News Classification and Its Techniques

Aman Gaur (10), Sagar Rathore (49), Satyam Sahu (41)
Department of Computer Science University of Delhi

Under the supervision of

Dr. Ankit Rajpal
Assistant Professor
Department of Computer Science
University of Delhi

1. Abstract

Our project is based on News Classification, in which we aimed to classify a multiclass news dataset in a properly explainable way. Essentially, our project utilizes a few techniques to classify news, which we will discuss further. However, this report will focus on outlining the overall workflow of news classification, without including any experimental results.

2. Introduction

In today's digital age, a massive amount of information is stored electronically. Extracting meaningful insights from this data is crucial for decision-making. Machine Learning can help us to get useful information in faster way.

Until recently, accessing news was not straightforward. However, with the rise of online news services, news has become easily accessible. Classification, a challenging field in text mining, involves

converting unstructured data into structured information. Then classifying it with the help of various machine learning model is somewhat challenging.

As the volume of news grows, users struggle to find relevant content quickly. News categorization becomes essential. Grouping articles allows users to navigate and access news of interest in real-time. However, classifying news remains challenging due to the constant influx of information, including entirely new categories.

So, we review various approaches to news classification based on content and headlines.

In order to do news classification, we have studied few different techniques after which some of them we have implemented in our original project.

3. News Classification Process Workflow

There are different steps involved in news classification. Classification is a difficult activity as it requires pre-processing steps to covert the textual data into structured form from the un-structured form. Text classification process involves following main steps for classification of news article.

3.1 News Collection

The first step of news classification is accumulating news from various sources. In our original project we have used a news dataset of BBC News.

3.2 News Pre-Processing

News preprocessing involves converting unstructured text data into a structured format suitable for analysis. This process typically includes steps such as text cleaning, tokenization, stop-word removal, stemming or lemmatization, and possibly feature extraction. Cleaning involves removing irrelevant characters, symbols, and formatting inconsistencies.

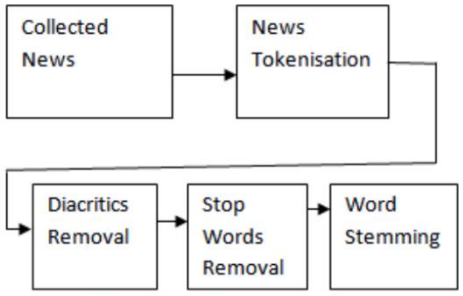

Fig: News Pre-processing

3.2.1 News Tokenisation

News tokenization is the process of breaking down a news article into individual words or tokens. This step is essential for further analysis and classification. Tokenization helps in standardizing the text data and enables the extraction of meaningful information. It involves identifying word boundaries, handling punctuation, and dealing with special characters.

3.2.2 Stop Word Removal

Stop word removal is a crucial preprocessing step in natural language processing, including news analysis. Stop words are common words like "the," "is," "and," etc., that occur frequently in a language but carry little semantic meaning. Removing stop words helps reduce noise and improve the efficiency of text analysis algorithms by focusing on words that convey significant information.

3.2.3 Word Stemming

Word stemming is a process in natural language processing that involves reducing words to their base or root form, known as the stem. This helps in standardizing variations of words and reducing the vocabulary size, which can improve the performance of text analysis algorithms. For example, words like "running," "ran," and "runner" would all be stemmed to the root "run." Stemming simplifies the analysis of text by treating different forms of the same word as equivalent, which is especially beneficial for tasks like sentiment analysis, text classification, and information retrieval.

3.3 Feature Selection

Feature selection in news classification involves choosing the most relevant and informative attributes or features from the raw data to improve the performance of classification algorithms. This process helps in reducing dimensionality, decreasing computational complexity, and enhancing the model's generalization ability. Features can include words, phrases, or other characteristics extracted from news articles, such as word frequencies, sentiment scores, or topic representations.

3.3.1 Boolean Weighting

Boolean weighting assigns a binary value to each term in the document-term matrix based on its presence or absence. In the context of news classification, Boolean weighting can be used to represent whether a specific word occurs in a news article or not. While simple, Boolean weighting may not capture the importance or relevance of terms accurately, as it treats all terms equally regardless of their frequency or significance.

