Research Review

A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification

2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18thMarch 2016, Coimbatore, TN, India

Presented by
EL HARRARI ANAS
Department of Artificial Intelligence, Ueuromed University
anas.elharrari@eidia.ueuromed.org

ABSTRACT

The rise of weblogs and social networks has led to many news headlines being shared across various websites and platforms. As a result, there is a growing interest in text-mining techniques that can help classify news into different categories. This study proposes a news classification method that utilizes TF-IDF and SVM algorithms to group news articles by topic and country.

The approach involves three steps: text preprocessing, feature extraction using TF-IDF, and classification using SVM. The method was evaluated using two datasets from BBC and five groups of 20Newsgroup datasets, achieving high classification precisions of 97.84% and 94.93% respectively. These results demonstrate the effectiveness of the proposed approach compared to other classification methods.

In this review , we implement the presented approach and we evaluate its results comparing to other ones

Problem Introduction

In today's digital age, the amount of textual data available is growing at an unprecedented rate. This data is largely unstructured, meaning it lacks a defined format or organization, making it difficult for computers to process and extract meaningful information from it. This is where text mining techniques come in. Text mining involves using specialized tools to analyze large sets of unstructured text data, with the goal of uncovering previously unknown information. This unstructured data can be found in a variety of electronic forms, such as emails, web pages, and electronic publications. However, in order to extract useful patterns and information from this data, specific processing and preprocessing methods are required. By performing indexing and retrieval on this rapidly growing text data, text mining tools allow us to gain valuable insights and knowledge that would otherwise be inaccessible.

One of the key challenges in text mining is dealing with the ambiguity and complexity of natural language. Words can have multiple meanings and connotations, and sentences can be structured in a variety of ways. To address this challenge, text mining techniques often involve preprocessing steps such as tokenization, stemming, and stop word removal. These steps help to standardize the text data and remove irrelevant or redundant information.

Objectives

In this research work, we essentially develop a news customization system based on SVM that recommends favourite articles to users based on their previously defined interests created in their profiles

One of the main objectives of this work is To implement a classifier that combines the nearest neighbour algorithm and the SVM classifier, resulting in a new hybrid method named SVM-NN.
To propose a hybrid SVM-KK method that minimizes the effect of parameters on classification precision, particularly in relation to the effective K parameter in the nearest neighbour algorithm

.we also aim, following the research purposes, to evaluate the performance of the SVM-KK method and compare it with the KNN method in terms of classification precision and the effect of parameter values

Research Methodology

Based on the presented work, the research involved a literature review to identify gaps in the literature and formulate research questions, followed by data collection and preprocessing to standardize the format, remove irrelevant information, and extract relevant features for use in the classifiers. The SVM-NN and SVM-KK classifiers were then trained and tested using the preprocessed data, and their performance was evaluated in terms of classification precision and the effect of parameter values. The results were analyzed and interpreted to draw conclusions and identify areas for future research and improvement

TECHNICAL APPROACH

In this part, we are going to present all the techniques used in this research work and implemented in the implementation attached, the team presenting the search work used specifically RapidMiner Studio which is designed for better data presentation and contains various pre-defined tools for data-pre-processing and visualization

We are going also to present in this part the approach used in this work and all the steps followed to achieve the results

Choosing the right software application is crucial for technical text mining, as it must meet the criteria of reliability, stability, and high computing speed. In this particular study, the researchers opted to use RapidMiner Studio Professional 6.5 to analyze the data.

Text Processing

Text processing is the process of analyzing, manipulating, and transforming textual data using a variety of techniques and methods. This type of data is typically unstructured or semi-structured and can include things like emails, documents, web pages, and social media posts. The goal of text processing is to extract valuable insights and information from this data, making it easier to analyze and draw conclusions. By performing text processing tasks, researchers and analysts can gain a deeper understanding of the data they are working with, and use this information to drive further analysis and decision-making. This research used a lot of common processing techniques that are going to be presented in this part:

Transforming cases and tokenizing

To ensure consistency in text data, the transform case operator is commonly used to convert all characters to lowercase. This is particularly useful for eliminating homologous words that differ only in their case, as it ensures that all instances of a particular word are treated as identical. This operator is often utilized in text-processing tasks and can be implemented in various programming languages using specific functions or methods. In our implementation, as we are working with Python language, we use a refined function to transfer all the characters to lowercase

text mining studies typically involve the separation of words or sentences for more efficient processing. This involves breaking down sentences into individual words and removing any punctuation marks, as they do not represent a meaningful group, which is known as tokenizing (sentence & word tokenizing). By separating text in this way, researchers and analysts can more easily analyze and manipulate the data for various purposes

Filtering Stopwords

Filtering out stopwords in text mining is an important step that provides various advantages, including noise reduction, greater memory and storage efficiency, higher performance, a concentration on content-bearing words, and improved interpretability. By deleting regularly occurring words with little significance, such as articles, prepositions, conjunctions, and pronouns, the analysis may concentrate on the more informative terms that add to the text's context. As a result, text-mining findings are more precise and effective, allowing researchers and analysts to acquire deeper insights and knowledge of the data.

