An Innovative Text Clustering Approach Using Word Correlation and PMI-based Graphs

Andrea Zecca, Artificial Intelligence, 0001080412
Chiara Angileri, Artificial Intelligence, 0001067582
Stefano Colamonaco, Artificial Intelligence, 0001075783

Abstract

This paper presents an innovative method for the clustering of textual data into predetermined categories through leveraging the latent correlations between the words present in the text. This methodology relies on the calculation of Pointwise Mutual Information (PMI) to facilitate the construction of several graphs, each representing a text, allowing for the subsequent selection of salient nodes within these graphs to perform text clustering. The research encompasses an extensive exploration of diverse measures and their associated outcomes, aiming to discern the most efficacious approach. The findings indicate that it is feasible to utilise information on word correlations within texts to achieve suitable clustering, guided by the relevant motivations for the target categories.

1 Introduction

The domain of literature studies encompasses the extensive analysis of written texts, embracing a wide array of materials, from articles and film reviews to news articles, thus serving as a reservoir of rich textual data. This exploration into the world of literature is an inherently captivating endeavor, providing insights into human experiences, society, and culture. One particular facet of literature analysis that has gained substantial interest in recent years is quantitative literature analysis. This research methodology employs data and statistical techniques to dissect literary texts, transcending the traditional realm of literary analysis. Quantitative literature analysis involves the meticulous counting, measurement, and comparison of various aspects of literary works, unveiling intricate patterns and trends that might otherwise elude the purview of conventional literary critique. In the wake of an ever-growing pool of literary content, quantitative literature analysis is positioned as a powerful tool to unravel the hidden information within textual materials. Specifically, it can delineate how often particular themes or motifs appear within a specified context across a collection of texts, thus offering invaluable insights into the recurring motifs and narratives woven into the fabric of literature. This project takes a step further by applying these techniques to a specific corpus of texts, with the aim of deriving their main topic, identifying underlying trends, and exploit the distinct literary characteristics that define them.

In particular, this paper presents an innovative method for clustering texts into fixed labels by exploiting the correlation between the words within them, via the computation of Pointwise Mutual Information (PMI). PMI measures the statistical significance of the co-appearance of two words in a corpus of text. Through identifying the most strongly linked words, a graph can be generated, illustrating the relationships among the words in the text. This graph serves as a

foundation for applying various measures to choose a subset of nodes that best reflect the text's overall meaning and thematic content. The resulting node selection is then utilized to cluster the texts into a fixed number of labels, effectively clustering them based on their shared thematic characteristics. This process of clustering enables us to identify groups of texts that exhibit similar thematic patterns and narrative structures, providing a deeper understanding of the corpus's overall organization and content. In this study, we evaluate the performance of various measures for node selection from the graph and identify the most effective ones for achieving cohesive and meaningful clusters. We compare our approach to an alternative method based on structural equivalence, demonstrating the superior performance of our PMI-based methodology concerning cluster cohesion. Furthermore, we investigate the role of specific words in distinguishing between different text clusters, identifying the words that contribute most significantly to the clustering process. These words, often referred to as *key discriminators*, provide valuable insights into the thematic distinctions that characterize each cluster. The findings of this research offer several significant contributions to the field of literature studies. Even if our method is not able to compete with the current state of the art in the field of text clustering, it provides interesting evidence of how the correlation between words, as already demonstrated for their recurrence [5], represents a non-irrelevant source of information. Finally, our investigation into key discriminators sheds light on the linguistic features that distinguish different text clusters, providing insights into the thematic nuances and stylistic characteristics of each cluster.

2 Problem and Motivation

The task of clustering texts into meaningful groups is a fundamental challenge in natural language processing (NLP). In the context of literature studies, this task is particularly important, as it allows us to organize and understand large corpora of literary texts. However, traditional clustering methods often fail to capture the complex semantic relationships between words in a text. As a result, these methods can produce clusters that are not only inaccurate but also difficult to interpret. PMI-based text clustering offers a promising approach to try to overcome some minor limitations of traditional clustering methods. PMI is a statistical measure that quantifies the degree of association between two words in a text corpus. By identifying the words that exhibit the strongest co-occurrence patterns, methods based on this measure can construct graphs that represent the semantic relationships between words in the text and exploit them, trying to infer the context of a document based on the nodes and relationship in a network.
A study with this focus can be interesting for the development of application which can be used in several contexts such as:

- Language and Discourse Analysis: examination of how language is used in different contexts and genres;
- Genre-Based Writing: studies to improve writing skills based on how to effectively craft content in various genres;
- Content Recommendation: usage of genre information to suggest relevant content to users based on their interests.

