A Study on an Explainable Paper Classification System Using Topic Modeling and XAI

Nakyung Shin

Department of Statistics

The Graduate School of Sungshin Women's University

A Study on an Explainable Paper Classification System Using Topic Modeling and XAI

A Master's Thesis
Submitted to the
Graduate School of Sungshin Women's University

in partial fulfillment of the requirements
for the degree of
Master of Statistics

Nakyung Shin

November, 2024

This is to certify that we have examined the
Master's Thesis of
Nakyung Shin
Submitted to Department of Statistics

Approved as to style and content:

Thesis Advisor Hohyun Jung

Committee Chairman Heewon Park

Committee Member Frederick Kin Hing Phoa

Committee Member Kwan-Young Bak

The Graduate School of Sungshin Women's University

Abstract

Accuracy and interpretability are two critical factors of document classification systems. While accuracy evaluates how well a classifier can correctly predict unseen data, interpretability focuses on how easily humans can understand the model and its rationale for assigning labels to instances. An effective classification system should not only maintain high accuracy but also provide users with intuitive and comprehensive insights to support decision-making. This study proposes an innovative explainable paper classification system that incorporates topic modeling and an explainable artificial intelligence (XAI) technique. The proposed system utilizes latent semantic analysis (LSA) for topic modeling and applies Shapley additive explanations (SHAP) to improve transparency and comprehensibility of the classification outcomes. The system offers three key advantages in the interpretability of classification results: corpus-level, document-level, and word-level. The effectiveness of the proposed system is validated using the Web of Science dataset, specifically focusing on the nanomaterial field. Its performance is further evaluated on a large-scale dataset from the Semantic Scholar database.

Keywords : paper classification, topic modeling, latent semantic analysis, explainable artificial intelligence, Shapley value

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Classification systems are essential tools for managing and accessing the growing body of academic research. As research activities expand globally, the volume of published papers has surged, driven by technological advancements, increased collaboration, and the emphasis on publications as a measure of academic achievement. This exponential growth underscores the need for effective systems to organize and navigate the vast and rapidly evolving research landscape.

Accuracy and interpretability are two critical factors for document classification systems (Van Linh et al., 2017). While the accuracy of a classifier measures the ability to correctly classify unseen data, interpretability is the ability of the classifier to be understood by humans and provide reasons why each data instance is assigned to a label. An effective classification system should not only maintain high accuracy but also provide users with intuitive and comprehensive insights to support decision-making. However, high-performing models such as neural networks often outperform comprehensible methods like boosting models, including random forest and XGBoost (Fernández-Delgado et al., 2014). The trade-off between accuracy and interpretability has led many researchers to prioritize accuracy over interpretability. Models that deliver high accuracy without providing explanations, however, can undermine trust in the system. This lack of transparency may hinder accountability in academic and research contexts. In this study, we propose an innovative explainable paper classification system that achieves high accuracy while offering explanations for classification outcomes.

Topic modeling is an essential unsupervised learning method used to identify latent

thematic structures within extensive collections of unstructured documents (Blei et al., 2003). This method relies on analyzing the distribution and co-occurrence of words across documents, thereby revealing underlying themes and enhancing the interpretability of large datasets by clustering related documents together. Such algorithms not only discern the fundamental topics embedded within texts but also provide human-readable labels, which facilitate further analysis and focused exploration of specific themes within a corpus. Consequently, topic modeling significantly augments the interpretability of document classification by organizing content into coherent themes and deepening the understanding of the underlying textual data, thus proving invaluable for paper classifiers.

In the specific application of topic modeling, this study seeks to advance our comprehension of document themes through the incorporation of latent semantic analysis (LSA) (Deerwester et al., 1990). LSA distills documents into a lower-dimensional space, capturing essential semantic relationships and thereby refining the differentiation between topics. This process generates keywords that epitomize each topic, enabling users to quickly identify papers corresponding to specific thematic areas. Employing LSA in classifying papers is poised to enhance the accuracy and comprehensiveness of topic detection in research papers.

Deep learning models often face challenges in interpretability, making it difficult to intuitively understand classification outcomes. Explainable artificial intelligence (XAI) techniques have emerged as effective approaches to enhance the interpretability of model predictions by clarifying the underlying mechanisms (Tjoa and Guan, 2020). Among these, Shapley additive explanations (SHAP) is a game-theoretic XAI method that quantifies the contribution of each feature to predictions, enabling the explanation of outcomes across various machine learning models (Lundberg and Lee, 2017).

This study integrates topic modeling with SHAP to interpret classification results at two primary levels: global explanations and local explanations. These interpretations are

further analyzed at three specific levels: corpus, document, and word. For corpus-level, we introduce a new metric, the Global Influence Factor (GIF), which identifies the topic that have significant impacts on classification outcomes. For document-level, Shapley values are utilized, while for word-level, word assignments derived from LSA are applied. This integration contributes to enhancing the reliability and transparency of classification systems, particularly in academic and research environments where clear decision-making rationales are essential. Such multi-level analysis is pivotal for understanding the varying influences that specific topics and words have on the classification results, providing a clear roadmap for adjustments and improvements in the classification process.

To validate the robustness and explainability of the proposed system, we apply it to the Web of Science (WoS) dataset in the field of nanomaterials. Additionally, we apply it to a large-scale Semantic Scholar database of statistics-related research papers. This application demonstrates the superior capability of the system in classifying research papers, thereby affirming its utility in real-world academic settings.

The key contributions of this study are as follows:

- We propose an innovative explainable paper classification system that integrates topic modeling and XAI technique. This system achieves both high accuracy and intuitive explanations for classification outcomes.

- The system provides a multi-level interpretability framework through three levels of analysis: corpus-level, document-level, and word-level. This framework enhances the understanding of classification results by identifying influential topics and words.

- We define the Global Influence Factor (GIF), a novel metric that quantifies the global impact of topics on classification outcomes.

The remainder of the paper is structured as follows. In Chapter 2, we describe the system architecture and provide a comprehensive explanation of the proposed methodology.

Chapter 3 reviews relevant literature on paper classification systems. In Chapter 4, we apply the system to the Web of Science (WoS) dataset in the nanomaterial domain, performing a comparative study to evaluate its performance, with a focus on the interpretability of classification outcomes. Chapter 5 presents the results of applying our model to a large-scale Semantic Scholar database. Finally, Chapter 6 concludes the study and outlines potential directions for future research.

Chapter 2

Proposed method

2.1 Overview

Figure 1 illustrates the system architecture of the proposed system, outlining the process from document input to the final classification and explanation stages. As a preliminary step, each document undergoes preprocessing to standardize and prepare the data. These preprocessed documents are then used to construct a word dictionary, which is subsequently transformed into a term frequency-inverse document frequency (TF-IDF) matrix. LSA is applied to the TF-IDF matrix to produce three outputs: topic assignments, topic importance, and word assignments. Topic assignments are used as input embeddings for a multilayer perceptron (MLP) classifier. Word assignments are utilized to interpret classification results and identify influential terms. The system provides explanations at two main levels: global and local. These are further divided into three levels of analysis: corpus-level, document-level, and word-level. SHAP is applied to the MLP classifier to identify the topics that influence classification outcomes at the corpus and document levels while word assignments are analyzed to determine which specific words impact the classification results. This approach offers a structured framework for producing transparent and interpretable results in document classification tasks, thereby addressing the growing need for explainable machine learning methodologies in academic settings.

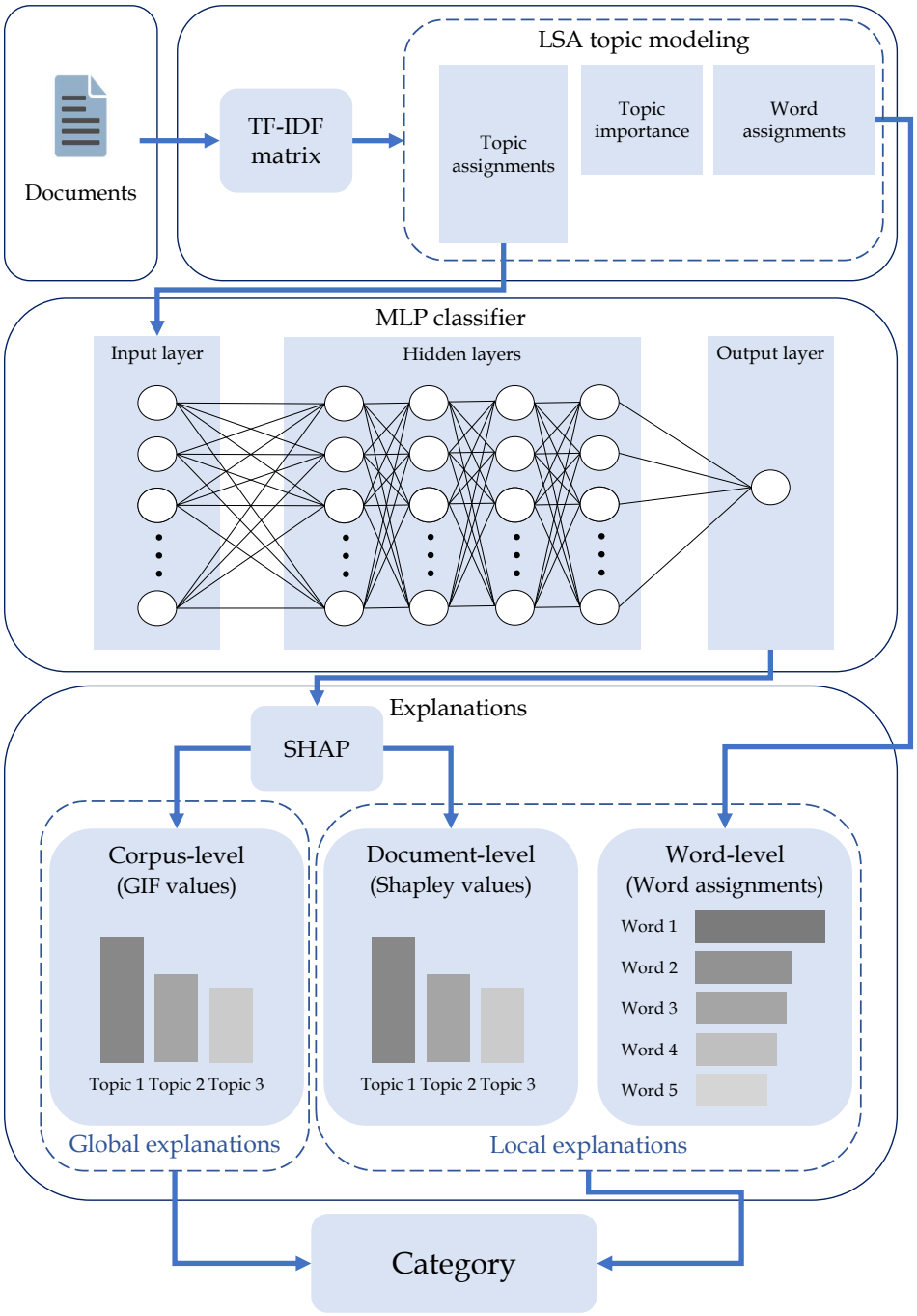


Figure 1: System architecture of the proposed system.

