Journal of Information Systems Engineering and Management
2025, 10(27s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Classification of Arabic Geographical Research Papers Using Machine Learning Techniques: A Comparative Analysis of TF-IDF and Word2Vec

Miaad Raisan Khudhair[1], Sarah Mohammed Abdulla[1], Iman Qays Abduljaleel[2], Zaid Ameen Abduljabbar[2,3], Vincent Omollo Nyangaresi[4,5], Ali Hasan Ali[6,7,8]

[1] *Directorate of Education in Basra Governorate of Basrah, Basrah, Iraq*
[2] *Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq*
[3] *Department of Business Management, Al-Imam University College, Balad 34011, Iraq*
[4] *Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, Bondo 40601, Kenya.*
[5] *Department of Applied Electronics, Saveetha School of Engineering, SIMATS, Chennai, Tami lnadu 602105, India*
[6] *Department of Mathematics, College of Education for Pure Sciences, University of Basrah, Basrah, 61004, Iraq.*
[7] *Technical Engineering College, Al-Ayen University, Thi-Qar 64001, Iraq.*
[8] *Institute of Mathematics, University of Debrecen, Pf. 400, H-4002 Debrecen, Hungary.*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 15 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

The classification of Arabic geographical research papers presents a unique challenge due to linguistic complexities and the absence of standardized datasets. In this study, we introduce a novel approach by creating a new dataset, comprising Arabic texts extracted from geographical research papers including research files, abstracts and geographical categories (human or physical geography). After preprocessing and text cleaning, TF-IDF and Word2Vec were employed as feature extraction techniques. Four machine learning models were tested: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest. Experimental results demonstrated that SVM a The classification of Arabic geographical research papers presents a unique challenge due to linguistic complexities and the absence of standardized datasets. In this study, we introduce a novel approach by creating a new dataset, comprising Arabic texts extracted from geographical research papers including research files, abstracts and geographical categories (human or physical geography).

After preprocessing and text cleaning, TF-IDF and Word2Vec were employed as feature extraction techniques. Four machine learning models were tested: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest. Experimental results demonstrated that SVM achieved the highest accuracy (84%) with TF-IDF while Random Forest performed best (81%) with Word2Vec effectively leveraging contextual word relationships. Other models exhibited lower performance underscoring the crucial role of a strategic selection of feature extraction techniques suitable for Arabic text classification.

The results underscore the importance of choosing the right feature extraction methods and the suitable model to improve classification accuracy for Arabic geographical studies. Moreover, the study paves the way for developing advanced models through deep learning and natural langua The classification of Arabic geographical research papers presents a unique challenge due to linguistic complexities and the absence of standardized datasets. In this study, we introduce a novel approach by creating a new dataset, comprising Arabic texts extracted from geographical research papers including research files, abstracts and geographical categories (human or physical geography).

After preprocessing and text cleaning, TF-IDF and Word2Vec were employed as feature extraction techniques. Four machine learning models were tested: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest. Experimental results demonstrated that SVM achieved the highest accuracy (84%) with TF-IDF while Random Forest performed best (81%) with Word2Vec effectively leveraging contextual word relationships. Other models exhibited lower performance underscoring the crucial role of a

strategic selection of feature extraction techniques suitable for Arabic text classification.

The results underscore the importance of choosing the right feature extraction methods and the suitable model to improve classification accuracy for Arabic geographical studies. Moreover, the study paves the way for developing advanced models through deep learning and natural language processing-based approaches.ge processing-based approaches achieved the highest accuracy (84%) with TF-IDF while Random Forest performed best (81%) with Word2Vec effectively leveraging contextual word relationships. Other models exhibited lower performance underscoring the crucial role of a strategic selection of feature extraction techniques suitable for Arabic text classification. The results underscore the importance of choosing the right feature extraction methods and the suitable model to improve classification accuracy for Arabic geographical studies. Moreover, the study paves the way for developing advanced models through deep learning and natural language processing-based approaches.

Keywords: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Classification Research Papers, TF-IDF, Word2Vec

INTRODUCTION

In recent years, the geography field has acquired greater significance since global concerns are basic to comprehending spatial elements, how individuals communicate with the environment, and normal asset administration [1]. Water deficiencies, desertification, quick advancement, and political and social elements that are affected by specific social and natural components are the pressing concerns that geographic inquiries within the Arab world handle. These concerns, such as climate change, urbanization, and resource management, necessitate nuanced and location-specific answers [2].

Considering these complexities, territorial geological investigation offers crucial viewpoints on maintainable improvement, Arab-specific policymaking, and natural administration [3].

Even in spite of the fact that inquiries about the geography of the Arabic dialect have made impressive commitments, one major deterrent to data availability is the nonattendance of a common categorization framework for investigating distributions on this subject [4]. Inquire about within research in the Arabic language dialect frequently needs a classification organization that encourages precise writing appraisal, comparison, and cross-regional examinations in differentiation from other dialects where careful categorization frameworks are more created [5]. This makes it troublesome for analysts to discover data on less considered geographic districts, inquire about patterns, or get germane paper [6].

