Document Length Variation in the Vector Space Clustering of News in Arabic: A Comparison of Methods

Abdulfattah Omar[1*]
College of Science & Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University

Wafya Ibrahim Hamouda[2]
Department of Foreign Languages
Faculty of Education
Tanta University, Egypt

***Abstract*—This article is concerned with addressing the effect of document length variation on measuring the semantic similarity in the text clustering of news in Arabic. Despite the development of different approaches for addressing the issue, there is no one strong conclusion recommending one approach. Furthermore, many of these have not been tested for the clustering of news in Arabic. The problem is that different length normalization methods can yield different analyses of the same data set, and that there is no obvious way of selecting the best one. The choice of an inappropriate method, however, has negative impacts on the accuracy and thus the reliability of clustering performance. Given the lack of agreement and disparity of opinions, we set out to comprehensively evaluate the existing normalization techniques to prove empirically which one is the best for the normalization of text length to improve the text clustering performance of news in Arabic. For this purpose, a corpus of 693 stories representing different categories and of different lengths is designed. Data is analyzed using different document length normalization methods along with vector space clustering (VSC), and then the analysis on which the clustering structure agrees most closely with the bibliographic information of the news stories is selected. The analysis of the data indicates that the clustering structure based on the byte length normalization method is the most accurate one. One main problem, however, with this method is that the lexical variables within the data set are not ranked which makes it difficult for retaining only the most distinctive lexical features for generating clustering structures based on semantic similarity. As thus, the study proposes the integration of TF-IDF for ranking the words within all the documents so that only those with the highest TF-IDF values are retained. It can be finally concluded that the proposed model proved effective in improving the function of the byte normalization method and thus on the performance and reliability of news clustering in Arabic. The findings of the study can also be extended to IR applications in Arabic. The proposed model can be usefully used in supporting the performance of the retrieval systems of Arabic in finding the most relevant documents for a given query based on semantic similarity, not document length.**

***Keywords*—Arabic; document length; news clustering; semantic similarity; TF-IDF; VSC**

I. INTRODUCTION

Variation in document length is widely considered an important factor in the validity of text clustering applications. It is essential in clustering applications that all documents within a collection corpus are equally represented [1-3]. Documents in any given corpus, however, can vary considerably in length. As a result, this characteristic can adversely affect the validity and thus reliability of clustering results. In document clustering applications, measuring the semantic similarity within texts can be greatly influenced by vectors that have the largest values. It is a tradition of all the proximity measurements to be dominated by longer documents. In vector space clustering (VSC), the distance between any two documents is determined by their length and the magnitude of the angle between the vectors. This means that if the length of the document increases, the number of times a particular term occurs in the document also increases. Consequently, length becomes an increasingly important determinant of vector clustering in the space. Vice versa, if the documents are short, the angles between the vectors become smaller and as a sequence, short documents will be clustered together [3].

The issue of document length variation has its implications to all text clustering applications including data organization, information retrieval (IR), document retrieval, information filtering, machine learning, text summarization, authorship detection and recognition, and even marketing purposes. In IR applications, for instance, documents that are longer have a higher number of words, hence the values or frequencies for those words are increased, and a document highly relevant for a given term that happens to be short will not necessarily have that relevance reflected in its term frequencies. So if length variation is not considered, longer documents come first irrespective of their relevance to the query. Longer documents have higher term frequency values and naturally, they have—for length reasons more distinct terms. The length factor results in raising the scores of longer documents, which is unnatural. So under the scoring scheme, longer documents are favored simply because they have more terms [4].

