Information systems. Computer sciences. Issues of information security
Информационные системы. Информатика. Проблемы информационной безопасности

UDC 025.4.03
https://doi.org/10.32362/2500-316X-2024-12-5-7-16
EDN CBEERK

CC BY

RESEARCH ARTICLE

Automating the search for legal information in Arabic: A novel approach to document retrieval

Kamel S. Jafar [1, @], Ali A. Mohammad [2], Ali H. Issa [3], Alexander V. Panov [1, @]

[1] *MIREA – Russian Technological University, Moscow, 119454 Russia*
[2] *Higher School of Economics, Moscow, 109028 Russia*
[3] *Russian Biotechnological University, Moscow, 125080 Russia*
[@] *Corresponding authors, e-mail: zhafar.k@edu.mirea.ru, panov_a@mirea.ru*

Abstract

Objectives. The retrieval of legal information, including information related to issues such as punishment for crimes and felonies, represents a challenging task. The approach proposed in the article represents an efficient way to automate the retrieval of legal information without requiring a large amount of labeled data or consuming significant computational resources. The work set out to analyze the feasibility of a document retrieval approach in the context of Arabic legal texts using natural language processing and unsupervised clustering techniques.

Methods. The Topic-to-Vector (Top2Vec) topic modeling algorithm for generating document embeddings based on semantic context is used to cluster Arabic legal texts into relevant topics. We also used the HDBSCAN density-based clustering algorithm to identify subtopics within each cluster. Challenges of working with Arabic legal text, such as morphological complexity, ambiguity, and a lack of standardized terminology, are addressed by means of a proposed preprocessing pipeline that includes tokenization, normalization, stemming, and stop-word removal.

Results. The results of the evaluation of the approach using a dataset of legal texts in Arabic based on keywords demonstrated its superior effectiveness in terms of accuracy and memorability. The proposed approach provides 87% accuracy and 80% completeness. This circumstance can significantly improve the search for legal documents, making the process faster and more accurate.

Conclusions. Our findings suggest that this approach can be a valuable tool for legal professionals and researchers to navigate the complex landscape of Arabic legal information to improve efficiency and accuracy in legal information retrieval.

Keywords: search for documents, NLP, Top2Vec, HDBSCAN, Arabic legal documents, word embeddings, cosine similarity

© K.S. Jafar, A.A. Mohammad, A.H. Issa, A.V. Panov, 2024

• **Submitted:** 05.05.2023 • **Revised:** 04.04.2024 • **Accepted:** 11.07.2024

For citation: Jafar K.S., Mohammad A.A., Issa A.H., Panov A.V. Automating the search for legal information in Arabic: A novel approach to document retrieval. *Russ. Technol. J.* 2024;12(5):7–16. https://doi.org/10.32362/2500-316X-2024-12-5-7-16

Financial disclosure: The authors have no financial or property interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Автоматизация поиска юридической информации на арабском языке: подход к поиску документов

Камел С. Жафар [1, @],
Али А. Мохаммад [2],
Али Х. Исса [3],
А.В. Панов [1, @]

[1] *МИРЭА – Российский технологический университет, Москва, 119454 Россия*
[2] *Национальный исследовательский университет «Высшая школа экономики», Москва, 109028 Россия*
[3] *РОСБИОТЕХ – Российский биотехнологический университет, Москва, 125080 Россия*
[@] *Авторы для переписки, e-mail: zhafar.k@edu.mirea.ru, panov_a@mirea.ru*

Резюме

Цели. Поиск юридической информации, например, информации, связанной с различными юридическими вопросами, такими как наказание за преступления, является сложной задачей. Предлагаемый авторами подход может быть эффективным и действенным способом автоматизации поиска юридической информации без необходимости использования большого количества размеченных данных или значительных вычислительных ресурсов. Целью статьи является анализ возможности использования подхода к поиску документов в контексте юридических текстов на арабском языке, с применением методов обработки естественного языка и неконтролируемой кластеризации.