Feature Extraction based on TF-IDF

A term's TF-IDF score in a document is derived by multiplying its TF value by its IDF value. This score measures the term's importance inside the document and the corpus as a whole. Terms with high TF-IDF scores are seen to be relevant or distinctive to the document, suggesting their potential significance to its content. Using Python language, the matrix of tf-idf frequencies can be presented by using a pre-defined model implemented in the implementation attached

Support Vector Machine

Joachims developed the support vector machine (SVM) as a strong classification tool for text classification. It is a supervised learning technique that, during training, constructs a hyperplane to distinguish positive and negative samples and then classifies fresh samples depending on where they fall on the hyperplane. We employed the nu-svc type for classification in this work, together with an RBF kernel type to translate training data nonlinearly to a higher dimensional space. The value of nu was set to its maximum of 0.5.

Previous research that employed SVM for text categorization used all terms in the text without regard for their importance. Other research, on the other hand, concentrated on keyword selection using set- or keyword-class-based approaches, which introduced new issues such as temporal complexity and the lack of standard datasets. Furthermore, SVM was not used as the classifier method in the majority of these investigations. Despite these obstacles, SVM remains a popular and successful text categorization method.

Results.

The research used three main tools for results evaluation; recall, precision and F-measure

Recall, often known as the true positive rate or sensitivity, is a parameter used to assess a classification model's performance, particularly in binary classification problems. It assesses a model's ability to accurately identify positive examples from a dataset's total number of positive occurrences.

Precision is a statistic used to assess a classification model's performance, notably in binary classification problems. It assesses the model's ability to make accurate positive predictions.

The F-measure, commonly known as the F1 score, is a metric that combines accuracy and recall into a single number, resulting in a fair assessment of a model's performance in binary classification tasks. It is a symmetrical mean of accuracy and recall.

The proposed method for news text classification demonstrated high precision levels, as shown in the review article for both the BBC and 20Newsgroup datasets. The precision levels were 97.84% and 94.93% respectively, indicating the high efficiency of the method. This impressive precision can be attributed to the use of the TF-IDF method for feature extraction. Also, the F values obtained for the BBC dataset, with the sports group achieving the highest F value of 99.22%, followed by entertainment with 98.57%. The business had the lowest F value of 96.70%. The F values were mostly close to 1, indicating the high efficiency of the proposed method

These values were resumed in two tables in the review article for a better representation of the results. Regarding our implementation, as we followed the same approach but using the common libraries of python , we were able to achieve the same results with less time complexity

work critiques

The suggested technique for news text categorization and feature extraction utilizing the support vector machine (SVM) algorithm and the term frequency-inverse document frequency (TF-IDF) method shows good efficiency and precision levels. Using SVM with nu-svc and rbf kernel types, as well as setting the parameter nu to its maximum value of 0.5, we were able to successfully classify news texts in both the BBC and 20Newsgroup datasets. Furthermore, the F values obtained for each group in the datasets were typically near 1, confirming the suggested method's excellent efficiency. One of the reasons for the high accuracy levels reached is the adoption of the TF-IDF approach for feature extraction. However, earlier research has revealed difficulties. with keyword selection for text classification, which may affect the effectiveness of the method. Overall, the suggested approach for news text categorization has shown to be successful and efficient.

However, challenges with keyword selection for text classification have been noted in previous studies, which may affect the effectiveness of the method.

Because of the large dimensions of the data involved, text categorization is a difficult process. This study developed a three-step strategy for identifying news texts to overcome this difficulty. The method included text preprocessing, TF-IDF feature extraction, and SVM classification. Because of its capacity to handle high-dimensional data, SVM was chosen as the classifier.

Conclusion

In the provided research and so as in our implementation, we presented a new approach for data mining and text classification using tf-idf feature. This approach shows a great interest by filteringthe data and proceeding the classification by taking in consideration the frequencyand the importance of words inside our text.

This research focuses on categorizing news using a mix of Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM). The proposed method consists of three steps: text preprocessing, TF-IDF feature extraction, and SVM classification. The approach was tested on two BBC datasets as well as five groups of 20Newsgroup datasets, with classification precisions of 97.84% and 94.93%, respectively. When compared to other categorization approaches, these findings are highly positive.

References

- 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18th March 2016, Coimbatore, TN, India.

- https://www.techtarget.com/searchbusinessanalytics/definition/text-mining

- https://scikit-learn.org/stable/modules/svm.html

- G. Krishnalal, S. B. Rengarajan, and K. Srinivasan, "A new text mining approach based on HMM-SVM for web news classification," in International Journal of Computer Applications, 2010, vol. 1, pp. 98-104.

- Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents International Journal of Computer Applications 181(1)