This study looks to bridge this hole by creating a classification system custom-made to inquire about papers in the geographical Arabic dialect. This strategy will help scholastics better comprehend the investigative environment, recognize understudied issues, and recognize repeating topics and regions of intrigue by systematically categorizing past works. In so doing, the classification of such work signals advanced engagement, enabling across-linguistic and -regional comparisons and ultimately enhancing the comparative import of Arabic dialect geographic inquiry in the international academic landscape. Additionally, this poses a great challenge for text classification in the field of Arabic-language geographical research papers as a result of the lack of publicly available datasets. To fill this gap, this study presents a new dataset that was created by systematically extracting the abstracts from Arabic geographical research papers and assigning these abstracts to two main groups: human geography and physical geography. This dataset provides a structured and reliable resource for training and evaluating machine learning models and lays a foundation for future research in Arabic text classification. Consequently this dataset has become a fundamental tool for training and evaluating machine learning models. The following section discusses the challenges associated with Arabic text classification and the motivation behind this study.

PROBLEM STATEMENT

A number of challenges have arisen in the midst of the severe categorization of Arabic dialect geographical examination articles, chief among them being the deficiency of easily open arrangement data for Arabic substance classification. Due to a collection of etymological and innovative issues, the Arabic tongue requests almost dispersions as often as possible. It requires a noteworthy course of action separate from other lingos with a wealth of resources.

Firstly, a huge number of the documents fundamental for this examination either included substances that required to be physically removed or were not opened in advance. It was troublesome to extricate a usable substance from these records, particularly when overseeing with unusual literary fashion sorts that have an impact on modified substance affirmation algorithms' precision. Moreover, the importance of Arabic dialect diacritical markings made content preparation indeed more troublesome since dialect models regularly apply or translate them inconsistently; coming about in contrasts in how the content is spoken. Furthermore, instead of carefully organized content, a noteworthy rate of the considered articles were scanned documents, numerous of which had destitute quality or vague determination. Since OCR frameworks frequently have inconveniences with Arabic script and moo filter quality, content extraction utilizing OCR was exceptionally challenging. Counting these troubles, a few original copies included decorative or aesthetic components around the content, which made exact content extraction and categorization much more troublesome. Collectively, these issues appear how challenging it is to create an orderly and reliable classification framework for Arabic geographical investigate papers, highlighting the require for advanced preprocessing strategies and customized arrangements to handle and normalize Arabic content information. To overcome these challenges, the study also introduces a new dataset of Arabic-language geographical research papers constructed by extracting and categorizing abstracts into human and physical geography. This dataset serves as a critical resource for training and evaluating machine learning models in the context of Arabic text classification.

OBJECTIVES AND CONTRIBUTIONS OF THE STUDY

This study aims to classify articles into two main branches—Physical geography and human geography. It attempts to provide an organized categorization system for geographical research papers written in Arabic. The research aims to accomplish the following particular goals in order to fulfill this overall objective:

1. Introducing a Novel Arabic Geographical Research Dataset, we aim to create a new dataset by systematically collecting abstracts of Arabic studies in geographic research. This unique dataset, which fills a critical gap in Arabic language resources for geographical text classification, consists of three key fields: the file name, the research abstract, and the geographical category.

2. Develop preprocessing techniques for Arabic text design and apply preprocessing techniques that tackle special cases such as diacritic normalization, font variation, lousy quality images, and ornaments in Arabic text. These special cases pose unique challenges in Arabic text processing, and our study provides effective solutions to overcome them. Data Preprocessing: This step is crucial when dealing with PDF files — these will be first converted to enhanced images, and then OCR performed on them for text extraction.

3. Evaluate Feature Extraction Techniques for Arabic Geographical Texts Investigate and compare the performance of different feature extraction techniques, specifically TF-IDF and Word2Vec, to identify the most effective approach for Arabic geographical text classification. This evaluation involves analyzing how these techniques influence model performance, considering accuracy, precision, recall, and F1-score.

4. Developing a Comprehensive Classification Framework, we have designed a thorough categorization scheme especially for research geographical in Arabic language geographical publications, classifying them into two categories: Human Geography (such as population studies, urban planning, and cultural geography) and Physical Geography (such as climate, hydrology, and geomorphology).Our study aims to accomplish these goals through four diverse classification models that provide a variation in analysis. The models include Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest. Since each model takes a different approach to text classification, we can also compare linear vs non-linear performance when using different features extraction methods such as TF-IDF and Word2Vec.

1. Naïve Bayes was selected because of its simplicity and performance. Its performance depends on the feature independence assumption [7]. Therefore, it is suitable for linear representations like TF-IDF. Since it has the potential to reach high speed as an all-in-one model, it can work as a good baseline model to evaluate the performance of different models.

2. Logistic Regression: This is a linear model that performs very well on high-dimensional text data. This method was chosen due to good compatibility with TF-IDF representations and the interpretability of coefficients, which is especially valuable for applications requiring clear and transparent decision-making [8].

3. Support Vector Machine (SVM): A powerful classifier that works well for high-dimensional text data. SVM aims to create hyper-planes in the feature space that separate the various classes with maximum margin. This quality renders SVM particularly efficient when assessing Arabic geographical texts, which frequently utilize technical language and intricate linguistic frameworks [9].

4. Random Forest was selected for its efficient management of non-linear relationships. It uses an ensemble approach, weaving together predictions from numerous decision trees with bagging and majority voting. This helps to make the model more stable and reduces the risk of overfitting. Word2Vec embedding significantly improves the performance of Random Forest due to the contextual parameters embedded in the random word [10].

