iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41554

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Comparing the Performance of SOM with Traditional Methods for Document Clustering Using Wordnet Ontologies

Abhishek Sawalkar[1], Mohit Mandlecha[2], Dnyanesh Kulkarni[3], Dr. Ratnamala S. Paswan[4]
[1, 2, 3, 4]*Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India*

Abstract: ***Retrieving useful information has become challenging due to the rapid expansion of web material. To improve the retrieval outcomes, efficient clustering methods are required. Document clustering is the process of identifying similarities and differences among given objects and grouping them into clusters with comparable features. We used WordNet lexical as an addition to compare several document clustering techniques in this article. The suggested method employs WordNet to determine the relevance of the concepts in the text, and then clusters the content using several document clustering algorithms (K-means, Agglomerative Clustering, and self-organizing maps). We wish to compare alternative ways for making document clustering algorithms more successful.***

Keywords: Document clustering, Clustering technique, Self-organizing maps, WordNet, K-means, Hierarchical Clustering.

I. INTRODUCTION

Document clustering has been studied for use in a variety of text mining and information retrieval applications. Document clustering was first researched as a way to improve the precision or recall of information retrieval systems and as a quick approach to discover a document's closest neighbors. Clustering has lately been proposed for use in viewing a collection of papers or organizing search engine results returned in response to a user's query. Document clustering has also been used to create hierarchical groupings of documents automatically. The hierarchical, partitive, and neural network algorithms are now the most used document cluster analysis approaches. In this study, a document clustering method based on Self Organizing Maps (SOM) is compared to K-means and Fuzzy c-means, two classic clustering methods. We've utilized two distinct parameters to evaluate the findings, and we've seen the benefits and drawbacks of each strategy. We'll also use WordNet to determine the text's relevance of concepts and examine which clustering technique takes advantage of WordNet the best. We'll also feed it into the clustering process using a variety of embeddings. For producing word embeddings, we'll utilise Doc2Vec and GloVe, and we'll compare the results to see which one performs better.

II. LITERATURE REVIEW

SR NO	PAPER NAME	AUTHORS	CONCLUSION
1	Document Clustering Based on Text Mining K-Means Algorithm Using Euclidean Distance Similarity	• Laxmi Lydia • P. Govindasamy • Lydia	This paper describes the document clustering process based on the clustering techniques, partitioning clustering using K-means and also calculates the centroid similarity and cluster similarity.
2	WordNet-based Text Document Clustering	• J Sedding • Dimitar Kazakov	In this research, naïve, syntax-based disambiguation is attempted by assigning each word a part-of-speech tag and by enriching the 'bag-of- words' data representation often used for document clustering with synonyms and hypernyms from WordNet.

3	Using the self-organizing map for clustering of text documents	• Dinolsa • V.P. Kallimani	In this paper fuzzy c-mean is used for document clustering with different values of clusters corresponds to different fuzzy partitions
4	The self-organizing map	1) T. Kohonen	The self-organizing map algorithm is reviewed, focusing on best matching cell selection and adaptation of the weight vectors.
5	Modern hierarchical, agglomerative clustering algorithms	2) Daniel Müllner	This paper presents algorithms for hierarchical, agglomerative clustering which perform most efficiently in the general-purpose setup that is given in modern standard software.

III. METHODOLOGY

The following methodology is used to cluster text documents. The nine phases are outlined in the methodology.. These phases are Dataset collection, preprocessing, Wordnet, Doc2Vec (DBOW), Doc2Vec (PV-DM), K-Means clustering, Self-organizing map, Agglomerative Clustering and Cluster evaluation. The flow of the paper is as follows:

