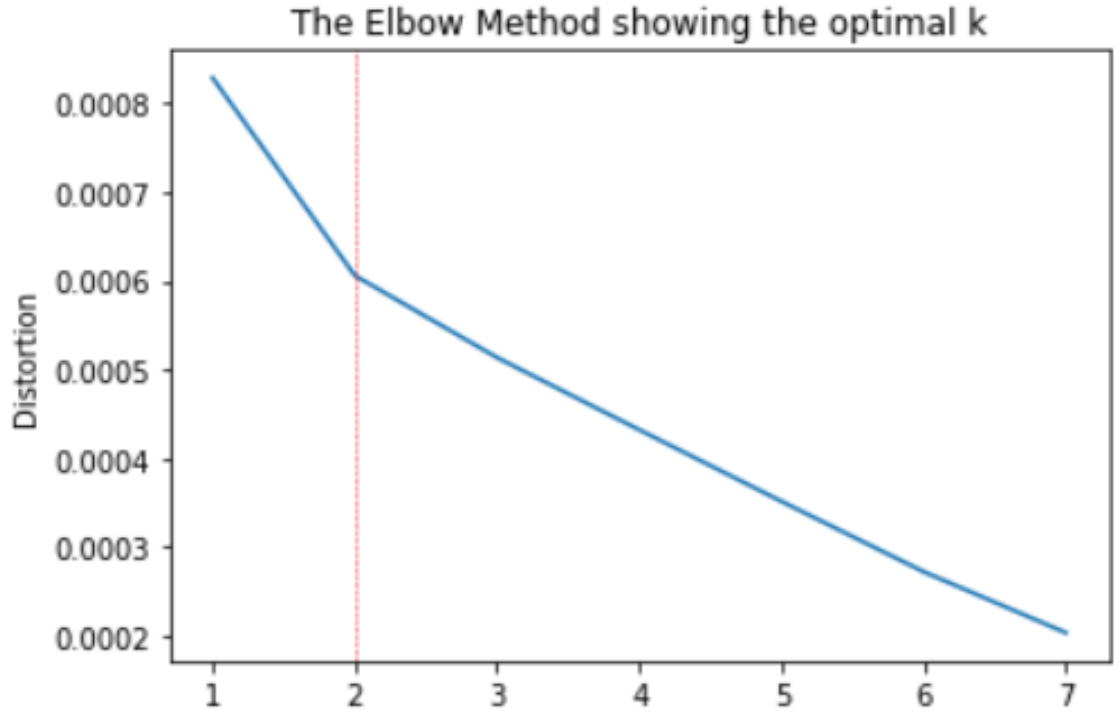


Figure 3.6: The Elbow with k=2 in Business Article 465

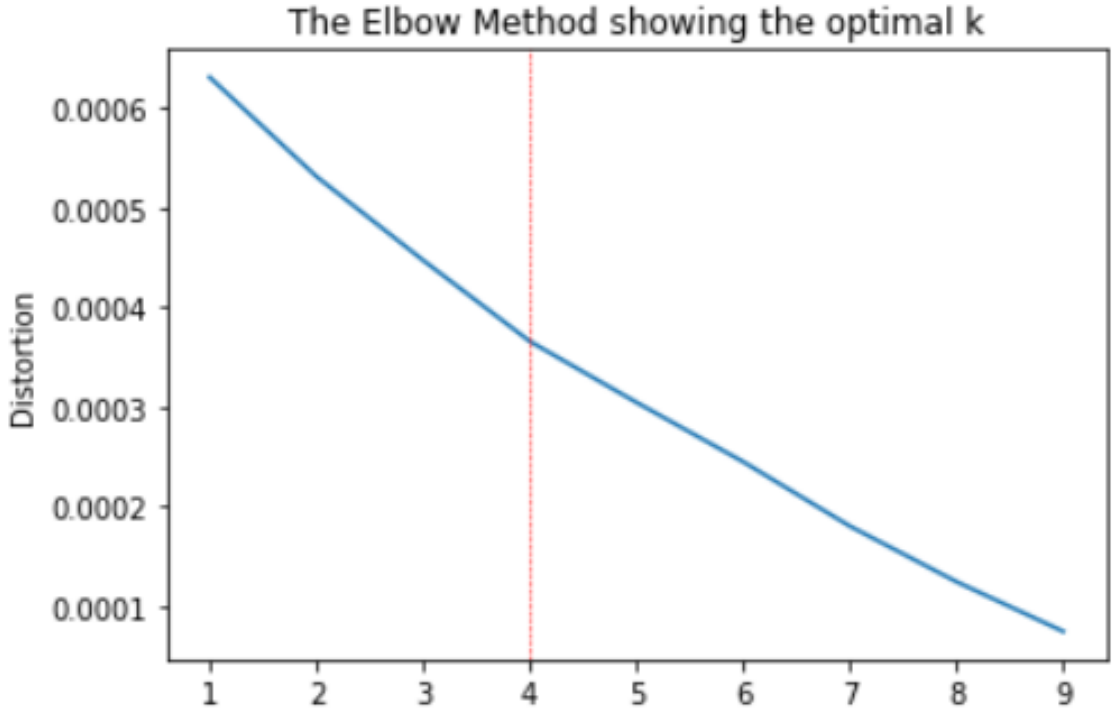


Figure 3.7: The Elbow with k=4 in Entertainment Article 338

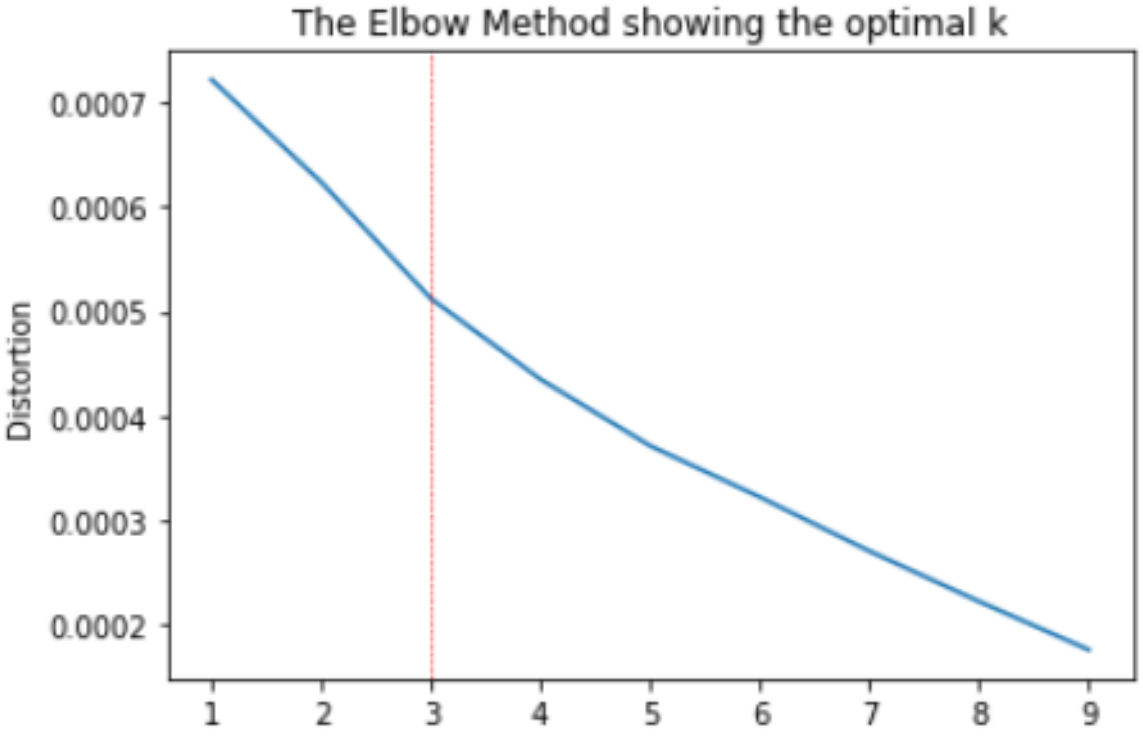


Figure 3.8: The Elbow with k=3 in Politics Article 172

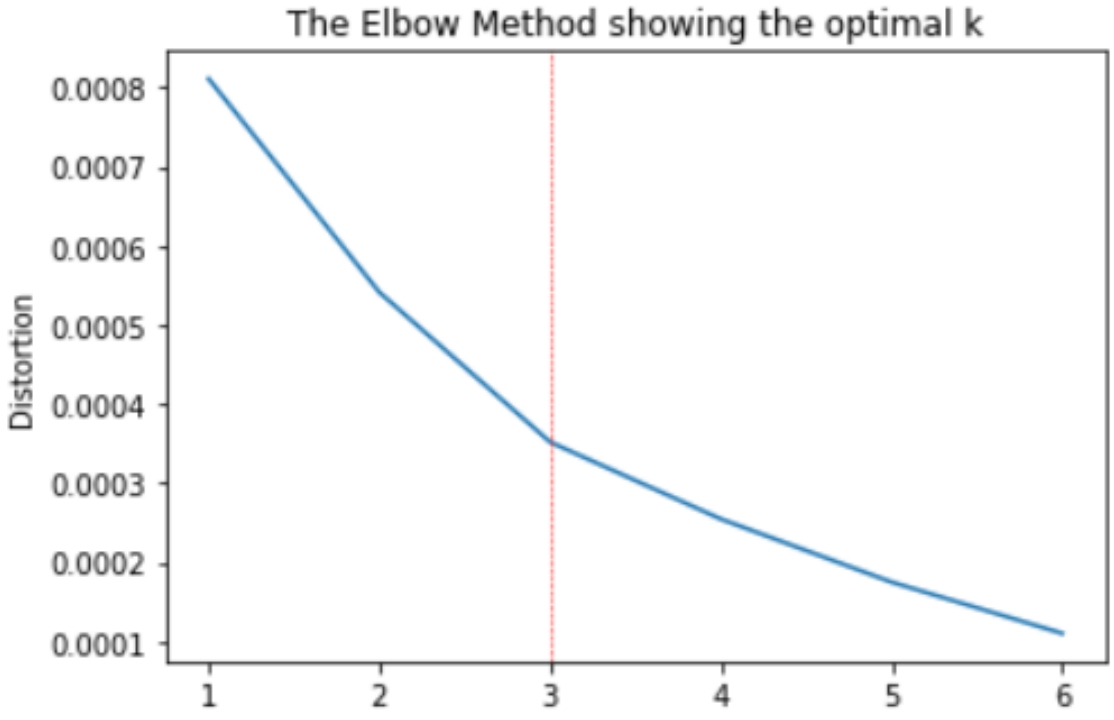


Figure 3.9: The Elbow with k=3 in Sports Article 256

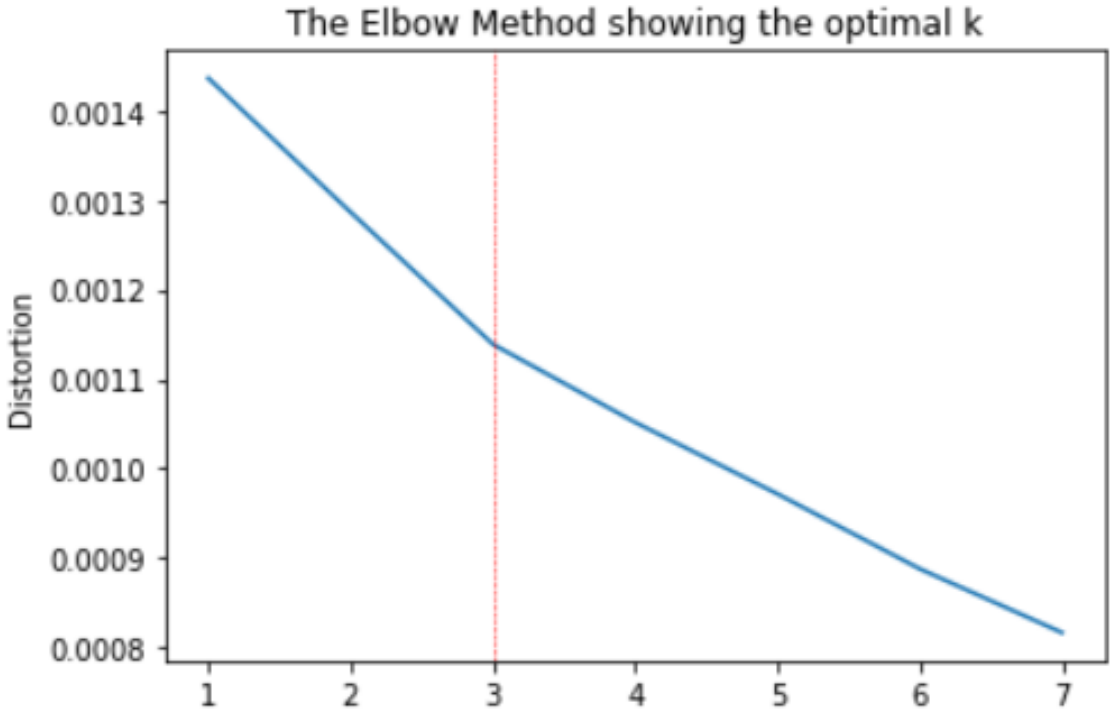


Figure 3.10: The Elbow with k=3 in Tech Article 226

3.1.6 Summary Extraction

Finally, the cluster having the maximum number of sentences has been selected as the higher frequency in cluster indicates the most valuable sentences of the text. All the sentences belong to that cluster have been scored based on mean similarity with the Word2Vec model and the appearance of numbers and nouns. For each number and noun, 1 and 0.25 will be added with the mean similarity of each sentence [1]. From that cluster of sentences, this model in pick n sentences (n is the number of one-third of total sentences). By joining these n sentences sorted based on their appearance on the text, the summary will be generated.

