MLKD 2024
The First International Conference on
Machine Learning and Knowledge Discovery
Amirkabir University of Technology, December 18-19, 2024

Impact of Feature Types on Boundary Detection Between Metadata and Body in Persian Theses

Nima Shadman[*] Jalal A. Nasiri[†]

Abstract

This paper addresses the challenge of distinguishing header metadata from body content in Persian electronic theses and dissertations. Accurate classification of these sections aids tasks such as metadata extraction from scientific documents and plays a crucial role in increasing the efficiency and retrieval of information in digital libraries. Several machine learning models were employed to achieve this goal. Additionally, five distinct feature types were utilized: Heuristic, Sequential, Lexical, Formatting, and Geometric. The dataset consisted of nearly 230,000 paragraphs extracted from 106 Persian ETDs, with the metadata class representing only 8.6%. After preprocessing, Random Forest slightly outperformed SVM and Naïve Bayes. Moreover, our findings indicate that features of sequential type notably impact the classification metrics.

Keywords: Paragraph Classification, Metadata Extraction, Persian Scientific Documents, Features Fusion

[*] Natural Language Processing Lab, Ferdowsi University of Mashhad, `nimashadman@alumni.um.ac.ir`

[†] Department of Applied Mathematics, Ferdowsi University of Mashhad, `jnasiri@um.ac.ir`