3.3.2 Information Gain

Information gain measures the importance of a term in distinguishing between different classes or categories of news articles. It quantifies the reduction in uncertainty about the class label of a document when the presence or absence of a term is known. In news classification, information gain helps identify terms that are most informative for discriminating between different topics or themes in articles, aiding in feature selection and model building.

3.3.3 Term Frequency – Inverse Class Frequency

TF-ICF combines the concepts of term frequency (TF) and inverse class frequency (ICF) to assign weights to terms based on their frequency in individual documents and their rarity across all documents in a particular class or category. In news classification, TF-ICF emphasizes terms that are frequent within a specific class but rare across other classes, potentially improving the discriminative power of features and enhancing the accuracy of classification models.

3.3.4 Class Frequency Thresholding

Class frequency thresholding involves setting a threshold on the frequency of terms within each class, beyond which the terms are considered significant for classification. Terms that exceed this threshold are retained as features, while others are discarded. In the context of news classification, class frequency thresholding helps filter out terms that are too common within a class and may not contribute much to distinguishing between different categories of news articles.

3.4 News Classification

After feature selection the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories. The most common news classification methods are Naive Bayes, Artificial Neural Networks, and Decision Trees, Support Vector Machines, Support Vector Machines, K-Nearest Neighbours.

3.4.1 Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that the presence of a particular feature in a class is independent of the presence of other features, hence the term "naïve." In news classification, Naïve Bayes is computationally efficient and can handle a large number of features well. However, it relies on the assumption of feature independence, which may not always hold true in real-world datasets, potentially affecting its accuracy.

3.4.2 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates different classes in the feature space. SVM is effective in handling high-dimensional data and can capture complex relationships between features. In news classification, SVM can achieve high accuracy, especially when dealing with linearly separable classes. However, SVMs can be computationally intensive, particularly with large datasets, and may require careful tuning of parameters.

3.4.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes organized in layers, including input, hidden, and output layers. ANNs can learn complex patterns and relationships from data, making them suitable for news classification tasks involving non-linear relationships. However, ANNs often require large amounts of data for training and are susceptible to overfitting, especially in cases of limited data availability.

3.4.4 Decision Tree

Decision Tree is a tree-like structure where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents a class label. Decision trees are interpretable and easy to understand, making them suitable for news classification tasks where transparency and interpretability are important. However, decision trees are prone to overfitting, especially when dealing with noisy or imbalanced data, and may not capture complex decision boundaries effectively.

3.4.5 K-Nearest Neighbour

K-Nearest Neighbors (KNN) is a simple yet effective classification algorithm that assigns a class label to a new data point based on the majority class among its k nearest neighbors in the feature space. KNN is non-parametric and does not make strong assumptions about the underlying data distribution, making it flexible and versatile. In news classification, KNN can be robust to noisy data and can handle multi-class classification tasks effectively. However, KNN's performance may degrade with high-dimensional data and large datasets due to computational complexity and the curse of dimensionality.

4. **Conclusion**

In conclusion, the implementation of a robust news classification workflow is pivotal in efficiently organizing and analyzing vast amounts of news data. Through the utilization of advanced machine learning algorithms and natural language processing techniques, such a workflow enables the automatic categorization of news articles into relevant topics or themes, facilitating streamlined access to information for users. In case of out BBC news dataset most of Machine Learning Models working well after a good cleaning of data.

References

3.1 Ramasubramanian, C., and R. Ramya. "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm.", International Journal of Advanced Research in Computer and Communication Engineering 2(12),2013

3.2 Gurmeet Kaur, Karan Bajaj. "News Classification and Its Techniques – A Review", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. III (Jan – Feb. 2016), PP 22-26

3.3 Mazhar Iqbal Rana, Shehzad Khalid, Muhammad Usman Akbar. "News Classification Based On Their Headlines: A Review" 17th IEEE International Conference on Multi-Topic Conference (INMIC), Karachi,Pakistan,,2014,211-216