Numerous techniques have been devised to account for the variation of length within documents. However, very little has been done in relation to the language processing of Arabic in general and Arabic news in particular. This study addresses this gap in the literature by proposing an integrated model that considers the linguistic peculiarities of Arabic. By way of illustration, a corpus of 693 stories representing different

Paper Submission Date: January 30, 2020
Acceptance Notification Date: February 12, 2020
*Corresponding Author

categories and of different lengths is designed. These represent different topics including politics, sports, family, environment, health, education, technology, and business. Seven normalization methods are compared to choose the best normalization method. These are byte length normalization, cosine normalization, maximum tf normalization, mean normalization, pivoted-cosine normalization, probability normalization, and Z-score. The remainder of this article is organized as follows. Section 2 defines the research problem. Section 3 is a brief survey of VSC and document length normalization methods. Section 4 outlines the data selection and creation processes, methods, and procedures. Section 5 is an analysis of the data using different document length normalization methods. Section 6 is the conclusion.

II. Statement of the Problem

With the explosion in the amount of news and journalistic content being generated in Arabic, there is an increasing need for more reliable clustering tools that can effectively classify raw texts to make it easier for users to identify topics, obtain the information that is relevant to their queries using content-driven groupings of articles. This has been done over recent years using different VSC methods. One problem with these methods, however, is document length variation which is a normal issue. In spite of the development of different techniques for addressing the problem of variation in document length, they cannot be universally applied to all languages. In other words, standard normalization systems of document length have traditionally ignored the issue of language peculiarities which has negative impacts on the validity and thus reliability of such methods. In natural language processing (NLP) of Arabic, the specific linguistic properties play a significant role in the success of NLP applications [5-8]. It is essential for NLP systems thus to consider the peculiarities of Arabic for more reliable results. Furthermore, there is no agreement on the best method to be selected. In VSC applications, different length normalization methods can yield different analyses of the same data set, and that there is no obvious way of selecting the best one. The choice of an inappropriate method, however, will have negative impacts on the accuracy and thus the reliability of clustering performance. The proposed solution is to analyze the data using different document length normalization methods and then to select the analysis on which the clustering structure agrees most closely with the bibliographical information of the news stories.

III. Literature Review

The literature suggests that recent years have witnessed the development of numerous text clustering methods ad algorithms. These include Explicit Semantic Analysis (ESA), Latent Semantic Analysis (LSA), Self-Organizing Maps (SOMs), Sensitive Text Clustering, Vector Space Clustering (VSC), and Word Sense Clustering. (VSC), however, remains among the popular and reliable methods in text clustering applications for its accuracy and effectiveness in different clustering applications. VSC is still widely used in different natural language processing (NLP) applications including data mining, information retrieval (IR), document organization and browsing, corpus summarization, and document classification [9-12]. It is used in different tasks and for different purposes including marketing, grouping similar documents (news, tweets, academic articles, etc.) and the analysis of customer/employee feedback, and discovering meaningful implicit subjects across all documents.

VSC is simply a technique where documents are compared with each other than indexed or classified in terms of their similarity or distance based on the words they contain. It can be defined as the organization of a collection of documents usually represented by a vector space model into distinct clusters based on similarity. The theory was first developed by Salton [13] essentially for IR purposes four decades ago and since then it has become a standard tool in IR systems. The underlying formula of VSC is initially to extract all useful information within a document collection and record it in an index known as a vector space. Then a proximity measurement is used to compute the semantic similarity among the documents with the purpose of grouping similar documents together.

In spite of its popularity and extensive use, VSC has many challenges that have negative impacts on the clustering performance and accuracy. In this regard, many studies have doubted the effectiveness of VSM as it is wholly based on lexical semantics with no regard to the importance of context in identifying intended meanings [14-17]. Likewise, some studies have argued that VSC is less effective in clustering and ranking web pages since these have some special features such as hyperlinks and structural information, which inevitably have additional information and these are ignored in VSC applications.

