Performance Evaluation of Manhattan and Euclidean Distance Measures For Clustering Based Automatic Text Summarization

*Shakirat A. Salihu, Ifeoma P. Onyekwere, Modinat A. Mabayoje and Hammed A. Mojeed

Department of Computer Science, University of Ilorin, Ilorin, Nigeria
salihusa1980@gmail.com | ifeomapauline4u@yahoo.com | {mabayoje.ma | mojeed.ha}@unilorin.edu.ng

Abstract— In the past few years, there has been an explosion in the amount of text data from a variety of sources. This volume of text is a valuable source of information and knowledge which needs to be effectively summarized to be useful. In this paper, automatic text summarization with K-means clustering techniques is presented by employing two different distance measurement methods (Euclidean and Manhattan). The dataset extracted from African prose was preprocessed using stopwords removal and tokenization. The preprocessed document is converted into vector representation using tf-idf technique and k-means clustering is applied using Euclidean and Manhattan distance measures to generate summary. There are different distance measures for k-means which has been used in several works. However, there is dearth of work on performance evaluation of these distance measures in text summarization. The experimental analysis was performed on Waikato Environment for Knowledge Analysis. The results obtained showed that the Euclidean variation produced an extractive summary of sentences amounting to 72% from three different clusters while the Manhattan variation produced an extractive summary of sentences that made up 94% of the total document all in one cluster using compression ratio as the performance metric.

Keywords— Text summarization, Euclidean distance, k-means clustering, Manhattan distance.

1 INTRODUCTION

The 21st century has welcome the deluge of data being generated via several sources which include humans and devices, due to the accelerated growth in communication and computing, information is now becoming the live stream of the society. However, the data are mostly in its raw state which is stored up in databases as mentioned by (Witten et al.,2011), which is now being explored for the sake of extracting potentially useful information that is once unknown and implicit in nature for the purpose of getting actionable insights for societal, business, and or governmental progress. This data is usually stored digitally and the search is being carried out by computer, either being automated or at least augmented, using various data mining tools and techniques in this regards. The immense importance of data mining and its value in our present society cannot be over-emphasized as it seeks to the betterment of the society by influencing lives, enterprises, governmental and nongovernmental organizations.

This expanding availability of documents has demanded exhaustive research in the area of automatic text summarization. The technology of automatic text summarization plays an important role in information retrieval and text classification, and may provide solution to the information overload problem (Zhang & Li, 2009). According to Allahyari, Trippe, and Gutierrez (2017) a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that. Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. Also, in the area of text summarization is the text reuse which is not only helpful in solving a new similar problem but can assist in authoring new experiences (Adeyanju et al., 2010).

In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. For example, search engines generate snippets as the previews of the documents (Andrew et al., 2007). Other examples include news websites which produce condensed descriptions of news topics usually as headlines to facilitate browsing or knowledge extractive approaches (Trippe, 2017). Among the many achievements of data mining application is the progress and benefits of Natural Language Processing (NLP), Intrusion Detection System (IDS), Natural Language Inference (NLI), Name Entity Recognition (NER), customer churn prediction, targeted marketing, predictive analytics, business analytics, sentimental analysis market-basket analysis and so on. The increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the NLP community (Das & Martins, 2007). Of importance to this research work is automatic text summarization (ATS), which is the application of data mining tools and techniques to the development of models that does text extraction or abstraction either for single or multiple documents (Kiser, 2016).

Machine learning is useful in extraction of information from any raw data in databases. This is the basis of data mining as it includes mathematical and or statistical theories or algorithms that are being used in the development of models. Its inferences are based on structures underlying it. Basically, machine learning works by accepting data as input, learns the underlying structure or patterns, and then develops a model that can be used for future purposes (Neto, Freitas, & Kaestner, 2002). Thus, this paper focuses on the application of machine learning on automatic text summarization system using two distance measurements on the clustering technique used. Automatic single document text summarization is addressed based on unsupervised learning techniques.

*Corresponding Author

2 Related Works

Recent research works on extractive-summary generation employ some heuristics such as butfewworks that indicate how to select the relevant features. Neto et al., (2002) present asummarization procedure based on the application of trainable MachineLearning algorithms which employs a set of features extracted directly from theoriginal text. These features are of two kinds: statistical – based on the frequency of some elements in the text; and linguistic – extracted from asimplified argumentative structure of the text. The author also presented somecomputational results obtained with the application of the summarizer to somewell known text analysis software such as WEKA, and these results was compared to some baseline of summarization procedures.

