UEMF – EIDIA

# Mini Project NLP Report
# News classification using SVM


Supervised by : **PR. ARALANE ZARGHILI**

Made by : **KHADIJA BAYOUD**
**MOUAD NECHCHAD**


**11/06/2023**

# Table des Matières

# Introduction

As part of our Artificial Intelligence Engineering studies, we had the opportunity to delve into Natural Language Processing (NLP) during a dedicated course. Our professor, Aralane Zarghili, assigned us a project centered around a scientific paper that introduces a classification system for news texts. Our task is to implement the methodology outlined in the paper.

A text classification system is a technology or approach used to automatically categorize or label text documents into predefined categories or classes. It is a fundamental task in natural language processing (NLP) and is commonly used in various applications, such as spam filtering, sentiment analysis, topic detection, and content categorization.
The goal of a text classification system is to accurately assign the most appropriate category or label to a given text document based on its content. The system typically relies on machine learning algorithms to learn patterns and relationships between the textual features and the corresponding categories from a training dataset. These algorithms can be supervised, unsupervised, or semi-supervised, depending on the availability of labeled data for training.

This mini-project served as an excellent opportunity to enhance our understanding and apply the concepts learned in both NLP and Machine Learning courses. It equipped us with the essential tools needed to develop more sophisticated and intelligent systems of a similar nature.

# Paper Problem and Objectives

The present paper, titled "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification," addresses the challenge of classifying news articles into distinct groups. The abundance of news headlines across various weblogs and websites has led to an exponential increase in their numbers. Consequently, users face the need to identify the most popular news group in their desired country at any given time.

To address this issue, the paper proposes a news classification method based on TF-IDF (Term Frequency-Inverse Document Frequency). The primary objective of this proposed method is to develop a highly precise system that automatically categorizes news articles into different groups. By doing so, it provides users with convenient access to desired and relevant information without the need to read through all the news articles.

# State of the Art that the present work is based on

As we know, the number of news articles is increasing every day on websites created by many newspapers and magazines which are targetting a huge number of humans. At this point, we note that these websites' users are looking for a method to retrieve quickly the relevant information from all available news. To achieve this goal, many scientific researches have been conducted to develop such a text classification system.

In the following, we cite several news classification methods proposed in the literature:

- Financial News Classification: used to classify news stories to estimate the direction of a stock market index.
- Classification of Short Texts: used for multi-value classification of very short texts.
- Automatic News Headlines Classification: used to classify automatically online news headlines.
- Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-Nearest Neighbor and Support Vector Machine.
- Classification of News Headlines for Providing User-Centered E-Newspaper.
- Emotions Extraction from News Headlines: used to classify the emotions extracted from news headlines using SVM.
- Short News Headlines Classification of Twitter: used for Twitter news classification using SVM.
- Automated document classification for news articles in Bahasa Indonesia based on TF-IDF.
- The Bayesian algorithm: An effective algorithm for improving the performance of Naïve Bayes for text classification.
- Supervised Feature Selection Approach: Effective text classification based on developing the similarity between a word and a class.

# The research methodology followed in the paper

The research methodology described in this paper is comprised of the following steps:

**1- Text Preprocessing**: This is an important step before moving to feature extraction and training model. It aims to clean the data from distortion and useless information such as punctuations, exclamations, semicolons, irrelevant sentences, quotations, dates, etc.

This preprocessing consists of:

- *Transforming Cases*: to transform all existing characters in lowercase.
- *Tokenizing*: to split sentences into words. Plus, we remove punctuation.
- *Filtering Stopwords*: delete unimportant words or words with no specific meaning such: a, an, the…

**2- Feature Extraction**: This step consists of calculating the weight of each word in the text using **TF-IDF algorithm.**

**3- Classification**: This step is implemented to classify the news articles into different groups using **SVM**.

**4- Performance Measures**: We measure our classification performance using the standard **recall, precision and F-measure** aggregated over all classes.

# The used techniques

The paper contains two techniques used in the proposed method for News Classification:

1- TF-IDF(Term Frequency-Inverse Document Frequency): This technique is used for the feature selection step. It aims to represent the relevance of a word in a document within a large corpus. It helps identify important terms or features that distinguish one document from another within a collection.

2- SVM(Support Vector Machine): This algorithm is used for the classification step.

