Proceedings of the International Conference IMCS-60, MSU
October 10–13, 2024, Chisinau, Republic of Moldova

A Comparative Analysis of Machine Learning Algorithms for Text Analysis

Alexandr Parahonco, Mircea Petic

Abstract

This article proposes a system of metrics for estimating text fetched from the Internet and selecting the one that should be further summarized. The research examines algorithms and software for determining user preferences, employing natural language processing (NLP) and supervised learning classification methods. An empirical assessment is conducted across academic, security, and non-security domains. The paper concludes with insights on the experimental results and the potential future of the implemented metric system.

Keywords: texts metrics, NLP, classification, supervised learning.

1 Introduction

Natural language processing and Machine learning algorithms are widely used in text analysis. These algorithms are able to analyze and interpret large amounts of text data, allowing for insights to be gained from unstructured information. Additionally, they can be used for tasks such as sentiment analysis, topic modeling, and text classification [1].

Learning is a fundamental human activity that entails the acquisition or change of behaviors, competences, and preferences. Diverse theories, including behaviorism and cognitive theory examine the mechanisms of individual learning. In contrast to humans, machines depend

©2024 by Alexandr Parahonco, Mircea Petic

on data for their learning processes. Machine learning (ML), a subset of artificial intelligence, enables computers to autonomously learn and adapt. Machine learning emphasizes empowering computers to adjust their activities to enhance precision, commonly assessed by the rate of accurate results [2].

ML algorithms can be trained on large datasets to recognize patterns and make predictions without being explicitly programmed. This ability allows machines to continuously improve their performance and efficiency over time. It is about programming computers to optimize performance using example data or past experience. The goal is to automatically detect patterns in data and use them to make predictions. Machine learning is also applied to databases, statistics, engineering, and signal processing, among other fields [3].

In our research, we aim to develop metrics for decoding text meaning, extracting valuable information, and identifying the most relevant source to recommend to the user. Our findings are presented in Fig. 1.

Regarding the **structure**, our metric system is made up of four main parts:

- Quality and depth of content
- Context of metrics
- Relevance of the topic
- Authority

The components illustrated in Fig. 1 employ parameters to evaluate content based on its characteristics. Given the complexity of these features, each parameter may encompass multiple sub-parameters representing more straightforward attributes. For instance, text difficulty can be quantified through factors such as lexical choice, stylistic elements, syntactic intricacy, and textual coherence. Concurrently, readability can be gauged by considering the target audience's age demographic and overall comprehension ease. In this paper, we will implement the module "Context of the Sources".