Nayeem (2017) developed several techniques for tackling both the extractive and abstractive text summarization tasks. A rank based sentence selection whichcan retain the most important and non-redundant contents to form the summary was implemented. For ensuringa pure sentence abstraction, several novel sentence abstraction techniqueswhich jointly perform sentence compression, fusion and paraphrasing at the sentence level was proposed. Also abstractive compression generation as a sequence-to-sequence (seq2seq) problem was modelled using an encoder-decoder framework. It is also a novel inclusion accordingto the state-of-the-art text summarization systems. A simple but yet effective solutionsto several common problems in neural seq2seq models such as redundant repetition and unknown token replacement was proposed. The sentence level models improve the informativity aswell as the grammaticality of the generated sentences.

Furthermore, sentence abstraction techniques to the multi-document text summarization were also applied. For the sentence level tasks, experiments on human generated abstractive compression datasets and system evaluation on several newly proposed Machine Translation (MT) evaluation metrics was conducted. In the case of the document level summary, experimentswas conducted on the Document Understanding Conference (DUC) 2004 datasets using ROUGE toolkit. The experiments demonstrate that the methods brings significant improvements overthe state-of-the-art methods. A new concept was introduced at the end of this work called "Reader Aware Summary" which can generate summaries for some critical readers(e.g. Non-Native Reader). Yousefi-Azar and Hamey (2017) proposed an algorithm that incorporates k-means clustering, term-frequency (tf) inverse-document-frequency (idf) and tokenization to perform extraction based text summarization.

3 Materials and Methods

Our implementation of automatic text summarization is an extractive based ATS model, which adopts basic text pre-processing procedure in natural language processing, making use of some related techniques in order to achieve the aim of this study as shown in figure 1. The Machine Learning algorithm selected for this research is the K-means algorithm – an unsupervised learning method. More so, several text pre-processing techniques were used, as it is known in NLP that texts have to be transformed into the form in which machine learning algorithms can understand and make use of in order to achieve our aims. More importantly, the implementation of the ATS model was carried using one of the popular data mining benchmark tool, "WEKA" – (Waikato Environment for Knowledge Analysis).

The dataset used was a prose titled "Forest of a thousand demons" written by D.O. Fagunwa. The novel was sourced in a text format and was at first processed into attribute-relation-file-format, ".arff" – the popular format used by WEKA. The whole novel containing texts in paragraph and segments were broken down into sentence line – one per line, and in turn transformed into ".arff" file format, making it ready for further pre-processing as related to natural language processing.
Vector space model representing the documents using various scheme are inverse document frequency (idf), term frequency (tf) or term frequency – inverse document frequency (tf-idf).

$$tf(t,d) = \frac{f_{td}}{\#terms}, \quad idf(t) = \log\left(\frac{\#B}{n_t}\right) \tag{1}$$

Where t denotes term, d is document, f_{td} is the frequency of a term in a document, #terms is the total number of terms and #B is the total number of documents. For this research, the tf-idf scheme is adopted as it provides a better representation of document as being highlighted and used by several authors. This work uses both Euclidean distance and Manhattan distance measurements techniques for finding similarity using k-means algorithm, this is done in order to achieve two distinct extracts. At the end of further pre-processing, the selected machine learning method – K-means, was applied to clustered the sentences into a fix number of clusters and the extraction of sentences were carried out.

3.1 Centroid Based Document Summarization Method

Clustering or cluster analysis is solely aimed at finding a set of correlating objects from a given set of objects and grouping them according to their peculiarity, i.e., finding and grouping similar objects from a set to form subset(s).

3.1.2 K-Means Algorithm

The algorithmic step of K-means in partitioning a dataset, where it's cluster centre is represented by the centroid of the data points in the cluster.

Algorithm: K-means
Input: k: the number of clusters
Output: A set of k clusters.

Method:
Step 1: Choose k numbers of clusters apriori.
Step 2: Choose Ck centroids randomly as the initial centers of the clusters.
Step 3: Repeat
3.1: Assign each object to their closest cluster center using selected distance measure.
3.2: Compute new cluster center by calculating mean points.
Step 4: Until

4.1: No change in cluster center OR
4.2: No object changes its clusters.

The result of this algorithm is the separation of a given dataset and the latter production of subsets, called clusters, of the dataset.