These accompanying models were chosen from an assortment of classification utilities to guarantee a comprehensive and exact assessment of geographic content highlights. Each show addresses a special perspective of investigation, allowing a deeper understanding of the data and accordingly facilitating the determination of the foremost proficient demonstration for future applications in this field.

LITERATURE REVIEW

The literature review for this study, a crucial component, delves into three main areas: existing classification systems in geographical research, the specific challenges associated with Arabic text processing, and prior studies on Arabic research publications. These insights are fundamental for understanding the methods and challenges pertinent to developing a classification framework for Arabic language geographical research studies.

1. Classification Systems in Geographical Research

Geographical research encompasses a wide range of fields that often require distinct classification approaches. Studies have appeared that partitioning geographical inquiries about into two primary branches Physical geographical and Human geographical is viable in organizing investigated topics, as each department addresses in a general sense distinctive sorts of spatial connections and techniques (Nagarale et al., 2022) [11]. Physical geography regularly incorporates considers common highlights such as climate, landforms, and biological systems, whereas Human geography analyzes human exercises and societal designs, counting urban advancement, populace elements, and social scenes (Carol. P. Harden, 2014)[12]. These categories serve as a foundational demonstration for classifying research geographically in the Arabic language, where comparable topical divisions can assist in recognizing patterns and research gaps inside the Arabic language setting.

2. Challenges of Arabic Text Processing

Arabic content presents special challenges in classification and analysis due to the language's unmistakable script, morphology, and composing traditions. Ponders on Arabic Normal Dialect preparation (NLP) highlight issues such as the complexity of diacritical marks (harakat), assorted textual style sorts, and varieties in composed dialects (Khaled et al., 2018)[13]. Diacritics can altogether change the meaning of words, driving potential blunders in content investigation, especially in optical character acknowledgment (OCR) and dialect modeling assignments (Mahmoud et al., 2023)[14]. Furthermore, many Arabic documents exist as low-quality scanned copies, which makes text extraction difficult. Numerous Arabic documents exist as low-quality scanned duplicates, which makes content extraction troublesome. Researchers have investigated the utilize of progressed OCR methods and pre-processing strategies to address these issues, but precise OCR for Arabic remains a creating field (Safiullah et al., 2023) [15].

3. Studies on Arabic Research Publication Analysis

Few studies have particularly tended to the organization and availability of Arabic inquiries about distributions. A survey by Jamal El-Ouahi (2022)[16] highlights the requirement for organized classification systems inside Arabic scholarly distributions to progress to quotation permeability. Arabic investigative papers, particularly within the social sciences and humanities, are regularly underrepresented in worldwide databases, somewhat due to conflicting classification and metadata measures.

Bilal Alharbi. (2021)[17] inspected classification approaches for Arabic scholarly writing, emphasizing that custom-made classification systems might incredibly move forward investigate discoverability and energize intrigue collaboration. This ponder builds on these bits of knowledge by proposing a system particular to Arabic geological

investigations, pointing to cultivating less demanding get to and perceivability of Arabic considers within the worldwide geographical field.

METHODS

In this study, we employ a systematic and thorough method to produce a new dataset of Arabic texts extracted from geographical research papers. The dataset comprises three features: file name, research abstract and geographical category. Our objective is to construct a machine learning model that can categorize papers into two categories based on their focus on either human or physical geography. The methodology we follow is comprehensive involving a step-by-step process that includes dataset creation, text preprocessing, feature extraction and model classification as detailed in the following sections.

To provide a clear and comprehensive overview of the classification workflow, Figure 1 illustrates the key stages involved in the study. The process begins with data collection and OCR processing followed by text preprocessing, feature extraction using TF-IDF and Word2Vec and finally a transparent model classification and evaluation process.

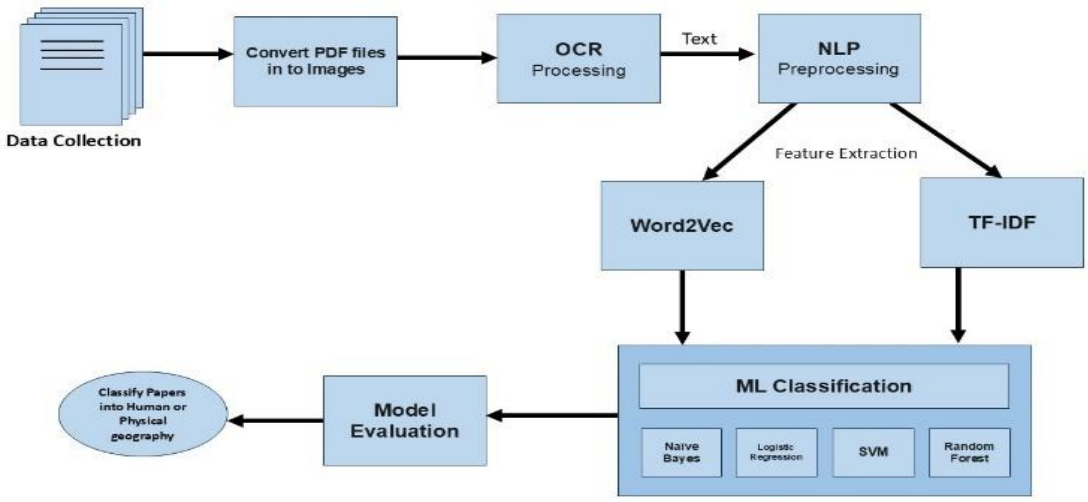


Figure 1: Flowchart of the Text Classification Process

I. Dataset Creation

No specific dataset in Arabic concerning geographic research has been developed before this work, which can be summarized in the following points:

A. Data Collection

This meticulous process resulted in the collection of 400 research papers from various well-known sources like Google Scholars. Equal representation was maintained in training bias data, where the 200 papers of human geography and 200 papers of physical geography were used in balanced data.