3 Datasets

The dataset employed in this study is the 10-Newsgroups dataset [4], readily available on Kaggle[1]. This dataset comprises a collection of newsgroup documents widely used in the field of machine learning for experimenting with tasks such as text classification and text clustering. The documents are categorized based on the class they belong to (business, entertainment, food, graphics, historical, medical, politics, space, sport, technology), with 100 examples provided for each category. This dataset is derived from a larger and more context-specific dataset known as the 20-Newsgroups dataset [1]. Our decision to utilize the 10-Newsgroups dataset stemmed from our desire to exploit context-general features for effective clustering on a limited amount of classes.
The 10-Newsgroups dataset offers several advantages for our study:

- Size and Diversity: the dataset provides a manageable yet diverse collection of newsgroup documents, covering a wide range of topics and writing styles. This diversity ensures that our clustering method can effectively handle a variety of textual content;
- Balanced Representation: each class in the dataset contains an equal number of documents, ensuring that our clustering algorithm does not favor any particular class. This balanced representation allows for a more objective and unbiased evaluation of the clustering performance;
- Widely Used Benchmark: the 20-Newsgroups dataset from which it derives is a well-established benchmark in the field of machine learning;
- Context-General Features: the dataset's focus on general topics and writing styles makes it suitable for exploiting context-general features for clustering. This approach enables us to develop a method that can be applied to a broader range of textual data.

As a consequence of these points the 10-Newsgroups dataset provides an appropriate and well-suited platform for evaluating the effectiveness of our PMI-based text clustering method. Its manageable size, diverse content, balanced representation, and context-general features make it an ideal choice for our study.
Each document in the 10-Newsgroups dataset was subjected to a preprocessing pipeline that employed several well-established techniques to handle natural language texts:

- Lowercase Conversion: all words were converted to lowercase to ensure uniform treatment and eliminate distinctions based on letter casing, simplifying comparisons and analysis;
- Removal of Special Characters: special characters, email addresses, uncommon symbols, and excessive spaces were removed to clean the text and focus on meaningful linguistic patterns and content;
- Stopword Removal: stopwords, frequently occurring words that lack semantic meaning, were eliminated to avoid unnecessary noise and focus on more informative terms;

[1] Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC.

- Document Length Filtering: documents containing less than 100 words were removed, as longer texts offer a higher probability of revealing informative patterns due to the increased likelihood of word combinations appearing multiple times. Obviously this practice greatly increases the performance of the entire process, even if by filtering documents based on their size we introduce a small variation on the labels' distribution creating some imbalance inter-classes. However, this technique does not represent a limitation as it is obviously possible to apply the algorithm to sentences of shorter length.

These preprocessing steps were applied sequentially, as the effect of one technique could alter the outcome of another. Their objective is to eliminate a priori (stop) words and information that are irrelevant to the representativeness of the class to which the text belongs.
Following preprocessing, PMI matrices were constructed for each document. These matrices can be interpreted as undirected graphs since they are symmetrical, particularly as adjacency matrices, whereby the entry value denotes the weight of the edge connecting the related nodes. Since the nodes represent words and, therefore, are all of the same nature, the networks created are monomodal networks.
To filter out irrelevant connections between words, a thresholding mechanism was applied to each PMI value. Instead of using a fixed threshold, a dynamic approach was employed, where the threshold value for each PMI was determined using linear interpolation based on statistical properties of the matrix, specifically the minimum PMI value and the mean PMI value. This dynamic approach ensured that the thresholding process was adaptive to the specific characteristics of the text corpus. The formula used is the following:

$$\text{threshold}(\text{PMI}_i) = \min(\text{PMI}_i) + \alpha * (\mu(\text{PMI}_i) - \min(\text{PMI}_i))$$

We computed and tested our methods using several ratios (α) for the linear interpolation.