3.2 Proposed Implementation Model

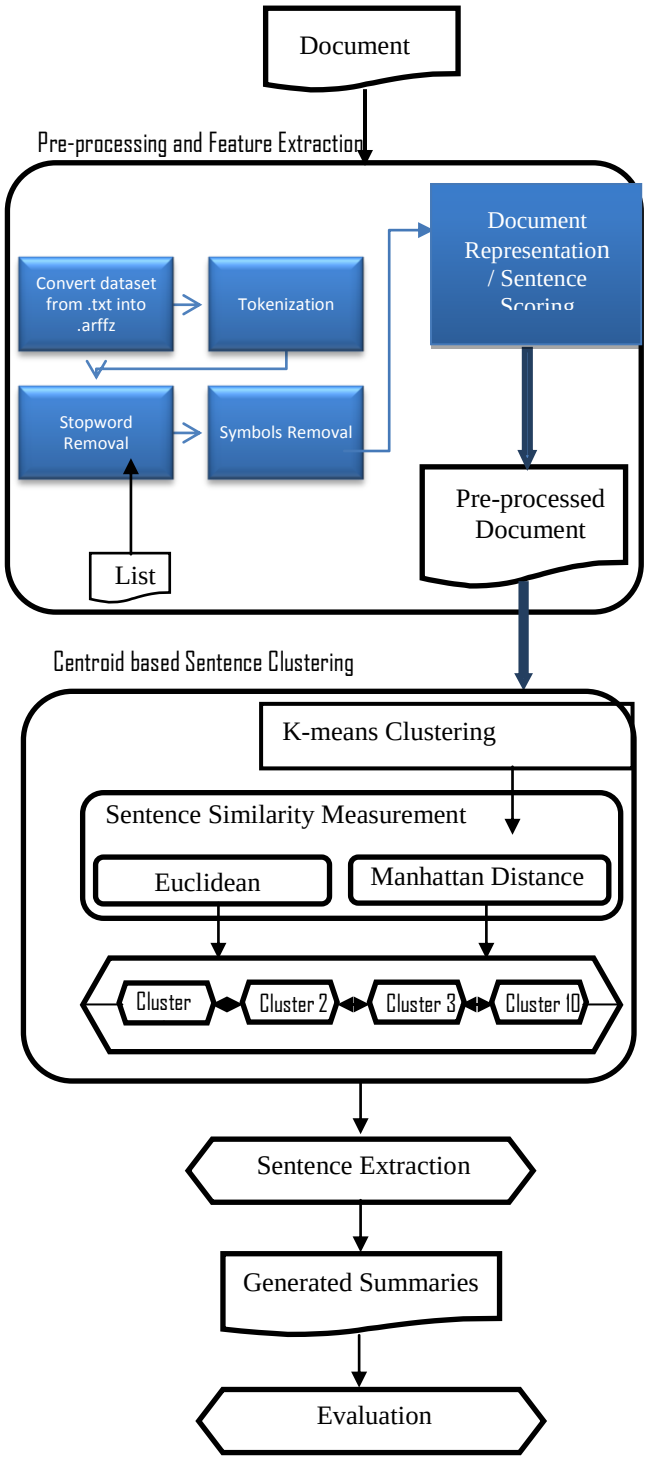

Fig. 1: Proposed Implementation Model

3.2.1 Dataset

The dataset used was an African prose titled "Forest of a thousand demons", written by D.O. Fagunwa downloaded from www.citylights.com. Typically, the text of the novel was contained in paragraphs and chapters which cannot be used directly by any machine learning technique. Thus, requiring a pre formatting. More so, the dataset contained 9913 words and was broken down into 378 sentences. The dataset was formatted as one sentence per line in its initial text file format in order to enforce order.

3.2.2 Tokenization

In this study, the sentences were tokenized at the word level using the WordTokenizer technique. This technique breaks down the stream of line sentences into words that are contained therein, and also maintain the instance order.

3.2.3 Stopwords Removal

Examples of those word, in this research, are "an", "the", "what"," when", "whether", "could", "that", "this" and "do" to mention a few. These stop words were compiled in a list and was used simultaneously during tokenization. As a result, the token generated does not contained the words in the stop word list.

3.2.4 Symbol Removal

This is the removal of signs, symbols and punctuation from the generated tokens. This was carried out after tokenization and the removal of stop words. Basically the symbols, signs and punctuations are of little or no significance in finding and group similar sentences, thus, the need for its removal.

