Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach

Conference Paper · December 2021

DOI: 10.1109/ICCCA52192.2021.9666391

Md. Mehedi Hasan
Daffodil International University
6 PUBLICATIONS 22 CITATIONS
SEE PROFILE

Sadia Tamim Dip
Daffodil International University
5 PUBLICATIONS 21 CITATIONS
SEE PROFILE

T. M. Kamruzzaman
Daffodil International University
2 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Imrus Salehin
Daffodil International University
16 PUBLICATIONS 64 CITATIONS
SEE PROFILE

Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach

Md. Mehedi Hasan
Department of CSE
Daffodil International University
Dhaka, Bangladesh
mehedi15-9021@diu.edu.bd

Sadia Tamim Dip
Department of CSE
Daffodil International University
Dhaka, Bangladesh
sadia15-8638@diu.edu.bd

T. M. Kamruzzaman
Department of CSE
Daffodil International University
Dhaka, Bangladesh
kamruzzaman15-9183@diu.edu.bd

Mst Sonia Akter
Department of CSE
Daffodil International University
Dhaka, Bangladesh
sonia25-901@diu.edu.bd

Imrus Salehin
Department of CSE
Daffodil International University
Dhaka, Bangladesh
imrus15-8978@diu.edu.bd

2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA) | 978-1-6654-1473-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCA52192.2021.9666391

***Abstract*—Since the evolution of digital and online text content, automatic document classification has become a significant research issue. There is a most commonly used machine learning approach to improve this task: an unsupervised approach, where no human interaction or labelling documents are required at any point throughout the whole procedure. This study addressed an approach for movie subtitle document classification using an unsupervised machine learning technique. The dataset has been created, collecting almost 500 English movie subtitle files based on the popular movies of IMDB. Two feature extraction methods have been used and combined with unsupervised machine learning algorithms and a dimension reduction technique has been used to reduce the dimensionality of this work. As unsupervised machine learning techniques, we used Bisecting K-Means, K-Means and Agglomerative Hierarchical Clustering Algorithm; Average link, Single Link and Double link. We assessed that K-means and Bisecting k-means are the best performers of the unsupervised techniques in the term of cluster quality. We addressed the reason for the outliers of the training set and recommended using unsupervised techniques to improve predefining categories and labelling the textual documents in the training set.**

Keywords—subtitle document classification, clustering, unsupervised learning, TF-IDF, BOW.

I. INTRODUCTION

In recent years, the dimensions of text documents are increasing tremendously on the digital libraries, internet sources, news sources and company-wide intranets. This leads to greater attention in developing ways to enable users to organize, navigate and summarize this information efficiently to help them find out what they want. High-quality algorithms for text clustering play key roles in this respect since they have been shown by organizing huge volumes of information into a few significant clusters to provide an spontaneous browsing mechanism [1]. Due to the linguistic richness, text data with unstructured form are more complicated and diverse than structured data in the database [2].

The movie industry has long used different types of data to address issues concerning client interest and proposed target categories. To make important decisions on the sorts of movies that are to be produced and which genres are favored by specific categories is entrenched in all kinds of filming and streaming decisions. A prominent example is Netflix's algorithms, which determine the types of movies and specific movies that its subscribers will most likely to watch next. The internet streaming service of Netflix inevitably relies on strong customer support to retain its customer base. This requirement involves collecting the appropriate population statistics and user preferences to estimate movie preferences correctly.

The technique of text classification can be performed using supervised and unsupervised machine learning algorithms. Clustering and association problems can be grouped as unsupervised machine learning. There are some common examples of unsupervised algorithms like Bisecting K-Means, Agglomerative clustering algorithm, K-Means, KNN (K Nearest Neighbors), etc. Nowadays, learning algorithms are more advanced for classifying or categorizing text by the development of deep learning algorithms such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) [3].

In this study, we developed a movie subtitle document-based classification system using several unsupervised machine learning approaches. These techniques focus on the raw movie subtitle document. In this proposed system, the unsupervised machine learning algorithms we used are Bisecting K-Means clustering, K-Means Clustering and Agglomerative Hierarchical Clustering. We combined the algorithms with TF-IDF and Bag of Word (BOW) as the feature extraction method [4]. To reduce the dimensionality of our proposed work, we used the Information Gain (IG) method. In order to get the best model, we evaluated the unsupervised algorithm using entropy, purity, Overall similarity and F1 measure.