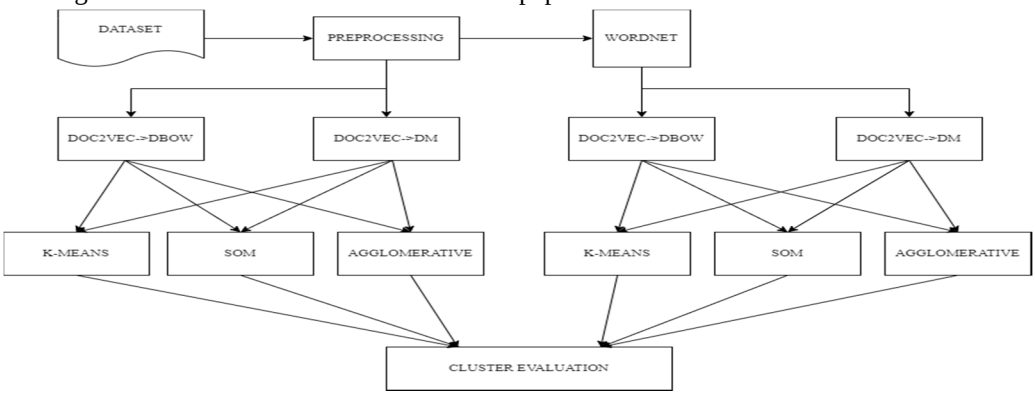


Figure 1: The proposed approach for clustering

A. Dataset Used

There are three different datasets used. The three datasets are:

1) 1490 recoded of BBC news articles downloaded from BBC website
2) Entertainment articles with 50000 records, and
3) Sports articles with 1000 records downloaded from UCI Machine Learning Repository.

We chose 1000 records per topic in this study to ensure that the entire dataset is balanced. The total number of records in the resulting dataset goes to 3000. It should be noted that some BBC news records may incorporate content pertaining to entertainment or sports, adding to the total dataset's complexity.

B. Text Preprocessing

Text Preprocessing is used for extracting information from unstructured data. A dataset is made up of a large number of text documents that have been collected from various sources. The following approaches are used to efficiently preprocess text documents. Stopwords are removed and tokenization is used.

1) *Tokenization:* The first step in any analysis is to tokenize the data. Tokenization is mostly used to identify significant keywords. The removal of stopwords minimises text data and improves system efficiency.
2) *Stopwords:* Stopwords are words like "also," "and," "or," "can," and "this" that appear repeatedly but have no meaning. After that, the preprocessed data is sent to WordNet for additional processing. \

We also pass this basic preprocessed input directly to the embedding layer, where it will be turned into numerical vectors, to see if the WordNet enhances the clustering procedure or not.

C. WordNet with Part of Speech tags

WordNet is a lexical database of semantic relations between words in more than 200 languages. WordNet links words into semantic relations including synonyms, hyponyms, and meronyms. The synonyms are grouped into synsets with short definitions and usage examples. WordNet can thus be seen as a combination and extension of a dictionary and thesaurus.

Part-of-Speech tags: The part-of-speech tagging (PoS) solves semantic ambiguity to some extent (40% in some of the tests). Based on this observation, the naïve word sense disambiguation by PoS tagging can help to improve clustering results.

We applied Wordnet with PoS tags to the dataset and then clustered using and without this strategy, attempting to determine whether Wordnet with PoS tags aided the clustering algorithm in general or only in certain types of clustering. Then this processed dataset is sent to the Doc2Vec embedding layer where the document is represented in vector form.

D. Doc2Vec

Doc2vec (also known as paragraph2vec or sentence embeddings) is a modification of the word2vec approach that allows for the unsupervised learning of continuous representations for bigger blocks of text, such as sentences, paragraphs, or complete documents. The basic purpose of doc2vec is to convert a document into a numerical representation, regardless of its length. Doc2Vec can be done in two ways: one utilising the Distributed Bag of Words (DBOW) technique, and the other using the PV-DM (Distributed Memory Version of Paragraph Vector) algorithm.

1) Doc2Vec using Distributed Bag of Words (DBOW)

The Distributed Bag Of Words (DBOW) model differs from the PVDM model in a few ways. The model does not use the context words in the input, instead attempting to predict words at random from the paragraph in the output. In the example above, let's imagine the model is learning by predicting two sampled words. As a result, the document vector is learned by sampling two words from the,span> cat, sat, on, the, sofa, as shown in the image.