3.2.5 Sentence Clustering

As mentioned earlier, the centroid based sentence clustering model is being adopted in this work, and the k-means clustering technique is used. Furthermore, for the purpose of robust and evaluation, this study adopted two variations of k-means, with respect to sentence similarity measurement. As it is known that a centroid based sentence clustering finds and group similar data points by finding the nearest data point to the centroid of each cluster, two similarity measurements were used respectively. More so, k was set to an arbitrary value of 10, with the desire of having at least 30 sentences per cluster.

4 Evaluation

As this study aim in the implementation of automatic text summarization using k-means algorithm, which is an extractive-based. The summaries generated by the variations of k-means, that is Euclidean and Manhattan distance measurements respectively. Compression Ratio (CR) was used as the performance metric to evaluate the performance of these two distance measures (Hassel, 2004).

$$CR = \frac{lenght\ of\ Summary}{lenght\ of\ full\ text}$$

4.1 Results and Discussion

This study is fixed on implementing an automatic text summarization model using k-means algorithm, and by further extension, implementing two variations of k-means algorithms (one with Euclidean distance measurement and the other using Manhattan distance measurement for finding similarities between sentences).

4.2 DATASET

After carrying out all pre-processing techniques, the streams of texts in the documents were tokenized and further pre-processed resulting in a group of features (a total of 2025 tokens) and 378 instances (sentences) as shown in figure 2. Having conducted the pre-processing, the stream of text was transformed into a vector form using the term frequency – inverse document frequency method.

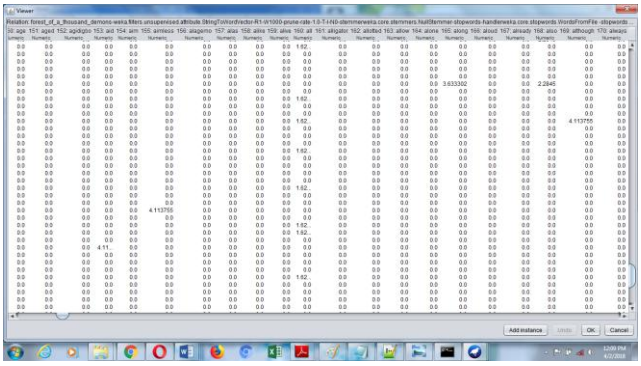

Fig. 2: Vector Representation of text (tf-idf method)

The vector representation above is the algebraic description of the textual data, keeping the order of the sentences, which was inputted into the k-means algorithm.

4.3 EXPERIMENTAL RESULT OF K-MEANS ALGORITHM: EUCLIDEAN DISTANCE SIMILARITY MEASUREMENT

Having inputted the document's vector representation into k-means algorithm and configured the algorithm to cluster the given dataset into ten (10) clusters, initialization method set to "kmeans++" and also setting 'Euclidean Distance" for finding the distance. The result of the K-Means clustering using Euclidean distance measure is presented in Table 1.

Table 1. K-means "Euclidean Distance" cluster results

Clusters	No of Sentences	Percentage (%)
0	1	0
1	26	7
2	88	23
3	104	28
4	80	21
5	1	0
6	1	0
7	6	2
8	1	0
9	70	19

As it can be seen, the algorithm divided the dataset into 10 clusters (0 – 9) and each cluster containing at least one instance. It is seen that clusters 2, 3, 4 had instances higher than 20% while others have lower. More so, the clusters were visualized and those meeting the evaluation criterion were saved. Figure 3 depicts the visualization of these results.

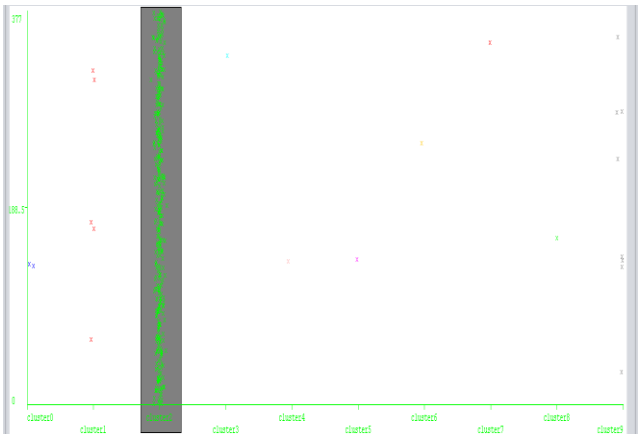

Fig. 3: K-means "Euclidean Distance" visualization result.

The *CR* for this method is computed thus:

$$CR = \frac{272}{378} = 0.719 \sim 72\%$$

Now, the save instances of the clusters which are the extracted summary have 272 sentences cumulatively for three clusters as highlighted in figure 3.