The outline of this paper is as follows, in chapter II we discussed the Related Work. In chapter III, we discussed how to preprocess and represent our movie subtitle documents. In chapter IV, we discussed Methodology and finally in chapter V, we discussed Model Evaluation.

II. RELATED WORK

Dharmadhikari S.C et al. have performed their research on various machine learning techniques for text classification. Their research is working with three types of machine learning techniques respectively unsupervised, supervised and semi-supervised. This proposed system has chosen KNN, SVM, and Decision Tree for supervised learning, K-Means for unsupervised learning, and graph-based algorithm for semi-supervised learning. Finally, they have shown that data source feature representation is crucial for performing better in automatic text classification in machine learning approaches [5].

Buying L. et al. experimented that using feature selection approaches can increase the accuracy and efficiency of text clustering. The validity of three clustering criteria is investigated and utilized to assess clustering outcomes. Due to the high level of feature space with the data sparsity, the efficiency of the clustering algorithm tends to decrease. This unsupervised feature selection method includes four methods; TC, DF, TV and TVQ. Here TV and TC are the two most effective of these four methods. As shown in this paper, the validity criteria of different clusters show different performances [6].

Taras Z. et al., devised a mechanism for classifying texts based on their objectivity and sentiment polarity. Also shown both as a continuum and translating classed documents into a coordinate system that shows the difference between neutral and balanced text. This technique handles both positive and negative sentiment and subjectivity and objectivity as a continuum but instead as discrete groups. They offer an unsupervised classification strategy utilizing the Chinese language to develop its sentiment vocabulary from scratch, starting with a very modest seed vocabulary and expanding it through repeated retraining [7].

Pimwadee C. et al., introduces two strategies for movie review: machine learning and the second of which is semantic orientation. The two methodologies are also used to compare the domains of movie reviews. Factual information is always blended with real-life data and sarcastic terms in composing movie reviews, making movie review mining problematic. Using web-mining techniques, this paper categorizes a vast number of thoughts as positive or negative. This work is helpful for movie review mining and other text classification problems, such as discrete "flames" signals [8].

Ji I. et al., examined the feasibility and limitations of document classification and conducted a case study to demonstrate its utility. The idea of this paper is built on typical modeling and literature approaches for document classification that uses unsupervised learning approaches. However, the accuracy of the prediction is directly proportional to the number of nearest neighbors. Unsupervised learning algorithms, when used Unsupervised learning algorithms correctly, may still be capable of delivering acceptable document classification results [9].

Antoine D. et al., has evaluated the XML clustering problem in this paper because XML clustering is evaluated against semantic categories in the INEX mining track. Using KNN, the need for tools to manage XML document collections has primarily been data-centric, but there is also an increasing number of text-centric document collections. They created a vector space model that was enriched with several sorts of textual and structural features that were combined at once or in a two-step manner, integrating structural characteristics first and then textual characteristics [10].

III. Document Preprocessing and Representation

In order to classify or cluster text documents using machine learning algorithms, Documents need initially to be preprocessed. The document should be turned into a representation appropriate for the application of learning algorithms in the preprocessing stage. The vector space model proposed by Salton et al. [11] is the most often used approach for text document representation, which we have also employed in our proposed model. Each of the movie subtitle documents is denoted as a vector d. Each component in the vector d indicates a unique term within the document collection term space.

A. Dataset

The dataset that we used in this study came from YIFY movie subtitle website [12]. The entire dataset includes about 500 movie subtitle files. The dataset collection has been done by choosing the most popular English movie subtitle files. The movie subtitle file has been downloaded as .srt format. Then we converted it into .txt format.

B. Parsing the Document and Case-folding

In this process, SGML markup tags, HTML tags and other non-alpha characters were removed from our movie subtitle dataset. Case-folding, which converts all characters to the same case in a subtitle document, is done by transforming all characters into lower cases. Tokens are extracted, which consist of alpha characters.

C. Stemming

Stemming is the process where the different forms of a word are conflated into a common representation [13]. For example, the words: "preparation", "prepared", "preparing" could all be reduced to a common representation of "prepare". This is a commonly used text processing technique for retrieving information based on the assumption that querying the term presentation involves an interest in text documents containing the words prepared and preparation. Stemming is applied to reduce the dimensionality of our proposed work [14].