Chapter 4

Result and Evaluation

There are multiple ways to compare two texts. One of them is BLEU Score which has been used in this model [20]. BLEU has been chosen because it is very easy to implement. It gives a result between 0 and 1 where 1 is the best similarity and 0 is the lowest similarity. The generated summary and the original summary of the article have been compared using BLEU. The maximum score from ten iterations has chosen as a BLEU score. Table 4.1, 4.2, 4.3, 4.4 and 4.5 represents the result of a few business, entertainment, politics, sports and tech articles summaries.

Business Article Number	K values With Elbow	BLEU Score for 1 gram	BLEU Scores for 2 gram	BLEU Scores for 3 gram	BLEU Scores for 4 gram	Cumulative BLEU
79	5	0.573	0.530	0.512	0.495	0.528
101	3	0.741	0.691	0.670	0.652	0.689
133	3	0.752	0.733	0.721	0.707	0.728
465	2	0.894	0.894	0.894	0.894	0.894
499	4	0.648	0.593	0.570	0.548	0.590

Table 4.1: BLEU Score of Business articles

Entertainment Articles Number	K values With Elbow	BLEU Scores for 1 gram	BLEU Scores for 2 gram	BLEU Scores for 3 gram	BLEU Scores for 4 gram	Cumulative BLEU
112	3	0.669	0.660	0.658	0.653	0.660
205	3	0.836	0.810	0.794	0.777	0.804
255	2	0.548	0.496	0.481	0.465	0.497
263	3	0.622	0.564	0.540	0.521	0.562
338	4	0.819	0.791	0.771	0.749	0.783

Table 4.2: BLEU Score of entertainment articles

Politics Articles Number	K values With Elbow	BLEU Score for 1 gram	BLEU Score for 2 gram	BLEU Scores for 3 gram	BLEU Scores for 4 gram	Cumulative BLEU
57	3	0.726	0.698	0.684	0.669	0.694
172	3	0.770	0.749	0.739	0.724	0.746
246	4	0.693	0.653	0.633	0.614	0.649
318	2	0.566	0.536	0.522	0.508	0.533
360	4	0.527	0.465	0.439	0.420	0.463

Table 4.3: BLEU Score of politics articles

Sports Articles Number	K values With Elbow	BLEU Scores for 1 gram	BLEU Scores for 2 gram	BLEU Scores for 3 gram	BLEU Scores for 4 gram	Cumulative BLEU
1	4	0.667	0.618	0.598	0.577	0.615
211	2	0.669	0.621	0.600	0.578	0.617
256	3	0.741	0.728	0.716	0.700	0.721
352	2	0.719	0.690	0.677	0.666	0.688
378	2	0.567	0.523	0.504	0.486	0.520

Table 4.4: BLEU Score of sports articles

Tech Articles Number	K values With Elbow	BLEU Scores for 1 gram	BLEU Score for 2 gram	BLEU Scores for 3 gram	BLEU Scores for 4 gram	Cumulative BLEU
74	3	0.755	0.705	0.676	0.650	0.696
91	6	0.481	0.443	0.425	0.406	0.439
152	3	0.744	0.711	0.693	0.675	0.705
226	3	0.670	0.596	0.560	0.532	0.589
297	3	0.557	0.481	0.451	0.427	0.479

Table 4.5: BLEU Score of tech articles

The highest score of the business article's maximum summarization score is 0.894 and the minimum score is 0.528. Fig 4.1 shows the maximum, minimum and average BLEU score of each category of the news article. Among all categories, the model worked better for the business articles as business articles contain more numerical values than other categories and numerical values got much priority in this model.