Методы. Использован подход Top2Vec – алгоритм моделирования темы, который создает вложения документов на основе семантического контекста, чтобы группировать юридические тексты на арабском языке в соответствующие темы. Использован алгоритм кластеризации на основе плотности для определения подтем внутри каждого кластера. Решаются проблемы работы с арабским юридическим текстом, такие как морфологическая сложность, двусмысленность и отсутствие стандартизированной терминологии. Предложен конвейер предварительной обработки, включающий в себя токенизацию, нормализацию, выделение корней и удаление стоп-слов.

Результаты. Результаты оценки подхода с использованием набора данных юридических текстов на арабском языке, основанного на ключевых словах, показали его эффективность и превосходство с точки зрения точности и запоминаемости. Предлагаемый подход обеспечивает точность поиска – 87% и полноту поиска – 80%. Применение этого подхода может значительно улучшить поиск юридических документов, сделав его более быстрым и точным.

Выводы. Предложенный подход может быть ценным инструментом для юристов и исследователей, которым необходимо ориентироваться в обширном и сложном ландшафте арабской юридической информации, повышая эффективность и точность ее поиска.

Ключевые слова: поиск документов, обработка естественного языка, Top2Vec, алгоритм кластеризации на основе плотности, арабские юридические документы, вложения слов, косинусное сходство

• **Поступила:** 05.05.2023 • **Доработана:** 04.04.2024 • **Принята к опубликованию:** 11.07.2024

Для цитирования: Жафар К.С., Мохаммад А.А., Исса А.Х., Панов А.В. Автоматизация поиска юридической информации на арабском языке: подход к поиску документов. *Russ. Technol. J*. 2024;12(5):7–16. https://doi.org/10.32362/2500-316X-2024-12-5-7-16

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

The search and analysis of legal information is associated with certain difficulties due to specificities pertaining to the field of law. In recent years, there has been a growing interest in using natural language processing techniques to automate the process of accessing such information [1]. In particular, significant efforts have been expended to develop question and answer systems that can extract specific answers from legal documents [2, 3]. However, building a good QA system requires a large amount of marked-up data and is often computationally intensive. The present work proposes an alternative approach to automate the retrieval of legal information related to crimes, including criminal offenses, in regulations and legal documents in Arabic. In practice, it is often sufficient to retrieve the most relevant legal documents related to the user's query without extracting specific answers from the documents. The proposed approach consists of several steps including data collection, data preprocessing, document indexing, query processing, and document retrieval. The source dataset comprises standard Arabic grammatical and legal documents related to crimes and felonies.

1. LITERATURE REVIEW

Legal information retrieval is a field with a rich history and an extensive body of research. In the present section, key works and developments in this field are reviewed with a special focus on approaches to legal information retrieval and document retrieval in Arabic. Sansone et al. investigate state-of-the-art Artificial Intelligence (AI) techniques used for legal information retrieval systems [4]. With the advent of information and communication technologies, legal practitioners have faced a dramatic increase in digital information, which makes efficient retrieval techniques essential. This article discusses various approaches to AI, including natural language processing, machine learning, and knowledge extraction techniques, explaining how they can aid legal information retrieval. As well as describing the challenges faced by legal practitioners, especially when searching for similar cases, statutes, or paragraphs, the authors discuss open questions regarding legal information retrieval systems. Overall, the study emphasizes the importance of AI in the legal field and the need for continued research and development in legal information retrieval systems. Sartor et al. review nine significant sources published in the last decade [5]. Four of the articles focus on analyzing legal cases, introducing contextual considerations, predicting outcomes from natural language descriptions of cases, comparing different ways of presenting cases, and formalizing precedential reasoning. One article introduces the argumentation scheme method for analyzing arguments, which has subsequently become very widely used in AI and law. Two of the reviewed articles refer to ontologies for representing legal concepts, while the remaining two take advantage of the increasing availability of legal datasets in this decade to automate document summarization and arguments retrieval.