The main problems with VCS are thus associated with the issue of selecting *appropriate features* of documents that should be used for clustering. Different studies have referred to the limitations of VSC methods in terms of extracting the most distinctive features within datasets [18-20]. For a better feature selection performance, however, some issues need to be addressed. These include document length variation. This issue represents a challenge to the accuracy of clustering performance. The problem is that in the representation of data, the same term usually occurs repeatedly in long documents and that the vocabulary of a long document is usually large. This has the effect that long documents are clustered together and in the same way, short documents are clustered together without any regard to thematic criteria [21]. In other words, clustering is generated based on document length, not semantic similarity. The literature suggests that different techniques have been developed in order to address the issue of document length variation in text classification. These are referred to as document length normalization (DLN) techniques. DLN is a way of penalizing the term weights for a document in accordance with its length. DLN has been one of the central topics of interest in IR and document clustering theory and applications for many years [2, 22, 23]. These include cosine normalization, relative frequency, maximum term frequency, mean term frequency, probability normalization, byte length normalization, and likelihood of relevance. The basic principle of all these techniques is that text length is adjusted so that long texts are not favored simply

because they have more terms. Here is a short review of some of the most commonly used length normalization techniques.

A. Mean Document Length Normalization

Mean document length normalization is one of the simplest and most straightforward normalization methods. It involves the transformation of the row vectors of the data matrix in relation to the average length of documents in the corpus using the function.

$$M_i = M_i(\frac{\mu}{length\ (C_i)})$$

Where

M*i* is the matrix row representing the frequency profile of any document collection C,

Length (C*i*) is the total number of letter bigrams in C*i*, and

μ is the mean number of bigrams across all documents in C:

$$\mu = \sum_{i=1}^{m} \frac{length\ (C_i)}{m}$$

The values of each row vector M*i* are multiplied by the ratio of the mean number of bigrams per document across the collection C to the number of bigrams in document c*i*. The longer the document, the numerically smaller the ratio is, and vice versa. This has the effect of decreasing the values in the vectors that represent long documents, and increasing them in vectors that represent short ones, relative to average document length [3, 24-26].

B. Cosine Normalization

Cosine normalization is the most commonly used technique in the vector space model. Cosine normalization was developed some decades ago with early information retrieval (IR) efforts; nevertheless, it remains one of the best normalization methods. The underlying principle of cosine normalization is that all documents in a given collection are represented equally. In this process, all row vectors of the matrix are transformed so as to have unit length and are made to lie on a hypersphere of radius 1 around the origin so that all vectors are equal in length [27-30]. Accordingly, variation in the lengths of documents and, correspondingly, of the vectors that represent them cannot be a factor [31]. One main problem however with cosine normalization is that it tends to be more biased towards shorter documents. This observation is quite obvious in IR applications where it tends to retrieve shorter documents more than longer documents [32].

C. Probability Normalization

This is a widely used method whereby the frequency values in each vector row are divided by the sum of frequencies in that row. This has the effect of replacing absolute frequency values, whose magnitudes are dependent on document size, with probabilities, which are not. In practice, probability normalization gives satisfactory results when dealing with reasonably small numbers of variables [33, 34].

Examination of the literature shows that there is no one strong conclusion recommending one approach. Besides, many of these have not been tested for the clustering of news in Arabic to find the best approach. Given the lack of agreement and disparity of opinions, we set out to comprehensively evaluate the existing normalization techniques to prove empirically which approach is the best for the normalization of text length to improve the text clustering performance of news in Arabic.

IV. Methodology

To address the research problem, this study is based on experimenting with different normalization techniques to propose a reliable normalization method for the text clustering of news in Arabic. In so doing, a corpus of 693 stories representing different categories and of different lengths is designed. Stories were derived from four different newspapers. These are *Al-Ahram* (Egypt), *Ash-Sharq Al-Awsat* (Saudi Arabia, located in London), *Al-Bayan* (United Arab Emirates), and *Al-Ghad* (Jordan). The selected stories represent different topics including politics, sports, family, environment, health, education, technology, and business as shown in Table I.

The size of the documents ranges from 01 KB to 480 KBs. This is shown in Table II.