Fig. 1. Stemming representation of movie subtitle documents

D. Remove Stopwords

There are several words in English as like conjunctions prepositions and pronouns that are employed instead of the substance to give structure in language. These words, which are often found without relevant information about the content and hence the document category, are called stopwords. Stopwords removal is mostly used to retrieve information from the documents. We eliminated the stopwords from the movie subtitle documents that leads to a significant decrease in dimensionality reduction of the feature space [15].

E. Term Weighting

We presented each of our movie subtitle document vector d as:

$$d = (w_1, w_1, \dots\dots\dots, w_T) \quad (1)$$

Where w_i denotes the weight of ith term of subtitle document d and T denotes the number of discrete terms in the subtitle documents. Salton and Buckley present a relative analysis of various term weighting techniques for automated text retrieval [16]. In the following sections we explained the term weighting approach we have employed in our experiment.

a) Term Frequency and Inverse Document Frequency (TF-IDF): TF-IDF is the furthermost frequent weighting method that takes term frequency across all the text documents in the corpus into consideration [17]. In our work, the term's weight i is assigned in subtitle document d as per a term arises in the document according to the number of occurrences and in contrary equivalent to the quantity of subtitle documents within the corpus where the term arises.

$$w_i = tf_i . log\frac{N}{N_i} \quad (2)$$

TF-IDF weighting technique measured a term's frequency in a document, if it arises in most documents, diminishes its significance.

b) Bag of Words (BOW): The Bag of Words (BOW) is a simplified representation used in retrieval of information and natural language processing. This approach represents a text like an unordered collection of words, which does not take into consideration grammar or even order of word [18]. For text categorization, A word in a document must be assigned as a weight by its frequency in the document. Words and their weights represent approximately BOW. This paper presents a novel element of the BOW model and examines how the scheme works in the classification task of text documents.

F. Dimensionality Reduction

We applied Information Gain (IG) methods for dimensionality reduction in our movie subtitle document classification. The number of bits of information obtained is measured by Information Gain for prediction of the category, when it is known that a term is present or absent in a document [19]. When there is a set of categories $c_1, c_1, \dots\dots\dots, c_m$, for each distinct term t, the IG is considered as follows:

$$TG(t) = \sum_{i=1}^{m} P(c_i). log\, P(c_i) + p(t). \sum_{i=1}^{m} P(c_i|t). logP(c_i|t) + P(\bar{t}). \sum_{i=1}^{m} P(c_i|\bar{t}). logP(c_i|\bar{t}) \quad (3)$$

As seen from the above equation, IG determines the decrease of entropy whenever the feature is provided vs. not provided. $P(c_i)$ denotes the aforementioned probability for the category c_i. It may be determined from a portion of subtitle documents within the category c_i of the training set. $P(t)$ denotes the aforementioned probability of the term t. It may be determined from a proportion of the training subtitle documents in which term t is present. $P(\bar{t})$ can be determined from a proportion of training subtitle document where term t is not present.

IV. Methodology

In this study, the model that is implemented is to group similar movie subtitle documents in clusters. For the method to be obtained, it is necessary first to pre-process the subtitle documents and generate the term documents matrix. The term-document matrix is the collection of terms calculated based on the weight of the words contained in the documents. These clusters finally identify the order of the subtitle documents. The initial step of movie subtitle document clustering is to read all the movie subtitle documents. All the unique terms are determined from the considered input movie subtitle documents. The input vectors are generated using BOW and TF-IDF feature extraction method in the n-dimensional term-document matrix. Then the Bisecting K-Means, K-Means and Agglomerative Clustering technique is employed on the document vector. After that, the clusters as well as measure performance, are observed. Finally, the cluster similarity and the centroid similarity is calculated for more effective result.