2.2 Latent Semantic Analysis

Latent semantic analysis (LSA) was initially introduced by Deerwester et al. (1990) as a method for analyzing the relationships between a set of documents and the terms they contain. It operates under the foundational principle that words that share similar meanings often occur in comparable contexts. LSA seeks to exploit this premise by utilizing mutual document constraints to deduce underlying topics. This conceptual framework posits that semantic structures can be discerned through patterns of word usage across texts, providing a robust mechanism for the induction and representation of knowledge.

Let $i = 1, 2, \cdots, M$ be the labeled documents, where M is the total number of documents. Let $j = 1, 2, \cdots, N$ be the words in the vocabulary, where N is the total number of unique words across the document set, selected based on word frequency. Let $C_{i,j}$ be the term frequency or the raw count of a word j in a document i, and let M_j be the document frequency or the number of documents containing a word j. The TF-IDF matrix combines term frequency and document frequency to express the importance of words in a matrix form, defined as $A = [A_{ij}]_{i=1,\cdots,M,\, j=1,\cdots,N}$. Each element A_{ij} is calculated as $C_{i,j} \cdot \log(M/M_j)$. This matrix is valuable in identifying word significance across documents but suffers from issues such as high dimensionality and noise (Deerwester et al., 1990). To address these limitations, singular value decomposition (SVD) is applied to the TF-IDF matrix, enabling the reduction of its dimensionality and facilitating the extraction of topic assignments. This process, referred to as LSA, employs truncated SVD to decompose the high-dimensional matrix $A \in \mathbb{R}^{M \times N}$ into a product of three smaller matrices as shown in the following equation (Kherwa and Bansal, 2017):

$$A \approx U \Sigma V'. \tag{2.1}$$

Let K be the number of latent topics. The matrix $U \in \mathbb{R}^{M \times K}$ is an orthogonal matrix

whose columns are the eigenvectors of AA', signifying the representation of documents in the latent topic space. The matrix $V' \in \mathbb{R}^{K \times N}$, also orthogonal, consists of the eigenvectors of $A'A$ and encapsulates the contribution of each word to the topics. Each row $V'_k \in \mathbb{R}^N$, corresponds to the contribution of all words in the vocabulary to topic k. The diagonal matrix $\Sigma \in \mathbb{R}^{K \times K}$ contains the singular values, indicating the relative importance of each topic.

In this context, the i-th row of U, $U_i \in \mathbb{R}^K$, serves as the embedding vector for document i with each element U_{ik} representing the contribution of topic k to that document. High absolute values in U_i suggest a strong association between document i and the respective topics. Similarly, high absolute values in V' indicate a significant influence of corresponding words on the topics. These embeddings, which are interpretable in terms of topics, are crucial for the interpretability of the classification, a feature that will be further explored in subsequent sections.

2.3 Multilayer Perceptron

The multilayer perceptron (MLP) is a neural network architecture consisting of multiple perceptron layers organized in a hierarchical structure (Popescu et al., 2009). This architecture typically includes an input layer, several hidden layers, and an output layer, with each layer containing multiple neurons. Neurons within these layers are interconnected through weights and biases, which are parameters adjusted during the learning process. Non-linearity in the network is introduced via activation functions, enhancing the capability of the model to capture complex patterns in the data.

We configure our MLP classifier with four hidden layers containing 128, 64, 32, and 16 neurons, respectively. Each hidden layer utilizes the rectified linear unit (ReLU) activation function, $g(x) = \max(0, x)$, to introduce non-linearity (Agarap, 2018). The Adam optimizer

is employed to update weights efficiently throughout the training process (Kingma and Ba, 2014).

The training of our MLP classifier involves up to 1,000 iterations with a batch size of 32, and the initial learning rate is set at 0.001. To mitigate overfitting and improve convergence, early stopping is implemented, halting training if no improvement in validation performance is observed over 10 consecutive iterations. The operational framework of the classifier for the topic embedding vector of a document U_i is expressed by the following equation:

$$f(U_i) = h\left(W_5 g(W_4 g(W_3 g(W_2 g(W_1 U_i + B_1) + B_2) + B_3) + B_4) + B_5\right), \tag{2.2}$$

where $h(x) = 1/(1+e^{-x})$ is the sigmoid activation function, used here to model the output as a probability (Alippi and Storti-Gajani, 1991). The variables W_l and B_l (for $l = 1,2,3,4,5$) denote the weights and biases associated with the four hidden layers ($l = 1,2,3,4$) and the output layer ($l = 5$), respectively. The function $f(U_i)$ thus predicts the probability that document i belongs to a particular category, based on its topic embedding. This detailed configuration and the mathematical representation of the MLP classifier underscore its robustness in classifying documents into thematic categories based on their latent semantic features.

2.4 Shapley Additive Explanations

SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee (2017), provides a game-theoretic framework for explaining the predictions of machine learning models. It assigns each topic a Shapley value, which quantifies its contribution to a prediction by calculating the average marginal impact of the topic across all possible subsets of features.

The core idea of SHAP lies in its ability to attribute importance to features both locally, for individual predictions, and globally, across the entire dataset. Locally, it explains why

a model predicted a certain outcome for a specific document by evaluating the contribution of each topic in isolation and in interaction with others. Globally, it aggregates these local explanations to provide an overall understanding of how features influence predictions on average.

Global importance, calculated by averaging Shapley values across all instances, measures the overall contribution of each topic. These aggregated contributions can be expressed using a simplified explanatory model:

$$f(z) \approx f'(z') = \phi_0 + \sum_{k=1}^{K} \phi_k z'_k, \tag{2.3}$$

where f denotes the MLP classifier previously defined in Eq. (2.2), and f' represents a simplified explanatory model for f. $z' \in \{0,1\}^K$ indicates the inclusion (1) or exclusion (0) of a specific topic in the estimation. The term $\phi_k \in \mathbb{R}$ represents the weight or importance of each topic within the local context, while ϕ_0 is the baseline value of the model when no topics are considered.

The local importance, or the Shapley value ϕ_k, for each topic k, assigns a value based on its contribution on the model prediction. Each Shapley value ϕ_k for a feature k is calculated to reflect its incremental impact when included in the learning process. The computation of the Shapley value for a topic k is formalized as follows:

$$\phi_k = \sum_{S \subseteq F \setminus \{k\}} \frac{|S|!\,(K-|S|-1)!}{K!} [f(S \cup \{k\}) - f(S)], \tag{2.4}$$

where $F = \{1,2,\cdots,K\}$ denotes the full set of topics, $S \subseteq F \setminus \{k\}$ is a subset excluding topic k, and $|S|$ represents the number of topics in the subset S. The difference $[f(S \cup k) - f(S)]$ measures the impact of including topic k in the prediction. The sum of the importance for topics ϕ_k yields an approximation of the prediction value for the original model f. Eq. (2.4)

represents that the Shapley value ϕ_k for topic k is computed as the weighted average of the changes in prediction values across all possible subsets of features, weighted by binomial coefficients.

The calculated Shapley values can be represented as a matrix:

$$\Phi = [\phi_{ik}] \in \mathbb{R}^{M \times K} \ , \ i = 1, 2, \cdots, M \ , \ j = 1, 2, \cdots, K, \tag{2.5}$$

where ϕ_{ik} denotes the Shapley value for topic k associated with document i. Positive values indicate a beneficial influence on the prediction, negative values suggest a detrimental effect, and values near zero imply minimal impact.

In the classification task, a particular class is expected to be involved with a relatively small number of topics compared to the total K topics. Topics that exhibit large Shapley values are interpreted as exerting substantial influence on model predictions. In this paper, we introduce the Global Influence Factor (GIF) ψ_k for topic k, defined as the average of the absolute Shapley values across each column:

$$\psi_k = \frac{1}{M} \sum_{i=1}^{M} |\phi_{ik}|. \tag{2.6}$$

A high GIF value indicates that the corresponding topic significantly influences the classification process. This metric allows us to systematically identify the most impactful topics at the corpus level, enhancing the explainability of the classification results.

In summary, the proposed system offers a three-fold advantage in enhancing the interpretability of classification outcomes:

- (Corpus-level) The system identifies which topics significantly impact classification outcomes through the GIF values calculated as per Eq. (2.6).
- (Document-level) The system elucidates the reasons behind the classification of a

document into a particular category, based on the Shapley values defined in Eq. (2.4).

- (Word-level) The system provides insights into which words contribute to the classification of a document by examining the significant words within influential topics, utilizing the word assignments from the matrix V' in Eq. (2.1).

Chapter 3

Related work

The classification of research papers is a critical task for organizing scholarly materials, enabling efficient access to relevant research within various fields. Over time, numerous techniques have emerged to categorize papers based on content, keywords, and structure.

Initial paper classification strategies employed traditional embedding techniques such as Word2Vec (Mustafa et al., 2021), Doc2Vec (Kim et al., 2019), FastText (Yao et al., 2020), and TF-IDF (Kim and Gil, 2019). These were often paired with classical classification models like support vector machines (SVM) (Chowdhury and Schoen, 2020) and k-nearest neighbor (KNN) algorithms (Nguyen and Shirai, 2013). While these methods laid the groundwork, they generally lacked in performance and did not provide explanations for their classification decisions, limiting their utility for deeper research analysis.

Further enhancements in topic modeling have incorporated embedding methods to refine thematic clustering. Nguyen et al. (2015) proposed a sophisticated hybrid model that combines the probabilistic likelihoods from LDA with a log-linear model employing pre-trained word embeddings to enhance topic specificity. Similarly, Bunk and Krestel (2018) introduced an innovative approach termed WeLDA, which involves randomly substituting words associated with a particular topic with their corresponding embeddings, sampled from a Gaussian distribution, to enrich the semantic texture of the topics. Xu et al. (2018) explored a geometric method by utilizing Wasserstein distances to concurrently learn topics and word embeddings, providing a more mathematically grounded approach to topic discovery. Additionally, Keya et al. (2022) created the neural embedding allocation (NEA), which parallels the generative process of the embedded topic model (ETM) but optimizes

it using a pre-fitted LDA model, thereby enhancing topic accuracy and relevance (Dieng et al., 2020).