This study adopts the vector space model (VSM) for the mathematical representation of data. The reason is that it is conceptually simple as well as it is convenient for computing semantic similarity within documents. The model is usually referred to as a 'bag of words' where a text is represented as a string of words disregarding context and/or word order. Each document is represented by the number of occurrences of each word in the document in Euclidean vector space where each token in the vector corresponds to a unique/given word in the matrix [35, 36]. In VSM, a document is mathematically represented by a vector of index words extracted from the document, with associated weights representing the lexical frequency of these words in the document and within the whole corpus collection. A data Matrix M*ij* was built in which rows Mi represent the documents and columns M*j* the lexical type variables, and the value at the M*ij* is frequency of lexical type *j* in document *i*. The data matrix M*ij* was built out of the lexical variables representing the 693 texts.

TABLE. I. News Categories

Topic	**Number of Documents**
Business	113
Education	87
Environment	78
Family	81
Health	84
Politics	82
Sports	109
Technology	59
Total	693

TABLE. II. THE LENGTHS OF THE DOCUMENTS IN THE CORPUS

Size	Number of documents
From 01 KB to 10 KBs	97
From 11 KBs to 50 KBs	108
From 51 KBs to 100 KBs	102
From 101 KBs to 200 KBs	86
From 201 KBs to 300 KBs	84
From3 01 KBs to 440 KBs	111
From 401 KBs to 500 KBs	105
Total	693

V. ANALYSIS

In this section, the selected data is analyzed using different document length normalization methods using $\mathcal{K}$-means clustering, one of the simplest and most popular VSC methods. In this process, every data point (the news stories in our case) is assigned to the closest center or nearest mean based on their Euclidean distance. Then, new centers are calculated and the data points are updated. This process continues until there are no further iterations and changes within the clusters as seen in Figure 1.

Initially, the selected texts were clustered without the use of any normalization method. The matrix M693 was assigned into two main clusters, which can be called A and B as shown in Figure 2.

Examination of the two clusters, however, shows that the texts do not cluster coherently in terms of thematic criteria, and the clustering, in fact, makes no obvious sense in terms of anything one knows about them and their subject matters. The reason is that there is a progression from the longest texts at the top of the tree to the shortest at the bottom; when correlated with cluster structure, it is easily seen that they have been clustered by length, so that A contains the longest texts and B the shortest. The idea is that in vector space, the distance between any two vectors in a space is determined by the size of the angle between the lines joining them to the origin of the space's coordinate system, and by the lengths of those lines [3, 23, 37]. Using external criteria methods, the clustering structure generated herein was evaluated in terms of the prior knowledge and information obtained about the news stories. Clustering accuracy was estimated to be only 17%. This supports the hypothesis that the lack to address variation in document length in VSC applications has negative effects on the accuracy and reliability of clustering performance. Clustering performance is thus improved when a normalization method compensates for length in all documents so all lexical entries are equally represented. This will have the effect that documents will be clustered based on semantic similarity, not document length. The next step then is to try different normalization methods to choose the most appropriate normalization method for the text clustering of news in Arabic where documents can be clustered based on semantic similarity, not document length. Seven normalization methods are used. These are alphabetically ordered as follows: byte length normalization, cosine normalization, maximum tf normalization, mean normalization, pivoted-cosine normalization, probability normalization, and Z-score.

Using byte length normalization method, the row vectors of the data matrix M693 were normalized to compensate for the variation in length among the texts so that their lexical frequency profiles could be meaningfully clustered. Texts were assigned into five clusters as shown in Figure 3.

This process is repeated with cosine normalization, maximum tf normalization, mean normalization, and probability normalization methods. Accuracy rates are represented in Table 3.

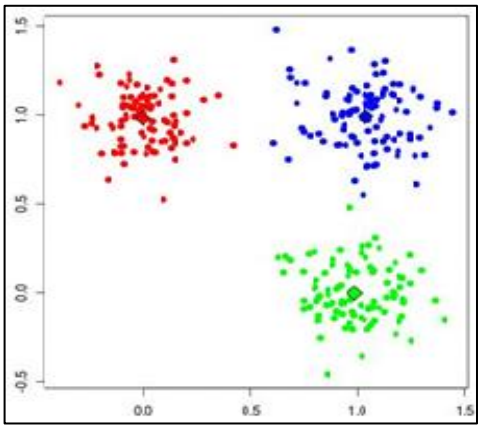

Fig. 1. Example of K-Means Clustering.