B. Text Extraction from PDF Files:

The objective of this step is primarily to extract the abstracts from the obtained research papers in order to construct a reliable dataset. There were major obstacles because directly reading text from PDF files would yield unreadable or incomprehensible results. These problems included improper character encoding, diacritics (tashkeel) problems, right-to-left alignment errors, unsupported fonts, decorative elements added around the words, and horizontal lines attached to the text. To address these matters, the PDF files had to be transitioned into images, and Optical Character Recognition (OCR) techniques were employed. This process involved three main stages: converting the PDF files into images, preprocessing the images for better quality, and finally, using OCR for character recognition. Here is the process one by one:

1- Convert PDF Pages to Images: We utilized pdf2image [18] Python library to transform PDF pages into images, which offers flexibility and the possibility to customize parameters, such as DPI (Dots per Inch). DPI is the

number of dots in an image per linear inch, which directly affects the sharpness as well as detail. In our study, we set a DPI of 600, which considerably improves the performance of OCR and is very efficient for complex scripts like Arabic [19]. Since, we would be most likely to find research's abstract in first pages in a document, we put a restriction on the conversion process of only the first five pages of each PDF. The methodology ensures the extraction of the necessary information without wasting time on the data that are not going to be used finally as an output as well as requiring lower computational resources.

2- Image Enhancement: Resolving issues like low resolution, noise and insufficient contrast is critical for improving text recognition accuracy in OCR systems. For this purpose, the OpenCV library, a powerful open-source library for computer vision and image processing, was used to preprocess images [20]. It provides advanced capabilities and efficient image processing making it ideal for this task. The conversion of original RGB images to grayscale at this stage is a step taken to reduce computational complexity by simplifying visual data streams while effectively mitigating color noise [21]. Then, we apply adaptive thresholding to get binary (black-and-white) images from grayscale. This technique works to increase text-background separation by exploiting the pixel intensity distribution of text regions. Its preservation of complex text features and handling of lighting and image quality differences assist in enhancing text readability and OCR text systems accuracy [22]. These targeted improvements were essential to ensure accurate text extraction and strong downstream analysis.

3- OCR Technique: The prepared images were used to extract text after image enhancement was done using the Tesseract OCR works in a methodical procedure that begins with segmenting into various components, such as characters or words, using advanced image segmentation techniques. These parts are processed by pretrained models to identify characters utilizing a database of known characters against each string test. The identified characters are then joined together to form words and sentences. The process ends with the use of error correction methods for misreported characters. It guarantees accurate and secure text extraction by utilizing image optimization and extensive correction techniques.

CSV file Creation:

1. Abstract Retrieval: The principal classification used in this study was to classify the abstract to determine the research paper category. Hence, the abstract is considered the essential part of the extracted text, identified by relying on a comprehensive list of all possible Arabic synonyms for the word "abstract." On the other hand some research papers may include only an introduction instead of an abstract. Therefore we considered this possibility and included alternative words for "introduction" to ensure comprehensive coverage.

2. Saving Extracted Data into CSV File: This step involves taking the extracted abstract or introduction and comparing it with three predefined lexicons: human geography terms, physical geography terms, and intersection terms. We include shared terms because certain keywords may represent both human and physical geography. To classify the extracted text, a counter calculates the frequency of keywords appearing in the abstract within each group (Human geography and Physical geography). We also consider common keyword frequencies and distribute them equally between the two groups to ensure balanced influence. The resulting classification is determined based on the group with the highest cumulative score. Finally, we save the file name, abstract, and its category into a CSV file, preparing the data for further processing before applying the machine learning model.

I. Preprocessing the data

Following a series of steps utilizing Natural Language Processing (NLP) techniques applied to the generated dataset enhances the quality of the extracted text. This improvement supports the training model in classifying the text with high accuracy and efficiency.

In this study the NLTK library [25] was employed to implement various NLP techniques as illustrated below.

Process Name	Objective of the Process
Remove numbers	Usually, numbers in the script do not contribute meaningful value to text analysis, so it is recommended to remove them.
Remove punctuation and special characters	The step of cleaning the text from punctuation and special characters is a critical measure to prevent their impact on the analysis process.

Remove extra whitespaces	To enhance text format, unnecessary spaces are trimmed, whether found at the beginning or the end of the text.
Normalization	Arabic text may contain multiple forms of the same letter. For example, “أ”, “إ”, “ا”,”آ ” all represent the same letter “ا” Normalization is the most suitable solution to reduce unnecessary variations and enhance performance.
Remove stop words	Words like “على“ ”,من,” and similar terms do not hold any significance for the geography category. Therefore, removing them from the text contributes to cleaning it and improving the accuracy of training. The Arabic stop word list used for this process was obtained from the NLTK library.
Stemming	To reduce redundancy in text and unify words derived from the same root, it is beneficial to use stemming, which simplifies the data and enhances the model's comprehension and performance. This trend is prominent in Arabic, a language with a plethora of derivations and formations. For our case, we want to analyze the data in general and not look for correct identification of its exact formats.
Tokenization	Tokenization is an essential stage of text analysis, especially at the word level of text processing, which are used for applications such as finding the most common words, keyword extraction, or the classification of text. This is typically referred to as text pre-processing, and it is one of the first steps in many Natural Language Processing (NLP) techniques. In geography terms, each type is defined by the terms in question, which help dictate their category. Thus, tokenization of words in the text is a pivotal part in the preprocessing stage of text analysis. Tokenization breaks the text into words so that the model can process one word at a time. Word frequency can be computed, making it easier to find patterns.