The advent of deep learning models, including MLP (Jindal et al., 2015), long short-term memory (LSTM) (Ranjan et al., 2017), and convolutional neural network (CNN) (Ech-Chouyyekh et al., 2019), significantly enhanced performance. Hybrid models like Bi-LSTM-CNN (Li et al., 2018) and C-LSTM (Zhou et al., 2015) combined LSTM and CNN capabilities to better capture complex patterns in text. However, these models continued to struggle with explaining results when using traditional embeddings.

A significant drawback of deep learning models, despite their enhanced performance, is their lack of interpretability. The inability to explain classification results remains a substantial hurdle, as understanding the reasoning behind decisions is crucial for validation and trust in automated systems. The concept of XAI has emerged to address this challenge, with numerous studies utilizing SHAP to provide insights into the decisions made by complex models. Vilone and Longo (2021) have proposed a system that classifies all scientific studies hierarchically using XAI technique. Kim et al. (2020) proposed the explaining and visualizing convolutional neural networks for text information (EVCT) framework using XAI technique, which effectively minimizes information loss while enhancing the explanations for predictions made by the algorithm. Ayoub et al. (2021) used explainable natural language processing models to counter the COVID-19 infodemic, employing SHAP to explain the outputs of the DistilBERT model.

Integrating XAI technique with embedding-enhanced topic modeling marks a significant advancement in paper classification. Such technologies elucidate the influence of specific topics on classification outcomes, improving decision-making transparency and fostering trust in automated systems. These methodologies not only elevate classification accuracy but also provide vital insights into the rationale behind model decisions (Alicioglu and Sun, 2022).

Chapter 4

Experiments

4.1 Dataset

We utilize a dataset comprising 456,472 research paper abstracts related to nanomaterials, sourced from the Web of Science (WoS) and spanning publications from 2012 to 2017. We focus on five specific fields within nanomaterial research: Carbon Nanotube, Quantum Dot, Graphene, Nanosilica, and Nanosilicon. The true labels of the dataset were generated by nanotechnology experts from the Nanotechnology Policy Center at the Korea Institute of Materials Science in the project "Study of Nanotechnology Policy and Information Analysis" (2017M3A7A7057113) of the National Research Foundation of Korea. It is worth noting that papers may be categorized under multiple nanomaterial fields, indicating the interdisciplinary nature of many studies. Table 1 provides a breakdown of the number of papers associated with each of the five fields.

Field	Carbon Nanotube	Quantum Dot	Graphene	Nanosilica	Nanosilicon	Total
Counts	130,526	57,030	194,063	53,037	50,498	456,472

Table 1: Distribution of research papers across five fields in the WoS dataset.

4.2 Experimental procedure

Figure 2 illustrates the experimental procedure for the WoS dataset. Before applying LSA, the dataset underwent a series of preprocessing steps. They were converted to lower-

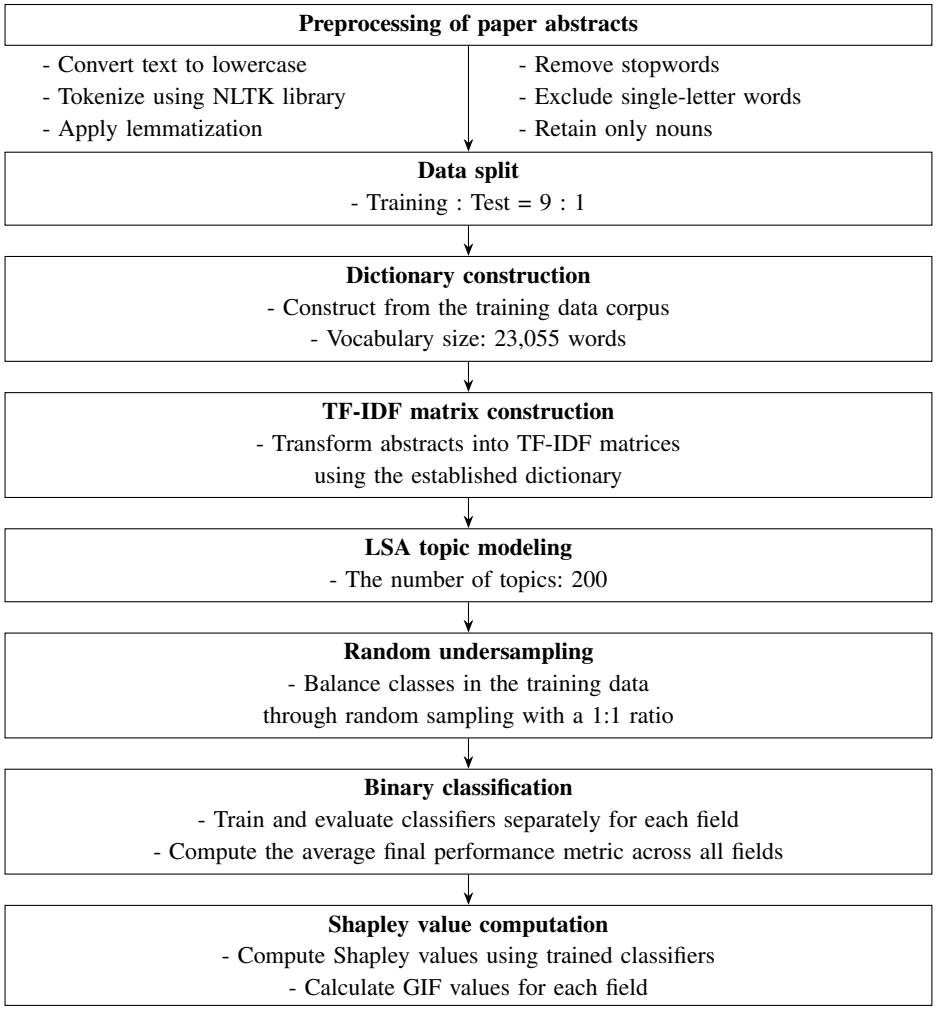


Figure 2: Experimental procedure for the WoS dataset.

case, and the tokenization of sentences into words was performed using the NLTK (version 3.8.1) library in Python. Lemmatization was applied to normalize different forms of words to their base forms, including converting plural to singular forms and standardizing verb tenses. Additionally, insignificant special characters, particles, articles, single-letter words, and words from the NLTK English stopword list were removed.

After preprocessing, the dataset was divided into training and test sets in a 9:1 ratio. We constructed a vocabulary set limited to words appearing in at least 30 documents using

the training data corpus. This resulted in a lexicon containing $N = 23,055$ words. The pre-processed abstracts were transformed into TF-IDF matrices. The TF-IDF matrix for the test data was generated based on the word importance derived from the training data.

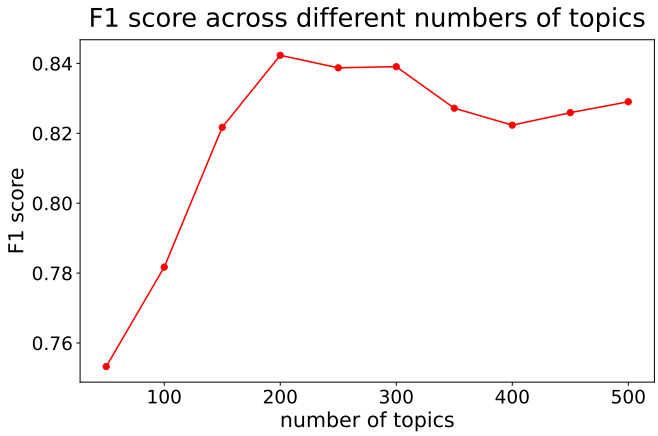


Figure 3: F1 score across different number of topics for the WoS dataset.

The optimal number of topics, K, was established at 200 after conducting iterative adjustments ranging from 50 to 500 topics during the training phase, as shown in Figure 3. This number was selected based on its contribution to maximizing classification F1 score.

According to Table 1, class imbalance is evident across all five fields. To address the imbalance, an undersampling technique was applied, which involved the random removal of indices of papers not relevant to each specific field. This strategy aimed to equalize the number of papers across fields, thereby enhancing model performance.

Due to the disparate amounts of training data available per field, individual models were trained and evaluated separately for each field. Binary classification models were developed accordingly, and an average of the individual evaluation results was calculated to derive comprehensive performance metrics across all five fields.

Finally, using the SHAP library (version 0.43.0), the Shapley values were calculated for each topic within each field. Furthermore, the GIF values were computed. These calculated

Shapley values and GIF values were subsequently utilized for interpreting the classification results.

4.3 Comparative study

We conducted a comparative analysis between our proposed system and four alternative embedding methods coupled with three different classification models. This comparison was designed to demonstrate the enhanced performance of our system in the task of paper classification.

4.3.1 Embedding methods

We utilized four embedding methods to analyze the effectiveness of different textual representations: Word2vec (Mikolov et al., 2013a,b), Doc2vec (Le and Mikolov, 2014), LDA (Blei et al., 2003), and BERTopic (Grootendorst, 2022).

Word2vec

Word2vec is a prominent word embedding method that represents words as vectors based on their meanings and contextual usage (Mikolov et al., 2013a,b). This model assesses the context of a word by considering up to five words preceding and following it within a sentence, including only those words that appear with a frequency of 40 or more. Notably, the skip-gram (SG) model was selected over the continuous bag of words (CBOW) model due to its superior performance in predicting the center word from its surrounding context. Each word is represented by a 200-dimensional vector. For document embedding, the average of word vectors within the document is used to form a document vector, providing a consolidated representation that encapsulates overall semantic content.

Doc2vec

Doc2vec is an extension of Word2vec that encodes entire documents into unique vectors (Le and Mikolov, 2014). Like Word2vec, it examines up to five words before and after the target word within a sentence, including only those that meet a minimum frequency threshold. The distributed memory version of the paragraph vector (PV-DM) model was utilized for our Doc2vec implementation.This model captures the semantic meanings of words within their contexts, enhancing the overall document representation. Each document is thereby represented by a 200-dimensional vector, which ensures a rich, context-aware embedding that encapsulates the thematic essence of the text.

Latent Dirichlet Allocation

LDA is a topic modeling method that probabilistically infers the topic structure within documents (Blei et al., 2003). To train the LDA model, the same dictionary generated via LSA was utilized. The batch size for document processing was set to 2,000, and the model was configured to discern 200 topics. The parameters of the Dirichlet distribution for each topic distribution were set equally, resulting in $\alpha = 0.005$. The resulting probability distributions for the topics provided the embedding values for each document.

BERTopic

BERTopic leverages clustering technology and class-based TF-IDF to model topics effectively, generating discernible topic representations and enhancing the granularity of topic detection (Grootendorst, 2022). Documents are embedded into a vector space using the Sentence-BERT (SBERT) framework, specifically employing the all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019). This version of SBERT has been trained on a vast corpus of 1,170,060,424 training tuples, including sources like Reddit comments and Wikipedia pages. It is known for its state-of-the-art performance on various sentence embedding tasks, facilitating effective semantic comparisons.