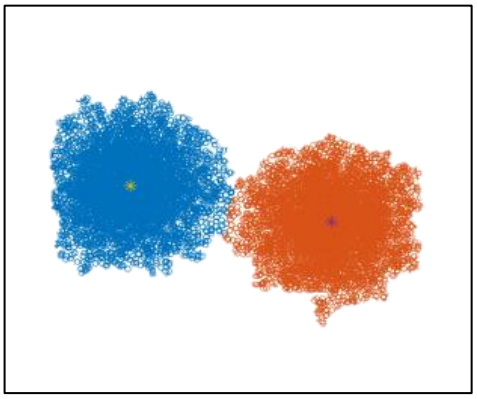

Fig. 2. K-Means Clustering of the Data Matrix M693 without the use of any Normalization Method.

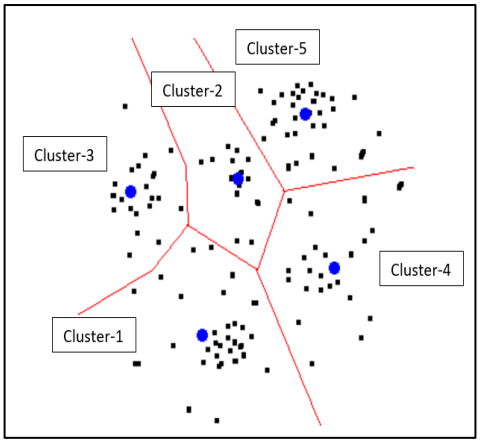

Fig. 3. K-Means Clustering of the Data Matrix M693 based on the Byte Length Normalization Method.

TABLE. III. ACCURACY RATES OF THE SELECTED DOCUMENT LENGTH NORMALIZATION METHODS

Normalization Method	Accuracy Rate
Byte length/size normalization	81%
Cosine normalization	73%
Maximum tf normalization	66%
Mean normalization	73%
Pivoted-cosine normalization	78%
Probability normalization	65%
Z-score	69%

The analysis indicates that byte normalization is the best method in terms of representing the terms within all the documents equally. One advantage of this method is addressing the issue of variation without distorting the byte size of documents. However, the analysis pointed to a major limitation with this method. It represents the documents equally without ranking of the lexical variables within the data set. For improving the document length normalization performance, thus, it is suggested that term frequency-inverse document frequency (TF-IDF) is used alongside the byte normalization method. The hypothesis is that TF-IDF will have the effect of ranking the words within all the documents so that only those with the highest TF-IDF values will be retained [20, 38-40].

Given that the highest TF-IDF variables are the most important, each column was calculated using the function:

$$tfid(t_j) = tf(t_j) log_2(\frac{M}{df_j})$$

Where $tf(t_j)$ is the frequency of *term* t_j across all documents in the data matrix M693. Using the above formulation, the TF-IDF of some lexical type A that occurs once in a single document is 1 x $\log_2$ (1000 / 1) = 9.97, and the TF-IDF of a type B that occurs 400 times across 3 documents is 400 x $\log_2$ (1000 / 3) = 3352, that is, B is far more useful for document differentiation than A, which is more intuitively satisfying than the alternative. The variables are sorted in descending order as shown in Figure 4 and only the highest 1500 lexical variables within the data corpus were retained.

As a final step, a K-means clustering based on the byte normalization method and TF-IDF analysis was carried out. The documents were assigned to clearly define six groups (as seen in Figure 5) which correspond to a great extent to the information obtained about these documents with an accuracy rate of around 95.6%.

It can be thus claimed that the use of a single normalization method is not effective in terms of the document clustering of news in Arabic. The performance of normalization performance; however, can be improved with the use of TF-IDF alongside the byte normalization method.