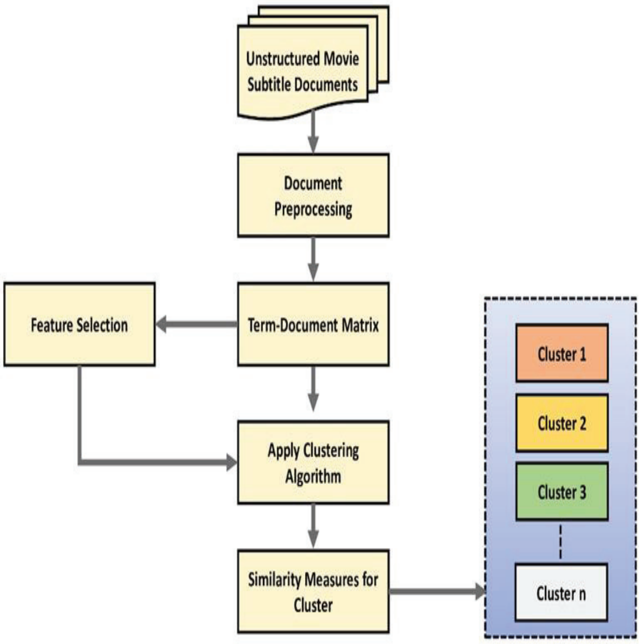


Fig. 2. Proposed model for movie subtitle document classification

A. K-Means Clustering

The principle of the K-Means clustering algorithm is that the mean of the documents allocated to the cluster can be represented by each of the k clusters assumed to be the centroid of that cluster, discussed by Hartigan [20]. In our experiment, the second variant of K-Means clustering algorithm is used which is known as the incremental or online version [21]. In the field of subtitle text document assortments, online K-Means is more efficient than the batch version. Initially, k subtitle documents are selected randomly from the corpus as the initial centroids. After that, the subtitle documents are allocated iteratively and after each of the allocation of a subtitle document to its adjacent centroid, centroids are updated incrementally. When no relocations of subtitle documents arise, the iteration stops. The centroid vector c from cluster C of subtitle documents has defined as follows:

$$c = \frac{\sum d \epsilon C^{d}}{|C|} \tag{4}$$

Where c is determined by the average weight of terms of the subtitle documents in C. The resemblance between a centroid vector c and a subtitle document D has defined by the cosine similarity as:

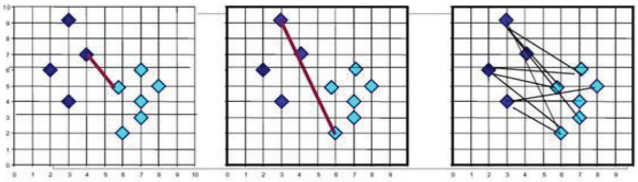
$$cos(d, c) = \frac{d \bullet c}{||d||\ ||c||} \tag{5}$$

B. Bisecting K-Means Clustering

Bisecting K-Means is the algorithm of clustering that attains a cluster of hierarchy through the repeated application of the basic k-means algorithm [22]. In each step, a cluster is selected to be split by applying basic K-Means for k = 3 in Bisecting K-Means. The clusters with the minimum overall similarity or the cluster with the maximum number of subtitle documents may be assigned to separate that might preferred to be split. In both cases, we conducted experiments and recognized the similar performance. Therefore, only when the largest cluster is chosen for split, we reveal the result.

C. Agglomerative Hierarchical Clustering

Agglomerative clustering algorithms starts in a separate cluster with each document and combine the most related clusters at each iteration until the stop criteria is encountered [23]. They primarily categorize as single ties, complete links and decent links and they define their inter-cluster similarity depending on this classification. The principle is clarified in figure 3.



single-link max. cos(d_i,d_j)

complete-link min. cos(d_i,d_j)

average-link average pairwise cos(d_i,d_j)

Fig. 3. Inter cluster similarity

- The single link technique determines the similarity among two different clusters C_i and C_j represents the similarity within two subtitle documents $d_i \in C_i$ and $d_j \in C_j$ which are most similar.

$$Similarity\ \mathcal{Y}_{single-link}(C_i, C_j) = \max_{d_i \in C_i, d_j \in C_j} |cos(d_{i,} d_j)| \tag{6}$$

- The complete link technique determines the similarity between two different clusters C_i and C_j represents the similarity within two subtitle documents $d_i \in C_i$ and $d_j \in C_j$ which are least similar.

$$Similarity\ \mathcal{Y}_{complete-link}(C_i, C_j) = \min_{d_i \in C_i, d_j \in C_j} |cos(d_{i,} d_j)| \tag{7}$$

- The average link technique determines the similarity between two different clusters C_i and C_j as the pairwise average similarities of the subtitle documents from each cluster where n_i and n_j are sizes of clusters C_i and C_j respectively.

$$Similarity\ \mathcal{Y}_{average-link}(C_i, C_j) = \frac{\sum d_i \in C_i, d_j \in C_j |cos(d_{i,} d_j)|}{n_i n_j} \tag{8}$$

V. Model Evaluation

In unsupervised machine learning, we used two kinds of measurement to evaluate the cluster quality, measurement of internal quantity and measurement of external quality [24]. The quantification of internal quality does not use external resources, including information about the class label, to assess the clustering solution that is produced. On the opposite, the measurement of external quality depends on the labelled test documents. We used purity as an external quality measure, overall similarity as an internal quality measure, and two most commonly used external quality measurement in text mining: F-measure and entropy [25] to assess the quality of the clustering algorithms.