II. Feature Extraction

Feature extraction techniques are instrumental in effectively classifying Arabic geographical research papers into physical and human geography. Then feature extraction was done on geoliterature using two methods — TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec. They provide unique perspectives on how textual information can be utilized and performed across selected classification models. TF-IDF: TF-IDF [26] is a statistical method for text vectorization, which numerically represents words or \textit{documents} in a \textit{corpus}. It measures the significance of a word in a given document against its presence in the rest of the documents in the corpus. This is accomplished by determining two things:

1. Term Frequency (TF): Shows how often a term is in a document. This is computed as the number of times that the term appears in the document, divided by the total number of terms in the document.

2. Inverse Document Frequency (IDF): A metric that gives greater priority to infrequent terms (i.e., those appearing in a limited number of documents) and low scores to very common words [27]. The IDF is computed as:

$$IDF(t) = \log(\frac{Total\ number\ of\ documents\ in\ the\ corpus}{Number\ of\ documents\ that\ contain\ the\ term\ \boldsymbol{t}})$$

On one hand, Word2Vec is a technique that stands out in creating word embedding from large-scale datasets like text and speech. It generates word vectors from usage frequency with neural networks [28] for example, the skip-gram or continuous bag-of-words (CBOW) models. What sets Word2Vec apart is its unique ability to automatically capture semantic relationships and contextual similarity between words allowing for a better semantic understanding of textual data. This is a feature that is not present in TF-IDF representation. Word vectors, unlike TF-IDF, have the ability to encode fine-grained relationships, capturing analogies (e.g., "king–man+woman = queen") thus making it a very versatile approach for text interpretation and analysis [29]. Note that in this study, we did not use any pre-trained Word2Vec model, but we trained our own Word2Vec model from scratch from our constructed Arabic Geographical research dataset. The training was conducted with the Gensim [30] library and the input was tokenized abstracts. By doing so, the model was able to learn domain-specific relationships between different geographical terms like "climate change," "topographic analysis," and "spatial distribution" which may not have a good representation in more general-purpose pretrained models.

By employing these two separate methods we gained a unified understanding of the text structure. TF-IDF was successful in finding the class-specific terms that often appeared together with each label of geography. At the same time Word2Vec helped to understand the semantics of geography relations and improved the capacity of the model to generalize with datasets in different research topics. The features that were extracted using TF-IDF and Word2Vec were then passed to the four adopted classification models, namely: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest. These models were crucial in measuring classifier performances on diverse sets of geographical research texts thereby validating the effectiveness of our statistical-semantic representation.

Geographic Dataset Classification

The classification task was a careful choice of available machine learning models depending on their theoretical concepts and sensitivity to the complexity of both language and context in Arabic. Then, I fit four models on the data: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest. Both of these models were then applied along with features extracted through TF-IDF and Word2Vec to the dataset in order to identify the patterns, relationships, and characteristics that distinguish human and physical geography abstracts.

The Naïve Bayes model, which is used as a probabilistic approach and is quite simple to compute, is the most commonly used model for the classification of high-dimensional text. The model-like structure of TF-IDF assigns more importance to less frequent terms due to their rarity, and Naive Bayes only assumes feature independence which works well with the sparse nature of TF-IDF. Utilizing GaussianNB from the scikit-learn package, the model was trained on TF-IDF as well as Word2Vec embeddings. The TF-IDF was expected to provide better performance, as it directly represented the importance of the terms, while the Word2Vec embeddings were used to add semantic richness to the model. In contrast, Logistic Regression was chosen for its transparent, interpretable nature and aptitude for handling sparse datasets. The reliance of the model on logistic functions to distinguish between classes based on term importance makes it particularly relevant for Arabic geographical abstracts where context plays a crucial role. Through the use of L2 regularization and a maximum iteration limit of 1000, the model was applied to the dataset after a 60/40 stratified split. Its linear structure was expected to respond favorably to the word frequency patterns captured by TF-IDF, with Word2Vec embeddings potentially providing semantic richness but introducing complexity.

An SVM model was also implemented in the study as it excels at distinguishing overlapping classes by maximizing the margin between decision boundaries. Geographical texts have overlapping terminology— (borders) and (transport) can both appear in human and physical geography contexts. We used the linear kernel of SVC from scikit-learn to cope with the high-dimensional TF-IDF representation, whereas Word2Vec embeddings were used to investigate semantic relationships across the dataset.

Finally, Random Forest was applied to unearth non-linear patterns found in Arabic documents. With 100 decision trees allowed at a maximum depth of 20, its ensemble learning mechanism offered a sound structure to reveal associations in this domain. The decision to use Word2Vec embeddings alongside TF-IDF was informed by the model's potential to identify word co-occurrences that are contextually significant but might be overlooked by term frequency alone.