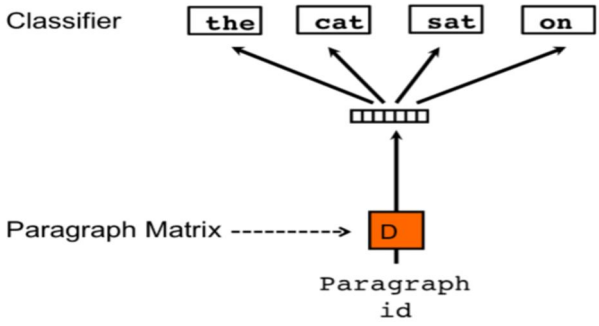


Figure 2: Distributed Bag of Words (DBOW) Model

2) Doc2Vec using paragraph Vector Distributed Memory (PV-DM)

The basic idea behind PV-DM is inspired from Word2Vec. In the CBOW model of Word2Vec, The CBOW model of Word2Vec tries to anticipate the centre word from context. A sentence “The cat sat on sofa”, CBOW model would try to predict the word “on” given the context words — the, cat, sat and sofa. Similarly, in PV-DM, the idea is to randomly sample consecutive words from a paragraph and predict a word that is in the center or close to center from the set of words from the sentence by taking as input — the context words and a paragraph id.

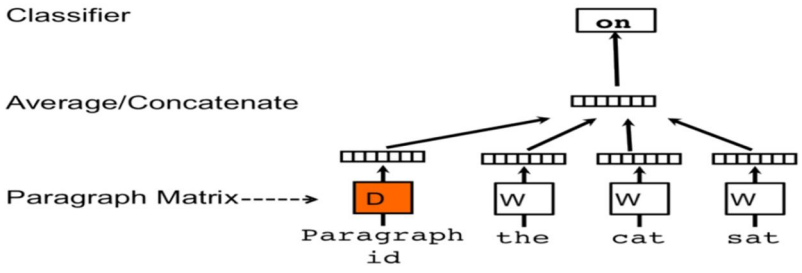


Figure 3: Paragraph Vector Distributed Memory (PV-DM) Model

E. K-means Clustering Algorithm

The k-means algorithm have a input which specifies the number of clusters to be formed as 'k' and divides a set of n-objects into K-clusters, resulting in high intra-cluster similarity and low inter-cluster similarity. The overall mean value of the objects in the cluster, which may be thought of as the cluster's centre of gravity, is used to determine cluster similarity. The Euclidean distance is used to determine the phrases' syntactical similarity.

Steps of K-Means clustering

1) Choose k observations to serve as the cluster's initial centroids (seeds).
2) Assign each observation to the cluster with the most centroids that is closest to it (for example, in Euclidean sense).
3) Recalculate the coordinates of the k centroids once all of the observations have been assigned.
4) Continue until the cluster centroids do not change any longer.

F. Self-Organizing Map

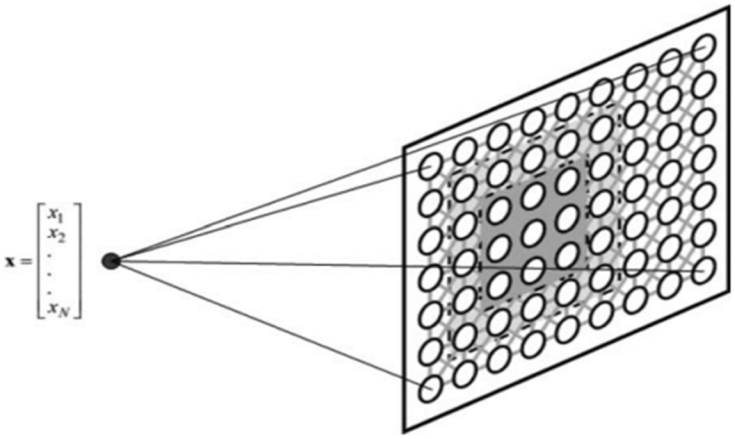


Figure 4: Kohonen's self-organizing map model. The input is connected to every cell in the neural network sheet.

The self-organizing map (SOM) is a clustering algorithm that organises data using a similarity measure derived from Euclidean distance calculations. The goal of this approach is to identify a winner-takes-all neuron that will find the case that is the most similar. Kohonen proposed the SOM, which is based on the assumption that the neurons in the human brain are connected. Collectivism can be implemented by feedback, and therefore in the network, where several surrounding neurons react collectively when events are stimulated. Neighboring neurons are influenced when a neuron is engaged during the learning process.

G. Agglomerative Clustering Algorithm

Agglomerative clustering is a frequent used clustering algorithm for hierarchical clustering. AGNES is another name for it (Agglomerative Nesting). Each data point is a singleton cluster at first for the agglomerative algorithm. Following that, clusters are merged together one by one until all clusters have been merged into a single large cluster holding all items. The output is represented as a dendrogram(tree-based representation of the objects).

