International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 08, ISSN No. 2455-2143, Pages 94-96
Published Online December 2023 in IJEAST (http://www.ijeast.com)

ENHANCING TEXT CLASSIFICATION: INSIGHTS FROM FEATURE SELECTION RESEARCH

Don C Delish
Student-Department of IT 2020-24
Amal Jyothi College Of Engineering, Kanjirappally, Kottayam, India

Abstract—**This research paper, "ENHANCING TEXT CLASSIFICATION: Insights from Feature Selection Research," explores the pivotal role of feature selection methodologies in optimizing text classification processes. Text classification is a cornerstone of natural language processing, aiding in the efficient organization and comprehension of large textual datasets. However, the inherent high dimensionality of text data presents challenges in processing efficiency and model performance. The paper delves into a comprehensive survey of feature selection techniques, shedding light on their methodologies and implications. The paper covers diverse feature selection methods, including probabilistic approaches, hybrid strategies that combine filter and wrapper methods, and systematic literature reviews categorizing and analyzing existing methods. It underscores the significance of feature selection as a fundamental step in improving text classification efficiency and effectiveness. The evolving trends in this field are also a central focus, revealing a growing preference for wrapper strategies over traditional filter-based methods, the emergence of metaheuristic approaches, and the maturation of feature selection research within text classification. The paper delves into the evaluation of feature selection methods, discussing the usage of various datasets, particularly English text corpora, and the application of machine learning algorithms like Support Vector Machines (SVM) and Naive Bayes (NB). A range of metrics is examined to assess the effectiveness of feature selection techniques, including accuracy, precision, recall, F1 scores, and dimension reduction rates. The paper concludes by emphasizing the dynamic nature of feature selection research, highlighting the opportunities for ongoing advancements in text classification**

Keywords— **High dimensionality, Filters, Wrappers, Hybrid Feature Selection, Feature Selection**

I. INTRODUCTION TO DOMAIN

Feature selection is a critical step in text classification that improves accuracy and reduces processing time by identifying the most relevant features from a large set. Each distinct term in a text document corresponds to a feature. The high dimensionality of the feature space poses a challenge due to its impact on processing time and accuracy. Feature selection techniques fall into three categories: filters, wrappers, and embedded methods. Filters evaluate features independently of a learning model or classifier using various scoring frameworks. Wrappers evaluate features using a specific learning model and search algorithm, considering feature dependencies. Embedded methods integrate feature selection into classifier training, making them specific to the learning model. Commonly preferred feature selection methods in text classification are filters due to their relatively low processing time. Filters, wrappers, and embedded methods can be combined in hybrid approaches.

Benefits of feature selection include improved classification accuracy, reduced processing time, reduced dimensionality of the feature space, and improved model interpretability.

II. DEFINITIONS

A. High dimensionality

The large number of features in a text classification problem that can lead to processing time and accuracy issues.

B. Filters

Feature selection techniques that assess feature relevancies using various scoring frameworks that are independent of a learning model or classifier and select top-N features attaining the highest scores. Examples of filter methods include term strength, odds ratio, document frequency, mutual information, chi-square, information gain, improved Gini index, measure of deviation from Poisson distribution, a support vector machine-based feature selection algorithm, ambiguity measure, class discriminating measure, and binomial hypothesis testing.

C. Wrappers

Feature selection techniques that evaluate features using a specific learning model and search algorithm. Wrapper techniques consider feature dependencies, provide interaction between feature subset search and choice of the learning model, but are computationally expensive with respect to the filters.

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 08, ISSN No. 2455-2143, Pages 94-96
Published Online December 2023 in IJEAST (http://www.ijeast.com)

III. RELEVANCE OF THE TOPIC

The topic of feature selection in text classification is highly relevant in today's world due to the increasing amount of textual data generated every day. With the rise of social media, e-commerce, and other online platforms, there is a vast amount of unstructured data that needs to be analyzed and classified. Feature selection techniques help to reduce the dimensionality of the feature space, which is a critical concern in text classification problems due to processing time and accuracy considerations. The relevance of this topic can be seen in real-world applications such as spam filtering, sentiment analysis, and news classification. For example, spam filtering is a common problem that requires the classification of emails as either spam or not spam [Zareapoor et.al [1]]. Feature selection techniques can help to improve the accuracy of this classification by selecting the most relevant features from a large set of features. Similarly, sentiment analysis involves classifying text as positive, negative, or neutral, and feature selection techniques can help to improve the accuracy of this classification by selecting the most relevant features.