4.4 EXPERIMENTAL RESULT OF K-MEANS ALGORITHM: MANHATTAN DISTANCE SIMILARITY MEASUREMENT

This technique also followed the previous configuration of k-means except for changing the distance function from Euclidean to Manhattan Distance function. Thus, it yielded an entirely different result (extractive summary) from that of Euclidean's. Table 2 presents the result of the clusters formed from Manhattan distance measure.

Table 2: K-means "Manhattan Distance" cluster results

Clusters	No of Sentences	Percentage (%)
0	2	1
1	5	1
2	357	94
3	1	0
4	1	0
5	1	0
6	1	0
7	1	0
8	1	0
9	8	2

The *CR* for this method is computed thus:

$$CR = \frac{375}{378} = 0.944 \sim 94\%$$

Also, it can be seen that this measurement threaded a different path from its counterpart, having provided ten clusters but with only one being densely contained with instances – cluster 2, with 357 out of 378 instances, yielding 94%. Notwithstanding, the clusters were also visualized and the instances in only cluster 2 was saved as others failed to meet up the set criterion as shown in figure 4.

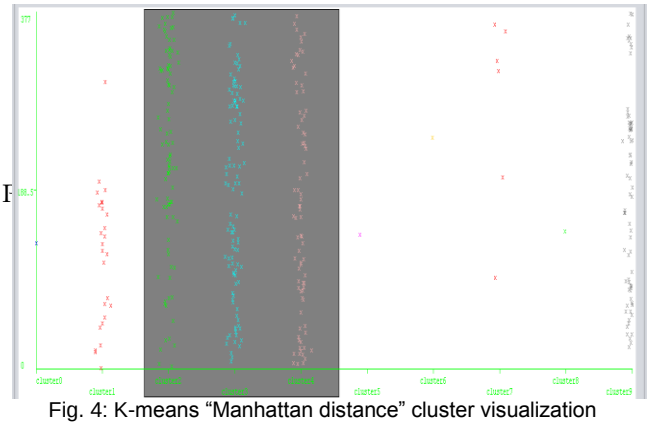

Fig. 4: K-means "Manhattan distance" cluster visualization

Having saved the cluster instances, and thus serving as the summarized text for the original dataset, as shown in figure 4.

5 CONCLUSION

This paper presents an automatic text summarization model using an unsupervised machine learning technique k-means, being variated by employing two different distance measurement method (Euclidean and Manhattan). Both variations produced an extractive summary having pre-processed the original textual dataset following several state-of-art textual data pre-processing method which produced a 'tf-idf" vector form document representation.

The Euclidean variation produced an extractive summary of sentences amounting to 72% from three different clusters while the Manhattan variation produced an extractive summary of sentences that made up 94% of the total document all in one cluster. This implementation is a proof of the fact that different distance measures of K-means can be used to determine which of them will produce a better summary as far as ATS is concerned. The future work can consider automatic determination of number of clusters to enhance K-means performance. Also, other distance measures can be considered for comprehensive evaluation.

REFERENCES

Adeyanju, I., Wiratunga, N., Recio-Garcia, J.A., & Lothian, R. (2010).Learning to Author Text with textual CBR.In ECAI (pp. 777-782).

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Das, D., & Martins, A.F. (2007). A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU, 4, 192-195.

Hassel, M. (2004). Evaluation of automatic text summarization. *Licentiate Thesis, Stockholm, Sweden*, 1-75

Kiser, M. (2016). Introduction to Natural Language Processing (NLP) - Algorithmia Blog. Retrieved from https://blog.algorithmia.com/introduction-natural-language-processing-nlp/

Nayeem, M. T. (2017). *Methods of sentence extraction, abstraction and ordering for automatic text summarization* (Doctoral dissertation, Lethbridge, Alta.: Universtiy of Lethbridge, Department of Mathematics and Computer Science).

Neto, J. L., Freitas, A. A., & Kaestner, C. A. A. (2002). Automatic Text Summarization using a Machine Learning Approach. *SBIA '02 Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence*, (i), 205–215. https://doi.org/10.1007/3-540-36127-8_20

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yousefi-Azar, M., & Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications*, *68*(October), 93–105. https://doi.org/10.1016/j.eswa.2016.10.017

Zhang, P.Y., & Li, C.H. (2009). Automatic text summarization based on sentences clustering and extraction. In Computer Science and Information Technology, *2nd IEEE Internation Conference ICCSIT 2009* (167-170)