A. Evaluation Metrics

Overall consistency is a metric of internal quality that usages the weighted similarities of internal cluster to estimate the consistency of the clusters produced. For cluster C_j, the similarities of internal cluster I can be defined as:

$$I_j = \frac{1}{n^2} \sum_{d \in C_j, d' \in C_j} \cos(d, d') \tag{9}$$

$$Overall\ Similarity = \sum_j \frac{n_j}{N} I_j \tag{10}$$

Where, n_j denotes the total number of subtitle documents within cluster j and N refers to the entire subtitle documents in the corpus.

Purity measures the dimension of each single cluster comprises subtitle documents from a single class. Purity for a specific cluster j of size n_j is defined to be:

$$p_j = \frac{1}{n_j} \max_i n_{ji} \tag{11}$$

Where, n_j denotes the number of subtitle documents of class i which are assigned to cluster j. Therefore, P

represents the portion of the aggregated cluster size is assigned to the cluster by the primary class of subtitle documents. The absolute purity is accomplished by the weighted sum of each cluster purity.

$$P = \sum_j \frac{n_j}{N} P_j \tag{12}$$

Where N denotes the entire number of subtitle documents among the whole movie subtitle dataset. In general, the higher the purity values, the better is the clustering solution.

Entropy measures the homogeneity of the clusters. The optimal solution for clustering conducts to clusters consisting of subtitle documents from a single class. The entropy is zero in this case. Generally, the clusters are more homogenous when the entropy is lower. The overall entropy E is the sum of the entropies E of each cluster j for a set of clusters.

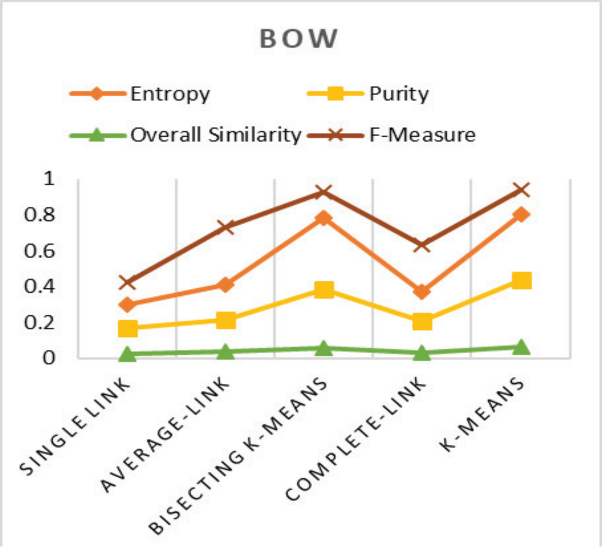
$$Ej = -\sum_i P(i,j).logP(i,j) \tag{13}$$

$$E = \sum_j \frac{n_j}{N} E_j \tag{14}$$

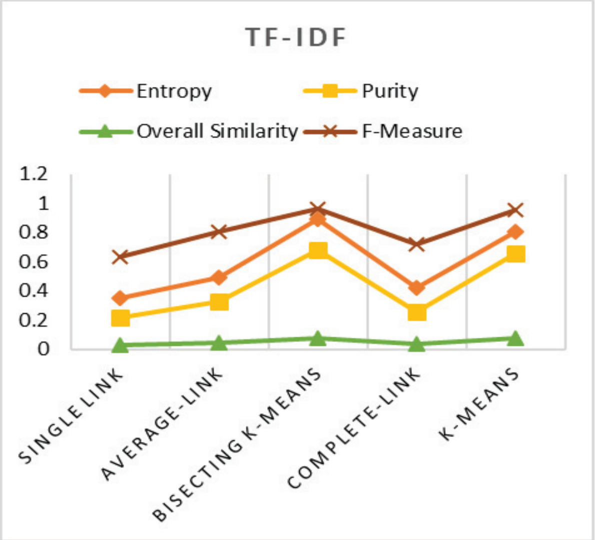
$P(i,j)$ denotes the probability of a subtitle document with the class label i which also assigned to j, n_j refers to the size of cluster j and N represents the entire subtitle documents in the corpus.