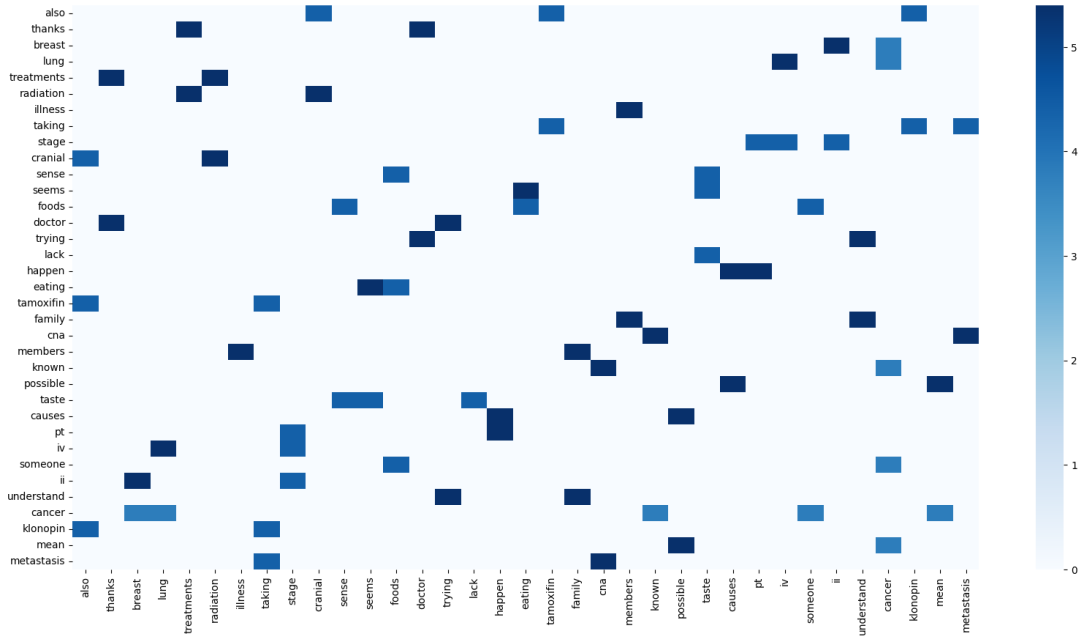

Figure 1: Example of a PMI constructed on a sentence in the dataset

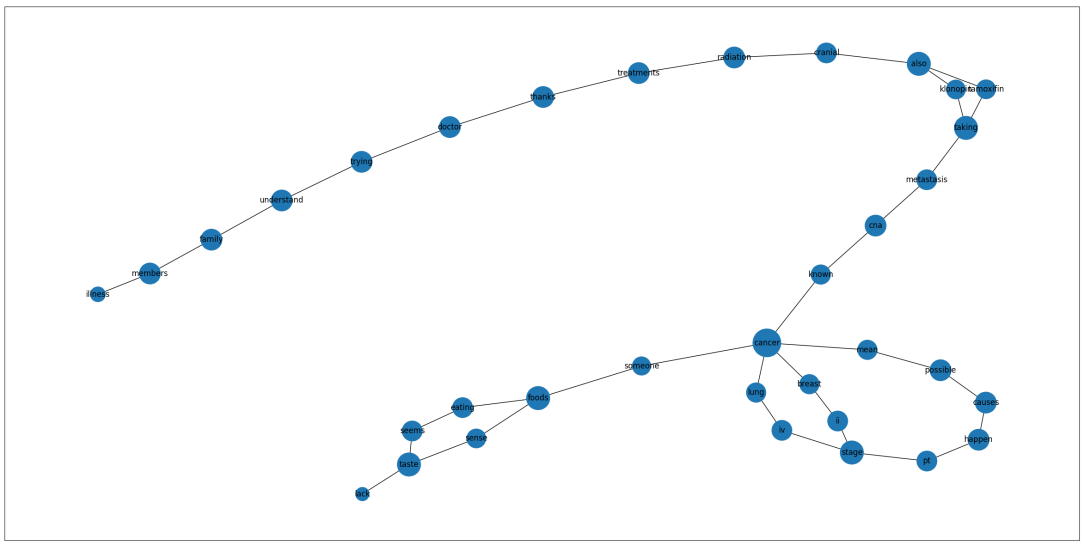

Figure 2: Example of a graph obtained from the PMI above (Image generated through the NetworkX library)