To manage the high dimensionality of the data, uniform manifold approximation and projection (UMAP) is applied, reducing dimensions to a three-dimensional space that preserves the semantic similarity of documents. This step involves setting the number of neighboring points to 20 and the minimum distance between points to 0.1, using the Euclidean distance metric for calculations.

Following dimensionality reduction, hierarchical density-based spatial clustering of applications with noise (HDBSCAN) is employed to identify dense clusters and segregate noise. Additionally, clustering incorporates Prim's algorithm, which utilizes k-dimensional (KD) trees to enhance clustering efficiency. The minimum cluster size is set to 200, with a minimum of 30 neighbors required to form a cluster. The Euclidean distance metric and an excess mass method are used to select and validate clusters.

Finally, class-based TF-IDF (C-TFIDF) is implemented to refine topic representations. Unlike traditional TF-IDF, C-TFIDF assesses word importance within clusters, facilitating the generation of specific topic-word distributions for each document cluster. To optimize topic coherence, the representation of the least prevalent topic is iteratively merged with the most similar one, reducing the number of topics to 200.

4.3.2 Classification models

As classification models, we utilized logistic regression (LR) (Kirasich et al., 2018), randomforest (RF) (Ali et al., 2012), extreme gradient boosting (XGB) (Chen and Guestrin, 2016).

Logistic Regression

Logistic regression (LR) is widely used for binary classification, leveraging the logistic function to predict probabilities (Kirasich et al., 2018). This method compares the output probability against a fixed threshold, set at 0.5, to categorize observations into binary

classes (0 or 1). In our implementation, the model incorporates an L2 regularization penalty to prevent overfitting. We optimize the model using the limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm, renowned for its effectiveness in large-scale applications. The optimization process is controlled with a maximum of 1,000 iterations and a convergence tolerance of 0.001.

RandomForest

Randomforest (RF) is an ensemble method that enhances predictive accuracy by aggregating outputs from multiple decision trees (Ali et al., 2012). This method not only increases the robustness of predictions but also helps control overfitting through its ensemble approach. We configured our RF with 200 trees, employing the Gini index as the criterion for optimizing splits. We capped the maximum depth of each tree at 20, allowing for complex pattern recognition while preventing overfitting. Each internal node in the trees requires at least 5 samples to split, ensuring sufficient data for reliable decision-making and maintaining an equilibrium between bias and variance.

Extreme Gradient Boosting

Extreme gradient boosting (XGB) builds upon traditional gradient boosting frameworks by iteratively correcting the errors of previously built trees (Chen and Guestrin, 2016). We use 200 trees, and each tree can grow to a maximum depth of 20 to capture complex interactions, and the minimum child weight is set at 10 to control overfitting by making the algorithm more conservative. The learning rate is fixed at 0.3 to moderate the impact of each individual tree and prevent rapid convergence to suboptimal solutions. Additionally, the gamma parameter, which specifies the minimum loss reduction required to make further splits on a leaf node, is set at 0.1.

4.3.3 Evaluation metrics

The key evaluation metrics considered in this study are Accuracy, F1 score, and area under the receiver operating characteristic curve (AUROC). These metrics are explained with reference to the confusion matrix shown in Table 2.

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Table 2: Confusion matrix for binary classification.

First, Accuracy represents the proportion of correctly classified observations among all observations. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{4.1}$$

Second, the F1 score is computed as the harmonic mean of precision and recall. Precision indicates the ratio of true positives among the instances predicted as positive, and recall represents the ratio of true positives among the actual positive instances. They are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4.2}$$

The F1 score achieves a high value when precision and recall are balanced, and a high F1 score signifies better model performance:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.3}$$

Finally, AUROC is the area under the receiver operating characteristic (ROC) curve. The ROC curve visualizes the relationship between the false positive rate (FPR) and the true positive rate (TPR). The FPR represents the proportion of actual negatives incorrectly predicted as positive, while the TPR represents the proportion of actual positives correctly predicted as positive. The TPR and FPR are defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}. \tag{4.4}$$

To calculate AUROC, the ROC curve is constructed by calculating the TPR and FPR at various threshold levels, and AUROC is calculated using the Trapezoidal Rule. The AUROC can be computed as follows:

$$\mathrm{AUROC} = \sum_{i=1}^{n-1} (\mathrm{FPR}_{i+1} - \mathrm{FPR}_i) \cdot \frac{\mathrm{TPR}_i + \mathrm{TPR}_{i+1}}{2}. \tag{4.5}$$

where n is the number of threshold levels, and TPR_i and FPR_i are the true positive rate and false positive rate at the i-th threshold level, respectively.

In this study, we use the F1 score as the key metric to compare model performance. When class 1 is frequently misclassified as class 0, or vice versa, accuracy can be misleading, especially in imbalanced datasets. In contrast, the F1 score provides a high value only when the model performs well in both precision and recall, even in cases with high misclassification rates. Additionally, AUROC is a metric that assesses predictive ability across various threshold values but does not provide information on performance at a specific threshold. In comparison, the F1 score evaluates predictive outcomes at a fixed threshold, allowing for a more intuitive assessment of the final model performance.

4.4 Results

4.4.1 Comparison results

Embedding	Classifier	Accuracy	F1	AUROC
W2V	LR	0.8667	0.6970	0.9300
	RF	0.8074	0.6130	0.8905
	XGB	0.8665	0.6990	0.9373
	MLP	0.8792	0.7215	0.9468
D2V	LR	0.8283	0.6247	0.8967
	RF	0.7983	0.5684	0.8606
	XGB	0.8339	0.6355	0.9062
	MLP	0.8471	0.6566	0.9184
LDA	LR	0.7684	0.5520	0.8466
	RF	0.7695	0.5312	0.8225
	XGB	0.7578	0.5349	0.8324
	MLP	0.7745	0.5626	0.8575
LSA (proposed)	LR	0.9077	<u>0.7792</u>	0.9490
	RF	0.8746	0.6968	0.9227
	XGB	0.8977	0.7568	0.9510
	MLP (proposed)	**0.9080**	**0.7816**	**0.9574**
BERTopic	LR	<u>0.9079</u>	0.7781	0.9481
	RF	0.8748	0.6975	0.9226
	XGB	0.8973	0.7558	<u>0.9522</u>
	MLP	0.8983	0.7576	0.9503

Table 3: Comparative results for twenty combinations in the WoS dataset; **bold** for best, <u>underline</u> for second-best.

Table 3 summarizes the classification performance for twenty combinations of embedding methods and classification models applied to the WoS dataset. The proposed system, utilizing LSA for embedding and an MLP classifier, achieved the highest performance across all metrics, with an Accuracy of 0.9080, an F1 score of 0.7816 and an AUROC of 0.9574. The second-best combination was the LSA-LR configuration, highlighting the robustness of the LSA embedding method.

The results also reveal that the choice of embedding method has a more significant impact on classification performance than the classifier selection. For instance, despite BE-

Topic being a sophisticated topic modeling approach built on pre-trained models, it did not surpass the performance of the LSA-MLP configuration. This underscores the efficiency of LSA as a lightweight yet highly effective embedding technique when combined with appropriate classifiers.

4.4.1.1 Comparative study with other datasets

We employ two additional datasets for our comparative study: the Web of Science (WoS2) dataset (Kowsari et al., 2017) and the Medical-Abstracts-TC-Corpus (MATC) dataset (Schopf et al., 2022). The WoS2 dataset is categorized into seven parent categories: Medical science, Psychology, Computer science, Biochemistry, Electrical and computer engineering, Civil engineering, and Mechanical and aerospace engineering. The MATC dataset encompasses five different classes of patient conditions: General pathological conditions, Neoplasms, Cardiovascular diseases, Nervous system diseases, and Digestive system diseases.

Following the experimental procedure described in Section 4.2, we applied our proposed system and baseline models to both datasets. As shown in Table 4, our proposed system achieved the highest performance on the MATC dataset, with an F1 score of 0.7518. For the WoS2 dataset, the proposed system attained the second-highest F1 score of 0.6814, closely trailing the top-performing BERTopic-MLP configuration by a marginal difference of 0.0022.

These results validate the competitive performance of the proposed system, particularly when compared with models utilizing pretrained BERT embeddings. This comparative analysis further confirms the robustness and effectiveness of our model across diverse datasets and classification tasks.

Dataset	Embedding	Classifier	Accuracy	F1	AUROC
WoS2	W2V	LR	0.8871	0.6689	0.9532
		RF	0.8700	0.6403	0.9487
		XGB	0.8750	0.6588	0.9513
		MLP	<u>0.8899</u>	0.6796	0.9436
	D2V	LR	0.8226	0.5439	0.8953
		RF	0.8021	0.5150	0.8806
		XGB	0.8342	0.5640	0.9096
		MLP	0.8361	0.5716	0.9132
	LDA	LR	0.7678	0.4859	0.8713
		RF	0.7304	0.4352	0.8288
		XGB	0.7512	0.4550	0.8416
		MLP	0.7647	0.4817	0.8687
	LSA (proposed)	LR	0.8841	0.6687	0.9432
		RF	0.8756	0.6582	0.9502
		XGB	0.8800	0.6550	<u>0.9592</u>
		MLP (proposed)	**0.8932**	<u>0.6814</u>	**0.9674**
	BERTopic	LR	0.8863	0.6596	0.9401
		RF	0.8757	0.6557	0.9501
		XGB	0.8896	0.6536	0.9586
		MLP	0.8802	**0.6836**	0.9347
MATC	W2V	LR	0.8695	0.7405	0.9212
		RF	0.8525	0.7004	0.9012
		XGB	0.8404	0.6593	0.8855
		MLP	<u>0.8719</u>	<u>0.7507</u>	**0.9278**
	D2V	LR	0.8241	0.6658	0.8845
		RF	0.8310	0.6459	0.8721
		XGB	0.8383	0.6438	0.8710
		MLP	0.8487	0.6863	0.8968
	LDA	LR	0.7091	0.5139	0.7589
		RF	0.6918	0.4721	0.7361
		XGB	0.7084	0.4020	0.6835
		MLP	0.7936	0.5169	0.7627
	LSA (proposed)	LR	0.8863	0.7462	0.9125
		RF	0.8625	0.6998	0.9042
		XGB	0.8524	0.6539	0.8852
		MLP (proposed)	**0.8736**	**0.7518**	<u>0.9255</u>
	BERTopic	LR	0.8639	0.7154	0.9039
		RF	0.8425	0.6798	0.8842
		XGB	0.8348	0.6378	0.8757
		MLP	0.8476	0.7086	0.9054

Table 4: Comparative results for twenty combinations in the WoS2 dataset and MATC dataset; **bold** for best, <u>underline</u> for second-best.