B. Result and Discussions

Figure 4 displays the performance of Bisecting K-Means, K-Means and agglomerative hierarchical algorithms: single link, average link and complete link in the terms of overall similarity, purity, entropy and F-measure evaluation metrics using TF-IDF and BOW feature representation method over movie subtitle dataset. Amongst the agglomerative clustering algorithms, the single-link algorithm performs significantly worse for both feature representation methods. Each subtitle document is assigned to the cluster of its nearest neighbor by this algorithm. Yet, any two subtitle documents might share several of the same terms and they can be the nearest neighbors even if not included in the same class. In our movie subtitle dataset, a large number of subtitle documents are nearest neighbors included with different topics. These characteristics make our movie subtitle dataset less complicated. In the TF-IDF representation method, the performance of the average link is a little bit improved rather than the BOW representation method. For both of these methods, the average-link performs best out of agglomerative hierarchical clustering algorithms. We intuitively explained the reasons for the unsatisfactory performance of the single link algorithm. The complete link is built on the presumption that all the subtitle documents are very similar in the cluster. The high dimensional diversity of the text document domain does not take this assumption into consideration, where each individual word is regarded as a prominent feature and framework knowledge as like hypernyms, hyponyms and synonyms are not appraised. By relying on more global properties, the average-like algorithm overcomes these problems in measuring cluster similarity. The similarities of the two clusters are evaluated by taking all the subtitle documents in both clusters into consideration in this algorithm. When overall similarity, purity and entropy metrics are appraised, the Bisecting K-Means and online K-Means achieves better than complete link and single link clustering algorithms for both TF-IDF and BOW representation methods. Their performance is either similar or better to the average-link algorithm. With respect to F-measure, Bisecting K-Means achieves better than K-Means on TF-IDF representation. In the case of BOW representation, the performance of K-Means achieves better than the Bisecting K-Means and in both cases, Bisecting K-Means achieves better than average-link clustering algorithms.



(a)



(b)

Fig. 4. Performance measurement of unsupervised learning for (a) BOW and (b) TF-IDF method

VI. Conclusion

The breakthroughs in computer, electronic technology and the rising popularity of the internet have led to significant growth of information on electronic text information. An essential research issue was established to browse the text databases and retrieve useful information quickly, efficiently, and adequately. In this study, we have presented an experimental evaluation of this paradigm. We proposed a system for the classification of movie subtitle documents using an unsupervised machine learning approach. We used two different alternatives for movie subtitle document representation: BOW and TF-IDF, where each distinctive stemmed word is considered a term. The stopwords and other markup tags were removed, stemmed by using the stemming process, and the dimensionality was reduced by Information Gain (IG). After preprocessing stage, we implemented the most leading unsupervised machine learning algorithms on our movie subtitle dataset. We evaluated the most prevalent clustering techniques Bisecting K-Means, K-Means and Agglomerative Clustering Algorithm; Average link, Single Link and Double link. We conclude that the average-link algorithm performs best out of Agglomerative Hierarchical Clustering Algorithm. The cause for the improper performance of the complete link and the single link clustering algorithms is that the nature of document collections is far from reality. The single-link clustering algorithm supposes that the nearest neighbors correspond to the similar class and the complete link approach presupposes the similarity of subtitle documents in a cluster. On the other hand, the Bisecting K-Means and the K-Means is more satisfactory than Agglomerative Clustering Algorithm. The performance of Bisecting K-Means and the K-Means mostly depends on the value of parameter k and the preliminary centroid selection. Finally, we can say that Bisecting K-Means and the K-Means algorithms are more compatible than Agglomerative Clustering Algorithm in the quality of clusters they produced. Agglomerative Clustering Algorithm generally produces inhomogeneous and unbalanced clusters.

ACKNOWLEDGMENT

The author acknowledges the support provided by Daffodil International University Natural Language Processing (NLP) Lab and Innovation Lab to develop the proposed model. We also acknowledge the supervision of our supervisor and honorable faculty to develop this system.

REFERENCES

[1] G X. Dongkuan & T. Yingjie, (2015), “Y. A Comprehensive Survey of Clustering Algorithms”, Annals of Data Science, 165–193.

[2] K. R. Larsen, E. David, S. Dirk, H. Christopher & N. Bailey, (2008), “Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems”, Decision Support Systems, Volume 45, Issue 4, Pages 884-896.