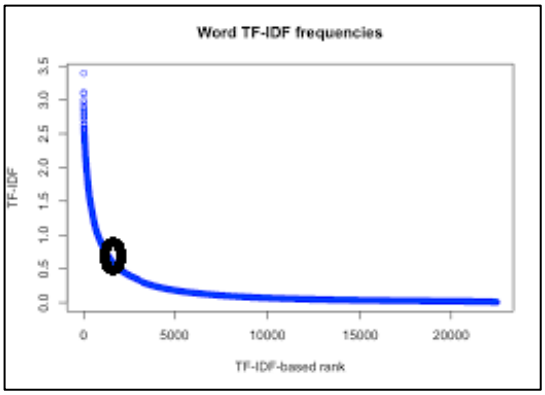

Fig. 4. Term Ranking using TF-IDF.

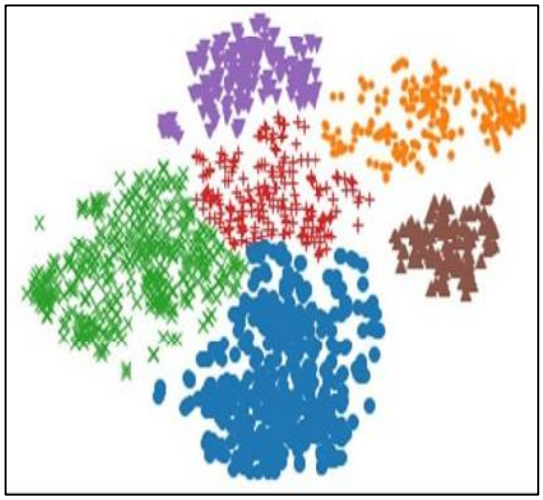

Fig. 5. K-Means Clustering of the Data Matrix M693 based on Byte Length Normalization Method and TF-IDF.

VI. CONCLUSION

This study addressed the issue of the effect of document length variation on the accuracy of the news clustering in Arabic. Different normalization methods were used and compared. It was found out that the byte length normalization method despite its limitations is the most appropriate for clustering applications of news in Arabic. In order to address these limitations, this study proposed the use of TF-IDF alongside this normalization method. The proposed model had the effect of improving the function of the byte normalization method and thus increasing the accuracy rate of the clustering performance. It can be finally concluded that the use of a single normalization method is not sufficient in addressing the issue of document length variation. The findings of the study can also be extended to IR applications in Arabic. The proposed model can be usefully used in supporting the performance of the retrieval systems of Arabic in terms of finding the most relevant documents for a given query based on semantic similarity, not document length.

ACKNOWLEDGMENTS

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

REFERENCES

[1] Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[2] C. X. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Association for Computing Machinery and Morgan & Claypool Publishers, 2016.

[3] H. Moisl, Cluster Analysis for Corpus Linguistics. De Gruyter, 2015.

[4] B. Mitra and N. Craswell, An Introduction to Neural Information Retrieval. Now Publishers, 2018.

[5] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, no. 4, 2009.

[6] I. Guellila, H. Saâdaneb, F. Azouaoua, B. Guenic, and D. Nouve, "Arabic natural language processing: An overview," Journal of King Saud University - Computer and Information Sciences, 2019.

[7] S. L. Maire-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic Natural Language Processing and Machine Learning-Based Systems " IEEE Access, vol. 7, pp. 7011-7020, 2019.

[8] N. Y. Habash, Introduction to Arabic Natural Language Processing (Synthesis Lectures on Human Language Technologies). Morgan and Claypool Publishers, 2010.

[9] H. Lane, H. Hapke, and C. Howard, Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Manning Publications, 2019.

[10] H. Saggion and G. Hirst, Automatic Text Simplification. Morgan & Claypool Publishers, 2017.

