

Report title

First and last name

Politecnico di Torino

Student id: s123456

email@studenti.polito.it

Abstract—In this report we address the problem of automatic classification of online news articles into predefined thematic categories. The proposed approach combines textual information extracted from titles and article bodies with selected metadata, using a TF-IDF representation and a linear Support Vector Machine classifier. A preprocessing strategy is designed to handle class imbalance, heterogeneous article lengths, and feature relevance. Experimental results, evaluated using the Macro F1-score, show that the proposed pipeline achieves robust performance across multiple news categories.

I. PROBLEM OVERVIEW

The proposed task is a multi-class classification problem on a large-scale dataset of online news articles. Each article must be automatically assigned to one of seven predefined thematic categories. The dataset is divided into two parts:

- a development set, containing 79,997 articles with associated labels;
- an evaluation set, consisting of 20,000 unlabeled articles.

The development set is used to train and validate the classification models, while predictions on the evaluation set are submitted for final assessment.

An initial exploration of the development set reveals that the target labels are not uniformly distributed. Some categories are significantly more represented than others, resulting in a class imbalance that must be taken into account during model training and evaluation. Since the official evaluation metric is the Macro F1-score, balanced performance across all classes is required.

The dataset includes both textual content and metadata. The textual component is composed of the article title and body, whose lengths vary considerably across samples. While most articles have comparable sizes, a small number of outliers exhibit extremely long textual content. Additionally, a correlation between article length and target label can be observed, suggesting that different news categories tend to follow distinct editorial styles. Furthermore, many articles contain hyperlinks, image references, or residual HTML elements, which introduce noise into the textual signal and require careful preprocessing.

Regarding metadata, the *source* attribute shows a strong correlation with the target label, as many publishers predominantly focus on specific types of news. On the other hand, the *page_rank* feature exhibits very limited variability across samples, making it only weakly informative for the classification task. Finally, the *timestamp* attribute is characterized by a substantial proportion of missing values and does not

provide significant discriminative power with respect to the article topic and is therefore not considered a relevant feature.

These observations motivate the design of a classification pipeline that primarily leverages textual information, while selectively incorporating metadata and addressing issues related to noise, class imbalance, and feature relevance.

II. PROPOSED APPROACH

A. Data Preprocessing

The preprocessing phase focuses on the construction of a meaningful textual representation and the selection of informative metadata. The main source of information is obtained by combining the article title and the full text. Specifically, the title is given higher importance by duplicating its contribution in the final textual field, in order to emphasize its summarizing role. For articles belonging to the *General News* category, a different weighting scheme is applied, assigning a larger weight to the article body to better capture its broader content. All textual data are cleaned by unescaping HTML entities and processed using standard text normalization steps. The resulting text is transformed into a numerical representation through a TF-IDF vectorization, with stop-word removal and configurable n-gram ranges.

In addition to textual content, selected metadata are incorporated into the model. The *source* attribute is encoded using one-hot encoding, as it exhibits a strong correlation with the target label due to publisher specialization. Although characterized by low variability, the *page_rank* feature is also encoded categorically and included in the pipeline. The *timestamp* attribute is discarded, as it does not provide significant information for the classification task. Class imbalance is addressed during training by adopting balanced class weights in the classifier.

B. Model selection

The classification model is based on a linear Support Vector Machine. This choice is motivated by its effectiveness in high-dimensional sparse spaces, which naturally arise from TF-IDF representations of text. Moreover, linear SVMs offer a good trade-off between computational efficiency and classification performance for large-scale text classification problems.

The model is implemented within a unified pipeline that combines preprocessing and classification steps, ensuring consistency between training and evaluation phases. A stratified train-test split is employed to preserve the original class distribution in both sets.

C. Hyperparameters tuning

Hyperparameters are optimized using a randomized search strategy on the development set. The search space includes parameters related to the TF-IDF vectorization, such as vocabulary size, n-gram range, term frequency scaling, and document frequency thresholds. Classifier-specific hyperparameters, including the regularization strength, penalty type, and class weighting scheme, are also explored.

The optimization process is carried out using cross-validation and Macro F1-score as the selection criterion, in order to favor balanced performance across all news categories. The final model configuration is selected as the one achieving the best cross-validated Macro F1-score.

III. RESULTS

Here you will present your results (models & configurations selected, performance achieved)

IV. DISCUSSION

Any relevant discussion goes here.

REFERENCES