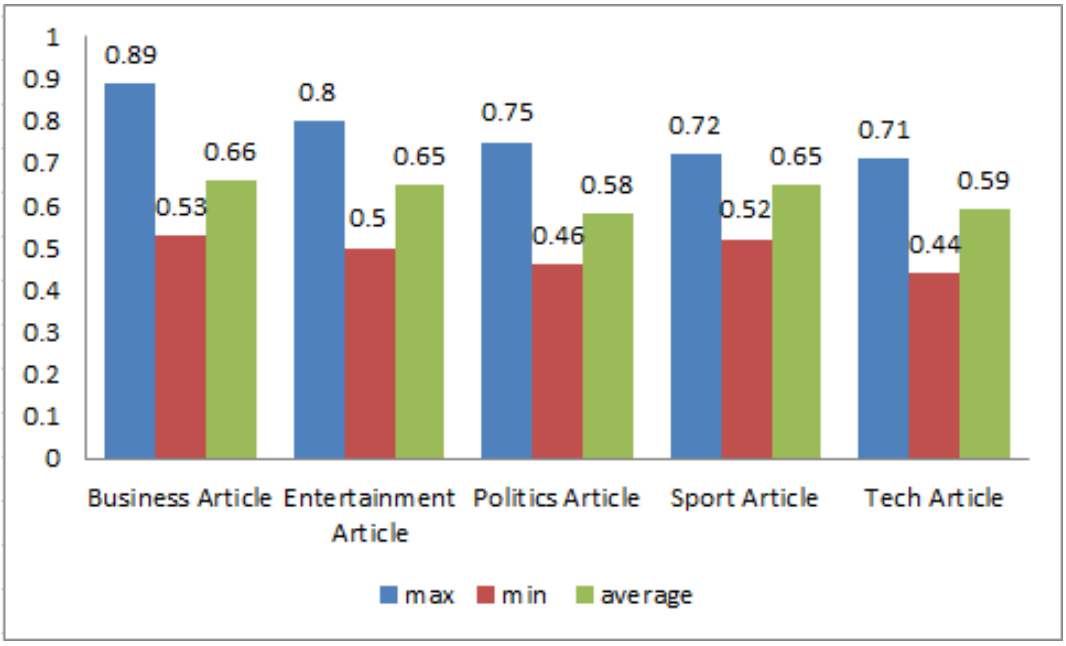


Figure 4.1: Summary score comparison between news article categories

Text summarization accuracy may vary depending on the type of dataset. A similar type of approach has been applied by R. Khan, Y. Qian, and S. Naeem [20] where they have used the TF-IDF score instead of Gensim Word2Vec. In our research we have used BBC news articles whereas they have worked on a different dataset, their highest BLEU score was 0.503984. On the other hand, our highest BLEU score was 0.894 for the business atricle, 0.804 for the entertainment atricle, 0.746 for the political article. Most of our articles showed better BLEU scores than other research papers. Our model has been modified and another sentence scoring algorithm has been used to improve the performance of the system. From the data, it is quite obvious that if we use Gensim Word2Vec instead of TF-IDF it shows a better score.

Chapter 5

Sample Result

The following figures demonstrate the research findings relevant to business, entertainment, politics, sports, and tech articles.