Zhong et al. present an overview of the development, current state and future directions of legal AI [6]. Legal AI applies natural language processing to assist lawyers in their work with the potential to increase efficiency by automating repetitive tasks. Illustrating perspectives of lawyers and natural language processing researchers through experimentation and analysis of existing work, the authors identify knowledge modeling, legal reasoning, and interpretability as three major problems in legal tasks that require further research. As a means of addressing these challenges, the paper proposes a combination of embedding-based and character-based methods to create large-scale and high-quality datasets, as well as to address ethical concerns such as gender bias and racial discrimination. Ultimately, legal AI should play a supporting role in the legal system, with professionals making decisions first and foremost. Shamma et al. describe the development of an Arabic system for extracting information from legal documents, which utilizes a hybrid approach of machine learning and rule-based methods [7]. The system is designed to extract important information from documents and present it in a structured form for complex queries. The described approach, which has been tested on a limited class of Arabic legal documents, has demonstrated good results. The authors

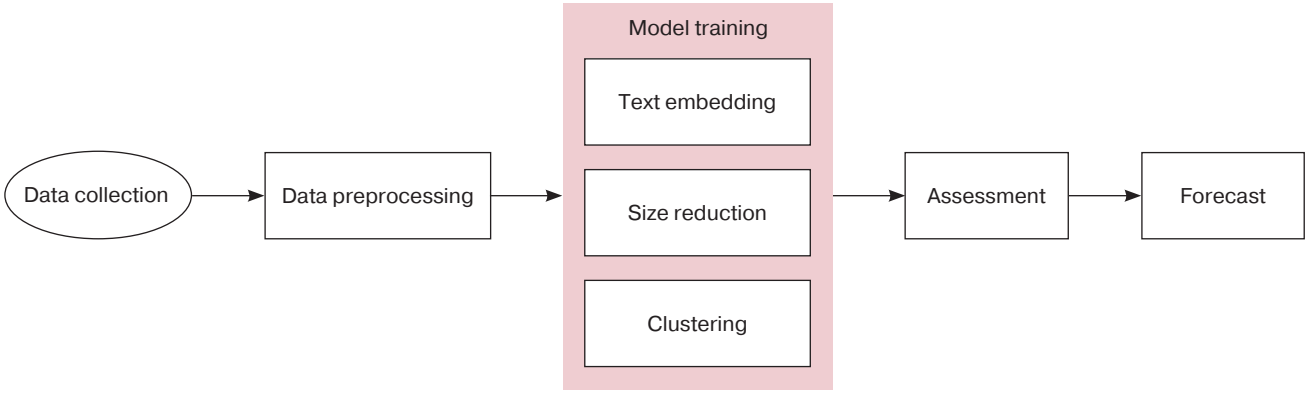

Fig. 1. Top2Vec approach to legal information retrieval

suggest several possible extensions to the system, such as considering different types of cases in a given legal system, using more advanced Arabic natural language processing techniques tools, exploring the use of deep learning, extracting more relationships, improving the presentation of results, and extending the system to other domains such as healthcare and finance.

2. PROPOSED APPROACH

The proposed Topic-to-Vector (Top2Vec) methodology for legal information retrieval comprises several steps (Fig. 1). The first step is data preparation, which involves collecting legal documents related to crimes and criminal offenses described in Arabic legal documents. In order to prepare it for the model, the data is then preprocessed using a pipeline that includes tokenization, normalization, root extraction, and stop word removal.

The next step is the model training, which involves creating document embeddings using the Top2Vec model. The embeddings are then reduced to a lower dimensional space using dimensionality reduction techniques such as a uniform manifold approximation and projection (UMAP)[1] in order to simplify document clustering. Clustering is performed using a density-based clustering algorithm such as the hierarchical density-based spatial clustering of applications with noise (HDBSCAN), which can efficiently identify document clusters based on their similarity [8].

Once the clustering is complete, the model can be evaluated with a set of real user queries. For each query, the model retrieves the most relevant document clusters, allowing the user to browse through the documents to find the relevant information. The performance of the model can be evaluated using metrics such as accuracy and completeness.

Finally, the model can be used for prediction, where the user enters a query and the model extracts the most relevant clusters of documents. The user then browses through the documents to find the relevant information. The prediction can be repeated for different queries, and the model can be continuously updated with new data to improve its performance. Overall, this methodology represents an efficient and effective means to automate the retrieval of legal information related to felonies, including criminal offenses, in Arabic legal documents.