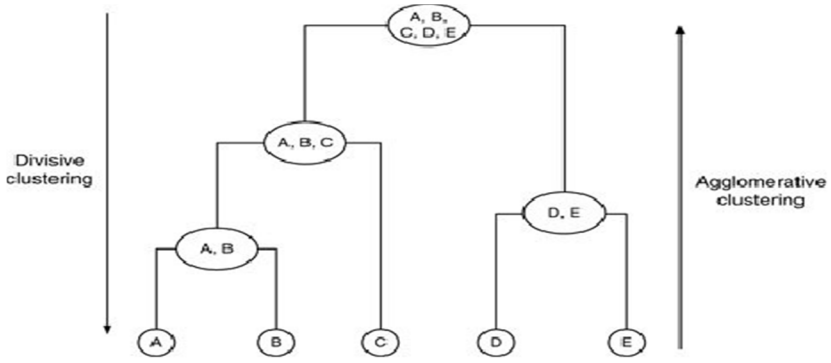


Figure 4: Dendrogram describing Agglomerative Clustering.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

IV. EVALUATION METRICS

The evaluation metrics used in this paper are precision, recall, f-measure and accuracy.
Some important definitions:

1) True positive (TP) = the total number of objects correctly identified as True
2) False positive (FP) = the total number of objects incorrectly identified as True
3) True negative (TN) = the total number of objects correctly identified as False
4) False negative (FN) = the total number of objects incorrectly identified as False

A. Precision

This measure retrieves the number of correct text documents out of the number of total text documents made by the system.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved (TP)}}{\text{Number of documents retrieved (TP+FP)}}$$

B. Recall

This measure retrieves the number of correct text documents made by the system, out of the number of all possible text documents.

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved (TP)}}{\text{Number of relevant documents (TP+FN)}}$$

C. Accuracy

The accuracy of a measurement is how close a result comes to the true value.Systemetic error or Inaccuracy is quantified by the average difference (bias) between a set of measurements obtained with the test method with a reference value or values obtained with a reference method.

$$\text{Accuracy} = \frac{TP + TN}{(TP+FP+TN+FN)}$$

D. F-MEASURE

This measure is a combination of the precision and recall measures used in machine learning.

$$\text{F-Measure} = \frac{2^* (\text{Precision}^*\text{recall})}{(\text{Precision} + \text{recall})}$$

V. RESULT

A. Results of Self-Organizing Map

TABLE I
Performance of Self-Organization Map using different approaches

SR NO	Model	Precision (%)	Recall (%)	Accuracy (%)	F-Score (%)
1	SOM using Doc2Vec PV-DM embeddings	79	80	79	79
2	SOM using Doc2Vec PV-DM embeddings + WordNet	87	91	87	87
3	SOM using Doc2Vec DBOW embeddings	93	93	93	93
4	SOM using Doc2Vec DBOW embeddings + WordNet	96	96	95	95

As we can see from the Table I, self-organizing map doesn't work as efficient with just basic pre-processing steps. Both DBOW and PV-DM embeddings standalone doesn't work as good as with WordNet, the performance of both types of embeddings gets boosted when WordNet is applied to the approach. Also, we can see that DBOW (Distributed Bags of Words) performs better than PV-DM (Distributed Memory).

B. Results of K-means Clustering Algorithm

TABLE II
Performance of K-means Clustering Algorithm using different approaches

SR NO	Model	Precision (%)	Recall (%)	Accuracy (%)	F-Score (%)
1	K-means using Doc2Vec PV-DM embeddings	94	93	93	93
2	K-means using Doc2Vec PV-DM embeddings + WordNet	95	95	94	95
3	K-means using Doc2Vec DBOW embeddings	99	99	98	99
4	K-means using Doc2Vec DBOW embeddings + WordNet	97	97	97	97

As we can see from the Table II, K-means clustering does a very good job doing clustering with and without the help of WordNet. But we do see a bit performance boost when we use PV-DM embeddings with WordNet, whereas there is a bit dip in the performance of K-means when we use DBOW embeddings with WordNet. Also, we can see that DBOW embeddings performs better than PV-DM embeddings.