[3] R. Du, S. Naini & W. Susilo, (2003), “Web filtering using text classification”, 11th IEEE International Conference on Networks.

[4] A. P. Pimpalkar & R. Raj, (2020), “Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features”, Advances in Distributed Computing and Artificial Intelligence Journal, 9(2), 49-68.

[5] S. C. Dharmadhikari, M. Ingle & P. Kulkarni, (2011), “Empirical Studies on Machine Learning Based Text Classification Algorithms”, Advanced Computing: An International Journal (ACIJ), Vol.2, No.6.

[6] L. Liu, J. Kang, J. Yu & Z. Wang, (2005), "A comparative study on unsupervised feature selection methods for text clustering," International Conference on Natural Language Processing and Knowledge Engineering, pp. 597-601.

[7] T. Zagibalov & J. Carroll, (2008), “Unsupervised classification of sentiment and objectivity in Chinese text”, In Proceedings of the International Joint Conference on NLP (IJCNLP).

[8] P. Chaovalit & L. Zhou, (2005), "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", 38th Annual Hawaii International Conference on System Sciences, pp. 112c-112c.

[9] J. He, C. Tan, H. Low & D. Shen, (2020), “Unsupervised Learning for Document Classification: Feasibility, Limitation, and the Bottom Line”.

[10] A. Doucet & M. Lehtone, (2007), “Unsupervised Classification of Text-Centric XML Document Collections”, Lecture Notes in Computer Science, vol 4518.

[11] A. Geron, Hands-On Machine Learning with Scikit-Learn & TensorFlow, 7th Edition, pp: 67-179.

[12] YIFY Subtitles for English Movies, Available at: https://yts-subs.com/, Accessed: 10 December 2020.

[13] A. K. Sangaiah, A. E. Fakhry & A. Basset, (2019), “Arabic text clustering using improved clustering algorithms with dimensionality reduction”, Cluster Comput 22, 4535–4549.

[14] V. Romano, J. J. Andrew & C. Sueur, (2020), “Stemming the Flow: Information, Infection, and Social Evolution”, Trends in Ecology & Evolution, Volume 35, Issue 10, Pages 849-853, ISSN 0169-5347.

[15] D. S. Maylawati, W. B. Zulfikar, C. Slamet & Y. A. Gerhana, (2018), "An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media", International Conference on Cyber and IT Service Management (CITSM), pp. 1-6.

[16] I. Alsmadi & G. K. Hoon, (2019), “Term weighting scheme for short-text classification: Twitter corpuses”, Neural Comput & Applic 31, 3819–3831. https://doi.org/10.1007/s00521-017-3298-8.

[17] C. Liu, Y. Sheng, Z. Wei & Y. Yang, (2018), "Research of Text Classification Based on Improved TF-IDF Algorithm," International Conference of Intelligent Robotic and Control Engineering (IRCE), pp. 218-222.

[18] T. Walkowiak, S. Datko & H. Maciejewski, (2019), “Bag-of-Words, Bag-of-Topics and Word-to-Vec Based Subject Classification of Text Documents in Polish”, Advances in Intelligent Systems and Computing, vol 761.

[19] S. Jadhav, H. He & K. Jenkins, (2018), “Information gain directed genetic algorithm wrapper feature selection for credit rating”, Applied Soft Computing, Volume 69, Pages 541-553, ISSN 1568-4946.

[20] Bedi P., et.al. (2022) A Framework for Personalizing Atypical Web Search Sessions with Concept-Based User Profiles Using Selective Machine Learning Techniques. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_23

[21] C. Zhu, C. U. Idemudia & W. Feng, (2019), “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques”, Informatics in Medicine Unlocked, Volume 17, ISSN 2352-9148.

[22] Goyal S.B., et.al. (2022) Multi-objective Fuzzy-Swarm Optimizer for Data Partitioning. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_25

[23] L. R. Emmendorfer & A. M. P. Canuto, (2021), “A generalized average linkage criterion for Hierarchical Agglomerative Clustering, Applied Soft Computing”, Volume 100, 106990, ISSN 1568-4946.

[24] Rajawat A.S., et.al. (2022) Efficient Deep Learning for Reforming Authentic Content Searching on Big Data. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_26

[25] F. Liu, G. Zhang & J. Lu, (2020), "Heterogeneous Domain Adaptation: An Unsupervised Approach", IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5588-5602.