TEXT CLASIFICATION USING TF-IDF

Michael H. Shimanyula

University of San Diego

Masters in Applied Artificial Intelligence

Abstract

This study aimed to classify news articles into distinct categories, leveraging Natural Language Processing and Machine Learning techniques. Specifically, we investigated the efficiency of TF-IDF vectorization and experimented with various machine learning classifiers, including Support Vector Machines (SVM), to determine the optimal model.

Keywords: Text Classification, TF-IDF, SVM, News Articles, NLP.

Methodology

Data Preprocessing

News articles underwent preprocessing, which included the removal of stopwords, punctuation, and the conversion of text to lowercase.

TF-IDF Vectorization

Scikit-learn's TfidfVectorizer was employed to convert the news articles into vectors. TF-IDF stands for Term Frequency-Inverse Document Frequency, a metric that quantifies the importance of a term within a document relative to a corpus.

Model Selection and Training

Several classifiers, including SVM, Naive Bayes, and Random Forest, were experimented with. Hyperparameters were fine-tuned using training data and tools like GridSearchCV.

Results

The SVM model, after tuning, demonstrated superior performance, achieving an accuracy of 98.4%. Furthermore, precision, recall, and F1-score were correspondingly high, suggesting a well-balanced model. The AUC-ROC score, used to evaluate the performance of the multi-class classification, was close to 1, indicating an excellent separability between the classes**.**Visualizations, such as PCA and t-SNE, were plotted to provide insights into data distribution and class separability in the vector space.

Conclusion

Automated classification of news articles is vital for modern-day content management. This study demonstrates the effectiveness of SVM classifiers combined with TF-IDF vectorization in accomplishing this task.

References

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Practical Natural Language Processing: A Comprehensive Guide to Building Real-world NLP Systems. O’Reilly Media. https://books.google.com/books?id=G40jywEACAAJ