4.4.2 Explanations of classification results

4.4.2.1 SHAP Evaluation

The fidelity measures the degree to which the explanations approximate the original predictions of the model (Zhou et al., 2021). Fidelity scores of SHAP are usually quantified using the mean squared error (MSE) between the sum of Shapley values and the predicted probabilities of each model. The smaller the fidelity score, the better the SHAP explanations align with the model output, indicating that the interpretations are more faithful to the decision-making process of the model. As shown in Table 5, the fidelity scores for each field and their overall average are remarkably low. These results indicate that SHAP provides highly reliable and consistent explanations that closely reflect the predictions of the MLP classifier, demonstrating its utility in enhancing the transparency and interpretability of the model.

Field	Carbon Nanotube	Quantum Dot	Graphene	Nanosilica	Nanosilicon	Average
Fidelity	4.2298e-6	2.3664e-5	4.4665e-6	2.3064e-5	3.6221e-5	0.00073

Table 5: Fidelity scores (MSE) for Shapley values.

4.4.2.2 Global explanations

We focus on corpus-level explanations to identify which topics significantly influence the classification results based on the defined GIF values in Section 2.4.

Table 6 lists the top three topics that best describe each field, along with the five most representative words for each topic. Topics were selected based on their highest GIF values, while the representative words were determined by their highest absolute word assignments in the V' matrix. The topic names were manually assigned based on the top five words.

Field	No.	Topic name	Top 5 words
Carbon Nanotube	11	Carbon-based materials	tio, go, nanotube, qds, rgo
	10	Graphene-oxide materials	graphene, go, rgo, oxide, carbon
	21	Metallic materials	cu, fe, ag, cell, cnts
Quantum Dot	5	Quantum Dot-based membranes	composite, membrane, quantum, go, dot
	7	Photocatalytic catalysts	tio, catalyst, photocatalytic, qds, cd
	12	Materials for film-making	qds, ag, np, film, co
Graphene	10	Graphene-oxide materials	graphene, go, rgo, oxide, carbon
	2	Electrochemical devices	film, electrode, catalyst, capacity, electrochemical
	11	Carbon-based materials	tio, go, nanotube, qds, rgo
Nanosilica	12	Materials for film-making	qds, ag, np, film, co
	10	Carbon-oxide composites	graphene, go, rgo, oxide, carbon
	38	Materials for fiber composites	al, sio, cd, fiber, silica
Nanosilicon	6	Membranes for applications	film, adsorption, membrane, tio, thin
	26	Materials for electronic devices	si, cu, rgo, zno, go
	2	Electrochemical devices	film, electrode, catalyst, capacity, electrochemical

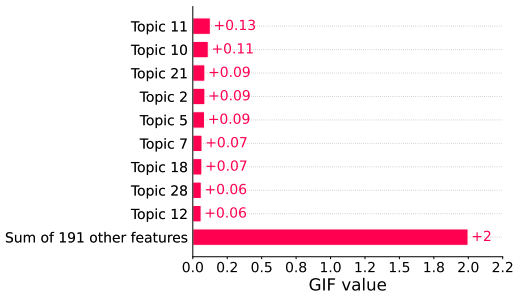
Table 6: Top three topics with the highest GIF values for each field, along with the top five words with the highest absolute assignments in V'.

The GIF values of the topics for each field are shown in Figure 4. In the figure, the horizontal length of each bar corresponds to the GIF value of the topic. Longer bars indicate a greater influence on the classification outcomes within the respective field.

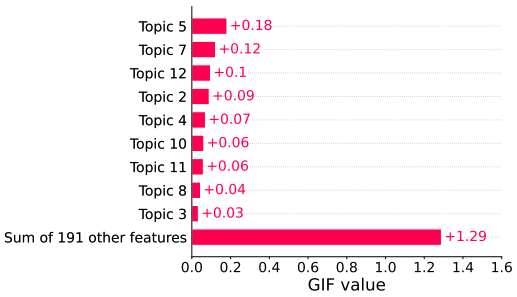
Carbon Nanotube

Carbon nanotubes are cylindrical allotropes of carbon known for their unique nanostructure. Renowned for their remarkable strength, high electrical conductivity, and low density, carbon nanotubes are highly valued in advanced applications such as batteries and composite materials. The topics significantly influencing the classification of papers within the Carbon Nanotube field are Topics 11 (Carbon-based materials), 10 (Graphene-oxide materials), and 21 (Metallic materials).

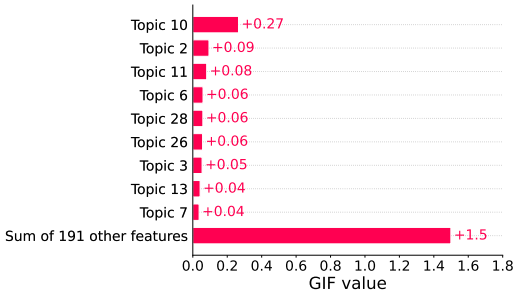
Topic 11 predominantly covers carbon-based materials including nanotubes, quantum dots (qds), reduced graphene oxide (rgo), and graphene oxide (go). Topic 10 is closely associated with graphene and its derivatives such as graphene oxide and reduced graphene



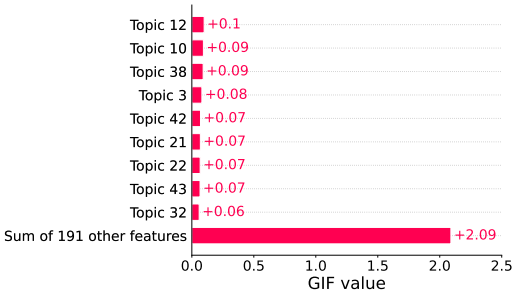
(a) Carbon Nanotube



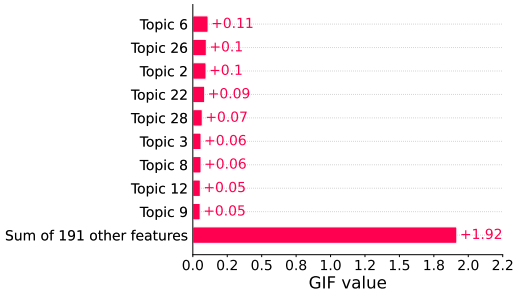
(b) Quantum Dot



(c) Graphene



(d) Nanosilica



(e) Nanosilicon

Figure 4: GIF values for the topics, ranked in descending order of GIF values for each field.

oxide, reflecting the relevance of graphene-related research in this field. Meanwhile, Topic 21 captures the classification of papers focusing on metals like copper (cu), iron (fe), and silver (ag), known for their high electrical conductivity, alongside their interaction with carbon nanotubes.

Figure 4a presents the GIF values for these topics, where $\psi_{11} = 0.13$ and $\psi_{10} = 0.11$,

both notably higher than $\psi_{21} = 0.09$ for Topic 21. This disparity indicates that Topics 11 is approximately 1.5 times more influential in classifying papers within the Carbon Nanotube category compared to Topic 21, highlighting their significant impact on the field.

Quantum Dot

Quantum dots are ultrafine semiconductor particles that are pivotal in developing display devices, offering vivid colors, longer lifespans, and greater cost-effectiveness compared to traditional LEDs and OLEDs. The significant influence on the classification of Quantum Dot is chiefly guided by Topics 5 (Quantum Dot-based membranes), 7 (Photocatalytic catalysts), and 12 (Materials for film-making).

Topic 5 pertains to the use of membranes comprising quantum dot composites, essential for various nanotechnology applications. Topic 7 includes research on photocatalytic substances such as titanium dioxide (tio2), which is often abbreviated as 'tio' in processed texts, qds, and cadmium (cd) used in photocatalytic reactions. Topic 12 deals with materials employed in the manufacturing of displays, prominently featuring qds, silver (ag), and cobalt (co).

According to Figure 4b, Topic 5, which is directly related to quantum dots, demonstrates an impact that is approximately 1.5 times greater than that of Topic 7 and twice that of Topic 12: $\psi_5 = 0.18$, $\psi_7 = 0.12$, and $\psi_{12} = 0.1$.

Graphene

Graphene is a two-dimensional nanomaterial comprised of a single layer of carbon atoms arranged in a hexagonal lattice, celebrated for its remarkable electrical and thermal conductivities. Its versatility makes it a pivotal material in the development of advanced biocomposites for dental and medical applications. The classification of Graphene-related papers is predominantly influenced by Topics 10 (Graphene-oxide materials), 2 (Electrochemical devices), and 11 (Carbon-based materials).

Topic 2 includes words associated with electrochemical applications, underscoring its relevance to devices that capitalize on the exceptional conductive properties of graphene. Topics 10 and 11 effectively describe both Carbon Nanotubes and Graphene. This is attributed to their similar properties to carbon allotropes.

As illustrated in Figure 4c, the GIF value for Topic 10 is $\psi_{10} = 0.27$, which is significantly higher than those for Topics 2 and 11, at $\psi_2 = 0.09$ and $\psi_{11} = 0.08$, respectively. This disparity underscores the paramount importance of graphene-oxide materials in the classification within the Graphene field, demonstrating that Topic 10 is over twice as influential as the other topics examined.

Nanosilica

Nanosilica refers to silica synthesized on the nanometer scale, which consists primarily of silicon and oxygen. The classification of Nanosilica-related papers is significantly influenced by Topics 12 (Materials for film-making), 10 (Graphene-oxide materials), and 38 (Materials for fiber composites).

Topic 38, in particular, is characterized by its focus on fiber composite materials that often incorporate elements like aluminum (al) and cadmium (cd), demonstrating direct applications in Nanosilica technology. Notably, the term 'sio' within this topic represents silicon dioxide (SiO2), further emphasizing the connection to Nanosilica. This topic provides a comprehensive description of how Nanosilica is utilized within various composite materials, highlighting its widespread application.

Figure 4d presents the GIF values, with Topic 12, 10, and 38 showing values of $\psi_{12} = 0.1$, $\psi_{10} = 0.09$, and $\psi_{38} = 0.09$. The similar GIF values across these topics indicate that while Topic 38 plays a crucial role in classifying Nanosilica, it operates within a context where several topics collectively contribute to the classification of Nanosilica.

Nanosilicon

Nanosilicon refers to silicon at the nanoscale, with its properties and applications extensively explored in the realms of bio- and energy-related materials. The classification of Nanosilicon-related papers is notably influenced by Topics 6 (Membranes for adsorption applications), 26 (Materials for electronic devices), and 2 (Electrochemical devices).

Topic 6 is particularly centered on research pertaining to advanced membrane materials that are pivotal in adsorption applications, often employing thin-film technologies. The focus of this topic reflects the innovative use of nanoscale materials in enhancing the functionality and efficiency of adsorption processes. Topic 26 delves into materials used in electronic device applications, encompassing a range of essential components such as silicon (si), zinc oxide (zno), cu, rgo, and go. The inclusion of diverse materials underscores the broad application spectrum of Nanosilicon in modern electronics.