Main Article	Machine Summary
The trial of Bernie Ebbers, former chief executive of bankrupt US phone company WorldCom, has started in New York with the selection of the jury. Mr Ebbers, 63, is accused of being the mastermind behind an $11bn (Â£6bn) accounting fraud that eventually saw the firm collapse in July 2002. His indictment includes charges of securities fraud, conspiracy and filing false reports with regulators. If found guilty, Mr Ebbers could face a substantial jail sentence. He has firmly declared his innocence. Under Mr Ebbers' leadership, WorldCom emerged from Mississippi obscurity to become a $160bn telecoms giant and the darling of late 1990s investors. Yet as competition intensified and the telecoms boom petered out, WorldCom found itself under growing financial stress. When WorldCom finally collapsed, shareholders lost about $180bn and 20,000 workers lost their jobs. Mr Ebbers' trial, which is expected to last two months, is the latest in a series of attempts by US prosecutors to pursue senior executives for fraud. It will coincide with the retrial of former Tyco International chief Dennis Kozlowski and his top lieutenant, accused of looting the industrial conglomerate to the tune of $600m. Trail preparations are also preparing for former executives of shamed US energy firm Enron.	The trial of Bernie Ebbers, former chief executive of bankrupt US phone company WorldCom, has started in New York with the selection of the jury. Mr Ebbers, 63, is accused of being the mastermind behind an $11bn (Â£6bn) accounting fraud that eventually saw the firm collapse in July 2002. Under Mr Ebbers' leadership, WorldCom emerged from Mississippi obscurity to become a $160bn telecoms giant and the darling of late 1990s investors. Mr Ebbers' trial, which is expected to last two months, is the latest in a series of attempts by US prosecutors to pursue senior executives for fraud.
	Original Summary
	The trial of Bernie Ebbers, former chief executive of bankrupt US phone company WorldCom, has started in New York with the selection of the jury. Mr Ebbers, 63, is accused of being the mastermind behind an $11bn (Â£6bn) accounting fraud that eventually saw the firm collapse in July 2002. Under Mr Ebbers' leadership, WorldCom emerged from Mississippi obscurity to become a $160bn telecoms giant and the darling of late 1990s investors. Mr Ebbers' trial, which is expected to last two months, is the latest in a series of attempts by US prosecutors to pursue senior executives for fraud. If found guilty, Mr Ebbers could face a substantial jail sentence.

Figure 5.1: Business Article 456

Main Article	Machine Summary
The British producers of US Wife Swap are taking legal action against a show they claim is "a blatant and wholescale copycat" of their programme. RDF Media, which makes the show for US network ABC, has filed a damages claim for $18 million (Â£9.25 million) against Fox's Trading Spouses. ABC bought the rights to the British show, which was first aired in 2003 and became a hit on Channel 4. The US network is not part of the claim, but has supported RDF's action. "We respect our producing partners' right to protect their intellectual property in whatever manner they deem most appropriate," said ABC in a statement. A spokesman for Fox said it had not seen the details of the legal action and could not comment. Their show was first screened in June, and was criticised in the press for its similarities to Wife Swap. ABC originally planned to call their programme Trading Moms, but changed it to avoid confusion with the Fox version. Earlier this year, the NBC network claimed that Fox's boxing show The Next Great Champ had been hurriedly produced to ensure its programme was the first to be screened. NBC alleged that boxing regulations had been violated, but failed in their attempt to have the show pulled. The Fox show proved a ratings flop, while NBC's The Contender is due to begin in February.	The British producers of US Wife Swap are taking legal action against a show they claim is "a blatant and wholescale copycat" of their programme. RDF Media, which makes the show for US network ABC, has filed a damages claim for $18 million (Â£9.25 million) against Fox's Trading Spouses. ABC bought the rights to the British show, which was first aired in 2003 and became a hit on Channel 4. "We respect our producing partners' right to protect their intellectual property in whatever manner they deem most appropriate," said ABC in a statement. Earlier this year, the NBC network claimed that Fox's boxing show The Next Great Champ had been hurriedly produced to ensure its programme was the first to be screened.
	Original Summary
	The British producers of US Wife Swap are taking legal action against a show they claim is "a blatant and wholescale copycat" of their programme. ABC bought the rights to the British show, which was first aired in 2003 and became a hit on Channel 4. Earlier this year, the NBC network claimed that Fox's boxing show The Next Great Champ had been hurriedly produced to ensure its programme was the first to be screened. RDF Media, which makes the show for US network ABC, has filed a damages claim for $18 million (Â£9.25 million) against Fox's Trading Spouses. Their show was first screened in June, and was criticised in the press for its similarities to Wife Swap.