The synergy between these models and the feature extraction techniques lays the groundwork for the classification process. The subsequent section introduces the performance metrics with the evaluation of their application which is discussed later in the results.

III. Evaluation Metrics

To evaluate the performance of the classification models, this study employs the following key metrics:

1. Accuracy measures the proportion of correct predictions among the total number of predictions:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

While useful for assessing overall model performance, it may be less informative when the dataset is imbalanced [31].

2. Precision evaluates the proportion of correctly predicted positive samples among all predicted positives:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

This metric is critical in scenarios where minimizing false positives is essential [32].

3. Recall calculates the ability of the model to correctly identify all relevant positive samples:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

It is particularly important in applications where false negatives are costly [33].

4. F1-Score: which is the harmonic mean of precision and recall, is a useful metric when both are equally important [34]:

$$F1 - score = 2.\frac{Precision.Recall}{Precision + Recall}$$

5. Confusion Matrix: breaks down the accurate and incorrect predictions for each class, helping to provide insights into the performance of a classification model using true positives, true negatives, false positives, and false negatives [35].

This in conjunction ensures a thorough understanding of model performance, both in terms of how accurately it is overall, as well as the trade-off between precision and recall, which is needed to go deeper into the analysis of geo text classification.

RESULTS AND DISCUSSION

The study evaluated four classification models, a Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest using two feature extraction techniques: TF-IDF and the main aim was to find the best possible extraction and classification model for the classification of Arabic geographical research papers into human geography and physical geography. The models were evaluated through a comprehensive process using several performance metrics (accuracy, precision, recall and F1-score) to ensure a thorough and robust evaluation. The classification results demonstrated clear differences depending on the feature extraction method. TF-IDF-based models generally exhibited superior performance with linear classifiers, whereas Word2Vec performed better with non-linear classifiers, particularly Random Forest. The confusion matrices provide deeper insights into the misclassification patterns as illustrated in Figures 2 and 3. Below are the confusion matrices illustrating the misclassification patterns for each model.

Performance of Models with TF-IDF The results indicate that SVM, with its impressive 84% accuracy, emerged as the most effective model when using TF-IDF features. This aligns with expectations since SVM is well-suited for high-dimensional sparse text data making it an ideal classifier for term-frequency representations. Logistic

Regression followed closely with 82% accuracy confirming its effectiveness as a linear classifier for numerical word representations.

In contrast Naïve Bayes showed moderate performance (76% accuracy), reflecting its strong dependence on the assumption of feature independence which may not hold in complex geographical texts. Random Forest, while slightly outperforming Naïve Bayes (78% accuracy) exhibited instability particularly in distinguishing between overlapping terminologies, as indicated by its lower precision and recall scores.

Confusion Matrix Insights for TF-IDF Models The confusion matrices for TF-IDF models (Figures 2a–2d) highlight key misclassification patterns:

- SVM (Figure 2a) exhibited the lowest misclassification rate with only five misclassified human geography papers and 17 misclassified physical geography papers confirming its effectiveness.
- Logistic Regression (Figure 2b) showed similar performance to SVM, with just two errors in human geography and 22 in physical geography.
- Naïve Bayes (Figure 2c) displayed nine errors in human geography and 23 in physical geography reflecting its limited ability to handle contextual similarities.
- Random Forest (Figure 2d) had the highest misclassification rate for physical geography (24 errors) indicating difficulty distinguishing between domain-specific words.