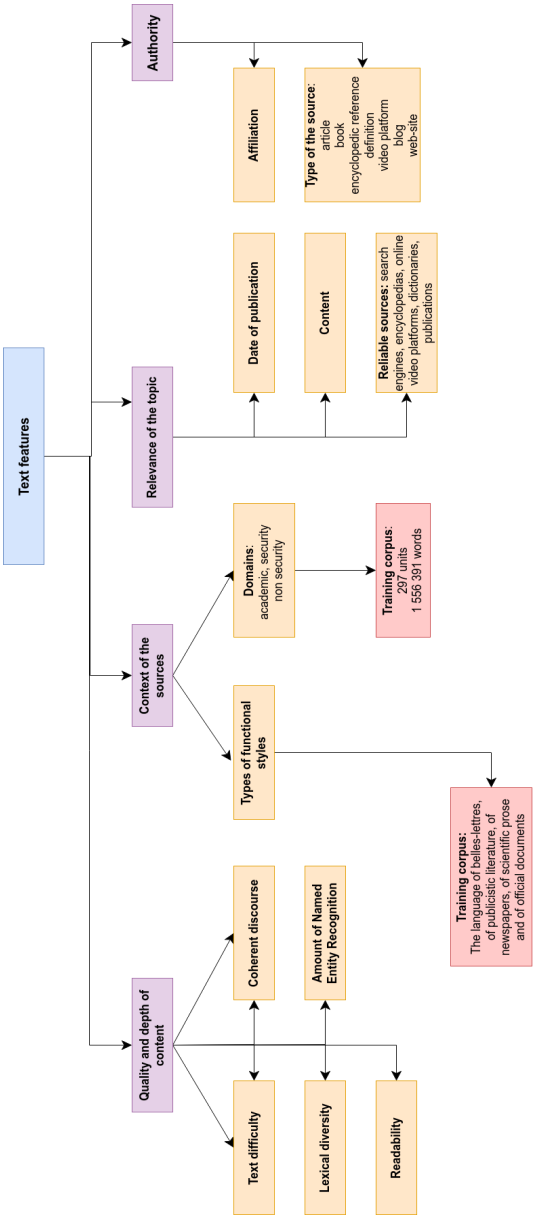


Figure 1. The metric system architecture

Our study focuses on three distinct domains: **academic, security, and non-security** (encompassing areas outside the security field). While our primary aim was to differentiate between academic and non-academic texts, we anticipate potential future expansion to accommodate various university faculties. To assess the adaptability of our system, we opted to incorporate an additional domain: security. This inclusion allows us to evaluate the system's capacity to handle diverse subject areas and potentially scale to meet evolving educational requirements.

2 Description of the experiment

Our research methodology involves primary, quantitative experimental analysis, structured in two distinct phases:

- Phase 1: text preprocessing, feature extraction procedure with the subsequent splitting the corpus into two parts: for training and evaluation.
- Phase 2: comparison of the classification models.

Based on our literature review, the most prevalent and effective supervised learning classification models are **K Nearest Neighbors, Decision Tree, and SGD Classifier**. In addition to these algorithms, our study examines the efficacy of **counting vectors** and the term **frequency-inverse document frequency** (TF-IDF) approach for feature extraction.

For our research, we have compiled a **corpus** comprising **297 documents** with a total word count of 1,556,391. The corpus is segmented into three domains:

1. Academic: 920,032 words
2. Security: 136,359 words
3. Non-security: 500,000 words

This dataset will be partitioned into two subsets: one for training our models and another for subsequent evaluation. This division allows us to assess the performance of our classification algorithms on previously unseen data, ensuring the robustness of our results.

3 Discussion and results

The results indicate that the TF-IDF feature extraction model outperforms Counting vectors. As shown in Table 1, 11 classification models achieved 100% prediction accuracy with TF-IDF, while only 7 models reached this level with Counting vectors.

Among the examined models (K-Nearest Neighbors, Decision Tree, and SGD Classifier), the SGD Classifier with TF-IDF feature extraction consistently demonstrated the highest accuracy. Counting vectors, on the other hand, proved to be less effective and reliable.

Table 1. The comparison of the feature extraction models

$Model$	$Accuracy100\%$	$Accuracy100-95\%$
$TF-IDF$	11	19
$Countingvectors$	7	17

The results obtained from our experiments validate the quality and suitability of the developed corpus for implementing our metric system for text classification. Specifically, the corpus effectively distinguishes between academic, security, and non-security domains. To further refine our findings, we recommend conducting a final validation using manually selected texts from diverse sources. This step will provide a more precise assessment of our results.

4 Conclusion

In conclusion, we have successfully developed a metric system and conducted initial testing of its "Context of the Sources" module. Our experiment involved two stages: text preprocessing and feature extraction, followed by corpus splitting for training and evaluation. By comparing various classification models, we determined that the SGD classifier using TF-IDF features is the most effective for domain identification.

Acknowledgments. This article was written within the framework of the research project 011301 "Information systems based on Artificial Intelligence".

References

[1] I. H. Sarker. *Machine Learning: Algorithms, Real-World Applications and Research Directions*. SN COMPUT. SCI., vol. 2, Article No. 160, 2021. https://doi.org/10.1007/s42979-021-00592-x.

[2] I. El Naqa and M. J. Murphy. *What Is Machine Learning?*. In: I. El Naqa, R. Li, and M. Murphy, (eds) Machine Learning in Radiation Oncology. Springer, Cham., 2015. https://doi.org/10.1007/978-3-319-18305-3_1.

[3] E. Alpaydin. *Machine Learning, revised and updated edition*. MIT Press. 2021.

Alexandr Parahonco[1], Mircea Petic[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University; Alecu Russo Balti State University
E–mail: alexandr.parahonco@math.usm.md
ORCID: https://orcid.org/0009-0007-3486-5597

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University
E–mail: mircea.petic@math.usm.md
ORCID: https://orcid.org/0000-0001-6044-7646