A systematic literature review [Pintas 2021] assessed 1376 unique papers from journals and conferences published in the past eight years (2013–2020) and found that feature selection methods have received a great deal of attention from the text classification community due to their strength in improving retrieval recall and computational efficiency. The review also identified which datasets, languages, machine learning algorithms, and validation methods have been used to evaluate new and existing techniques. By mapping issues and experiment settings, the review helps researchers to develop and position new studies with respect to the existing literature. The relevance of feature selection in text classification can be seen in the increasing amount of textual data generated every day and the need to analyze and classify this data accurately and efficiently. Real-world applications such as spam filtering and sentiment analysis demonstrate the importance of feature selection techniques in improving classification accuracy. Systematic literature reviews also show that feature selection methods have received a great deal of attention from the text classification community and can help researchers to develop and position new studies with respect to the existing literature

IV. IMPLEMENTATION DETAILS

The hybrid feature selection scheme proposed in the paper by [Gunal 2018]consists of both filter and wrapper selection stages. The hybrid system is implemented in two stages, with the second stage being a wrapper approach that uses a genetic search algorithm (GS) to optimize the feature set obtained from the first stage. The combined feature set, which is obtained during the first stage of the hybrid selection scheme, is fed into the GS in the second stage. Optimal GS parameters are empirically obtained, with a population size of 50, 30 generations, a crossover probability of 0.8, and a mutation probability of 0.08. The fitness values are Micro-F1 and Macro-F1 measures obtained by two different classification algorithms.In the first stage, the features are selected using DF-, MI-, CHI2-, and IG-based filter methods. Next, the features selected by the filters are combined together and fed into the GS in the second stage. This 2-stage hybrid feature selection process is repeated under particular conditions, including different feature set sizes. The chromosomes are encoded with a {0, 1} binary alphabet. In a chromosome, the indices represented with “1” indicate the selected features, whereas “0” indicates the unselected ones. The fitness value corresponding to a chromosome is determined by a particular success measure that is obtained with the selected features.

The study uses two datasets, Reuters and Newsgroups, with different distributions. The highest Micro-F1 score (85.83%) in the Reuters dataset is achieved by the SVM, whereas the highest Macro-F1 score (66.19%) is attained using DT. However, in the Newsgroups dataset, both the highest Micro-F1 (98.48%) and Macro-F1 (98.44%) scores are obtained by the DT classifier.Furthermore, there is a considerable reduction in the size of the feature sets after the selection in all of the cases. While the maximum dimension reduction rate reaches as high as 56% in the Reuters dataset, this value goes up to approximately 54% in the Newsgroups dataset. In spite of this amount of reduction in the feature set size, both the Micro-F1 and Macro-F1 values after hybrid selection are even higher with respect to their corresponding values before the selection. Therefore, the hybrid feature selection scheme can significantly reduce the size of the feature sets while improving the classification performance..

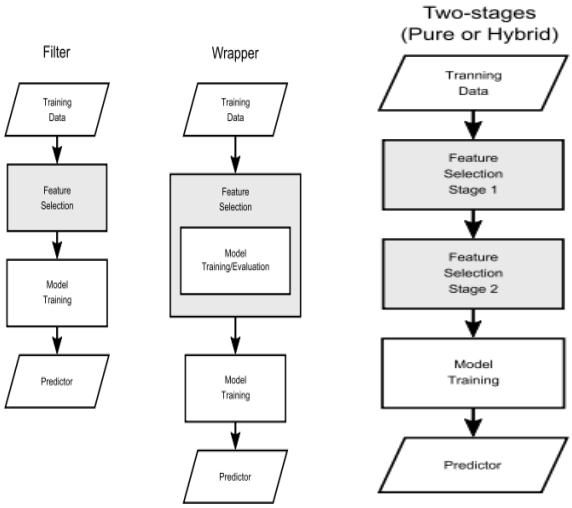

Fig. 1. Block diagram for Filter Wrapper and Hybrid Methods

The filter method and wrapper method can be used independently, or they can be combined to form a hybrid method. In a hybrid method, the filter method is used to select