C. Results of Agglomerative Clustering Algorithm

TABLE III
Performance of Agglomerative Clustering Algorithm using different approaches

SR NO	Model	Precision (%)	Recall (%)	Accuracy (%)	F-Score (%)
1	Agglomerative using Doc2Vec PV-DM embeddings	95	94	94	94
2	Agglomerative using Doc2Vec PV-DM embeddings + WordNet	96	96	96	96
3	Agglomerative using Doc2Vec DBOW embeddings	99	99	99	99
4	Agglomerative using Doc2Vec DBOW embeddings + WordNet	33	33	35	33

As we can see from the Table III, Agglomerative clustering does a very great job doing clustering. The WordNet gives a slight increase in efficiency for PV-DM embeddings whereas, it gives a very bad results when used with DBOW embeddings. Also, Agglomerative clustering performs better with standalone DBOW embeddings rather than adding WordNet component to the approach or using PV-DM embeddings with or without the support of WordNet.

D. Overall Observations

Looking at all the tables we can say that K-means and Agglomerative Clustering algorithms are more capable than self-organizing map for clustering when only basic preprocessing steps are performed. When we add WordNet to self-organizing map, the results that the SOM shows are almost at the level of K-means clustering algorithm, whereas the other two clustering algorithms doesn't show that much of increase in performance. This means self-organizing map uses a lot of help from WordNet with PoS tags. We can also see that in all of the clustering algorithms works better in Distributed Bags of Words (DBOW) than Distributed Memory (PV-DM).

VI. CONCLUSION

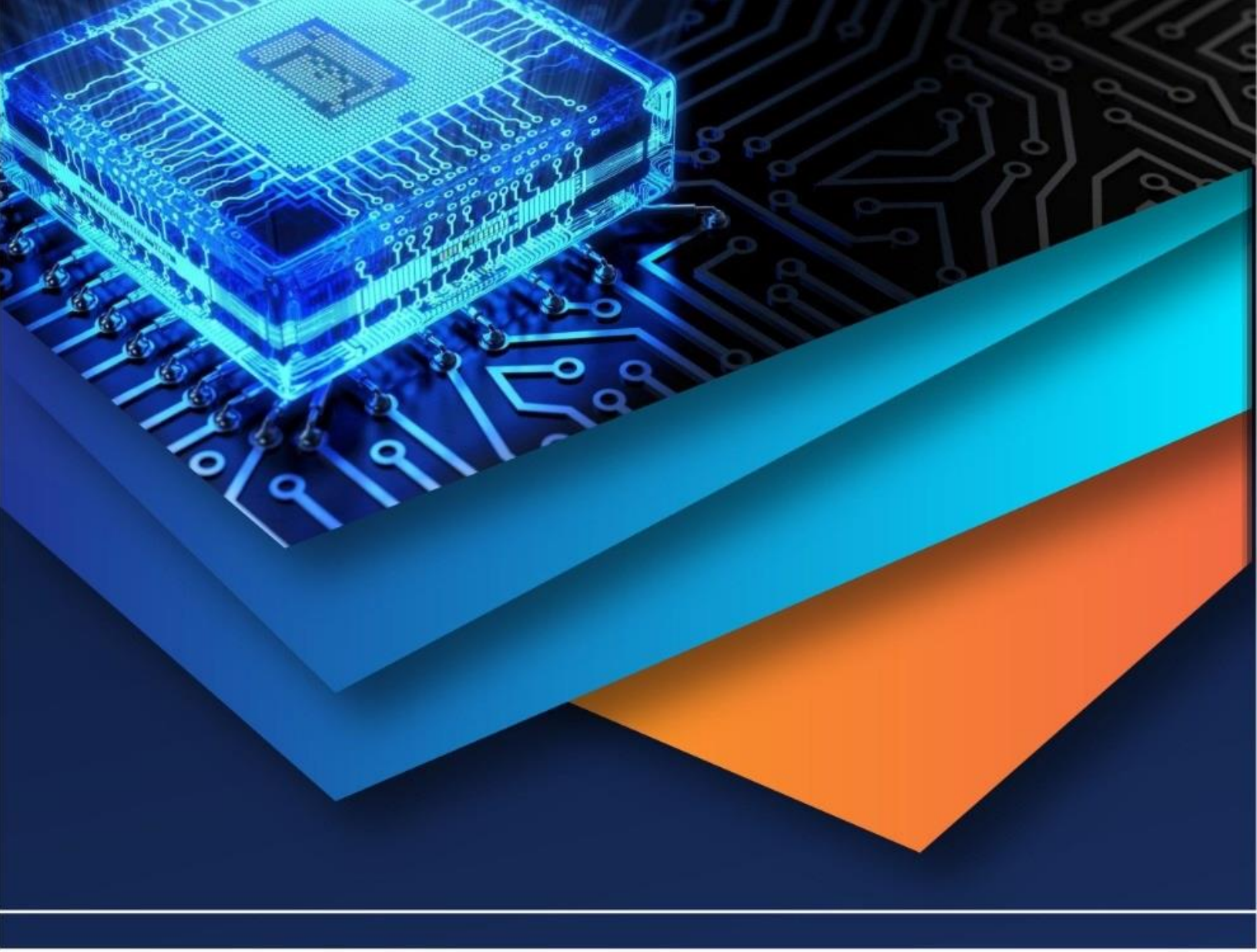
In this paper, we analyzed various clustering algorithms with and without using WordNet as an advance preprocessing unit. We also analyzed how different types of embeddings perform with different clustering algorithms. Then we compared these results with self-organizing map to see how self-organizing map behaves with respect to other clustering algorithms.

