Article

Not peer-reviewed version

POS-HC: A Part-of-Speech Hierarchical Clustering Approach for Normative Texts Partition

Wanyi Li , Yu Liu , Keqing Deng , Xiao-kun Wu *

Posted Date: 28 February 2024

doi: 10.20944/preprints202402.1575.v1

Keywords: NLP; semantic partition; clustering techniques; POS

Article

POS-HC: A Part-of-Speech Hierarchical Clustering Approach for Normative Texts Partition

Wanyi Li [1], Yu Liu [2], Keqing Deng [3] and Xiao-kun Wu [1,*]

[1] School of Journalism and Communication, South China University of Technology, China; 202210188247@mail.scut.edu.cn
[2] School of Journalism and Communication, South China University of Technology, China; 201930520440@mail.scut.edu.cn
[3] School of Journalism and Communication, South China University of Technology, China; kristindeng@foxmail.com
* Correspondence: wuxiaokun@scut.edu.cn

Featured Application: The clustering methodology delineated in this research offers advantages in enhancing the efficacy of text classification and clustering processes. It is particularly adept at elucidating the intricate semantics embedded within normative texts, thereby rendering it applicable in the realms of natural language acquisition and social science inquiry.

Abstract: Chinese texts often feature a substantial volume of normative content, and their analysis has increasingly become a focal point in recent semantic analysis endeavors. Enhancing the effectiveness of text classification and clustering is pivotal in unveiling the intricate semantics within extensive textual data. Empirical research within the realm of computational social science has underscored the need for improved explanatory power when it comes to general text clustering methods applied to normative texts. This study introduces an innovative hierarchical clustering method that incorporates part-of-speech (POS) features and pioneers a novel semantic network model. It achieves this by amalgamating a POS-based feature weight calculation method with hierarchical clustering techniques. To assess its efficacy, the method is rigorously evaluated across three diverse text datasets, encompassing normative reports, news articles, and social media content. The experimental results unequivocally demonstrate the superior performance of the POS-HC method, surpassing the traditional TF and TF-IDF feature extraction methods in terms of clustering accuracy, and surpassing some existing clustering methods. Furthermore, its classification effectiveness exhibits notable advantages, particularly in the semantic classification of normative texts.

Keywords: NLP; semantic partition; clustering techniques; POS

1. Introduction

The issue of semantic partitioning within semantic networks represents a significant and nuanced challenge in the realm of natural language processing (NLP) research [1]. It involves the organization and segmentation of information within semantic networks to enhance their efficiency and accuracy. In essence, semantic partitioning serves as a technique to dissect, understand, and navigate the intricate world of semantic relationships within text. The applications of semantic partitioning delve into the intricate the relations within textual content, scrutinizing aspects like their contextual construction, textual coherence, and the structure of communication, among others [2]. This multifaceted approach serves as a foundational framework for evaluating complex issues, including but not limited to semantic projection [3], collective linguistic rationality [4] , and cognitive memory [5] within social contexts.

The primary objective of semantic partitioning is to attain a fine-grained segmentation of textual data within the semantic network. This process involves breaking down lengthy sentences or texts into smaller semantic units, which can encompass individual words or phrases. The ultimate aim is

 doi:10.20944/preprints202402.1575.v1

to unveil the profound meaning embedded within the text. By clustering words into distinct units, semantic information can be extracted with greater precision, leading to the enhancement of computational methods for measuring semantic relatedness. As a result, this technique empowers more effective exploration of various facets, such as topic clustering, the degree of correlation between topics within a given set of data, the identification of underlying text structures, the organization of domain-specific knowledge, and the efficient dissemination of information.

A central concern within the realm of semantic network analysis revolves around community segmentation. This issue encompasses the delineation of communities or segments within the network, which can be achieved through different approaches, including topic segmentation or segmentation based on relevant semantics. Presently, various established methods are available, such as .1) clustering-based approaches use algorithms such as K-means, hierarchical clustering, and spectral clustering to cluster nodes in the semantic web based on their semantic features [6]. 2) Community discovery methods identify communities by detecting the underlying structure and assigning nodes with strong connections to the same community using algorithms such as Louvain, GN, and modularity maximization [7]. 3) The graph partitioning approach considers the semantic network as a graph and divides it into subgraphs using algorithms such as spectral graph partitioning, max-flow min-cut, and modularity maximization [8]. 4) Labeling definition methods propagate labeling information by using algorithms based on co-occurrence relationships or semantic similarity to assign similar nodes to the same labels[9]. These methodologies offer avenues for the examination of semantic network structures, node interactions, and patterns of information dissemination within intra-group dynamics. They find extensive applications in the semantic partition of texts from social networks and news sources.

As the demand for uncovering latent semantic connections in computational social science research continues to rise, a fresh challenge emerges in enhancing the precision of semantic partitioning techniques tailored to diverse textual content. This challenge is particularly evident in scenarios such as the analysis of policy texts. These documents adhere to a standardized format and structure, comprising chapters, paragraphs, and precise punctuation. Such rigidity may pose an obstacle to conventional community segmentation methods, as they might struggle to capture subtle distinctions within the text [10]. Furthermore, the uneven distribution of information within policy texts can complicate matters. Key policy-related information may be concentrated in specific sections, while other parts primarily serve explanatory or background purposes. This uneven distribution can potentially impact the accuracy of partition techniques.

Given this concern, our research endeavors to tackle challenges prevalent in text partition approaches grounded in traditional TF-IDF features. Notably, issues such as an excessive dependence on document vocabulary, hindering the capture of intricate semantic associations, and the limitation of semantic knowledge due to vector sparsity are among the concerns[11]. In light of these considerations, our study seeks to introduce an innovative hierarchical clustering method, denoted as POS-HC, which capitalizes on part-of-speech division. This approach offers distinct advantages in categorizing policy texts characterized by normative features. The key innovations of our study encompass:

Introducing a mechanism to evaluate words with similar meanings but distinct expressions by quantifying their semantic similarity, thereby regulating ambiguity in Chinese by assigning weights.

Incorporating the POS tagging process, which supplements specific words in sentences with contextual grammatical knowledge. This empowers the application of POS feature processing within hierarchical clustering algorithms and the segmentation of semantic networks in policy texts.

Presenting a methodology for capturing intricate semantic relationships, particularly those rooted in part-of-speech (POS) associations within semantic structures.

This approach not only offers valuable insights and references for semantic network analysis across various domains but also contributes to the advancement of structural division within semantic networks in the field of Natural Language Processing (NLP).

 doi:10.20944/preprints202402.1575.v1

2. Related Works

Presently, the field of semantic partitioning has seen the development of numerous techniques, offering a rich array of avenues for delving deeply into the intricacies of network structure, node interactions, and information dissemination patterns. Among these, clustering-based approaches stand out as particularly popular. These methods uncover latent associations and thematic elements within networks by grouping nodes with similar attributes into cohesive categories. In a comprehensive analysis of journal papers pertaining to the Circular Economy (CE) spanning the years 2000 to 2019, a combination of k-means clustering and Multiple Correspondence Analysis was employed. This method yielded the identification of twelve distinct and meaningful clusters, shedding light on the overarching conceptual landscape that underpins research on the Circular Economy [12]. Simultaneously, another study aimed to construct a document clustering system focused on semantic similarity, leveraging the K-Means and Hierarchical Agglomerative Clustering (HAC) algorithms. This investigation entailed a comparative assessment of the clustering performance of these two methods concerning concise, real-time descriptions from online laboratory repositories. The results of the study indicated that, particularly on small datasets, HAC surpassed the K-Means algorithm in terms of clustering efficacy[13] . In addition, a study explored the latent semantic structure of the label set according to spectral clustering and presented a new evaluation function based on information theory [14]. Nonetheless, these methods are extensively applied in the realm of short text processing and are yet to find widespread utilization in the context of lengthy, normative texts.

The computation of features plays an important role in clustering methods. In the context of text document clustering, where text constitutes unstructured or semi-structured data spanning multiple dimensions, the optimal clustering outcome necessitates the structuring of text data to extract meaningful features from the words within. Among the various weighting schemes employed for this purpose, TF-IDF stands out as a widely adopted method. TF-IDF is a frequently used weighting factor in information retrieval and text mining, serving to quantify the significance of a word within a document set or corpus [15]. Numerous ongoing studies are dedicated to enhancing the accuracy of clustering by refining the TF-IDF algorithm. In one such study, built upon the conventional TF-IDF algorithm, additional weighting factors were introduced. These factors encompass the dispersion level of feature words, the dispersion within specific classes, and the association degree between feature words and their respective classes [16]. In a separate study, an innovative approach was adopted, incorporating part-of-speech weight coefficients and position weights (span weights) for characteristic words. This methodology assigns distinct weights to words based on their positions and parts of speech within the text. This modification not only enhanced the traditional TF-IDF algorithm but also found successful application in a public opinion analysis system[17].

In addition to clustering algorithms, community discovery methods also play an important role in semantic partition. These methods aim to detect the underlying structure in the network, identify nodes with strong connectivity, and assign them to different communities or groups. Some well-known community discovery algorithms, such as Louvain algorithm [18], GN algorithm [19] and modular maximization [20]and so on, have been widely used in social networks, news networks and other fields to help reveal community structures and social relationships.

Graph partitioning methods consider the semantic network as a graph and use algorithms such as spectral graph partitioning [21], max-flow min-cut [22], and modularity maximization [20] to divide the network into subgraphs or partitions. This simplification aids in the analysis of intricate networks, enabling a targeted exploration of various network components to gain a deeper insight into the network's inherent structure and correlations. Moreover, the label definition method leverages algorithms that make use of co-occurrence relations or semantic similarity to disseminate label information, thereby assigning similar nodes to shared labels. Besides, a notable approach was introduced that hinges on co-occurrence group similarity, utilizing the ternary relation among users, resources, and tags to gauge the semantic relevance between tags [23]. Based on this research, Hang et al. proposed to learn tag semantics and tag-specific features in a collaborative way [24]. While this method may be considered conventional and less efficient, it nevertheless offers valuable insights.

 doi:10.20944/preprints202402.1575.v1

The wide-ranging application of these techniques equips us with potent tools across various domains. However, certain fields, such as normative text clustering and the deep learning of semantic partitioning and modeling for policy texts, still lack a substantial body of research [25], [26], [27].

Furthermore, with regard to contemporary enhancement techniques, the majority of clustering units predominantly operate at the document level rather than at the finer-grained word level. Nonetheless, it's imperative to recognize that words, being the fundamental building blocks of any document, wield significant influence in the realm of semantic partitioning [28], [29]. Hence, it becomes imperative to explore ways to effectively integrate semantic and part-of-speech (POS) features within text for an enhanced comprehension and analysis of semantic partitioning.

Within this context, the present study introduces a hierarchical clustering technique for analyzing government texts. It incorporates word-level features and integrates them with the part-of-speech (POS) feature weight calculation method.

3. Experimental Design

3.1. Methodological Framework

Our POS-HC (Part-of-Speech Hierarchical Clustering) method presents an innovative approach that incorporates the Part-of-Speech Probability Weight (POSP) concept as an extension to the traditional TF-IDF technique. This augmentation significantly enhances the efficacy of the TF-IDF algorithm in text analysis. POSP assigns weighted values to part-of-speech features, harnessing part-of-speech probabilities, thereby highlighting the importance of these linguistic attributes in the clustering process. As a result, our approach yields clustering outcomes that clearly manifest the influence of part-of-speech characteristics. This innovation equips researchers with a more sophisticated toolset for exploring semantic structures within textual data. Please refer to Figure 1 for the visual representation of the POS-HC method's workflow.

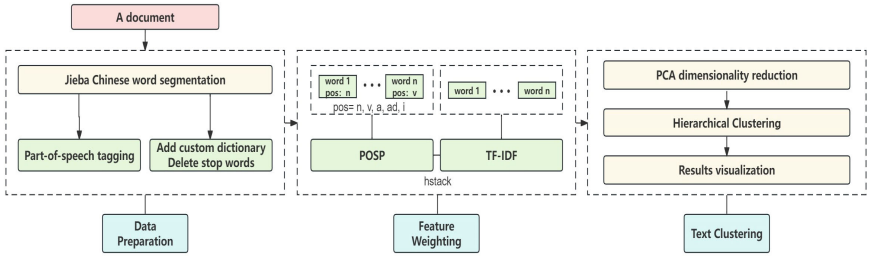

Figure 1. Flowchart of POS-HC method.

3.2. Words Segmentation

The methodology initiates with the utilization of Python's Jieba module for segmenting Chinese words within the dataset. The primary stages encompass lexical segmentation, the removal of stop words, and the implementation of part-of-speech tagging, along with other critical preprocessing steps. Following this, in accordance with the semantic structure and the unique characteristics of part-of-speech embedded in the Chinese text, a discerning process is executed to preserve terms based on their part-of-speech attributes. The selective retention process is exclusively focused on preserving nouns (n), verbs (v), adverbs (ad), adjectives (a), and idioms (i), while deliberately excluding any extraneous parts of speech. This meticulous approach is taken to improve the overall precision and relevance of subsequent analytical efforts by minimizing potential disruptions caused by linguistically irrelevant elements.

3.3. Improved POSP-TFIDF

This methodology presents an algorithm crafted to integrate TF-IDF and POS attributes, leveraging the POSP (Part-of-Speech Probability) to calculate the likelihood of diverse parts of speech

 doi:10.20944/preprints202402.1575.v1

occurrences. These outcomes are meticulously organized in a dictionary. Subsequently, we calculate the TF-IDF values for words using the TfidfVectorizer, and a horizontal stacking operation is performed through the hstack function. It combines the TF-IDF matrix with the part-of-speech probability data to create a novel feature matrix denoted as X. This approach aims to synergize semantic and statistical elements for an enhanced semantic analysis.

3.3.1. Subsubsection

TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used weighting technique in information retrieval and data mining. At its core, the fundamental principle of TF-IDF revolves around the concept that words with higher frequencies within a particular document, yet lower occurrences across other documents, should be attributed greater significance [30]. These words inherently possess greater informativeness and, as a result, offer increased utility for classification purposes. Consequently, TF-IDF enjoys widespread use in various domains, including keyword extraction, textual similarity assessment, and thematic categorization [31]. The formula is:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

In this formula, $tf_{i,j}$represents the number of times i appears in j, df_i represents the number of documents containing i, and N represents the total number of documents [32].

3.3.2. POSP-TFIDF

Our study introduces a novel approach termed POSP, designed to compute part-of-speech probabilities and subsequently integrates these probabilities with the TF-IDF weighting scheme to enhance the feature representation of words. POSP can be interpreted as the probability of a certain part of speech appearing in the entire text. Its formula is as follows:

$$POSP = p(n_k|n)$$

In the formula, n represents a word in the text, n_k represents the part-of-speech feature in which feature n appears, and k represents a certain part-of-speech.

Subsequently, the part-of-speech (POS) features and TF-IDF features are horizontally combined to create the new feature matrix for the vocabulary. The weighting formula for POSP-TFIDF is as follows:

$$w_{POSP-TFIDF}(n_k) = w_{tfidf}(n_k) * POSP$$

Where $w_{POSP-TFIDF}(n_k)$ represents the word weight under the POS-HC method, $w_{tfidf}(n_k)$ represents the operation value of TF-IDF and * represents the horizontal stacking operation of hstack function.

3.4. Principal Component Analysis

In our study, we utilize Principal Component Analysis (PCA) as a dimensionality reduction technique to condense the feature matrix into a reduced dimensionality consisting of two principal components. This approach significantly improves data visualization. Additionally, our analysis includes the calculation of the proportion of variance explained and cumulative variance explained through PCA. This quantifies the extent to which the original data's variance is retained, ensuring the preservation of meaningful information.

Initially, a Principal Component Analysis (PCA) object is instantiated with a specified target dimensionality of 2. Subsequently, a StandardScaler object is generated with the purpose of standardizing the dataset, thereby harmonizing the scales of individual features to ensure uniformity. A processing pipeline is then established to seamlessly link the normalization and PCA dimensionality reduction procedures. Thereafter, the dimensionality reduction process is executed on the original feature matrix denoted as X.

 doi:10.20944/preprints202402.1575.v1

3.5. Hierarchical Clustering

In our approach, we employ the Agglomerative Clustering method for hierarchical clustering of the dimensionally reduced data. The number of clusters, referred to as 'k' is explicitly specified, and we use the Euclidean distance metric in combination with the ward linkage criterion as the proximity measures. Before initiating the clustering process, we prioritize the crucial step of identifying the optimal 'k' value. We accomplish this by evaluating the internal clustering quality using two essential metrics: the Silhouette Score and the Calinski-Harabasz Score. These metrics work together to determine the most appropriate 'k' value, ensuring a well-informed and data-driven selection for the clustering process. Before conducting each text clustering experiment, we followed an iterative procedure to identify the optimal number of clusters. This procedure involved a range of 'k' values, spanning from 2 to 30, each assessed five times. The resulting metric values were then subjected to an averaging process. Subsequently, we carried out a comparative analysis to visualize and assess the Silhouette Score and the Calinski-Harabasz Score in relation to 'k'.

4. Evaluations and Results

4.1. Datasets

This study was conducted using Chinese text data, which comprises three distinct types of text: government reports, news articles, and social media content.

Normative report text: This part of the content mainly comes from the government work report. The specific content includes the political report text includes the report to the 20th National Congress of the Communist Party of China (32,490 characters), the report to the 19th National Congress of the Communist Party of China (20,016 characters), and the Report to 2023 Beijing Municipal Government Work (15,016 characters).

News reporting text: The news text comes from the paper reports of Nanfang Daily in 2011. There is no limit on the topic and the total number of characters.

Social media text: Social media text comes from the Chinese online social media platform: Weibo. Randomly select 300 items from a certain topic as experimental samples, totaling 33,268 characters.

The experimental environment is Windows 11 operating system, 8 GB memory, and developed using Python.

4.2. Experimental Method

We propose a hierarchical clustering method called POS-HC, which leverages part-of-speech information to enhance the accuracy and meaningfulness of semantic partition in policy texts. The specific steps of this method are as follows:

Text preprocessing: The text is preprocessed using the Jieba Chinese word segmentation tool, which includes word segmentation, stop word removal, and part-of-speech tagging. A keyword-based vocabulary is selected to divide the text, and nouns, verbs, adverbs, adjectives, idioms, and sayings are filtered out to reduce irrelevant information.

Weighting of part-of-speech features: The POSP algorithm is employed to calculate the probability of occurrence for different parts of speech, and the results are stored in a dictionary. The TF-IDF value of each word is calculated using the TfidfVectorizer, and the TF-IDF and POSP matrix are horizontally stacked using the hstack function to create a new feature matrix, denoted as X.

PCA dimensionality reduction: Principal component analysis (PCA) is applied to reduce the dimensionality of the feature matrix, transforming the data into two principal components for visualization purposes.

Hierarchical clustering: Agglomerative Clustering is employed to cluster the dimensionally reduced data, with the number of clusters set to K. Euclidean distance is used as the distance measure, and ward connection is utilized (Murtagh & Legendre, 2014). The silhouette score and Calinski-Harabasz score are computed to evaluate the quality of clustering and determine the optimal value of K.

 doi:10.20944/preprints202402.1575.v1

Comparative experiment: A comparative experiment is conducted to evaluate the performance of our proposed hierarchical clustering method based on part-of-speech. It is compared with other community division methods, namely TF-IDF-based hierarchical clustering and TF-based hierarchical clustering, to assess its relative performance.

Robustness testing: The robustness of our proposed method is tested using different text datasets, varying data sizes, and noise levels to determine its applicability in different scenarios.

4.3. Experimental Evaluation Metrics

This method primarily employs an unsupervised machine learning approach for text data analysis. It relies on clustering internal indices, with the Silhouette score being a key metric (Shahapure & Nicholas, 2020) and CH score (Baarsch & Celebi, 2012), are selected as the evaluation metrics to assess the effectiveness of the clustering results.

In our study, we utilize the Silhouette Score as an evaluation metric to assess the proximity and distinctiveness of the clustering results. It calculates the silhouette coefficient for each sample and then computes the average across all the clustering outcomes. A higher silhouette coefficient value indicates a more effective clustering. The Calinski-Harabasz Score serves as an additional index in our analysis, allowing us to measure the compactness and separation of the clustering results. It quantifies the ratio between the within-cluster variance and the between-cluster variance, resulting in an exponential value. A higher Calinski-Harabasz index suggests a more favorable clustering outcome.

4.4. Results

This study encompasses three distinct comparative experiments. First, it evaluates the clustering performance of three weighted clustering algorithms. Second, it assesses the clustering outcomes when the POS-HC method is applied to three different types of text. The results reveal that the POS-HC method outperforms the other two methods and exhibits superior clustering effectiveness, especially in the domain of government text clustering. Furthermore, the robustness of the POS-HC method was tested, demonstrating its resilience. In addition, the Rest-Mex 2023 clustering method proposed by Madera-Quintana et al(Madera-Quintana et al., 2023). was also compared, which uses the TF-IDF and LSA algorithms to convert text into vectors and performs text clustering by the deK-Means method. Finally, the POS-HC method was employed to conduct an initial analysis of the semantic structure of the normative report.

4.4.1. Comparative experiment 1: Evaluating the Clustering Performance of Three Weighted Clustering Algorithms

To assess the relative advantages of the POS-HC method, three commonly used weighted clustering algorithms (TF-IDF, TF, and POS-HC) were employed to analyze three distinct datasets. The experimental results demonstrate that the POS-HC method exhibits significantly enhanced clustering effects across all three dataset types, as depicted in the Figure 2.

 doi:10.20944/preprints202402.1575.v1

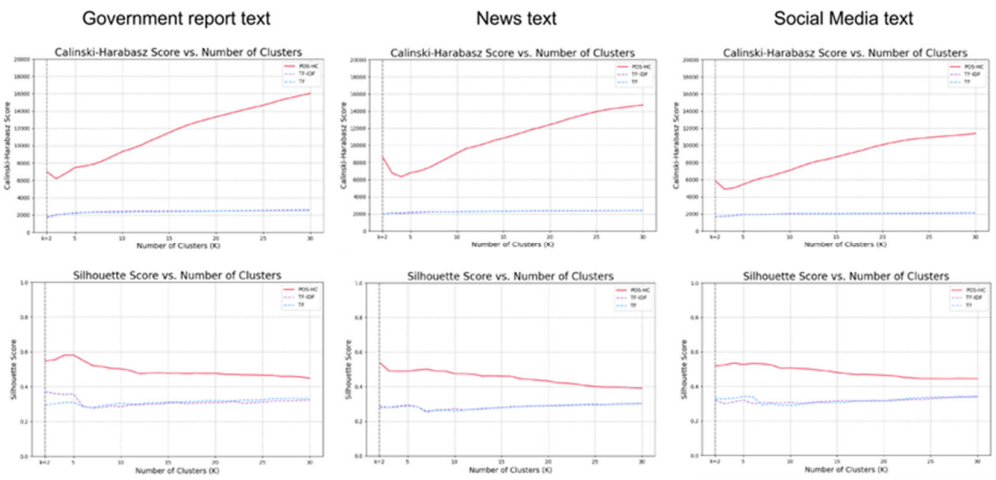

Figure 2. Clustering effect index of three clustering algorithms.

Specifically (As shown in Table 1), for government report text, the CH score interval [6191.52, 16037.06] and silhouette score interval [0.45, 0.58] achieved by the POS-HC method surpass those of the TF-IDF (CH:[1722.07, 2584.12], S:[0.28,0.37]) and TF(CH:[1797.80, 2525.87],S:[0.28,0.33]) methods. Similarly, for news text, the CH score interval [6339.23, 14718.59] and silhouette score interval [0.39, 0.54] under the POS-HC method outperform the TF-IDF(CH:[1946.66, 2396.15], S:[0.25, 0.30]) and TF(CH:[1926.99, 2394.22], S:[0.26, 0.30]) methods. Likewise, for social media text, the CH score interval [4885.20, 11387.22] and silhouette score interval [0.44, 0.54] obtained using the POS-HC method exceed those of the TF-IDF(CH:[1642.68, 2118.25], S:[0.30, 0.34]) and TF(CH:[1667.44,2139.47], S:[0.29,0.34]) methods.

Table 1. Clustering effects of three clustering algorithms on three different types of text.

Calinski-Harabasz score		
government report text		
TF-IDF [1722.07, 2584.12]	TF [1797.80, 2525.87]	POS-HC [6191.52, 16037.06]
news text		
TF-IDF [1946.66, 2396.15]	TF [1926.99, 2394.22]	POS-HC [6339.23, 14718.59]
social media text		
TF-IDF [1642.68, 2118.25]	TF [1667.44,2139.47]	POS-HC [4885.20, 11387.22]
Silhouette score		
government report text		
TF-IDF [0.28,0.37]	TF [0.28,0.33]	POS-HC [0.45, 0.58]
news text		
TF-IDF [0.25, 0.30]	TF [0.26, 0.30]	POS-HC [0.39, 0.54]
social media text		
TF-IDF [0.30, 0.34]	TF [0.29,0.34]	POS-HC [0.44, 0.54]

Regarding the observed trends of the two-evaluation metrics, the POS-HC method demonstrates an upward trend in the CH score and a downward trend in the silhouette score as the value of k (the number of clusters) increases. This pattern can be attributed to the dataset's inherent characteristics,

 doi:10.20944/preprints202402.1575.v1

which may lack a well-defined clustering structure. With an increase in the number of clusters, the number of samples within each cluster might decrease, resulting in reduced compactness within the clusters. However, the higher number of clusters can also enhance the separation between clusters, leading to a higher Calinski-Harabasz score. At the same time, the reduced number of samples within each cluster may decrease the similarity between individual samples and their respective clusters, consequently lowering the Silhouette score.

Therefore, when determining the optimal value of k, a balance between these two evaluation metrics is crucial. Additionally, the selection of the optimal k value should be guided by the visualization results of the clustering analysis. One approach to achieving a balance between the two indicators involves the utilization of Z-score normalization on the CH score and silhouette score values. This normalization process aims to standardize the data, enabling a fair comparison between the two indicators. Subsequently, visualization charts are generated depicting the normalized values. By examining Figure 3, one can pinpoint the intersection point and the highest point, which are critical in determining the optimal balance between the indicators.

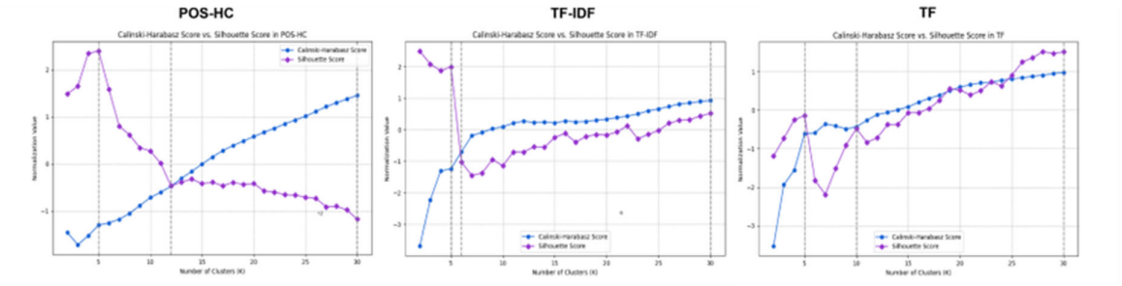

Figure 3. Combine CH score and Silhouette score to determine k.

Subsequently, the most suitable "k" values were selected for cluster visualization, as illustrated in Figure 4 below. Using the POS-HC method, "k" values of 5, 12, and 30 were chosen for clustering based on the two indicators. In the case of the TF-IDF method, "k" values of 5, 6, and 30 were selected for clustering, while for the TF method, "k" values of 5, 10, and 30 were utilized.

 doi:10.20944/preprints202402.1575.v1

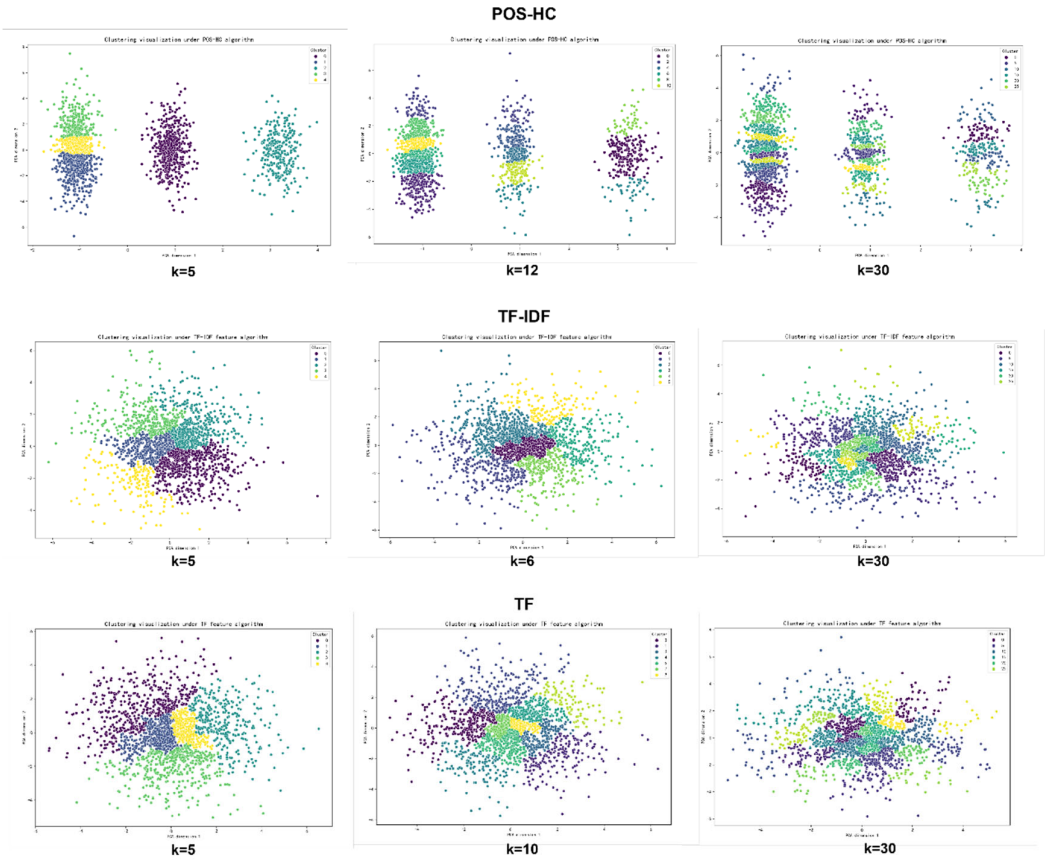

Figure 4. Clustering visualization of three algorithms under different k.

By comparing the visualization results, it was observed that the clustering effect was optimal when k was set to 5 under the POS-HC method, with the word clusters exhibiting the most distinct boundaries. Consequently, hierarchical clustering with k=5 under the POS-HC method was selected as the foundation for semantic partition.

4.4.2. Comparative experiment 2: Comparison of clustering effects of three types of text tests using POS-HC approach

To assess the high applicability of the POS-HC method to normative texts, we conducted a comparative analysis on three different types of datasets. The empirical findings demonstrate that, in terms of the CH score, government report texts (M= 11518.49) exhibit a superior performance compared to news texts (M= 10958.82) and social media texts (M= 8647.06). The application of ANOVA to analyze the differentiation among the three datasets yielded a p-value of 0.000126, which is less than the predetermined significance level of 0.05. Consequently, the observed p-value indicates a statistically significant separation among the datasets.

Regarding the silhouette scores, government report texts (M= 0.49) demonstrate a generally higher quality than news texts (M= 0.45) and social media texts (M= 0.47). The ANOVA analysis conducted to assess the separation degree of the three groups of data yielded a p-value of 0.000077, which is also less than the predetermined significance level of 0.05. This outcome further supports the conclusion that the observed separation among the datasets is statistically significant.

Figure 5 illustrates that, when k≥13, the performance of the POS-HC method on government report text surpasses that of other text types. This observation indicates that with an increase in the number of clusters (k), the POS-HC method exhibits improved performance in handling government report text. These outcomes collectively offer empirical evidence affirming the superior suitability of the POS-HC method for normative texts.

 doi:10.20944/preprints202402.1575.v1

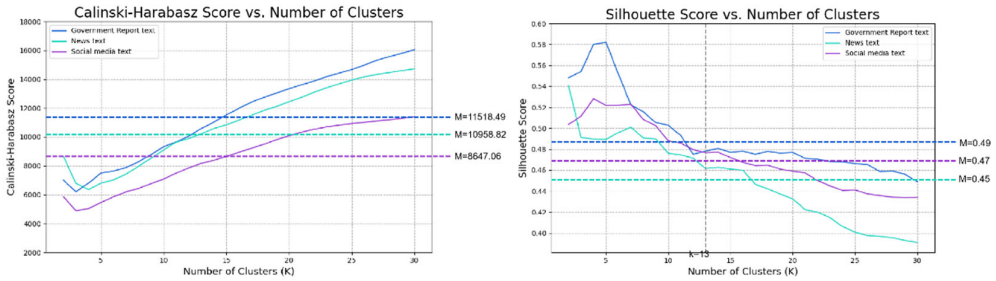

Figure 5. Comparison of three types of text clustering effects under POS-HC method. Note. The horizontal dashed line corresponds to the mean CH score and silhouette score of each type of text.

4.4.3. Comparative experiment 3: Comparison of Clustering Effects between Rest-Mex 2023 Method and Method on Normative Texts

The Rest-Mex 2023 method is a classification procedure for tourism texts in the Spanish language. This method is also unsupervised and applies the TF-IDF algorithm to convert the preprocessed text into a vector matrix. And reduce the dimensionality of the vector matrix by using latent semantic analysis (LSA) technology. Finally, K-Means clustering algorithm is used to obtain clusters. The results show that the proposed method is comparable to others in the same field [33].

To ascertain the superiority of the POS-HC method, a comparative analysis was conducted on the clustering parameters and results of the two methods.

This experiment uses the text of the 20th National Congress of the Communist Party of China as experimental data. The optimal k =5,8,30 using the Rest-Mex 2023 method is determined through the Silhouette score and CH, the silhouette coefficient interval is [2195.63, 2955.42], and the CH value interval is [0.24, 0.36], which is much lower than the POS-HC method(CH=[6191.52, 16037.06], Silhouette score =[0.45, 0.58]). According to the findings presented in Figure 6, it is evident that the clustering results obtained from the Rest-Mex method exhibit a higher degree of dispersion. Moreover, the boundaries between clusters are less distinct compared to those observed in the POS-HC method. Both quantitative indicators and visual representations consistently support the notion that the POS-HC method outperforms the Rest-Mex 2023 method.

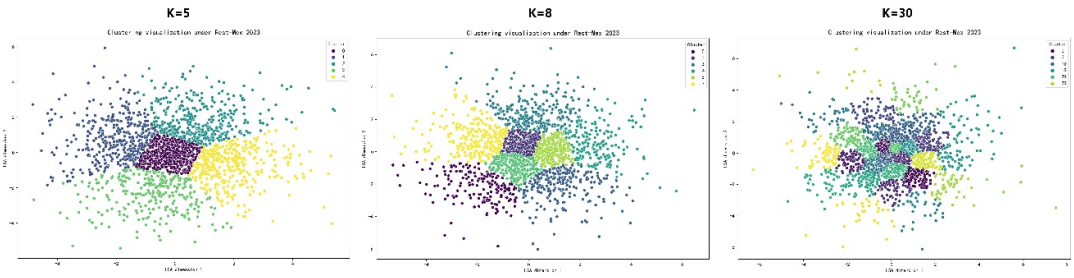

Figure 6. Clustering visualization of Rest-Mex 2023 method under different k.

4.4.4. Robustness Test: Evaluate the Clustering Effects Using POS-HC on Normative Texts

To evaluate the robustness of normative texts, we selected a sample of three report texts and employed the Mean Absolute Deviation (MAD) as an objective evaluation index to measure the stability of the clustering results. The experimental findings demonstrate the presence of clustering stability, as indicated by the calculation of the MAD averages for both the Calinski-Harabasz scores and Silhouette scores. Specifically, the average MAD of the Calinski-Harabasz score across the three report datasets was determined to be 0.27, while the average MAD of the Silhouette score was found to be 0.034. Notably, the separation with p=0.78, which is greater than 0.05, indicates that the separation is not statistically significant.

 doi:10.20944/preprints202402.1575.v1

Additionally, as depicted in the boxplot in Figure 7, a striking similarity in trends is observed among the three text types concerning these two indicators. Upon careful outlier examination, it's worth noting that only three outliers were detected in the silhouette score index across the three text types, and their number is relatively low. This visual representation further substantiates the robustness of the normative semantic partition approach employed in this study.

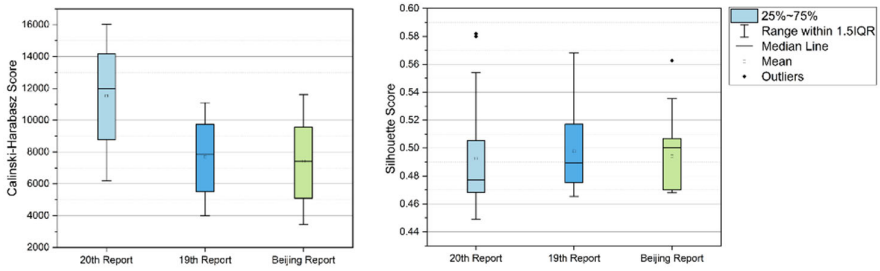

Figure 7. Comparison of the clustering effects of three government report texts under the POS-HC method.

4.4.5. Exploratory Semantic Clustering of Normative Texts with POS-HC

Following the insights gathered from Comparative Experiment 1, we opted for the POS-HC method to perform semantic clustering of normative texts. Specifically, we employed hierarchical clustering with a parameter "k" set to 5. We then executed the algorithm, and the top twenty words from each cluster were meticulously arranged into the table provided below:

Table 2. Clustering results - the top 20 words with highest frequency.

Cluster 0	pos	count	Cluster 1	pos	count	Cluster 2	pos	count	Cluster 3	pos	count	Cluster 4	pos	count
persist	v	169	politics	n	54	sound	a	52	subject	n	10	people	n	118
						considerable			administration			nation		
advance	v	105	strategy	n	45		a	34	on	n	8		n	105
strengthen						Firm						entireness		
	v	92	ability	n	38		a	27	vitality	n	7		n	100
complete						revived			judiciary			system		
	v	73	policy	n	19		a	22		n	7		n	88
						implemented			comrade			institution		
promote	v	61	level	n	17		a	22		n	6		n	74
accomplish						firmly			target			socialism		
	v	55	ecology	n	16		ad	21		n	6		n	60
innovate	v	47	era	n	16	stable	a	17	power	n	6	society	n	52
promote	v	39	focus	n	15	important	a	16	group	n	6	question	n	51
						Fundamental						world		
enhance	v	36	nationality	n	15		a	15	sovereignty	n	5		n	48
						actively			party&people			mechanism		
maintain	v	34	way	n	14		ad	14		n	5		n	34
improve	v	34	military	n	14	powerful	a	11	rectify	n	5	economy	n	33
assure	v	33	environment	n	13	extensive	a	11	compatriots	n	5	rule of law	n	32
accelerate	v	33	green	n	13	unity	a	10	overall	n	5	history	n	31
implement	v	32	Urban and rural	n	13	Effective	a	10	decision making	n	5	culture	n	31
			strong			respected			farmer			internationality		
govern	v	32	country	n	13		a	10		n	5		n	31

 doi:10.20944/preprints202402.1575.v1

construc						superior			diplomatic			Marxism		
t	v	29	specification	n	12		a	9		n	4		n	29
			party			unshakabl			creativity			technology		
planning	v	26	leadership	n	11	e	i	9		n	4		n	29
protect	v	24	area	n	11	healthy	a	9	theme	n	3	Spirit	n	26
practice	v	24	feature	n	10	Huge	a	8	situation	n	3	lead	n	26
support	v	22	guide	n	10	Efficient	a	8	patriot	n	3	theory	n	26

Upon analyzing the provided table, it becomes evident that verbs are consolidated within a single cluster, while adjectives, adverbs, and idioms constitute another cluster. In contrast, nouns are distributed across three distinct clusters. Notably, both the verb and noun clusters exhibit higher word frequencies, with the verb "persist" emerging as the most frequently occurring word. This observation underscores the importance of both nouns and verbs in government report texts, highlighting the fixed nature of verbs, which are often paired with various nouns to convey specific meanings. The prevalence of the word "persist" further emphasizes the consistent presence of substantial content in normative reports, reinforcing its recurring significance. Upon scrutinizing the high-frequency words within the three noun clusters, it becomes clear that "politics," "subject," and "people" are the most prominently employed terms. Furthermore, the clustering of adjectives, adverbs, and idioms into a single cluster aligns with semantic coherence, given the potential overlap between these parts of speech in the Chinese language.

A comprehensive examination of these clustering results yields valuable insights into the characteristics and importance of various parts of speech within government report texts. In future research endeavors, it is advisable to integrate both the word co-occurrence matrix and the entity relationship matrix for a more extensive and profound analysis.

5. Conclusion

In this study, we address the need for text clustering to achieve semantic partitioning. Building upon the TF-IDF clustering approach, we augment it with the inclusion of POS (Part-of-Speech) features, culminating in the development of a novel text clustering method called POS-HC (Part-of-Speech Hierarchical Clustering). Through the application of the POS-HC clustering algorithm to a diverse set of normative text datasets, we unearth concealed layers of information and intricate patterns within these texts. Our empirical comparative analysis underscores the method's superior clustering efficiency when compared to traditional feature calculation techniques. This work aims to offer an insightful perspective on mining structures and underlying themes in normative texts, particularly in the realm of social science research.

Our study broadens the scope of this assessment to encompass three distinct categories of textual data: normative report texts, news articles, and social media content. The findings reveal that POS-HC excels in clustering government report texts in comparison to the other two text categories. Moreover, to gauge the robustness of the approach, we conducted a comparative analysis involving report texts from three different sources. This analysis revealed that the average Mean Absolute Deviation (MAD) for the Calinski-Harabasz score across the three report datasets was 0.27, while the average MAD for the Silhouette score was 0.034, indicating the presence of clustering stability.

These findings underscore the effectiveness and robustness of the POS-HC method when applied to texts with varying degrees of normality, affirming its suitability for normative text datasets. The enhanced feature weight algorithm of POS-HC contributes to the refinement of semantic partition models for normative text, thus expanding the toolkit of social computing.

However, it's important to acknowledge certain limitations of this study. Firstly, the stability of the hierarchical clustering algorithm used may not reach an optimal level of robustness. To address this, our study implemented a strategy of averaging the results from multiple iterations to mitigate this instability. Future research should explore alternative clustering algorithms with higher stability to reduce uncertainty [34].

 doi:10.20944/preprints202402.1575.v1

Moreover, it's essential to acknowledge that the method's applicability may exhibit potential biases across various domains, especially in the realm of government-authored normative texts. Consequently, a comprehensive validation and fine-tuning process becomes imperative in practical applications to ensure the method's broader utility. Furthermore, it's worth noting that the algorithm's performance remains contingent on data quality and size. Therefore, parameter adjustments and optimization are necessary when dealing with datasets of varying quality and size.

Funding: This research was supported by the National Social Science Foundation of China under Grant 23&ZD215.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. M. L. Doerfel, "What Constitutes Semantic Network Analysis? A Comparison of Research and Methodologies," 1998.
2. A. Miani, T. Hills, and A. Bangerter, "Interconnectedness and (in) coherence as a signature of conspiracy worldviews," *Sci Adv*, vol. 8, no. 43, p. eabq3668, 2022.
3. G. Grand, I. A. Blank, F. Pereira, and E. Fedorenko, "Semantic projection recovers rich human knowledge of multiple object features from word embeddings," *Nat Hum Behav*, vol. 6, no. 7, pp. 975–987, 2022, doi: 10.1038/s41562-022-01316-8.
4. Marten. Scheffer, Ingrid. Van De Leemput, Els. Weinans, and Johan. Bollen, "Reply to Sun: Making sense of language change," *Proc Natl Acad Sci U S A*, vol. 119, no. 26, p. e2206616119, 2022.
5. H. Lee and J. Chen, "Narratives as Networks: Predicting Memory from the Structure of Naturalistic Events," *Cold Spring Harbor Laboratory*, 2021.
6. N. M. Salih and K. Jacksi, "Semantic Document Clustering using K-means algorithm and Ward's Method," in *2020 International Conference on Advanced Science and Engineering (ICOASE)*, 2021.
7. S. Ghosh *et al.*, "Distributed louvain algorithm for graph community detection," in *2018 IEEE international parallel and distributed processing symposium (IPDPS)*, IEEE, 2018, pp. 885–895.
8. J. Geller, Y. Perl, M. Halper, Z. Chen, and H. Gu, "Evaluation and application of a semantic network partition," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 2, pp. 109–115, 2002, doi: 10.1109/TITB.2002.1006297.
9. Z. Xu, F. Zhichen, Y. Ting, Mao, S. Jianchang, and Difu, *Towards the Semantic Web: Collaborative Tag Suggestions*. 2006.
10. J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, vol. 21, no. 3, pp. 267–297, 2013.
11. A. Abdi, N. Idris, R. M. Alguliev, and R. M. Aliguliyev, "Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems," *Inf Process Manag*, vol. 51, no. 4, pp. 340–358, 2015.
12. J.-P. Schöggl, L. Stumpf, and R. J. Baumgartner, "The narrative of sustainability and circular economy-A longitudinal review of two decades of research," *Resour Conserv Recycl*, vol. 163, p. 105073, 2020.
13. S. H. Haji, K. Jacksi, and R. M. Salah, "A Semantics-Based Clustering Approach for Online Laboratories Using K-Means and HAC Algorithms," *Mathematics*, vol. 11, no. 3, p. 548, 2023.
14. P. Zhang, W. Gao, J. Hu, and Y. Li, "Multi-label feature selection based on the division of label topics," *Inf Sci (N Y)*, vol. 553, pp. 129–153, 2021.
15. P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, 2016, pp. 61–66.
16. C. Liu, Y. Sheng, Z. Wei, and Y.-Q. Yang, "Research of text classification based on improved TF-IDF algorithm," in *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, IEEE, 2018, pp. 218–222.
17. Y. Yang, "Research and realization of internet public opinion analysis based on improved TF-IDF algorithm," in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*, IEEE, 2017, pp. 80–83.
18. D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and A. G. Dunn, "Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection," *J Med Internet Res*, vol. 18, no. 8, p. e232, 2016.
19. N. Arhab, M. Oussalah, and M. S. Jahan, "Social media analysis of car parking behavior using similarity based clustering," *J Big Data*, vol. 9, no. 1, pp. 1–29, 2022.

 doi:10.20944/preprints202402.1575.v1

20. Z. Akbar, J. Liu, and Z. Latif, "Mining social applications network from business perspective using modularity maximization for community detection," *Soc Netw Anal Min*, vol. 11, pp. 1–19, 2021.
21. H. Zhang, L. Lin, L. Xu, and X. Wang, "Graph partition based privacy-preserving scheme in social networks," *Journal of Network and Computer Applications*, vol. 195, p. 103214, 2021.
22. C. Chan, A. Al-Bashabsheh, and C. Zhao, "Finding better Web communities in digraphs via max-flow min-cut," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019, pp. 410–414.
23. H. Li, X. Hu, Y. Lin, W. He, and J. Pan, "A social tag clustering method based on common co-occurrence group similarity," *Frontiers of Information Technology & Electronic Engineering*, vol. 17, no. 2, pp. 122–134, 2016.
24. J.-Y. Hang and M.-L. Zhang, "Collaborative learning of label semantics and deep label-specific features for multi-label classification," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 12, pp. 9860–9871, 2021.
25. S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, pp. 1–21, 2019.
26. D. Tian, M. Li, J. Shi, Y. Shen, and S. Han, "On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach," *Advanced Engineering Informatics*, vol. 49, p. 101355, 2021.
27. M. Devaraj, "Analyzing News Sentiments and their Impact on Stock Market Trends using POS and TF-IDF based approach," in *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, 2020, pp. 1–6.
28. T. Wang, L. Liu, N. Liu, H. Zhang, L. Zhang, and S. Feng, "A multi-label text classification method via dynamic semantic representation model and deep neural network," *Applied Intelligence*, vol. 50, pp. 2339–2351, 2020.
29. O. Savic, L. Unger, and V. M. Sloutsky, "Exposure to co-occurrence regularities in language drives semantic integration of new words.," *J Exp Psychol Learn Mem Cogn*, vol. 48, no. 7, p. 1064, 2022.
30. M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "LSTM, VADER and TF-IDF based hybrid sentiment analysis model," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.
31. L. Zhu, G. Wang, and X. Zou, "A study of Chinese document representation and classification with Word2vec," in *2016 9th International symposium on computational intelligence and design (ISCID)*, IEEE, 2016, pp. 298–302.
32. R. Gopalakrishnan and A. Venkateswarlu, Machine Learning for Mobile: Practical guide to building intelligent mobile applications powered by machine learning. Packt Publishing Ltd, 2018.
33. J. Madera-Quintana, A. Hernández-Gónzalez, and Y. Martínez-López, "Thematic Unsupervised Classification of Tourist Texts using Latent Semantic Analysis and K-Means," *Proceedings http://ceur-ws. org ISSN*, vol. 1613, p. 0073, 2023.
34. L. Abualigah *et al.*, "Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis," *Algorithms*, vol. 13, no. 12, p. 345, 2020.