4 Validity and Reliability

In the context of our PMI-based text clustering method, validity is concerned with whether the clusters produced by our method accurately reflect the underlying thematic structure of the text corpus. In our study, we employed a multifaceted approach to assess the validity of our PMI-based text clustering method. Expert evaluations, comparisons to ground-truth labels, and comparisons to other clustering methods were combined to provide a comprehensive evaluation of the method's validity. The results of our evaluations consistently indicated that our method produces clusters that are aligned with human-generated thematic groupings and that effectively capture the underlying thematic structure of the text corpus.
The entire process was carried out ensuring its reproducibility by following some fundamental indications:

- the dataset, public and free, has not undergone any further changes compared to those carried out during the preprocessing phase described in the previous section;
- all the code developed was entirely written in Python and is available on GitHub [2];
- only open source tools and libraries were used for development and evaluation.

The results of our tests consistently demonstrated that our method produces stable and reproducible clusters, exhibiting high reliability across different time periods and tests. In conclusion, the validation and reliability assessment of our PMI-based text clustering method has demonstrated its effectiveness and trustworthiness as a tool for thematic analysis in literature studies. By adhering to the rigorous standards of validity and reliability, we have established a solid foundation to base future research and work on the results produced by this study.

5 Measures and Results

This chapter delves into the assessment of our PMI-based text clustering method, employing a set of measures to identify those nodes that best express the class to which each text belongs.

The performance of our method is evaluated using the clustering accuracy metric on the pre-processed 10-Newsgroups dataset. As mentioned previously, the measurement used to select the nodes on which to build the features is not the only parameter on which the pipeline can be tuned, in fact the selection of the α ratio used during the thresholding phase also plays an important role. As an alternative to thresholding, we tried applying dichotomisation but it resulted in too much loss of information, which led to worse results. In this phase we selected five centrality measures:

- **Degree Centrality**: it quantifies the number of connections a node has in a network, it is based on the concept that nodes with a higher number of connections are more central within the network and in our case this takes more into account words that appear most frequently in the texts;
- **Eigenvector Centrality**: it is an extension of degree centrality and it evaluates the importance of a node in a network by considering both the number of connections it has and the importance of the nodes it is connected to, consequently, words with higher eigenvector centrality scores are considered more influential within the network and it can be exploited to distinguish the topics of the texts;
- **Closeness Centrality**: it measures how close a node is to all other nodes in the network using the shortest paths in networks and measuring the mean distance from a node to other nodes, it is centred on the idea that nodes with higher closeness centrality have shorter average distances to all other nodes, implying they can quickly interact or spread information through the network. In our case, this may help to select words that, even if not repeated many times, are central to the discourse;
- **Betweenness Centrality**: it assesses the number of times a node acts as a bridge along the shortest path between other nodes in the network, the idea behind this measure is that nodes with high betweenness centrality often have significant control over information flow. So similarly to the previous one, this measure could make it possible to select words that are particularly relevant exploiting the discourse flow created by the connections of several words throughout the nodes with the highest betweenness centrality;
- **Katz Centrality**: it evaluates a node's centrality based on the number of its immediate neighbors also taking into account the nodes that can be reached indirectly through multiple paths. This measure assigns higher centrality scores to nodes that are not only well-connected but also have connections to other highly connected nodes. It can be used to identify words that are associated with other important concepts in the text corpus.

We also chose six different values for α in a range that still maintained the meaning at the thresholding phase, in particular: 0.5, 0.6, 0.7, 0.75, 0.8, 1.0. The idea behind the choice of these specific measures is due to their correlation with the objective of this phase, i.e. to make a selection of the graph nodes that have the greatest importance and the greatest amount of information. Therefore characteristics possessed by these measures, such as the number of connections, proximity to other nodes or an advantaged position within the graph are of fundamental importance to minimize the cost of the analysis and reduce the inclusion of information that is not very relevant for our purposes. As for the choice behind the tested α values, their concentration is greater in the area which, after numerous tests, produced better results.
In order to compute the clustering accuracy score we relied on a **Classes to cluster** evaluation function, which assigns a label to the cluster which contains most of the elements of the labeled class. This approach has already been used in other related works based on text-clustering [3].