Figure 4e illustrates the GIF values for these topics, with $\psi_6 = 0.11$, $\psi_{26} = 0.1$, and $\psi_2 = 0.1$, respectively. The close proximity of these values highlights a balanced impact of these diverse topics on the classification within the Nanosilicon field, indicating that while each topic contributes significantly, they do so to a similar degree.

Various topics related to the characteristics and applications of nanomaterials have been identified for each nanomaterial and provide insights into diverse research trends in the nanomaterial field. Carbon Nanotubes, Quantum Dots, and Graphene, which are all nanostructures primarily based on carbon, exemplify the versatility and wide-ranging utility of carbon in nanotechnology. Conversely, Nanosilica and Nanosilicon, which are derived from silicon, showcase the diverse applications of silicon-based materials. The shared utilization of these elemental nanomaterials across similar application domains not only underscores their comparable properties but also highlights their integral role in advancing the nanotechnology field.

4.4.2.3 Local explanations

Local explanations provide insight into why individual documents are classified into specific fields, offering a detailed understanding of the underlying classification mechanisms. To explore these explanations, we randomly selected abstracts from each field and interpreted the classification results at both the document-level and word-level.

At the document-level, topics with higher Shapley values are more influential in determining classification into specific fields. Figure 5 visualizes these results using randomly selected abstracts for each field. Topics are listed in descending order of the absolute values of their Shapley values, with red bars representing positive values and blue bars representing negative values.

At the word-level, words with higher absolute word assignments in the V' matrix are considered influential for their corresponding topics. Table 7 presents the five most influential words for each topic, along with the frequency of these words in the actual documents. As shown in Figure 5, the Shapley values of the top two topics are significantly higher compared to the others. Thus, we focus on explaining the results using these two topics.

Carbon Nanotube

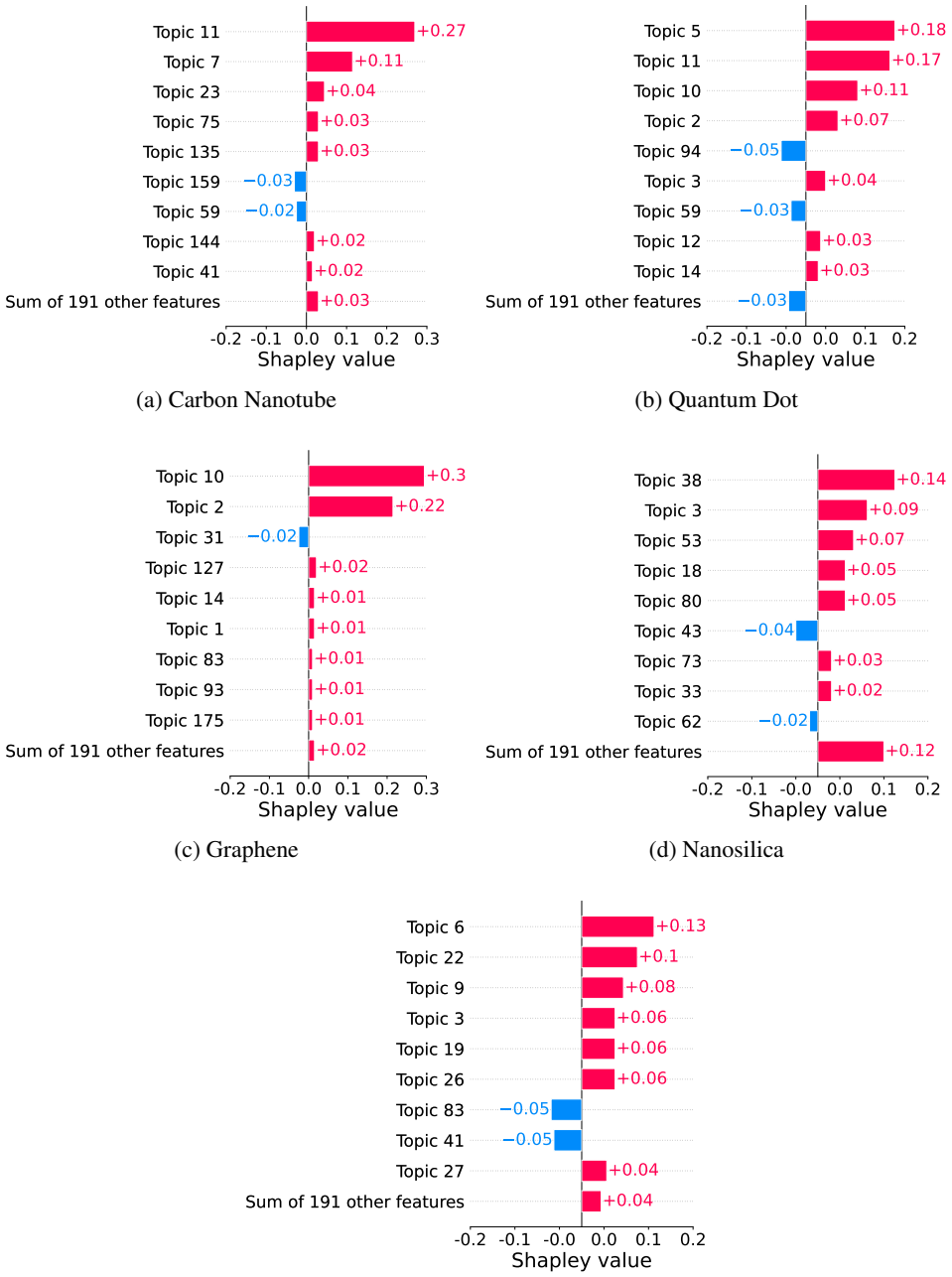
At the document-level, as shown in Figure 5a, the Shapley values demonstrate a strong association with Topic 11, with a value of 0.27. This value is more than twice that of Topic 7 and over six times that of Topic 23. At the word-level, the document contains keywords central to the Carbon Nanotube field: ‘nanotube’ appears 6 times. Consequently, the document is classified based on the frequent occurrence of ‘nanotube,’ which contributes to the high Shapley value for Topic 11.

Quantum Dot

As shown in Figure 5b, the Shapley values of Topic 5 and Topic 11 are 0.18 and 0.17, respec-

Field	Topics			
	Topic 11		Topic 7	
	Word	Freq	Word	Freq
Carbon Nanotube	tio	0	tio	1
	go	0	catalyst	1
	nanotube	6	photocatalytic	0
	qds	0	qds	0
	rgo	0	cd	0
	Total	6	Total	1
	Topic 5		Topic 11	
	Word	Freq	Word	Freq
Quantum Dot	composite	0	tio	0
	membrane	0	go	0
	quantum	4	nanotube	0
	go	0	qds	6
	dot	3	rgo	0
	Total	7	Total	6
	Topic 10		Topic 2	
	Word	Freq	Word	Freq
Graphene	graphene	3	film	0
	go	1	electrode	2
	rgo	2	catalyst	0
	oxide	1	capacity	0
	carbon	4	electrochemical	3
	Total	11	Total	5
	Topic 38		Topic 3	
	Word	Freq	Word	Freq
Nanosilica	al	0	catalyst	0
	sio	2	electrode	0
	cd	0	activity	0
	fiber	0	battery	2
	silica	4	capacity	0
	Total	6	Total	2
	Topic 6		Topic 22	
	Word	Freq	Word	Freq
Nanosilicon	film	2	go	0
	adsorption	0	laser	0
	membrane	0	si	2
	tio	0	cell	1
	thin	1	zno	0
	Total	3	Total	3

Table 7: Frequency of the top five words for the two topics with the highest Shapley values in randomly selected paper abstracts across each field.



(a) Carbon Nanotube

(b) Quantum Dot

(c) Graphene

(d) Nanosilica

(e) Nanosilicon

Figure 5: Shapley values for the topics, ranked in descending order of absolute Shapley values for each field.

tively. These two topics exhibit higher Shapley values compared to others, as this document includes the words ‘quantum,’ ‘dot,’ and ‘qds’ four, three, and six times, respectively. In total, words directly representing the Quantum Dot appear 13 times, which ultimately leads to the classification of the document as related to Quantum Dot.

Graphene

Figure 5c illustrates that the Shapley values for Topic 10 and Topic 2 are 0.3 and 0.22, respectively, more than 10 times higher than those of other topics. This indicates that the influence of other topics is minimal, while these two topics have a dominant impact. This significant influence is likely due to the frequent occurrence of words associated with these topics, which appear a total of 16 times. Specifically, the top five words in Topic 10, representing graphene and its compounds, each occur at least once. Additionally, ‘electrode’ and ‘electrochemical’ appear 2 and 3 times, respectively, further emphasizing graphene’s role in electrode materials.

Nanosilica

As shown in Figure 5d, the Shapley values for Topic 38 and Topic 3 are 0.14 and 0.09, respectively. Notably, Topic 38 has a value approximately twice as large as Topic 53, the third-highest topic. The dominance of Topic 38 can be attributed to its inclusion of words such as ‘sio’ and ‘silica,’ which directly reference Nanosilica and appear six times in total. Meanwhile, the term ‘battery’ in Topic 3 suggests that this document likely pertains to the use of Nanosilica in battery-related materials.

Nanosilicon

As shown in Figure 4e, the GIF values of the top nine topics are relatively similar, suggesting that the appearance of specific words within topics has a significant impact on Shapley values. However, Figure 5e shows that the two topics with the highest Shapley values, Topic

6 and Topic 22, rank first and fourth in terms of GIF values, respectively. The second-highest Shapley value for Topic 22 can be attributed to the inclusion of ‘si,’ a fundamental element of Nanosilicon, which appears twice. Meanwhile, Topic 6 highlights discussions on the use of Nanosilicon in thin-film fabrication for display applications.

Overall, we can say that papers are typically classified into specific fields based on the prominence of elemental symbols or abbreviations that directly represent the nanomaterial, as well as words associated with its applications, within the abstract. This trend underscores the importance of targeted vocabulary in accurately categorizing research papers according to their focus within the realm of nanomaterials.

Chapter 5

Applications to Semantic Scholar

5.1 Dataset

Semantic Scholar is a free, AI-powered academic search engine developed by the non-profit Allen Institute for Artificial Intelligence. From the Semantic Scholar database, we extracted academic papers related to the field of statistics. We randomly selected fifteen statistics-related keywords each year, each consisting of either a single word or a two-word compound. Papers containing these keywords were considered relevant to the respective topics. Using the paper IDs of the selected papers, we retrieved their abstracts via the Semantic Scholar API. Over the eight years from 2016 to 2023, a total of 4,602,097 paper abstracts were collected. Table 8 and 9 present the distribution of papers across the selected keywords for each year, highlighting the number of publications associated with each keyword. Notably, each article is associated with at least one keyword and may be linked to multiple keywords. Table 10 provides an illustrative example of the data structure by showing the presence of keywords in abstracts using binary indicators, where a value of 1 denotes that an abstract includes a specific keyword and a value of 0 indicates its absence. The example data is taken from the 2023 dataset.

5.2 Comparison results

The experimental procedures are described similarly to Section 4.2, with one key difference: the number of topics, K, was set to 500. This choice was based on the fact that the F1 score reached its peak at $K = 500$ among values ranging from 50 to 700. Figure 6

	2016		2017		2018		2019	
No.	Keyword	Counts	Keyword	Counts	Keyword	Counts	Keyword	Counts
1	distribution	45,992	average	45,222	community	49,747	function	57,130
2	hypothesis	44,774	empirical	44,470	parameter	48,719	programming	54,395
3	parameter	44,350	sampling	42,913	regression	46,436	population	52,812
4	algorithm	44,333	hypothesis	42,774	correlation	45,900	correlation	51,851
5	dimension	43,729	confidence	42,727	spatial	45,504	algorithm	48,906
6	empirical	42,784	simulation	42,524	optimization	44,239	significance	48,706
7	p values	41,457	correlation	42,197	randomized	43,149	parameter	46,884
8	significance	41,141	distribution	40,918	noise	41,895	randomized	45,121
9	score	38,977	estimation	39,507	distribution	41,810	network	44,500
10	robust	38,046	studentization	38,199	variable	41,224	exchangeability	42,369
11	partial	37,629	parameter	37,774	dimension	40,738	stability	38,871
12	mean	35,970	regression	37,358	transform	38,479	hypothesis	37,918
13	exchangeability	34,736	p values	36,239	hierarchy	33,191	studentization	37,878
14	dependence	34,591	bias	33,550	time series	33,153	classification	36,030
15	time series	31,008	heterogeneous	31,507	diffusions	31,518	convergence	33,794
Total	508,017		511,208		528,977		570,620	

Table 8: Distribution of research papers across each keyword from the Semantic Scholar database for the years 2016 to 2019.

	2020		2021		2022		2023	
No.	Keyword	Counts	Keyword	Counts	Keyword	Counts	Keyword	Counts
1	sampling	67,195	sampling	77,317	network	66,989	sampling	75,371
2	mean	56,999	parameter	61,602	regression	61,679	predictor	68,014
3	scale	56,862	algorithm	59,973	optimization	58,571	parameter	66,107
4	parameter	52,682	function	59,256	distribution	58,180	scale	66,087
5	simulation	52,412	variable	58,064	correlation	57,725	optimization	62,299
6	distribution	50,447	random	54,883	variable	55,916	coefficient	54,859
7	coefficient	49,896	significance	49,897	classification	52,145	quantitative	53,649
8	error	49,276	regression	49,511	error	50,930	condition	53,074
9	rate	48,443	population	49,304	algorithm	50,690	p values	49,360
10	correlation	43,077	adaptive	48,089	significance	49,862	probability	44,741
11	bias	42,420	distribution	46,681	cluster	45,756	hypothesis	44,412
12	partial	42,217	dimension	44,742	dimension	44,851	convolution	43,547
13	multivariate	38,412	prediction	44,646	spatial	44,851	threshold	38,775
14	standardization	36,173	bias	43,685	hypothesis	42,224	convergence	36,626
15	matrix	34,648	domain	40,462	multivariate	34,152	boosting	33,989
Total	582,320		619,100		627,249		654,606	

Table 9: Distribution of research papers across each keyword from the Semantic Scholar database for the years 2020 to 2023.

No.	abstract	keyword			
		sampling	predictor	⋯	boosting
1	This article aims to find ⋯	1	0	⋯	0
2	The ability to predict ⋯	0	1	⋯	1
⋮	⋮	⋮	⋮	⋱	⋮
654,606	A significant challenge ⋯	1	0	⋯	0

Table 10: Example of the 2023 data structure showing abstracts with associated keywords.

presents the F1 score of the classification model with varying topic numbers.

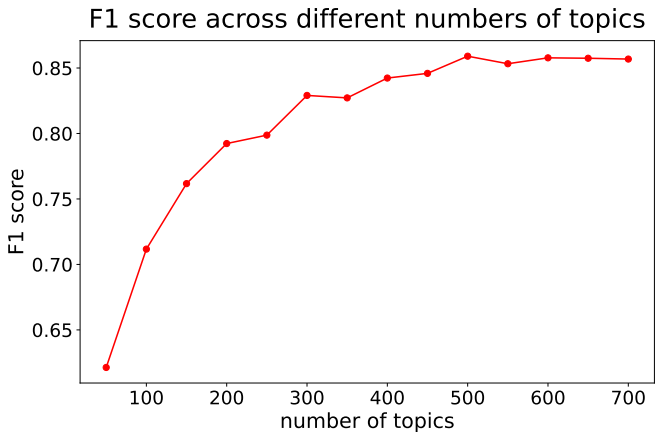


Figure 6: F1 score across different number of topics for the Semantic Scholar dataset.

Figure 7 compares the performance of four classifiers using LSA as the embedding method. Among these classifiers, MLP demonstrated the highest performance in terms of F1 score while maintaining reasonable computational efficiency, indicating its suitability for high-quality and timely predictions.

Figure 8 illustrates the performance comparison of four embedding methods when MLP is used as the classifier. The results indicate that LSA consistently outperforms the other embedding methods, achieving the highest F1 scores and the shortest processing times. This finding underscores the effectiveness of our model as the optimal solution for document classification, delivering superior performance while minimizing computational

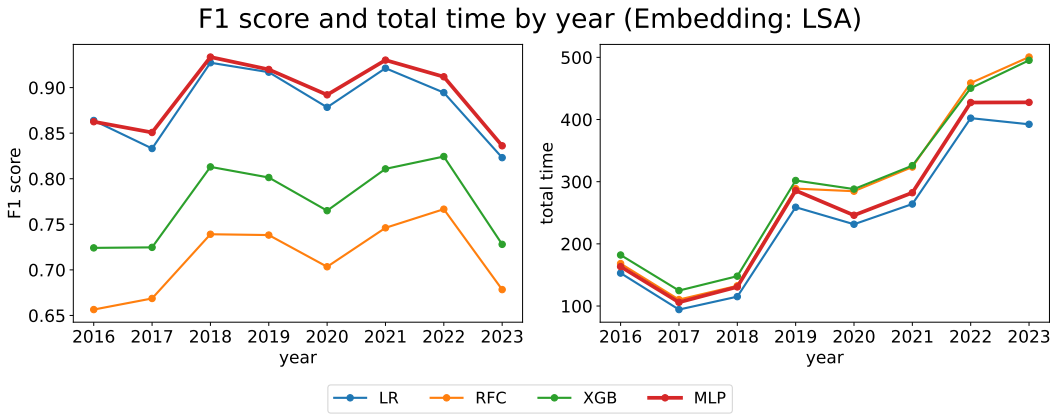


Figure 7: Comparison of F1 score and total time across four classification models using LSA embedding over eight years.

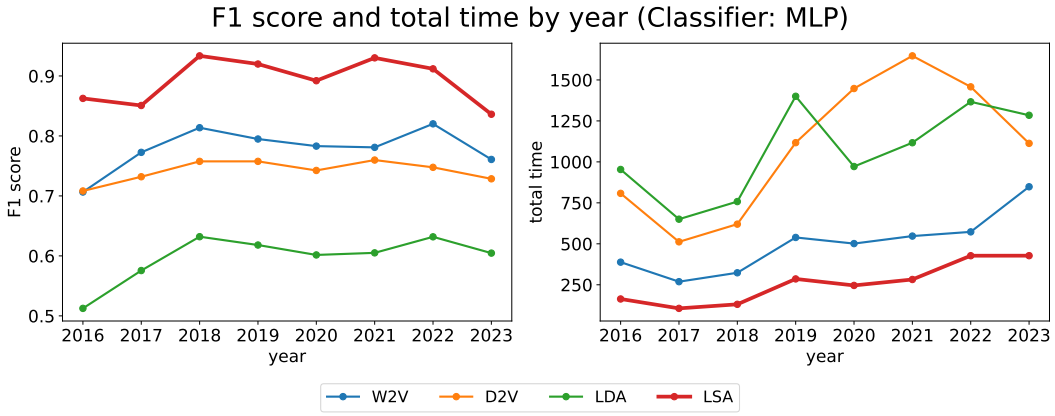


Figure 8: Comparison of F1 score and total time across four embedding methods using MLP classifier over eight years.

costs.

Figure 9 shows the F1 scores for different keyword counts from 5 to 15 over eight years. Dotted lines represent the F1 scores for each year, while the solid line indicates the overall average across all years for each keyword count. The results reveal that classification performance is highest when using only five keywords. This outcome likely occurs because fewer keywords reduce ambiguity, leading to clearer distinctions between classes and higher

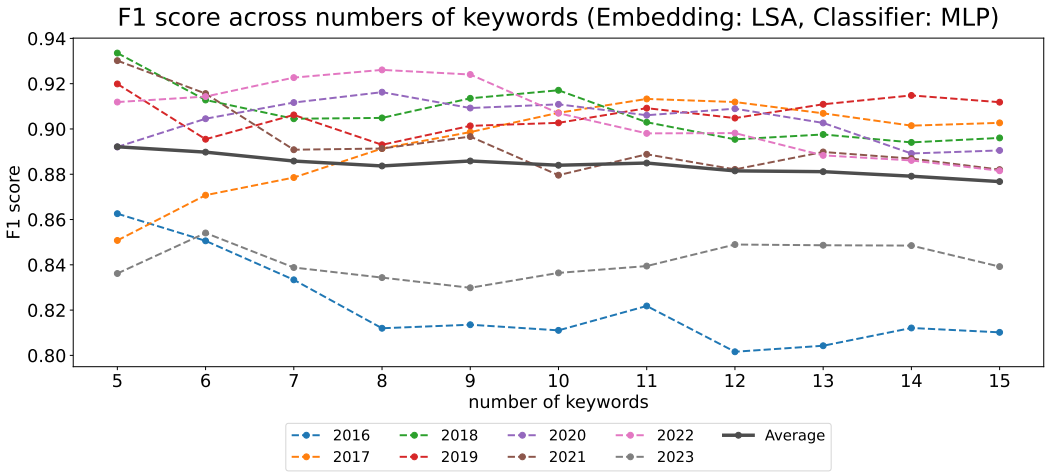


Figure 9: Average F1 scores for number of keywords ranging from 5 to 15 over eight years and overall average for all years.

F1 scores. As the number of keywords increases, F1 scores gradually decrease, suggesting that additional keywords may add noise that affects classification accuracy.

Chapter 6

Conclusion

In this study, we proposed an explainable paper classification system that balances high accuracy with intuitive explanations, leveraging LSA and SHAP. The system outperformed nineteen combinations of embeddings and classifiers across three datasets and provided enhanced interpretability at corpus, document and word levels.

Despite its effective classification performance and interpretability, LSA can sometimes result in oversimplifications and misinterpretations due to its reliance on linear assumptions. Recent advancements in natural language processing, such as transformer-based pretrained models like BERT, have demonstrated superior capabilities for complex text interpretation. However, in our comparative analysis, the proposed LSA-MLP combination outperformed the BERTopic-MLP combination, suggesting that simpler models can still excel under specific circumstances.

The significance of this system spans multiple perspectives, offering valuable support to various stakeholders. For journal editors, it streamlines the editorial decision-making process by helping determine whether submitted papers align with the journal's scope. For researchers, it serves as a practical tool for identifying the most appropriate journals for paper submission, thereby increasing the likelihood of publication success. Additionally, for readers, the system enhances the user experience by leveraging keywords for accurate topic identification, enabling quick and efficient retrieval of relevant articles.

Future research will focus on expanding the applicability of this system to diverse datasets beyond the nanomaterial domain, including interdisciplinary and multilingual datasets. Investigating the integration of transformer-based models like BERT or GPT for improved

topic modeling and classification accuracy is another promising avenue. Additionally, optimizing computational efficiency to handle real-time or large-scale applications, such as dynamic updates in journal databases or large repositories, will be a key area of development.

References

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues*, 9(5):272.

Alicioglu, G. and Sun, B. (2022). A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520.

Alippi, C. and Storti-Gajani, G. (1991). Simple approximation of sigmoidal functions: realistic design of digital neural networks capable of learning. In *IEEE International Sympoisum on Circuits and Systems*, pages 1505–1508. IEEE.

Ayoub, J., Yang, X. J., and Zhou, F. (2021). Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58(4):102569.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Bunk, S. and Krestel, R. (2018). Welda: Enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 293–302.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chowdhury, S. and Schoen, M. P. (2020). Research paper classification using supervised

machine learning techniques. In *Intermountain Engineering, Technology and Computing*, pages 1–6. IEEE.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Ech-Chouyyekh, M., Omara, H., and Lazaar, M. (2019). Scientific paper classification using convolutional neural networks. In *International Conference on Big Data and Internet of Things*, pages 1–6.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jindal, R., Malhotra, R., and Jain, A. (2015). Techniques for text classification: Literature review and current trends. *Webology*, 12(2).

Keya, K. N., Papanikolaou, Y., and Foulds, J. R. (2022). Neural embedding allocation: Distributed representations of topic models. *Computational Linguistics*, 48(4):1021–1052.

Kherwa, P. and Bansal, P. (2017). Latent semantic analysis: an approach to understand semantic of text. In *International Conference on Current Trends in Computer, Electrical, Electronics and Communication*, pages 870–874. IEEE.

Kim, B., Park, J., and Suh, J. (2020). Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134:113302.

Kim, D., Seo, D., Cho, S., and Kang, P. (2019). Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information sciences*, 477:15–29.

Kim, S.-W. and Gil, J.-M. (2019). Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9:1–21.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirasich, K., Smith, T., and Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3):9.

Kowsari, K., Brown, D. E., Heidarysafa, M., Jafari Meimandi, K., , Gerber, M. S., and Barnes, L. E. (2017). Hdltex: Hierarchical deep learning for text classification. In *IEEE International Conference on Machine Learning and Applications*. IEEE.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR.

Li, C., Zhan, G., and Li, Z. (2018). News text classification based on improved bi-lstm-cnn. In *International Conference on Information Technology in Medicine and Education*, pages 890–893. IEEE.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Mustafa, G., Usman, M., Yu, L., Afzal, M. T., Sulaiman, M., and Shahid, A. (2021). Multi-label classification of research articles using word2vec and identification of similarity threshold. *Scientific Reports*, 11(1):21900.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Nguyen, T. H. and Shirai, K. (2013). Text classification of technical papers based on text segmentation. In *International Conference on Applications of Natural Language to Information Systems*, pages 278–284. Springer.

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., and Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588.

Ranjan, M. N. M., Ghorpade, Y., Kanthale, G., Ghorpade, A., and Dubey, A. (2017). Document classification using lstm neural network. *Journal of Data Mining and Management*, 2(2):1–9.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Schopf, T., Braun, D., and Matthes, F. (2022). Evaluating unsupervised text classification:

zero-shot and similarity-based approaches. In *International Conference on Natural Language Processing and Information Retrieval*, pages 6–15.

Tjoa, E. and Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813.

Van Linh, N., Anh, N. K., Than, K., and Dang, C. N. (2017). An effective and interpretable method for document classification. *Knowledge and Information Systems*, 50:763–793.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Xu, H., Wang, W., Liu, W., and Carin, L. (2018). Distilled wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Yao, T., Zhai, Z., and Gao, B. (2020). Text classification model based on fasttext. In *IEEE International Conference on Artificial Intelligence and Information Systems*, pages 154–157. IEEE.

Zhou, C., Sun, C., Liu, Z., and Lau, F. (2015). A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.

국문 초록

토픽모델링과 XAI를 활용한 설명 가능한 논문 분류 시스템에 관한 연구

성신여자대학교
일반대학원
통계학과
신나경

정확도와 해석 가능성은 문서 분류 시스템의 두 가지 핵심 요소이다. 정확도는 분류기가 보지 못한 데이터를 얼마나 잘 예측할 수 있는지를 평가하며, 해석 가능성은 모델의 작동 방식을 인간이 얼마나 쉽게 이해할 수 있는지와 각 데이터가 특정 레이블로 할당된 이유를 설명하는 능력을 의미한다. 효과적인 분류 시스템은 높은 정확도를 유지함과 동시에 사용자에게 직관적이고 포괄적인 통찰을 제공하여 의사결정을 지원해야 한다. 본 연구는 토픽 모델링과 설명 가능한 인공지능 기법을 결합한 새로운 설명 가능한 논문 분류 시스템을 제안한다. 본 시스템은 토픽 모델링을 위해 잠재 의미 분석을 활용하며, 분류 결과의 투명성과 이해도를 높이기 위해 SHAP(Shapley additive explanations)를 적용한다. 본 시스템은 세 가지 주요 해석 수준(말뭉치 수준, 문서 수준, 단어 수준)에서 분류 결과의 해석을 제공한다. 본 시스템의 유효성은 Web of Science 데이터셋을 사용하여, 특히 나노소재 분야를 중심으로 검증되었으며, 추가로 Semantic Scholar 데이터베이스의 대규모 데이터셋을 활용하여 성능을 평가하였다.

핵심용어 : 논문 분류, 토픽 모델링, 잠재 의미 분석, 설명 가능한 인공지능, 샤플리 값

감사의 글 (Acknowledgements)

석사 과정과 학위논문을 마무리하며 감사의 글을 쓰니 감회가 새롭습니다. 대학원 생활은 학문적 성장뿐만 아니라 다양한 경험과 소중한 인연을 남겨준 뜻깊은 시간이었습니다. 이 감사함을 이 지면을 통해 전하고자 합니다.

가장 먼저, 지도교수님이신 정호현 교수님께 진심으로 감사드립니다. 학부 연구생으로 연구의 첫걸음을 내디딜 때부터 석사 졸업을 앞둔 지금까지 제 부족한 점을 세심하게 지도해주시고 언제나 귀 기울여 주셨습니다. 교수님의 따뜻한 격려와 조언 덕분에 한 걸음 더 성장할 수 있었습니다. 대만에서 값진 경험의 기회를 주신 Academia Sinica의 Frederick 선생님께도 깊이 감사드립니다. 날카롭고 의미 있는 조언 덕분에 연구를 한층 더 발전시킬 수 있었습니다.

여러 가지 경험을 할 수 있도록 좋은 기회를 주신 이성건 교수님, 대학원 생활 전반을 세심하게 돌봐주신 박만식 교수님, 학교 선배로서 늘 따뜻하게 이해해 주시고 든든한 격려를 보내주신 박희원 교수님, 냉철하고 객관적이지만 따뜻한 조언을 아끼지 않으신 박성오 교수님, 항상 모든 질문에 친절하게 답해주시며 열정적으로 가르쳐주신 김동하 교수님, 대학원 생활뿐만 아니라 인생에 있어서도 많은 조언을 해주신 박관영 교수님, 찾아뵐 때마다 유익한 말씀을 들려주신 최태화 교수님, 늘 환하게 웃으며 맞아주신 신준호 교수님께 깊은 감사를 드립니다.

같은 길을 걷고 있는 사람이 있다는 사실이 제게 얼마나 큰 힘이 되었는지 모릅니다. 자랑스러운 SDM Lab 구성원을 비롯한 성신여대 통계학과 선후배와 동기 여러분께 감사드립니다. 대학 입학부터 대학원 졸업까지 저의 성신을 함께한 서연이, 기쁜 일과 슬픈 일을 함께 나누며 공감해 준 든든한 윤진이, 밤새 공부하며 서로 의지한 매운 음식 메이트 수지, 힘든 여정을 함께 걸어준 동갑 서영이와 항상 저를 귀여워해 준 혜민 언니, 함께 졸업하게 되어 더욱 뜻깊습니다. 연구실 적응을 도와준 졸업 선배 지우, 세리, 윤아 언니, 큰 즐거움을 준 지혜 언니, 진주, 주이, 경민이와 914호에서 함께 동고동락한 민서에

게도 고마움을 전합니다. 앞으로 대학원 생활을 시작할 세영이와 미영이도 응원합니다. 그리고 대만에서 함께 연구하며 많은 힘이 되어준 Vincent, Jennifer, Gigi, Alice와 다른 인턴들에게도 고마움을 전합니다.

사랑하는 가족들에게 깊은 감사를 전합니다. 항상 저를 믿어주시고 아낌없는 사랑과 응원을 보내주신 부모님께 진심으로 감사드립니다. 부모님의 넘치는 사랑 덕분에 대학원 생활을 무사히 마칠 수 있었습니다. 항상 저를 믿고 따르는 두 동생 지현, 영준에게도 고마운 마음을 전합니다. 동생들에게 의지와 본이 되는 사람이 되겠습니다. 큰 손녀가 최고라며 언제나 꼭 안아주시는 외할머니, 제가 좋아하는 음식을 가득 준비해 주신 친할머니, 조카 걱정에 조언을 아끼지 않으신 이모를 비롯한 친척 가족분들 모두 사랑하고 감사합니다. 지면에 다 담지 못했지만, 언제나 힘이 되어준 친구들에게도 고마운 마음을 전합니다.

이 소중한 경험과 사랑을 마음에 새기며, 겸손하고 감사한 마음으로 앞으로도 열심히 살아가겠습니다. 감사합니다.