Figure 2a: Confusion Matrix for SVM Using TF-IDF

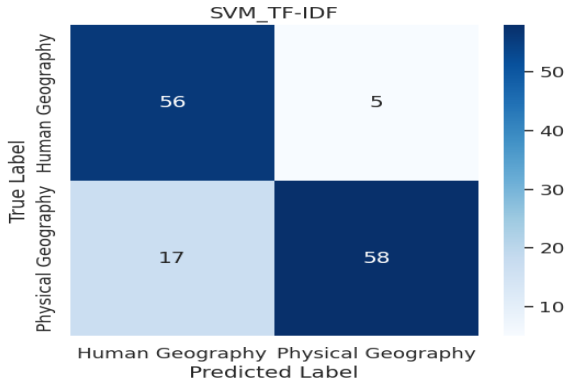


Figure 2b: Confusion Matrix for Logistic Regression Using TF-IDF

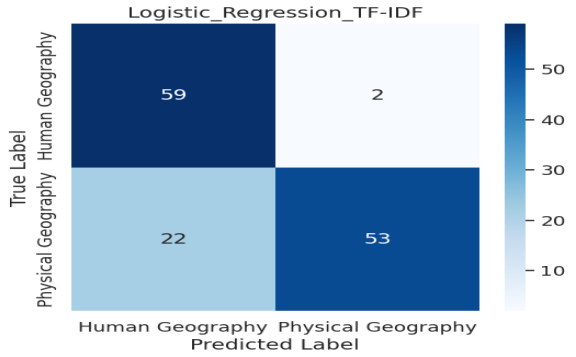


Figure 2c: Confusion Matrix for Naïve Bayes Using TF-IDF

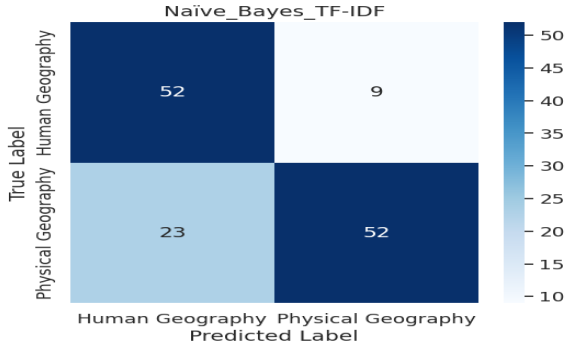
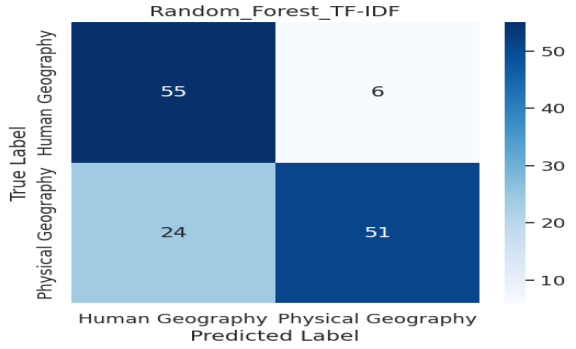


Figure 2d: Confusion Matrix for Random Forest Using TF-IDF



Comparison of Model Performance The following figure presents a comparison of model accuracy between TF-IDF and Word2Vec.

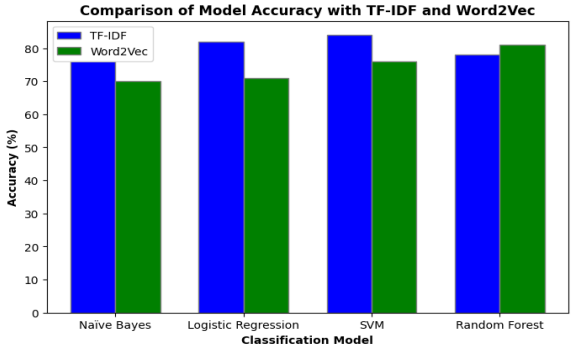


Figure 4: Comparison of Model Accuracy with TF-IDF and Word2Vec

The following tables present a comprehensive comparison of the precision, recall, F1-score and accuracy for each model using TF-IDF and Word2Vec.

Table 1. Classification Performance of Machine Learning Models Using TF-IDF Features

Model	Category	Precision	Recall	F1-Score	Accuracy
Naive Bayes	Human Geography	0.69	0.85	0.76	0.76
	Physical Geography	0.85	0.69	0.76	
Logistic Regression	Human Geography	0.73	0.97	0.83	0.82
	Physical Geography	0.96	0.71	0.82	
SVM	Human Geography	0.77	0.92	0.84	0.84
	Physical Geography	0.92	0.77	0.84	
Random Forest	Human Geography	0.70	0.90	0.79	0.78
	Physical Geography	0.89	0.68	0.77	

Table 2 . Classification Performance of Machine Learning Models Using word2Vec Features

Model	Category	Precision	Recall	F1-Score	Accuracy
Naive Bayes	Human Geography	0.72	0.54	0.62	0.70
	Physical Geography	0.69	0.83	0.75	
Logistic Regression	Human Geography	0.65	0.79	0.71	0.71
	Physical Geography	0.79	0.65	0.72	
SVM	Human Geography	0.73	0.72	0.73	0.76
	Physical Geography	0.78	0.79	0.78	
Random Forest	Human Geography	0.75	0.85	0.80	0.81
	Physical Geography	0.87	0.77	0.82	

Key Findings

• SVM was the best model with TF-IDF (84%) while Random Forest was the best with Word2Vec (81%).

• TF-IDF was most effective with linear classifiers (SVM & Logistic Regression) while the adaptability of Word2Vec to non-linear classifiers (Random Forest) was a fascinating discovery.

• Naïve Bayes struggled significantly with Word2Vec, underscoring the need for further research to understand its limitations with contextual embeddings.

• Random Forest was inconsistent with TF-IDF but excelled with Word2Vec emphasizing the importance of feature extraction selection.

CONCLUSION

This study, with its unique approach, aims to classify geographical research papers into two main branches: Human and Physical geography using four machine learning algorithms. The construction of an original dataset a significant undertaking consisting of Arabic texts containing geographical research publications played a crucial role in the analysis. Feature extraction was performed using TF-IDF and Word2Vec. The results showed that TF-IDF combined with SVM produced the best accuracy of 84% while Word2Vec with Random Forest yielded 81% accuracy. The key takeaway from these results is the importance of selecting feature extraction techniques based on the dataset properties and the chosen model. For future work it is essential to use more sophisticated models such as BERT or other deep learning-based models to further enhance classification performance. Expanding the dataset and incorporating domain-specific knowledge can significantly contribute to improving results and increasing the efficiency of models in geographical text applications.

REFRENCES

[1] A. Srivastava and R. Maity, "Assessing the Potential of AI–ML in Urban Climate Change Adaptation and Sustainable Development," Sustainability, vol. 15, no. 23, Article 16461, 2023

[2] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," arXiv preprint arXiv:1702.07835, 2017.

[3] N. Castree, A. Rogers, and R. Kitchin, A Dictionary of Human Geography. Oxford University Press, 2013.