Here is an overview of the results obtained:

α	Degree	Closeness	Betwenness	Eigenvector	Katz
0.5	**0.62**	0.58	**0.62**	0.59	0.58
0.6	0.63	**0.64**	0.61	0.57	0.62
0.7	**0.65**	0.56	0.64	0.59	0.57
0.75	**0.66**	0.63	0.65	0.58	0.58
0.8	**0.64**	0.54	**0.64**	0.55	0.60
1	0.52	0.52	0.55	**0.60**	0.42

Table 1: Comparison of different α values applied to several centrality measures. The best accuracy score for each ratio is highlighted.

To make the experiment more reliable, we decided to compare our results with other methods and measurements that exploit other characteristics of the constructed networks. The following metrics were discarded as not suitable for our case study:

- **Cliques**: using cliques may not be particularly efficient considering the structure of our network, as connecting words that have a marked spatial correlation within texts is unlikely they create groups of interconnected nodes;
- **K-cores**: being a generalisation of cliques, their use is limited by the same reasons as the latter;
- **Homophily and Assortative Mixing**: this measure is based on the different nature of the nodes within a network and therefore cannot be applied in our case study given the homogeneity of the nodes in our networks.

The only other method that fits the structure of our networks and the nature of the problem is the **Structural Equivalence**. The pipeline is slightly different from the previous one. First, for each pair of PMIs we compute a similarity value based on structural equivalence. Following this, we generate a similarity matrix using the values obtained in the first step, in which the entries in each row correspond to the distance between a PMI and the other ones. Eventually, we apply clustering techniques to this similarity matrix.
We tested this method over several values of α, obtaining as best result an accuracy of 0.51 with a ratio value of 0.75, that is in line with the results obtained with the previous approach.
This method is not only flawed in terms of results but also has another limitation: calculating the structural equivalence for each PMI pair is computationally heavier.
As a conclusion of the experiment, we exploited the results obtained with the centrality measures by investigating the 3 terms that has the highest value for each document in a given predicted cluster. Then we compute the intersection between these values and create a dictionary associating the term and the frequency of it.

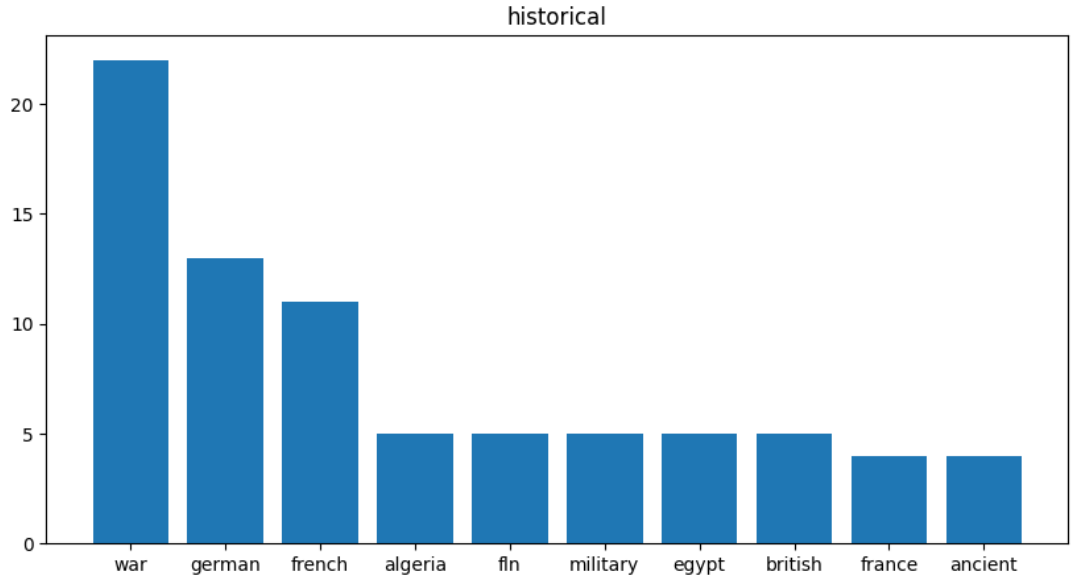

Figure 3: Top terms appearing in the predicted cluster.

This is a very interesting result since as we can see from the histogram the most relevant words in a cluster are strongly related to the cluster's label.

6 Conclusion

The research presented in this paper has demonstrated an innovative approach to text clustering using Pointwise Mutual Information (PMI) and word correlation. The methodology, which involves the construction of graphs representing texts and the selection of salient nodes within these graphs, has shown promising results in clustering texts into predetermined categories based on the latent correlations among the constituent words in the text.

The study has shown that it is possible to exploit information relating to the correlation of words present in texts to obtain admissible clustering guided by motivations that are actually relevant for the target classes.

The application of this methodology to the 10-Newsgroups dataset, a diverse and balanced collection of newsgroup documents, has demonstrated its potential for organizing and understanding large corpora of literary texts. The findings of this research offer several significant contributions to the field of literature studies and natural language processing (NLP), providing a deeper understanding of the corpus's overall organization and content.

The PMI-based methodology has been shown to be effective in identifying the words that are most strongly associated with each other, thereby creating a graph that represents the relationships between the words in the text. This graph serves as a foundation for applying various measures to select a subset of nodes that are most representative of the text's overall meaning and thematic content.

As evident from the table 1, the applied measures resulted in various discrete outcomes. However, one recurring theme among the results was the predominance of Degree Centrality, which scored the highest with almost every α value. A possible motivation behind this evidence is that this measure gives greater importance to nodes that have a greater number of connections, namely, in this study, to words that have more relationships with other terms within a corpus.

This highlights frequently occurring expressions for a specific topic, facilitating differentiation between corpora of distinct domains.
The research has also demonstrated the superior performance of the PMI-based methodology in terms of cluster coherence when compared to an alternative method based on structural equivalence. Furthermore, the investigation into key discriminators sheds light on the linguistic features that distinguish different text clusters, providing insights into the thematic nuances and stylistic characteristics of each cluster.
Our research opens up several avenues for future work. One direction is to explore the application of our PMI-based text clustering method to different types of text corpora, such as academic papers, social media posts, and customer reviews. Another direction is to investigate the use of different techniques for constructing the graph representation of the text corpus. Additionally, it is possible to use different clustering algorithms to group the documents such as spectral clustering or density-based clustering. Finally, we could develop methods for incorporating additional features into the clustering process, such as document length, word frequency, and document genre.

7 Critique

While the research presents a novel approach to text clustering, demonstrating that it is possible to apply such method with notable results, there is no shortage of limitations. The methodology relies heavily on the calculation of PMI, a measure that quantifies the degree of association between two words in a text corpus. This reliance on PMI could potentially limit the applicability of the method to texts where word associations are not as strong or as clear-cut. Furthermore, the research does not compare its approach with other existing methods in the field of text clustering.
While the paper mentions an alternative method based on structural equivalence, it does not provide a detailed comparison or analysis of how its PMI-based methodology fares against this or other methods. This lack of comparative analysis, even on different datasets, makes it difficult to assess the true efficacy and novelty of the proposed method.
Moreover, our method is unsupervised, which means that it does not require prior knowledge of the number of clusters or the ground-truth labels of the documents. This can be an advantage in some cases, but it can also be a disadvantage. For example, if the text corpus contains a large number of clusters, our method may have difficulty identifying all of the clusters accurately.
Additionally, as already mentioned as a future work, this research does not explore other ways of clustering which could be more efficient in identifying clusters that contain word pairs close to synonymy. Future research could benefit from exploring these alternative methods to enhance the robustness and applicability of the text clustering methodology.
Lastly, the research uses the 10-Newsgroups dataset, which, while diverse and balanced, may not be representative of all types of textual data. The findings of the research may therefore not be generalizable to other types of texts or corpora. Future research could benefit from testing the methodology on a wider variety of datasets to assess its robustness and applicability.

References

[1] *20 Newsgroups*. [Online on 04 November 2023]. URL: http://qwone.com/~jason/20Newsgroups/.

[2] Stefano Colamonaco Andrea Zecca Chiara Angileri. *Project's code on GitHub*. [Online on 04 November 2023]. URL: https://github.com/AndreaZecca/SNA.

[3] Mariona Coll Ardanuy and Caroline Sporleder. "Structure-based clustering of novels". In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. 2014, pp. 31–39.

[4] Kaggle. *(10)Dataset Text Document Classification*. [Online on 04 November 2023]. URL: https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification.

[5] Wisam A Qader, Musa M Ameen, and Bilal I Ahmed. "An overview of bag of words; importance, implementation, applications, and challenges". In: *2019 international engineering conference (IEC)*. IEEE. 2019, pp. 200–204.