# Our Implementation Results

After training the SVM model on BBC and 20Newsgroup datasets, we have observed high test accuracy scores for both datasets when applying the SVM model. These scores affirm the effectiveness of the method proposed in the paper for news text classification.

Table 1 shows the accuracy scores achieved for the classification of news texts using techniques presented in the paper.

| Method | BBC (%) | 20 newsgroup (%) |
|---|---|---|
| Our implementation of the paper approach | 97.3 | 94.95 |

**Table 1**

In Table 2, the F-measure values obtained for the BBC dataset indicate that Sport, Entertainment, and Tech were the most accurately classified groups, achieving a score of 99%. Following closely was Politics with a score of 96%. The lowest F-measure value was observed for Business, with a score of 95%.

| Group Name | F-measure |
|---|---|
| Entertainement | 0.99 |
| Tech | 0.99 |
| Sports | 0.99 |
| Politics | 0.96 |
| Business | 0.95 |

**Table 2**

# Our Implementation Results

Moving to **Table 3**, the **F-measure** values obtained for the **20Newsgroup dataset** reveal that Baseball achieved the highest accuracy among the groups, with a score of 0.98. Conversely, Forsale had the lowest F-measure score of 0.92.

| Group Name | F-measure |
|---|---|
| Sport.baseball | 0.98 |
| Politics.guns | 0.96 |
| Religion.christian | 0.95 |
| Graphics | 0.93 |
| Forsale | 0.92 |

**Table 3**

# Comparison with the results presented in paper

In this section, we will compare the results obtained from our implementation with those presented in the paper.

Regarding the **accuracy scores**, we have achieved similar results to the paper for both the BBC and 20Newsgroup datasets, approximately **97%** and **94.93%**, respectively. These matching accuracy values indicate the realism of our implementation. Moreover, the higher accuracy values reinforce the efficiency of the text classification system.

However, differences are observed in the **F-measure** values between our implementation and the paper (refer to table2 and table 3 ). Notably, variations are evident in the ranking of groups within their respective anf table classifications:

**For the BBC dataset:** In our implementation, Entertainment, Sport, and Tech are the top-performing groups with an F-measure of 99%, while Business exhibits the lowest F-measure of 95%. Contrarily, the paper only identifies Sport as the best group, and Business as the lowest, albeit with different scores (95% and 96%).

**For the 20Newsgroup dataset:** In our implementation, Baseball group achieves the highest F-measure of 98%, whereas Forsale group obtains the lowest F-measure of 92%. Conversely, the paper identifies Politics as the best-performing group with an F-measure of 97.34% and Graphics as the lowest with an F-measure of 91.65%.

These disparities highlight the variations between our implementation and the paper in terms of F-measure values and the ranking of groups in their respective classifications.

# Critical Evaluation of the Presented Work

Despite the good performance obtained in classifying the news contained in the two datasets BBc and 20Newsgroup into five groups, the proposed method present some limitations that can be improved through different methods. These limitations could be summarized in these following points:

1. Limited coverage of topics: This study mentioned that the researchers have used two datasets that include just five topics for each one. Although, other domains of news articles are not present. Consequently, we do not have an idea of how the model will be efficient to classify news articles from other categories.

2. Generalization to different languages: The news collected should be in other different languages rather than English in order that the model will be able to classify well news in other languages.

3. Dependency on TF-IDF feature extraction: TF-IDF is a widely used technique, but it has not the ability to represent the relationships between words and the meaning of complex sentences.

4. Limited consideration of context: The proposed method focuses on individual news articles and utilizes features extracted from them. However, news articles often exist within a broader context, and their accurate classification may require considering the relationships and dependencies between different articles or topics. Ignoring contextual information may result in misclassification or incomplete understanding of news articles.

5. Lack of comparative analysis: In this paper, they do not present other studies results to compare with that obtained.

# Conclusion

To conclude this mini-project, we are very proud to have successfully completed our work, which involves developing news articles classification system using two datasets: BBC and 20Newsgroup. To build this system, we have gone through different steps, such as: Text Preprocessing operations, Feature Extraction based on TF-IDF, Applying SVM model, Performance Measures.

At the end, we have learned interesting information during this mini-project about the field of classifying news articles, and we hope that this opportunity will be our key to success in future similar projects.