[4] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. Abdel Monem, "Challenges in Arabic Natural Language Processing," in Computational Linguistics, Speech and Image Processing for Arabic Language, 1st ed., A. Mourad, A. T. Azar, and T. Gaber, Eds. Singapore: World Scientific, 2018, pp. 59–83.

[5] M. Alsaleem, A. Al-Sakran, and A. Alarifi, "Arabic Text Classification: A Review," Procedia Computer Science, vol. 141, pp. 108–114, 2018

[6] M. A. Rahman and M. S. Hossain, "Review on Integrating Geospatial Big Datasets and Open Research Issues," in Proceedings of the IEEE International Conference on Smart Computing and Communications (ICSCC), 2020

[7] Z. Quan and L. Pu, "An improved accurate classification method for online education resources based on support vector machine (SVM): Algorithm and experiment," Educ. Inf. Technol., vol. 28, pp. 8097–8111, 2023.

[8] T. Wu, H. Fan, H. Zhu, C. You, H. Zhou, and X. Huang, "Intrusion detection system combined enhanced random forest with SMOTE algorithm," EURASIP J. Adv. Signal Process., vol. 2022, no. 39, 2022.

[9] X.-Q. Liu, X.-C. Wang, L. Tao, F.-X. An, and G.-R. Jiang, "Alleviating conditional independence assumption of naive Bayes," Statistical Papers, vol. 65, pp. 2835–2863, 2024.

[10] QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms," HECA J. Appl. Sci., vol. 1, no. 2, 2023.

[11] V. Nagarale, S. Anand, and P. Telang, "Research Methods and Techniques in Physical Geography," in Methodological Approaches in Physical Geography, F. B. Mustafa, Ed. Cham: Springer, 2022, pp. 113–126.

[12] C. P. Harden, "The human-landscape system: challenges for geomorphologists," Physical Geography, vol. 35, no. 1, pp. 76–89, 2014.

[13] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. A. Monem, "Challenges in Arabic Natural Language Processing," in Computational Linguistics, Speech and Image Processing for Arabic Language, 2018, pp. 59–83.

[14] M. S. Kasem, M. Mahmoud, and H.-S. Kang, "Advancements and Challenges in Arabic Optical Character Recognition: A Comprehensive Survey," arXiv preprint arXiv:2312.11812, 2023.

[15] S. Faizullah, M. S. Ayub, S. Hussain, and M. A. Khan, "A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges," 2023.

[16] J. El-Ouahi, "The Arabic Citation Index - Toward a better understanding of Arab scientific literature," 2022.

[17] A. Alharbi and M. Lee, "Arabic Text Classification: A Literature Review," in Proc. 2021 Int. Conf. Electr., Comput. Energy Technol. (ICECET), 2021.

[18] "pdf2image: Convert PDF to image with Python," PyPI, 2023. [Online].Available: https://pypi.org/project/pdf2image/
[19] J. E. Tapia, M. Russo, and C. Busch, "Generating Automatically Print/Scan Textures for Morphing Attack Detection Applications," arXiv preprint arXiv:2408.09558, 2024.
[20] P. Mudjirahardjo, "The Effect of Grayscale, CLAHE Image and Filter Images in Convolution Process," Int. J. Adv. Multidiscip. Res. Stud., vol. 4, no. 3, pp. 943–946, 2024.
[21] R. L. Kshetry, "Image Preprocessing and Modified Adaptive Thresholding for Improving OCR," arXiv preprint arXiv:2111.14075, 2021.
[22] OpenCV Documentation, "OpenCV: Open Source Computer Vision Library," OpenCV, 2024.
[23] M. Ponnuru and A. Likhitha, "Image-Based Extraction of Prescription Information using OCR-Tesseract," Procedia Computer Science, vol. 235, pp. 1077-1086, 2024.
[24] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. https://www.nltk.org/
[25] H. D. Abubakar, M. Umar, and M. A. Bakale, "Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec," SLU Journal of Science and Technology, vol. 4, no. 1, pp. 27-33, 2022.
[26] U. Rani and K. Bidhan, "Comparative assessment of extractive summarization: TextRank, TF-IDF and LDA," Journal of Scientific Research, vol. 65, no. 1, pp. 304-311, 2021.
[27] A. S. Menon, R. Sreekumar, and B. J. Bipin Nair, "Character and word level recognition from ancient manuscripts using Tesseract," in 2023 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2023, pp. 1743-1749
[28] I. Q. Abduljaleel and I. H. Ali, “Deep learning and fusion mechanism-based multimodal fake news detection methodologies: A review,” Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 15665–15675, 2024.
[29] C. Allen and T. Hospedales, "Analogies explained: Towards understanding word embeddings," in Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2019.
[30] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," Proc. Natl. Acad. Sci. U.S.A., vol. 116, no. 29, pp. 14516–14525, 2019.
[31] Gensim, "Gensim: Topic Modelling for Humans," PyPI, 2024. [Online]. Available: https://pypi.org/project/gensim/.
[32] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," BioData Mining, vol. 14, no. 1, p. 13, 2021.
[33] M. A. Albahli, M. A. Masood, and M. A. Alzain, "End-to-End Deep Learning Model for Corn Leaf Disease Classification," IEEE Access, vol. 10, pp. 25927–25934, 2022.
[34] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, 2023.
[35] K. Tsagris and A. Tolosana-Delgado, "Variable selection in regression models with compositional covariates," Computational Statistics, vol. 38, no. 1, pp. 95–123, 2023.