IEEE Research Articles Classification using Machine Learning and NLP for Efficient Search

7COM1039 – 0109 – 2023 - Advanced Computer Science Masters Project

Student Name: Devanand Kalluvettukuzhiyil Satish Kumar

Student ID: 21080194

Problem statement

The growing quantity and variety of research publications in many fields provide a notable difficulty in effectively organizing and finding pertinent information. Researchers, practitioners, and students have challenges in organizing and accessing academic material due to the absence of an automated and precise categorization system (Shi, et al, 2021). This issue is made more difficult by the fact that multidisciplinary studies are constantly changing, making it difficult to correctly classify papers using just manual, conventional approaches. The lack of a reliable automated categorization system hinders researchers from efficiently navigating the extensive knowledge domain, resulting in inefficiencies in literature reviews, information retrieval, and multidisciplinary cooperation (Charbuty, et al, 2021). To tackle this problem, advanced ML and NLP techniques need to be developed and applied to automate the accurate classification of research articles. This will help streamline the exploration and use of scholarly content in various fields of study.

Aim

This research aims to revolutionize the accessibility and user experience of IEEE research papers by developing a novel, efficient, and user-friendly classification system that leverages machine learning and natural language processing (NLP) techniques. The proposed system combines feature extraction (TF-IDF, GloVe and n-gram) to capture the semantic meaning of keywords and abstracts, with clustering algorithms (K-means, DBSCAN, and hierarchical clustering) to group papers with similar thematic content. This approach seeks to overcome the limitations of existing keyword-based methods and offer accurate classification with minimal user input.

Research Questions

- How does the combination of TF-IDF, GloVe, and n-gram features compare to utilizing each technique individually for capturing the semantic meaning of keywords and abstracts?
- How do the clustering algorithms compare in terms of accuracy (e.g., silhouette coefficient) and efficiency (e.g., runtime) on the selected dataset and features?
- Can n-gram features, in combination with TF-IDF and GloVe, provide a more comprehensive representation of thematic content compared to using TF-IDF and GloVe alone?

Objectives

- To conduct a comprehensive review of existing literature on IEEE research paper classification methods, their strengths and weaknesses, the limitations of keyword-based approaches, recent advancements in machine learning and NLP techniques for text classification and clustering algorithms (K-means, DBSCAN, hierarchical, etc.).
- To collect a large and diverse dataset of IEEE research papers from various fields and time period and preprocess the data by cleaning and standardizing text formats, addressing missing values, and potentially removing irrelevant information.
- To implement TF-IDF to assess the importance of individual words within abstracts and keywords, capturing their significance within the document collection.
- To utilize GloVe word embeddings to represent keywords and abstract terms as vectors in a high-dimensional semantic space, capturing their inherent meaning and relationships.
- To implement n-gram feature engineering techniques to potentially improve the representation of thematic content.
- To implement and evaluate the performance of various clustering algorithms like K-means, DBSCAN, and hierarchical clustering on the prepared data with extracted features.
- To assess performance metrics like clustering accuracy, silhouette coefficient, and computational efficiency based on the chosen evaluation dataset.
- To select the most effective clustering algorithm based on its performance and suitability for the specific research paper classification task.

Dataset Description

The dataset used for this study (IEEE, 2018) consists of numerous IEEE research papers in different fields and times, with the aim to provide a holistic picture of academic knowledge in multiple domains. Each paper in the dataset is carefully selected to be related to the research aims and covers a variety of topics, methods, and results. The dataset is carefully prepared, which includes cleansing and sexualizing text forms, dealing with missing values and even removing from it what is not appropriate in order to support the integrity of data and improve the quality of analysis. It includes text data like titles, abstracts, keywords, and full texts of research papers that are particularly useful content for feature extraction and clustering analysis. This data set forms the base for assessing the efficiency and effectiveness of the introduced classification system allowing a deep analysis of ML and NLP techniques in improving the accessibility and user experience in browsing through research papers published in IEEE.

Short description of the project

The current approaches of IEEE research papers classification that rely on keyword-based searches in most cases do not reflect the entire semantic meaning and complexity of the relationships between the papers (Benkhaya, et al, 2020). This results in reduced search efficiency and user satisfaction, with the user having to exert a lot of effort in tuning their search to include as many relevant keywords as possible and navigating through a multitude of irrelevant returns (Hwang, et al, 2020). This problem is addressed by this work through offering an innovative classification system that uses machine learning (ML) and natural language processing (NLP) tools to eliminate these deficiencies.

The research methodology starts with an extensive literature review covering present IEEE research paper classification methods and their pros and cons, highlighting the limitations of keyword-based approaches, and modern developments in machine learning and NLP techniques for text classification and clustering algorithms such as K-means, DBSCAN, and hierarchical clustering. After this, a heterogeneous dataset of papers from the IEEE research is collected and made to undergo preprocessing to clean and standardize text formats, address missing values, and potentially remove irrelevant information. Then, TF-IDF is used to measure word relevance within abstracts and keywords, while GloVe word embeddings are applied to represent words in a high-dimensional semantic space, capturing intrinsic meaning and relationships. The thematic content representation is then extracted using N-gram feature engineering approaches. K-means, DBSCAN, and hierarchical clustering algorithms, among others, are used and tested upon the preprocessed data with features out of feature selection for performance metrics like clustering accuracy, silhouette coefficient, and computational efficiency. Finally, the best clustering algorithm is chosen according to its performance and fit for the research paper classification task.

Gantt chart

IEEE Research Articles Classification using Machine Learning and NLP for Efficient Search	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Background research												
Framin aim, RQ												
Framin methodologies and plan to conduct research												
Data collection												
Importing requird packages												
Data pre-processing												
Data visualization												
Implementing algorithms												
K-means												
DBSCAN												
Hierarchical												
TF-IDF												
GloVe												
Evaluation of the algorithms												
Comparison of results												
Conclusion drawn												
Overal results												
Research findings												
Future enhancement												
Documentation												

References

Benkhaya, S., M'rabet, S. and El Harfi, A., 2020. A review on classifications, recent synthesis, and applications of textile dyes. Inorganic Chemistry Communications, 115, p.107891.

Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), pp.20-28.

Hwang, R.H., Peng, M.C., Huang, C.W., Lin, P.C. and Nguyen, V.L., 2020. An unsupervised deep learning model for early network traffic anomaly detection. IEEE Access, 8, pp.30387-30399.

Shi, Y., Zhang, W., Yang, Y., Murzin, A.G., Falcon, B., Kotecha, A., van Beers, M., Tarutani, A., Kametani, F., Garringer, H.J. and Vidal, R., 2021. Structure-based classification of tauopathies. Nature, 598(7880), pp.359-363.

IEEE (2018). IEEE - The world's largest technical professional organization dedicated to advancing technology for the benefit of humanity. [online] @IEEEorg. Available at: https://www.ieee.org/.