[11] A. K. Luhach, D. Singh, P. A. Hsiung, K. B. G. Hawari, P. Lingras, and P. K. Singh, Advanced Informatics for Computing Research: Second International Conference, ICAICR 2018, Shimla, India, July 14–15, 2018, Revised Selected Papers (no. pt. 1). Springer Singapore, 2018.

[12] T.-U. Jang, W. Lim, Y.-M. Yang, and B. M. Kim, "Classification of the motor imagery EEG signal using vector quantization and K-nearest neighbors' algorithm," International Journal of Advanced and Applied Sciences, vol. 2, no. 12, pp. 72-77, 2015.

[13] G. Salton, The Smart retrieval system experiments in Automatic document processing. Englewood Cliffs: Prentice Hall Inc., 1971.

[14] S. Deerwester, T. D. Susan, W. F. George, K. L. Thomas, and H. Richard, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[15] T. K. Landauer, P. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," Discourse Processes, vol. 25, no. 2-3, pp. 259-84, 1998.

[16] E. Gabrilovich and S. Markovitch, "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proceedings of the Twenty-First National Conference on Artificial Intelligence, pp. 1301--1306, 2006.

[17] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis," Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606--1611, 2007.

[18] R. Kaspar and B. Horst, Graph Classification And Clustering Based On Vector Space Embedding. World Scientific Publishing Company, 2010.

[19] A. E. Hassanien, C. Grosan, and M. F. Tolba, Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems. Springer International Publishing, 2015.

[20] C. X. Zhai, Statistical Language Models for Information Retrieval. Morgan & Claypool, 2009.

[21] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," 19th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval pp. 21–29, 1996.

[22] L. M. Q. Abualigah, Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. Springer International Publishing, 2018.

[23] T. Strzalkowski, Natural Language Information Retrieval. Springer Netherlands, 2013.

[24] W. Verhaegh, E. Aarts, and J. Korst, Algorithms in Ambient Intelligence. Springer Netherlands, 2013.

[25] A. W. Santoso et al., "A fuzzy approach for speckle noise reduction in SAR images," International Journal of Advanced and Applied Sciences, vol. 3, no. 5, pp. 33-38, 2016.

[26] R. Al-Jabar, "MFCC features with kernel PCA for speaker verification system " International Journal of Advanced and Applied Sciences, vol. 1, no. 6, pp. 37-44, 2014.

[27] A. Albalate and W. Minker, Semi-Supervised and Unsupervised Machine Learning: Novel Strategies. Wiley, 2013.

[28] U. S. Tiwary and T. Siddiqui, Natural Language Processing and Information Retrieval. OUP India, 2008.

[29] H. M. Blanken, A. P. de Vries, H. E. Blok, and L. Feng, Multimedia Retrieval. Springer Berlin Heidelberg, 2007.

[30] T. Sing, S. Siraj, R. Raguraman, P. Marimuthu, and K. Nithiyananthan, "Cosine similarity cluster analysis model based effective power systems fault identification," International Journal of Advanced and Applied Sciences, vol. 4, no. 1, pp. 123-130, 2017.

[31] H. Moisl, "Using Electronic Corpora in Historical Dialectology Research: The Problem of Document Length Variation," in Studies in English and European Historical Dialectology, vol. 98, M. Dossena and R. Lass, Eds., 2009, pp. 67-90.

[32] A. K. Singhal, Term Weighting Revisited. Cornell University, 1997.

[33] W. J. Stewart, Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling. Princeton University Press, 2009.

[34] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Science, 2011.

[35] Y. Ozgur, "Empirical selection of nlp-driven document representations for text categorization," Syracuse University, 2006.

[36] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, 2002, p. 224.

[37] W. Wu, H. Xiong, and S. Shekhar, Clustering and Information Retrieval. Springer US, 2013.

[38] L. Azzopardi et al., Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings. Springer, 2009.

[39] T. Roelleke, Information Retrieval Models: Foundations and Relationships. Morgan & Claypool Publishers, 2013.

[40] T. W. Miller, Web and Network Data Science: Modeling Techniques in Predictive Analytics. Pearson Education, 2014.