WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Categorization using Machine Learning

Harsh Yadav, Dr. Vibhakar Phathak, Dr. Akhil Pandey[3]

[1]B.TECH. Scholar, [2,3]Professor, [4]Assistant Professor
Department of Information Technology, Arya College of Engineering & I.T. Jaipur, India
harshyadav2823@gmail.com, **[2]vibharkar@aryacollege.in,[3]akhil@aryacollege.in**

ABSTRACT –

With an explosion of digital content has become a major challenge. Machine Learning (ML) has a revolution in text categorization, allowing computers to automatically classify documents. This article examines different ml algorithms - such as naive Bayes, supported vector machines (SVM), artificial neural networks (Ann) and decision -making trees - and evaluate their strengths and weaknesses. We are also discussing key techniques such as selection of elements, vector cosmic models and dimension reduction to increase accuracy and efficiency.

Keywords: Text Categorization, Machine Learning, Naive Bayes, Support Vector Machines, Neural Networks, Feature Selection.

1. Introduction

With rapid growth in digital content on the web, the categorization of the text has appeared as a necessary tool for managing and processing large volumes of text data. The categorization of the text involves assigning predefined labels or categories to the document based on its content. This task contains applications in many areas such as information search, E -mail filtering, sentiment analysis and document indexing.

Machine learning techniques, especially the supervision method, have shown a remarkable efficiency in automating text categorization. These methods learn from the marked data set of training to develop models that predict the category of invisible documents. However, challenges of these models often affect challenges, such as high -dimensional functions and unbalanced data sets. This article provides an overview of the most modern machine learning techniques in categorizing text and emphasizing their advantages, restrictions and application in the real world.

2. Background and Technology

2.1 Naive Bayes Classifier

With rapid growth in digital content on the web, the categorization of the text has appeared as a necessary tool for managing and processing large volumes of text data. The categorization of the text involves assigning predefined labels or categories to the document based on its content. This task contains applications in many areas such as information search, E -mail filtering, sentiment analysis and document indexing. Machine learning techniques, especially the supervision method, have shown a remarkable efficiency in automating text categorization.

These methods learn from the marked data set of training to develop models that predict the category of invisible documents. However, challenges of these models often affect challenges, such as high -dimensional functions and unbalanced data sets. This article provides an overview of the most modern machine learning techniques in categorizing text and emphasizing their advantages, restrictions and application in the real world.

2.2 Support Vector Machines (SVM)

Support vector machines (SVM) are widely considered to be one of the most effective algorithms for text categorization. SVM works by finding Hyperplane, which best separates data points belonging to different categories in high -dimensional features. The power of SVM consists of its ability to handle high -dimensional data, so it is particularly suitable for text classification tasks with large areas of functions.

Recent studies have shown that SVM overcomes other algorithms such as naive Bayes in the categorization of text, especially in combination with techniques such as latent semantic indexing (LSI) to reduce its size. However, SVM requires careful tuning of its parameters, such as the selection of core and regularization, which can be more expensive.

2.3 Artificial Neural Networks (ANN)

Artificial neural networks (Ann), especially deep learning models, have gained significant attention to categorizing the text. One common approach is the use of a neural network of backward promotion (BPN), a type of neural network for forward. BPN is trained by weight adjustment to minimize the error between the expected and actual outputs.

Although the Anns is powerful, they require a large amount of training data and are susceptible to overfilling, especially with limited marked data. Techniques are used to deal with these challenges, such as premature completion of the study and premature stop to avoid excessive connection and improvement of generalization.

2.4 Decision Trees

Artificial neural networks (Ann), especially deep learning models, have gained significant attention to categorizing the text. One common approach is the use of a neural network of backward promotion (BPN), a type of neural network for forward. BPN is trained by weight adjustment to minimize the error between the expected and actual outputs. Although the Anns is powerful, they require a large amount of training data and are susceptible to overfilling, especially with limited marked data. Techniques are used to deal with these challenges, such as premature completion of the study and premature stop to avoid excessive connection and improvement of generalization.

3. Feature Selection

Selection of functions The selection of functions includes the selection of the subset of the most informative functions from the original data set, thereby reducing the size. Common methods for selecting functions include mutual information, chi-quadrate tests and information gain. The aim is to maintain the most important functions while discarding irrelevant or unnecessary, which improves the performance of the model and reduces computing costs.

4. Challenges in Text Categorization

1. **High Dimensionality**: Text data often has tens of thousands of functions, which can lead to computational inefficiency and excessive amount. Techniques such as choosing elements and decreasing dimensions help relieve this problem, but require careful tuning.
2. **Imbalanced Datasets**: In many text categorization tasks, the number of samples in each category can be highly imbalanced. This can lead to biased classifiers that perform poorly on the underrepresented categories. Solutions like cost-sensitive learning, resampling, and ensemble methods are often employed to address this issue .
3. **Context and Semantics**: Machine learning models often struggle with understanding the context and semantics of words, which can affect the accuracy of text categorization. Advanced models such as deep learning and transformers (e.g., BERT) show promise in capturing contextual relationships between words.

5. Conclusion

Machine learning techniques revolutionized the field of text categorization and provided powerful tools to automate the document classification process. While algorithms like Naive Bayes, SVM, Ann and decision -making trees showed considerable success, challenges such as high dimensionality, unbalanced data sets and context awareness, still persist. Future research is likely to focus on increasing the interpretability of the model, improving performance in unbalanced environments and the development of more sophisticated techniques to capture semantic meaning of the text. Incorporating the selection of elements, reducing dimension and advanced models such as Deep Learning, the accuracy and efficiency of text categorization systems can be significantly improved, preparing a way for more robust and scalable applications across different domains.

References

1. Dasari, D. B., & Rao, V. G. (2012). Text Categorization and Machine Learning Methods: Current State of the Art. *Global Journal of Computer Science and Technology*.
2. Mahalakshmi, B., & Duraiswamy, K. (2012). Categorization of Text Using Machine Learning Algorithms. *International Journal of Modern Engineering Research*.
3. Yadav, B. P., Ghate, S., & Kumar, K. S. (2020). Text Categorization Performance Examination Using Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*.