Figure 5.2: Entertainment Article 205

Main Article	Machine Summary
The first children's commissioner for England has been appointed. Great Ormond Street Hospital professor of child health, Al Aynsley-Green, was chosen by the government and will start the Â£100,000-a-year job immediately. He will oversee a Â£2.5m annual budget and have the power to look into "any matter relating to the interests and well-being of children". Prof Aynsley-Green has also been the national clinical director for children in the Department of Health. He promised to make sure that children's opinions "count". "I will be drawing on my experience of working with children and young people to help ensure that those with the power to improve children's lives do live up to their responsibilities. "I want all children and young people to know that they can approach me to discuss any matter that affects them, knowing that I will value their opinion." Education Secretary Ruth Kelly said Prof Aynsley-Green would "strengthen the voice of children and young people". Prof Aynsley-Green was a lecturer at Oxford University, trained at Guy's Hospital Medical School, University of London; Oriel College, Oxford; and in Switzerland. He is described as "a proud grandfather" of four. Scotland, Wales and Northern Ireland already have children's commissioners.	Great Ormond Street Hospital professor of child health, Al Aynsley-Green, was chosen by the government and will start the Â£100,000-a-year job immediately. He will oversee a Â£2.5m annual budget and have the power to look into "any matter relating to the interests and well-being of children". "I will be drawing on my experience of working with children and young people to help ensure that those with the power to improve children's lives do live up to their responsibilities. "I want all children and young people to know that they can approach me to discuss any matter that affects them, knowing that I will value their opinion."
	Original Summary
	"I will be drawing on my experience of working with children and young people to help ensure that those with the power to improve children's lives do live up to their responsibilities. Education Secretary Ruth Kelly said Prof Aynsley-Green would "strengthen the voice of children and young people". Prof Aynsley-Green has also been the national clinical director for children in the Department of Health." I want all children and young people to know that they can approach me to discuss any matter that affects them, knowing that I will value their opinion. "Great Ormond Street Hospital professor of child health, Al Aynsley-Green, was chosen by the government and will start the Â£100,000-a-year job immediately.

Figure 5.3: Politics Article 172

Main Article	Machine Summary
Everton defender David Weir has played down talk of European football, despite his team lying in second place in the Premiership after beating Liverpool. Weir told BBC Radio Five Live: "We don't want to rest on our laurels and say we have achieved anything yet. "I think you start taking your eye off the ball if you make statements and look too far into the future. "If you start making predictions you soon fall back into trouble. The only thing that matters is the next game." He said: "We are looking after each other and hard work goes a long way in this league. We have definitely shown that. "Also injuries and suspensions haven't cost us too badly and we have a lot of self-belief around the place."	Everton defender David Weir has played down talk of European football, despite his team lying in second place in the Premiership after beating Liverpool. Weir told BBC Radio Five Live: "We don't want to rest on our laurels and say we have achieved anything yet. "I think you start taking your eye off the ball if you make statements and look too far into the future. "Also injuries and suspensions haven't cost us too badly and we have a lot of self-belief around the place."
	Original Summary
	Everton defender David Weir has played down talk of European football, despite his team lying in second place in the Premiership after beating Liverpool. Weir told BBC Radio Five Live: "We don't want to rest on our laurels and say we have achieved anything yet. "I think you start taking your eye off the ball if you make statements and look too far into the future. "Also injuries and suspensions haven't cost us too badly and we have a lot of self-belief around the place."

Figure 5.4: Sports Article 256

Main Article	Machine Summary
Software giant Microsoft is taking the plunge into the world of blogging. It is launching a test service to allow people to publish blogs, or online journals, called MSN Spaces. Microsoft is trailing behind competitors like Google and AOL, which already offer services which make it easy for people to set up web journals. Blogs, short for web logs, have become a popular way for people to talk about their lives and express opinions online. MSN Spaces is free to anyone with a Hotmail or MSN Messenger account. People will be able to choose a layout for the page, upload images and share photo albums and music playlists. The service will be supported by banner ads. "This is a simple tool for people to express themselves," said MSN's Blake Irving. This is Microsoft's first foray into blogging, which has taken off as a web phenomenon in the past year. Competitors like Google already offer free services through its Blogger site, while AOL provides its members with journals. Accurate figures for the number of blogs in existence are hard to come by. According to blog analysis firm Technorati, the so-called blogosphere, has doubled every five and a half months for the last 18 months. It now estimates that the number of blogs in existence has exceeded 4.8 million, although some speculate that less than a quarter are regularly maintained.	Microsoft is trailing behind competitors like Google and AOL, which already offer services which make it easy for people to set up web journals. Blogs, short for web logs, have become a popular way for people to talk about their lives and express opinions online. People will be able to choose a layout for the page, upload images and share photo albums and music playlists. According to blog analysis firm Technorati, the so-called blogosphere, has doubled every five and a half months for the last 18 months. It now estimates that the number of blogs in existence has exceeded 4.8 million, although some speculate that less than a quarter are regularly maintained. Original Summary Microsoft is trailing behind competitors like Google and AOL, which already offer services which make it easy for people to set up web journals. It is launching a test service to allow people to publish blogs, or online journals, called MSN Spaces. Competitors like Google already offer free services through its Blogger site, while AOL provides its members with journals. Blogs, short for web logs, have become a popular way for people to talk about their lives and express opinions online. It now estimates that the number of blogs in existence has exceeded 4.8 million, although some speculate that less than a quarter are regularly maintained. Accurate figures for the number of blogs in existence are hard to come by.

Figure 5.5: Tech Article 149

Chapter 6

Conclusion

Text summarization is one of the most renowned buzz words in the area of research in natural language processing as the textual data is increasing day by day. The proposed model introduces Gensim Word2Vec with the combination of the K-Means clustering algorithm and some new sentence scoring procedure which enables a new dimension of research in text summarization. In this model, all the sentences were clustered using the K-Means clustering algorithm. Sentence scoring algorithm rates a sentence based on the occurrence of numerical values and nouns. These techniques were implemented on BBC news article datasets. The proposed model showed the best performance on the business articles because the business article contains more numerical values and the sentence scoring algorithm gives priority to numerical values. In the future, the same idea can be also implemented on the extractive based multiple text document.

Bibliography

[1] M. M. Haider, *"Sentence Scoring Based on Noun and Numerical Values"*, https://towardsdatascience.com/sentence-scoring-based-on-noun-and-numerical-values-d7ac4dd787f2, Feb 1, 2020.

[2] D. Karani, *"Introduction to Word Embedding and Word2Vec"*, https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa, Sep 1, 2018.

[3] S. O'Dea, *"Volume of data/information created worldwide from 2010 to 2025"*, https://www.statista.com/statistics/871513/worldwide-data-created/, Feb 28, 2020.

[4] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters", 1973.

[5] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer", in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995, pp. 68–73.

[6] W. T. Chuang and J. Yang, "Text summarization by sentence segment extraction using machine learning algorithms", in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2000, pp. 454–457.

[7] D. Radev, "Centroid-based summarization of multiple documents: Sentence extration, utility-based evalutation, and user studies", in *Proc. ACL/NAAL Workshop on Summarization, Seattle, WA.(2000)*, 2000.

[8] J. L. Neto, A. A. Freitas, and C. A. Kaestner, "Automatic text summarization using a machine learning approach", in *Brazilian symposium on artificial intelligence*, Springer, 2002, pp. 205–215.

[9] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering", in *Proc. 23rd International Conference on Machine learning (ICML'06)*, ACM Press, 2006, pp. 377–384.

[10] D. Das and A. Martins, "A survey on automatic text summarization. literature survey for language and statistics", *II Course at CMU*, 2007.

[11] R. A. Garcıa-Hernández, R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh, and R. Cruz, "Text summarization by sentence extraction using unsupervised learning", in *Mexican International Conference on Artificial Intelligence*, Springer, 2008, pp. 133–143.

[12] J. P. Ortega, M. Del, R. B. Rojas, and M. J. Somodevilla, "Research issues on k-means algorithm: An experimental trial using matlab", in *CEUR workshop proceedings: semantic web and new technologies*, 2009, pp. 83–96.

[13] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques", *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[14] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", English, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, http://is.muni.cz/publication/884893/en, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[15] I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive text summarization", in *2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*, IEEE, 2012, pp. 132–138.

[16] A. R. Deshpande and L. Lobo, "Text summarization using clustering technique", *International Journal of Engineering Trends and Technology*, vol. 4, no. 8, pp. 3348–3351, 2013.

[17] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond", *arXiv preprint arXiv:1602.06023*, 2016.

[18] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization", *arXiv preprint arXiv:1705.04304*, 2017.

[19] S. A. Anam, A. M. Rahman, N. N. Saleheen, and H. Arif, "Automatic text summarization using fuzzy c-means clustering", in *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, 2018, pp. 180–184.

[20] R. Khan, Y. Qian, and S. Naeem, "Extractive based text summarization using k-means and tf-idf", *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 3, p. 33, 2019.

[21] B. Li, A. Drozd, Y. Guo, T. Liu, S. Matsuoka, and X. Du, "Scaling word2vec on big corpus", *Data Science and Engineering*, vol. 4, no. 2, pp. 157–175, 2019.

[22] J. Xu and Q. Du, "A deep investigation into fasttext", in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2019, pp. 1714–1719.