2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh

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

Mofiz Mojib Haider
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
haider4069@gmail.com

Md. Arman Hossin
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mdarmanhossin99@gmail.com

Hasibur Rashid Mahi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mahialhasib@gmail.com

Hossain Arif
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
hossain.arif@bracu.ac.bd

***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.**

***Index Terms*—Text summarization, Extractive, Single Document, NLP, Gensim, Word2Vec, K-Means.**

I. Introduction

The subfield of text summarization has increased over the half-century past. DR Radev [1] 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 [2]. On the other hand, abstractive text summarization systems create new sentences, likely rephrasing by using terms, not in the original document [3].

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.

The remainder of the report is sorted as follows. Part II reflects the literature review in the document overview the sector, part III describes the proposed model, the K-Means clustering model and Gensim Word2Vec, part IV presents the result analysis and evaluation finally in part V the conclusion has been portrayed including future ideas.

II. Literature Review

Rafael Ferreira et al. [4] 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.

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

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. [6] 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. [7] showed a single document summarization and multi-document summarization using extractive and abstractive text summarization approach. Where there are various algorithms has been applied like Na¨ıve-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. [3] 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.

Bofang li et al. [8] 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.

Rene´ Arnulfo Grc´ıa-Hernandez´ et al. [9] 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.

III. 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 [10]. At the very primary stage, a news article has been selected from the dataset and undergoes numerous pre-processing procedures. During pre-processing, the model will perform sentence tokenize, remove special characters, word tokenize, duplicate word remove and finally lemmatization to get the root word.

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.

A. Dataset

BBC news article dataset [11] contains 2225 news articles along with sample summaries for each article which are 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.

B. Preprocessing

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

1) Sentence Tokenization: It is the process to split the text into sentences [12]. 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 [12].
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 [12].
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 [12].

C. 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 [13].

$$sin(X,Y) = cos(\theta) = \frac{(X \cdot Y)}{||X||||Y||} \quad (1)$$

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

D. 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.

E. 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 [15]:

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:

$$v_i = \frac{(1)}{(c_i)} \sum_{j=1}^{c_i} m_i \tag{2}$$

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.

F. 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 [12]:

- 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.

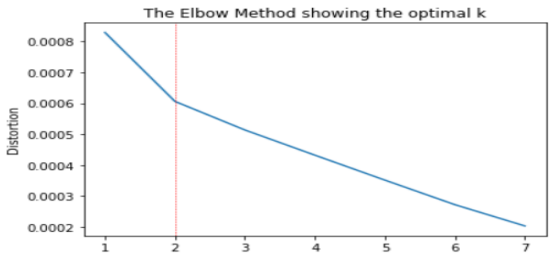

Fig. 1. The Elbow with k=2 in Business Doc_465

G. 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 [16]. 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.

IV. Result and Evaluation

There are multiple ways to compare two texts. One of them is BLEU Score which has been used in this model [12]. 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 1, 2, 3, 4 and 5 represents the result of a few business, entertainment, politics, sports and tech articles summaries.

TABLE I
BLEU Score of Business articles

Business Doc 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 II
BLEU Score of entertainment articles

Entertain-ment Doc 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 III
BLEU Score of politics articles

Politics Doc 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 IV
BLEU SCORE OF SPORTS ARTICLES

Sports Doc 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 V
BLEU SCORE OF TECH ARTICLES

Tech Doc 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

Fig.2 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.

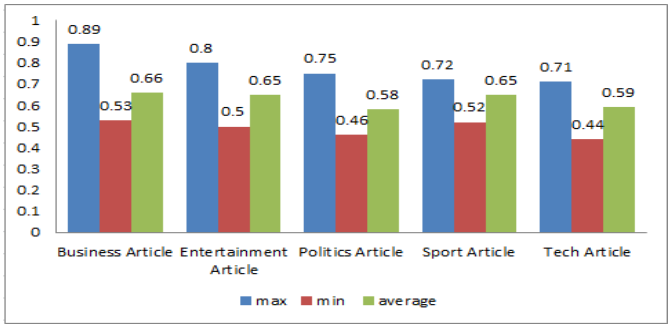

Fig. 2. Summary score comparison between news article categories

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 [12] 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. From the data, it is quite obvious that if we use a different sentence scoring algorithm and Gensim Word2Vec instead of TF-IDF it shows a better score.

V. 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.

REFERENCES

[1] 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.

[2] 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.

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

[4] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. Freitas, G. D. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization," *Expert systems with applications*, vol. 40, no. 14, pp. 5755–5764, 2013.

[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] 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.

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

[8] 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.

[9] R. A. García-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.

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.