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 08, ISSN No. 2455-2143, Pages 94-96
Published Online December 2023 in IJEAST (http://www.ijeast.com)

a subset of features, and then the wrapper method is used to further refine the feature set as shown in Figure 1

V. PERFORMANCE ANALYSIS

D. Comparative study

Filter methods are computationally efficient but less accurate (Deng 2019 [2]), wrapper methods are more accurate but computationally expensive, hybrid methods combine filter and wrapper methods to improve performance (Gunal 2012 [4]), and DFS is a filter-based probabilistic feature selection method that offers competitive performance in terms of accuracy, dimension reduction rate, and processing time (Uysal 2012 [5])

E. Performance analysis

An analysis is carried out by [Serkan GUNAL 2018] it is clear that hybrid feature selection models have improved performance over filter and wrapper models. From the graph and models given below it is clear that the novel Hybrid TF-IDF Term Frequency-Inverse Document Frequency
Feature selection method is effective and has the improved performance over filter and wrapper method

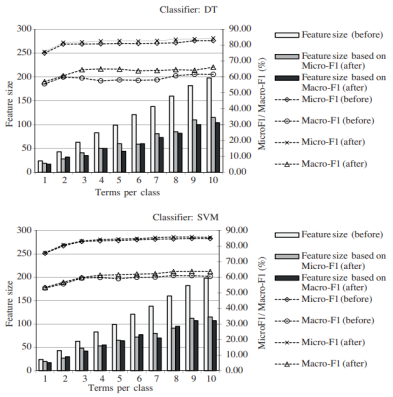

Fig. 2. Success measures and feature set sizes before and after hybrid feature selection in the Reuters dataset

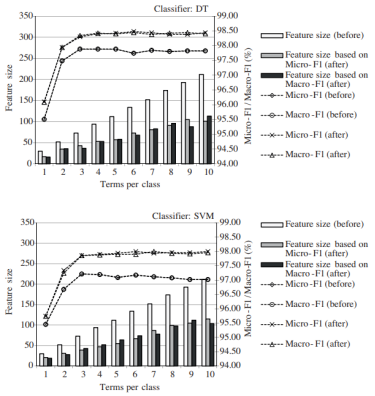

Fig. 3. -Success measures and feature set sizes before and after hybrid feature selection in the Newsgroups dataset

VI. ADVANTAGES/LIMITATIONS OF THE PROPOSED METHOD

F. Advantages

Hybrid methods combine the strengths of different feature selection techniques to improve performance. Hybrid methods can be customized to suit the specific needs of a problem. Wrapper methods can overfit the data because they use the learning algorithm to evaluate the subsets of features. Hybrid methods can reduce overfitting by using filter methods to reduce the feature space before applying the wrapper method. Filter methods are computationally efficient and can handle high-dimensional data. Hybrid methods that use filter methods can reduce the feature space and speed up processing times. Hybrid methods are more robust than multiple methods. This can lead to more reliable results that are less affected by noise or outliers in the data.

G. Limitations

Hybrid methods can be computationally expensive due to the use of wrapper methods, which evaluate feature subsets using a specific learning algorithm and search algorithm. This can lead to longer processing times compared to filter methods, which are computationally efficient and can handle high-dimensional data. Additionally, the effectiveness of a hybrid method can depend on the specific problem and dataset being analyzed, as well as the choice of feature selection techniques and learning algorithms. Therefore, it is important to experiment with different techniques and compare their performance to determine the most effective approach for a given problem.

VII. REFERENCES

[1] Masoumeh Zareapoor, Seeja K. 2015. R,"Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection", IJIEEB, vol.7, no.2, pp.60-65, DOI: 10.5815/ijieeb.2015.02.0

[2] Günal, S., 2012. Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering and Computer Science, 20(Sup. 2), pp.1296-1311.

[3] Deng, X., Li, Y., Weng, J. and Zhang, J., 2019. Feature selection for text classification: A review. Multimedia Tools and Applications, 78, pp.3797-3816

[4] Pintas, J.T., Fernandes, L.A. and Garcia, A.C.B., 2021. Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review, 54(8), pp.6149-6200

[5] Uysal, A.K. and Gunal, S., 2012. A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, pp.226-235