VII. ACKNOWLEDGEMENT

The authors of this paper would like to thank Dr. Ratnamala S. Paswan for her support and guidance in making this work possible.

REFERENCES

[1] Lydia, Laxmi & Govindasamy, P. & Lakshmanaprabu, S.K. & Ramya, D. (2018). "Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity", Journal of Advanced Research in Dynamical and Control Systems. 10.
[2] Müllner, Daniel. (2011). "Modern hierarchical, agglomerative clustering algorithms. "
[3] Sedding, Julian & Kazakov, Dimitar. (2004). "WordNet-based Text Document Clustering"
[4] Isa, Dino & Kallimani, Vish & Lee, Lam Hong. (2009). "Using the self organizing map for clustering of text documents. Expert Systems with Applications." 36. 9584-9591. 10.1016/j.eswa.2008.07.082.
[5] T. Kohonen, "The self-organizing map," in Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480, Sept. 1990, doi: 10.1109/5.58325.
[6] R. Kumbhar, S. Mhamane, H. Patil, S. Patil and S. Kale, "Text Document Clustering Using K-means Algorithm with Dimension Reduction Techniques," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 1222-1228, doi: 10.1109/ICCES48766.2020.9137928.
[7] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[8] Pacella, Massimo & Blaco, Marzia. (2016). "On the Use of Self-Organizing Map for Text Clustering in Engineering Change Process Analysis: A Case Study. Computational Intelligence and Neuroscience." 2016. 1-11. 10.1155/2016/5139574..
[9] Kamvar, Sepandar & Klein, Dan & Manning, Christopher. (2002). "Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach."
[10] Goutte, Cyril & Gaussier, Eric. (2005). "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes in Computer Science." 3408. 345-359. 10.1007/978-3-540-31865-1_25..
[11] Bouras, Christos & Tsogkas, Vassilis. (2010). "W-kmeans: Clustering News Articles Using WordNet". 379-388. 10.1007/978-3-642-15393-8_43..
[12] Quoc Le and Tomas Mikolov. 2014. "Distributed representations of sentences and documents". In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–1188–II–1196..
[13] Budiarto, Arif & Rahutomo, Reza & Putra, Hendra & Cenggoro, Tjeng Wawan & Kacamarga, Muhamad & Pardamean, Bens. (2021). "Unsupervised News Topic Modelling wiith Doc2Vec and Spherical Clustering". Procedia Computer Science. 179. 40-46. 10.1016/j.procs.2020.12.007..
[14] G. Wang and S. W. H. Kwok, "Using K-Means Clustering Method with Doc2Vec to Understand the Twitter Users' Opinions on COVID-19 Vaccination". 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2021, pp. 1-4, doi: 10.1109/BHI50953.2021.9508578..
[15] Bilgin, Metin & Senturk Izzet. (2017)."Sentiment analysis on Twitter data with semi-supervised Doc2Vec". 661-666. 10.1109/UBMK.2017.8093492.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY