Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Self-Constructing Feature Clustering for Text Classification: An Automated Approach

[1]Santoshkumar V. Chobe, [2]Swati Nikam
[1][2] Pimpri Chinchwad College of Engineering and Research, Pune, 412101, India

E-mail: [1]sanchobe@gmail.com, [2] swatinikam3@gmail.com

Abstract: Text classification is a pivotal aspect of natural language processing, requiring advanced techniques for feature extraction and representation. This paper presents a novel approach to feature clustering in text classification, employing a self-constructing algorithm enriched with statistical membership functions to address the challenge of efficient text classification. The proposed method efficiently reduces the dimensionality of the feature vector by grouping words into clusters. Each cluster is represented by a single feature, automatically generated through a process that considers the equality or dissimilarity of words. The clustering is driven by membership functions incorporating statistical mean and deviation, ensuring robust and representative feature grouping. The automatic creation of clusters enhances adaptability to diverse textual datasets. The integration of self-constructing feature clustering with statistical membership functions contributes to a scalable and adaptive solution for text classification tasks. Experimental results demonstrate the effectiveness of the proposed method, showcasing its ability to enhance text classification performance through efficient feature representation.

Keywords: Feature Clustering, Feature Selection, Natural Language Processing, Text Classification.

1. Introduction

In the ever-expanding realm of Natural Language Processing (NLP), the accurate classification of textual data remains a fundamental challenge. The accuracy and efficiency of text classification heavily rely on the representation of features within the data, demanding innovative approaches for dimensionality reduction and enhanced model interpretability. This paper introduces a pioneering methodology to address these challenges. The proposed methodology seeks to revolutionize the conventional feature clustering techniques by incorporating self-construction principles and statistical membership functions, aiming to provide a dynamic and efficient solution to the demands of text classification tasks.

Traditional feature clustering techniques often face limitations in adaptability and efficiency, particularly in handling diverse and evolving textual datasets. In response to this, the proposed approach harnesses a self-constructing feature clustering algorithm enriched with statistical membership functions. The key objective is to automate the process of creating representative features by dynamically clustering words based on their similarity or dissimilarity. These clusters are then characterized by statistical measures, including mean and deviation, ensuring a robust and interpretable representation of the underlying textual features.

As we investigate the intricacies of this methodology, we aim to showcase its capacity to streamline the feature selection process, enhance adaptability to varied datasets, and ultimately improve the accuracy and efficiency of text classification tasks. Through a series of experiments and evaluations, we demonstrate the advantages of our automated approach over conventional methods, thereby contributing to the evolution of scalable and adaptive solutions for the challenges posed by text classification in current natural language processing applications. This research not only contributes to the advancement of text classification techniques but also underscores the broader impact of automated feature construction in the evolving landscape of NLP applications.

The remaining of this paper is organized as follows. In section 2, we present the literature review where various other techniques are discussed. In section 3, we present our proposed methodology. Finally, in section 4 we present the future scope of proposed methodology.

2. Literature Review

Text classification finds wide-ranging applications in practical scenarios, encompassing tasks such as automated categorization of webpages or documents based on predefined labels [1]. Its utility extends to diverse domains, including but not limited to sorting new patents into relevant categories, analysing user sentiment in multimedia content on social networks [2], filtering spam emails, delivering targeted information to subscribers, identifying document genres, tagging videos [3], and recommending multimedia content [4].

However, the challenge arises from the fact that a typical webpage or text document can comprise hundreds or even thousands of distinct terms. Employing all these terms for text classification can yield suboptimal results due to the presence of uninformative terms and the potential for misguiding classifiers [5,6]. In this survey, our objective is to furnish researchers and practitioners with a comprehensive grasp of feature selection theories, models, and techniques. We particularly focus on the cutting-edge feature selection methodologies specifically tailored for enhancing the efficacy of text categorization.

Over the past decade, a large amount of statistical and machine learning techniques have been developed for the automated classification of documents. These encompass diverse methodologies, including the k-Nearest Neighbours (kNN) [7, 8], Naïve Bayes [9], Rocchio algorithm [10], multivariate regression models [11,12], decision trees [13,14], Support Vector Machines (SVMs) [15], neural networks [16, 17, 18], graph partitioning-based approaches [19], and methods utilizing genetic algorithms [20, 21]. This section provides an overview of some of the most frequently employed classifiers in the realm of text categorization.

The k-Nearest Neighbours (kNN) classification stands out as a widely utilized non-parametric method in diverse fields, including data mining, machine learning, information retrieval, and statistics [22]. In the context of document categorization, when confronted with a document d_i of unknown category, the kNN method relies on a user-defined parameter k and a dataset D containing documents, each associated with a category. This method computes the k nearest documents for d_i based on a specified similarity measure. Subsequently, the kNN method assigns to d_i the category that is most frequently observed among the k nearest documents.

The determination of the nearest neighbours involves assessing each document in D using a distance or similarity measure. Notably, when the parameter k is set to 1, the kNN method simplifies to the Nearest Neighbour (NN) method. Renowned for its simplicity and reasonably good performance, the kNN method has found application in real-world scenarios. However, despite its widespread use, it grapples with a significant drawback - high computational costs. This arises from its lazy learning nature, where, upon receiving an object, the kNN method must scrutinize the entire dataset to identify the k-nearest neighbours for the given object.

The Naïve Bayes (NB) classification method has been extensively explored in the realm of text categorization [9]. Typically, NB classifiers operate under the assumption that the value of a specific feature is independent of the value of any other feature. In the context of text classification, the Naïve Bayes assumption suggests that the probability of each word appearing in a document is independent of the occurrence of other words in the same document. There are two primary types of NB-based text classifiers. The first, known as multivariate Bernoulli NB, that utilizes a binary vector to represent a document. Each component of the vector signifies whether a term is present or absent in the document [23, 24]. The second type is multinomial NB, which also incorporates term frequencies within the document [24, 25]. In practical applications, multinomial NB classifiers often outperform their multivariate counterparts, especially in large document collections [24]. However, recent research has identified two drawbacks associated with multinomial NB classifiers. First, there are challenges related to rough parameter estimation. Second, there is a bias against rare classes that contain only a few training documents. To address these issues, researchers have proposed effective techniques aimed at further enhancing the prediction accuracy of multinomial NB classifiers [26].

The Decision Tree (DT) stands as a well-established machine learning algorithm widely utilized in diverse automatic classification tasks [13, 14]. In the context of text categorization, DT learning algorithms play a pivotal role in selecting informative words based on the information gain criterion. When tasked with classifying a document, the constructed decision tree is leveraged to predict the document's category by assessing the occurrence of word combinations within it. Studies have indicated that decision trees, in terms of prediction accuracy, often outperform Naïve Bayes classifiers and Rocchio's algorithm. However, they are reported to be marginally less effective than kNN methods [22, 27, 28]. This underscores the utility of decision

trees in text categorization tasks while acknowledging their comparative strengths and weaknesses in relation to other classification methodologies.

Support Vector Machines (SVMs), introduced by Vapnik et al. [15] for classification tasks, operate on the principle of structural risk minimization. This approach aims to construct an optimal hyperplane with the widest possible margin to effectively separate a set of data points comprising positive and negative examples. SVMs have emerged as a potent classification tool, finding widespread success in various applications, including object recognition [29], image classification [30], and text categorization [31, 32].

Joachims [31] was a pioneer in applying SVMs to text categorization tasks, driven by the compatibility of SVMs with key characteristics of textual data. Firstly, text data is inherently high-dimensional, often containing tens of thousands of terms. Remarkably, SVMs demonstrate an ability to learn independently of the dimensionality of the feature space. Secondly, despite the abundance of features in text data, there are typically few irrelevant features in a document. SVMs can consider all features, in contrast to conventional classification methods that often resort to feature selection techniques to manage the feature space. Thirdly, document vectors are sparse, with each document containing only a few non-zero entries. SVMs excel in handling such classification problems characterized by dense concepts and sparse instances. Through extensive experiments, Joachims demonstrated that SVMs consistently outperform traditional classifiers, including Naïve Bayes, Rocchio, decision trees, and kNN [31]. This substantiates the effectiveness of SVMs in text categorization and underscores their superiority in various dimensions compared to conventional classification approaches.

3. Proposed Methodology

Despite the abundance of classifiers designed for text categorization, a significant difficulty persists—the high dimensionality of the feature space [33]. Typically, a document encompasses hundreds or even thousands of distinct words, each considered a feature. However, a substantial portion of these features may be characterized as noisy, less informative, or redundant concerning class labels. This situation poses a risk of misleading classifiers, consequently compromising their overall performance [5, 6]. As a remedy, feature selection becomes imperative to weed out such noisy, less informative, and redundant features. This process is essential for condensing the feature space to a manageable level, thereby enhancing the efficiency and accuracy of the employed classifiers.

The proposed system introduces an innovative text classification method that leverages a self-constructing feature clustering algorithm. The primary objective is to streamline the text classification task by reducing the dimensionality of the feature vector while representing words as distributions. Figure 1 shows the architecture of the proposed system.

This pioneering approach not only optimizes the text classification task but also introduces a dynamic and adaptive element through self-constructing feature clustering. By representing words as distributions and strategically organizing them into clusters, this algorithm enhances the efficiency and accuracy of text classification, ultimately leading to more robust and nuanced results.

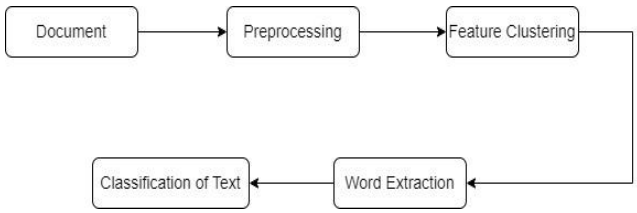

Figure 1. Proposed Architecture

The key components of the proposed system are as follows:

1. Pre-processing:

As shown in Figure 2 stop words are removed from the text document and generated the word pattern weightage from the feature vector.

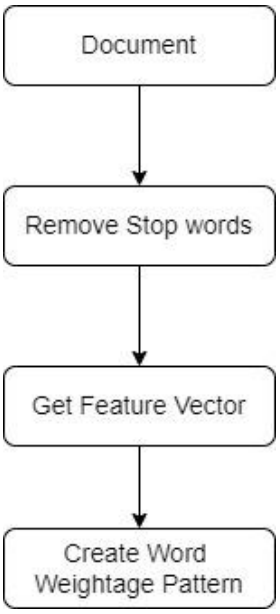

Figure 2. Pre-processing

2. Self-Constructing Feature Clustering Algorithm:

The heart of the system is a self-constructing feature clustering algorithm designed to dynamically group words in the feature vector. This algorithm operates iteratively, systematically forming clusters based on the similarity of words. The self-constructing nature of the algorithm ensures adaptability to varying linguistic patterns and data distributions. The detailed steps are shown in Figure 3.

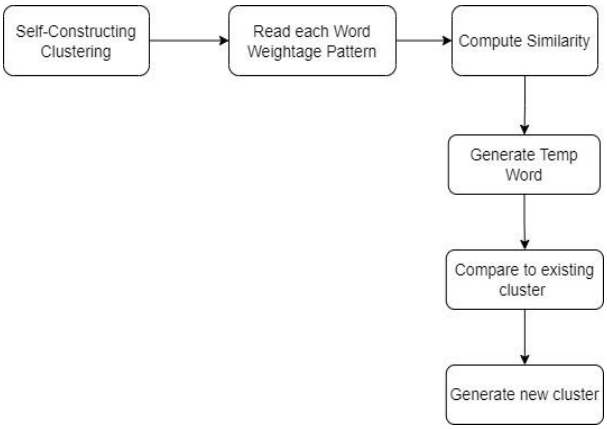

Figure 3. Feature Clustering

3. **Representation of Words as Distributions:**

In contrast to traditional methods, words in the feature vector are represented as distributions, capturing the probabilistic nature of word occurrences in documents. This representation allows for a detailed understanding of word relationships and contributes to the robustness of the clustering process.

4. **Cluster Characterization with Statistical Metrics:**

Words that are similar to each other are grouped into clusters, each characterized by statistical metrics such as mean and deviation. Statistical metrics provide a quantitative representation of the central tendency and variability within each cluster, enhancing the interpretability of the constructed features. If a word does not

match any existing cluster, a new cluster is dynamically created for that word. This dynamic creation process ensures that the algorithm can adapt to novel or evolving language patterns, making it well-suited for diverse and dynamic textual datasets. After getting the feature vector, the weighting matrix is found and then the text is classified as shown in Figure 4.

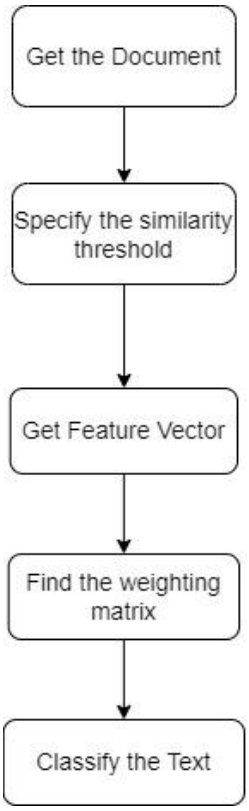

Figure 4. Text Classification

The proposed system stands out for its ability to automate the feature construction process, adapt to varying linguistic contexts, and represent words as distributions for a more comprehensive understanding of their contextual significance. Through the integration of self-constructing feature clustering and statistical characterization, the system aims to improve the efficiency and interpretability of text classification tasks.

4. Future Scope

The proposed methodology opens avenues for several promising directions in the realm of natural language processing and text analysis. The following are potential areas for future exploration and enhancement:

- **Dynamic Integration with Deep Learning Models:** Explore the seamless integration of the self-constructing feature clustering approach with deep learning models. Investigate how this automated feature construction method can complement the capabilities of neural networks, potentially enhancing the interpretability and performance of complex models.
- **Cross-Domain Adaptability:** Extend the methodology to explore its adaptability across various domains and industries. Investigate how the self-constructing feature clustering algorithm performs when confronted with diverse textual datasets, including those from specialized domains such as medical literature, legal documents, or technical reports.
- **Online Learning and Incremental Updates:** Develop strategies for incorporating online learning techniques, allowing the algorithm to adapt incrementally to changing data distributions. This would make the proposed approach more resilient to evolving language patterns and enable real-time adaptation to emerging trends.

- **Multimodal Text and Image Classification:** Extend the application of the methodology to multimodal tasks, combining textual and visual information. Investigate how the self-constructing feature clustering can be adapted to handle datasets where both textual and image-based features contribute to the classification task.
- **Optimization for Large-Scale Text Corpora:** Investigate optimization techniques to scale the proposed approach for large-scale text corpora. This includes exploring distributed computing frameworks and parallel processing strategies to handle vast amounts of textual data efficiently.
- **Human-in-the-loop Integration:** Explore the integration of human-in-the-loop feedback mechanisms to refine and improve the feature clustering process. Investigate how domain experts can provide feedback to enhance the interpretability and accuracy of the constructed features.
- **Real-time Text Classification:** Explore real-time applications of the proposed methodology, particularly in scenarios where timely classification decisions are crucial. Investigate optimizations and strategies to ensure the algorithm's responsiveness in processing and classifying incoming text streams.

By exploring these future directions, researchers can contribute to the ongoing evolution of automated feature construction techniques for text classification, advancing its capabilities and addressing emerging challenges in the field of text classification.

References

[1] Zhao S, Yao H, Zhao S, Jiang X, Jiang X (2016) Multi-modal microblog classification via multi-task learning. Multimed Tools Appl 75(15):8921–8938
[2] Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tool Appl 75(5):2507–2525
[3] Ballan L, Bertini M, Uricchio T, Del Bimbo A (2015) Data-driven approaches for social image and video tagging. Multimed Tool Appl 74(4):1443–1468
[4] Pappas N, Popescu-Belis A (2015) Combining content with user preferences for non-fiction multimedia recommendation: a study on ted lectures. Multi Tools Appl 74(4):1175–1197
[5] Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502
[6] Sebastiani F (2002) Machine learning in automated text cateogirzation. ACM Comput Surv 34(1):1–47
[7] Cunningham P, Delany SJ (2007) k-nearest neighbour classifiers. Multiple Class Syst 34:1–17
[8] Zhang S, Li X, Zong M, Zhu X, Wang R (2017) Efficient knn classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems. https://doi.org/10.1109/TNNLS.2017.2673241
[9] Domingos P, Pazzani M (1997) On the optimality of the simple bayesian classifier under zero-one loss. Mach Learn 29(2/3):103–130
[10] Rocchio JJ (1971) Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing
[11] Sch¨utze H, Hull DA, Pedersen JO (1995) A comparison of classifiers and document representations for the routing problem. In: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 229–237
[12] Yang Y, Chute CG (1994) An example-based mapping method for text categorization and retrieval. ACM Trans Inf Syst 12(3):252–277
[13] Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
[14] Quinlan JR (2014) C4. 5: programs for machine learning. Elsevier
[15] Vapnik VN, Vapnik V (1998) Statistical learning theory, vol 1. Wiley, New York
[16] Johnson R, Zhang T (2014) Effective use of word order for text categorization with convolutional neural networks. arXiv:1412.1058
[17] Ruiz ME, Srinivasan P (2002) Hierarchical text categorization using neural networks. Inf Retr

5(1):87–118
[18] Wiener E, Pedersen JO, Weigend AS et al (1995) A neural network approach to topic spotting. In: Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval, vol 317. Las Vegas, NV, p 332
[19] Gao B, Liu TY, Feng G, Qin T, Cheng QS, Ma WY (2005) Hierarchical taxonomy preparation for text categorization using consistent bipartite spectral graph copartitioning. IEEE Trans Knowl Data Eng 17(9):1263–1273
[20] Pietramala A, Policicchio VL, Rullo P, Sidhu I (2008) A genetic algorithm for text classification rule induction. In: Joint european conference on machine learning and knowledge discovery in databases. Springer, pp 188–203
[21] Uysal AK, Gunal S (2014) Text classification using genetic algorithm oriented latent semantic features. Expert Syst Appl 41(13):5938–5947
[22] Yang Y (1999) An evaluation of statistical approaches to text categorization. Inf Retr 1(1):69–90
[23] McCallum A, Nigam K (1998) Employing em in poll-based active learning for text classification. In: Proceedings of the 15th international conference on machine learning, pp 350–358
[24] McCallum A, Nigam K et al (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol 752. Madison, WI, pp 41–48
[25] Yang Y, Liu X (1999) A re-examination of text categorization methods. In: Proceedings of the 22nd ACM SIGIR, pp 42–49
[26] Kim SB, Han KS, Rim HC, Myaeng SH (2006) Some effective techniques for naive bayes text classification. IEEE Trans Knowl Data Eng 18(11):1457–1466
[27] Dumais S, Platt J, Heckerman D, Sahami M (1998) Inductive learning algorithms and representations for text categorization. In: Proceedings of the seventh international conference on Information and knowledge management. ACM, pp 148–155
[28] Lewis DD, Ringuette M (1994) A comparison of two learning algorithms for text categorization. In: 3rd annual symposium on document analysis and information retrieval, vol 33, pp 81–93
[29] Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol 3. IEEE, pp 32–36
[30] Lin Y, Lv F, Zhu S, Yang M, Cour T, Yu K, Cao L, Huang T (2011) Large-scale image classification: fast feature extraction and svm training. In: 2011 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1689–1696
[31] Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. Mach Learn: ECML-98:137–142
[32] Taira H, Haruno M (1999) Feature selection in svm text categorization. In: AAAI/ IAAI, pp 480–486
[33] Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: ICML, vol 97, pp 412–420