2.1. Top2Vec approach

Top2Vec [9, 10] is a novel unsupervised document clustering and topic modeling technique that can discover topics in large-scale datasets without any prior knowledge of their number. The basic idea of Top2Vec is to first embed documents and topics in the same space and then cluster the embedded documents using a density-based clustering algorithm [11]. Top2Vec can also automatically determine the number of topics, allowing the topics to be represented as a set of words and a connected vector. The algorithm outperforms traditional topic modeling methods such as latent Dirichlet distribution and nonnegative matrix factorization, both in terms of clustering quality and scalability for large datasets. The potential of Top2Vec has been demonstrated in various domains including document similarity search, visualization, and anomaly detection.

Figure 2 shows an example of a semantic space. The purple dots are documents, while the green dots represent words. Words are closest to the documents they best represent; similar documents are presented close to each other.

[1] Uniform manifold approximation and projection is a machine learning algorithm that performs nonlinear dimensionality reduction.

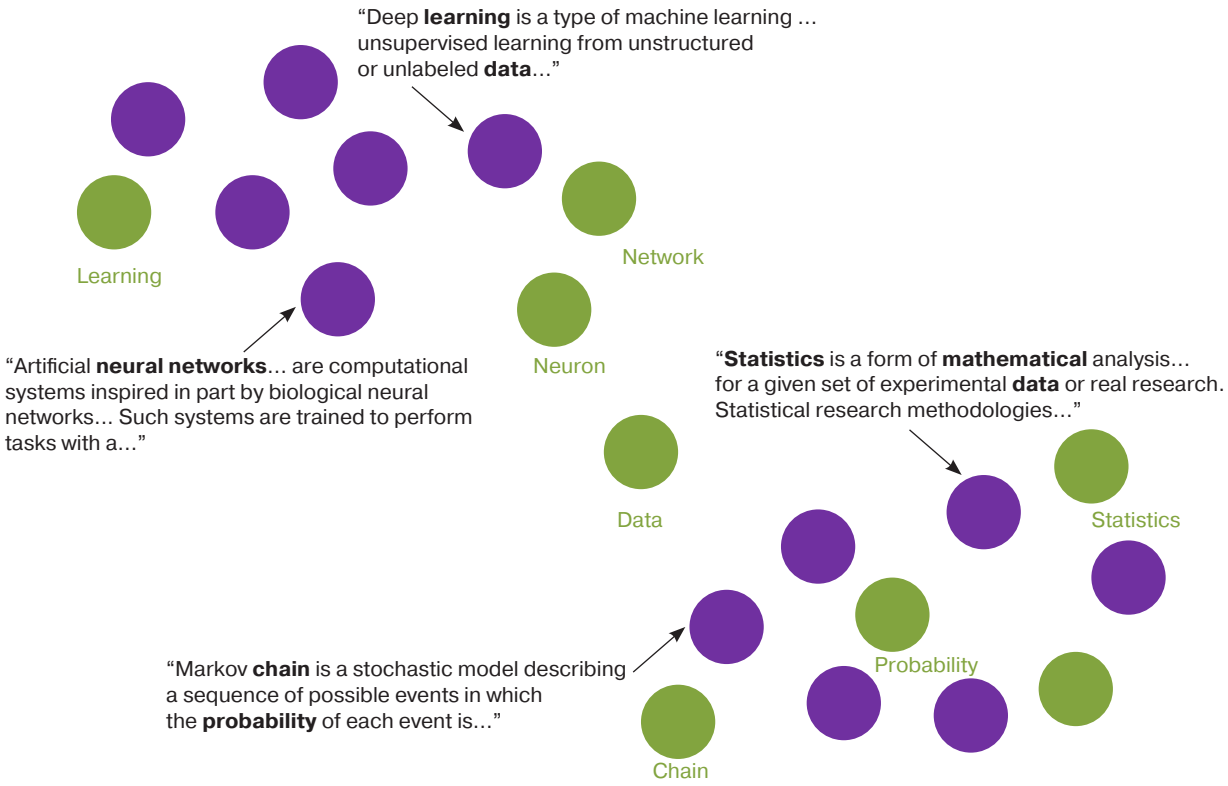

Fig. 2. Example of a semantic space

2.2. HDBSCAN clustering algorithm

The HDBSCAN application [12, 13] is a clustering algorithm that can identify clusters of different densities and shapes in multidimensional spaces. It uses a hierarchical approach to build a hierarchy of clusters and automatically determine the number of clusters. One of the advantages of HDBSCAN is that it can handle clusters of different sizes and shapes and identify noise points. The algorithm has several hyperparameters that can be tuned to optimize the clustering results. Some of the key hyperparameters are as follows:

- Minimum cluster size: this parameter sets the minimum number of points required to form a cluster. Increasing this parameter increases the number of clusters and decreases the number of clusters.
- Metric: this parameter defines the distance metric used to calculate the similarity between data points. Depending on the characteristics of the data, different metrics can be used.
- Cluster selection method: this parameter defines how the final set of clusters is selected from the hierarchy.

In the process of training the Top2Vec model, the HDBSCAN algorithm was used with the following hyperparameters: minimum cluster size—3, metric—Euclidean, cluster selection method—leaf.

These hyperparameters are selected based on the characteristics of the available dataset and tuned to optimize the clustering results. The Euclidean metric is a commonly-used metric for measuring the distance between points in multidimensional spaces; the leaf cluster selection method is suitable for large datasets due to providing a good balance between speed and accuracy. A minimum cluster size of 3 was chosen to prevent the formation of small clusters and the inclusion of noisy points.

The resulting visualization (Fig. 3) displays the clusters in two-dimensional space to permit a visual inspection of the relation of documents in each cluster. This approach is useful for exploring and navigating the complex landscape of legal information, helping legal professionals and researchers quickly identify relevant documents based on their content.

2.3. Data collection and preprocessing

In this phase, regulatory and legal documents in Arabic relating to various crimes are collected and pre-processed. The standard set of legal documents includes different types of legal texts, i.e., laws and regulations. These documents are collected in a single

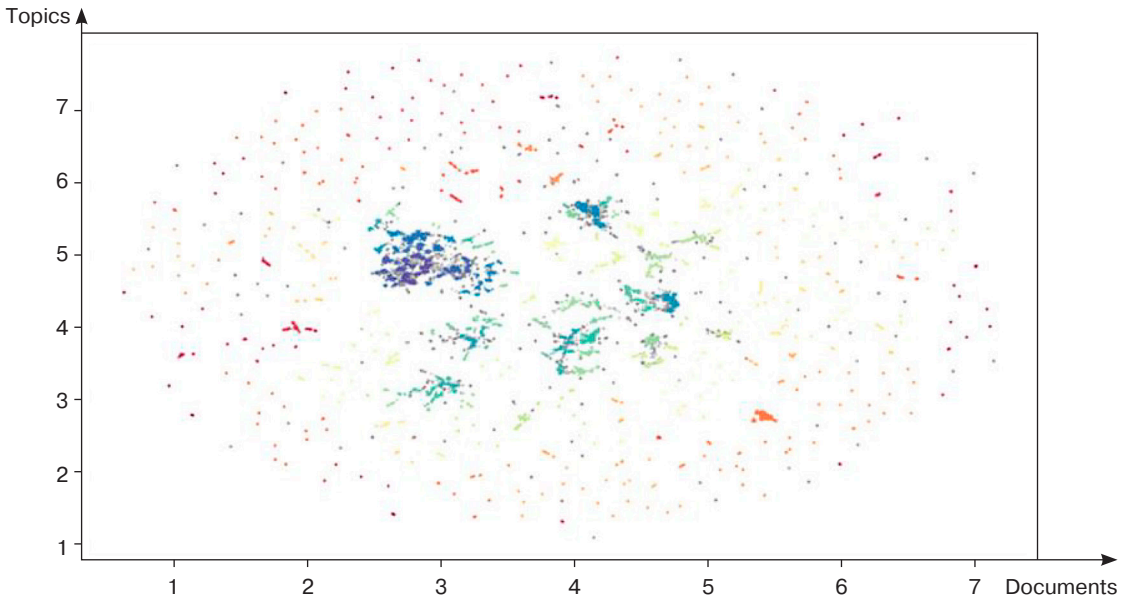

Fig. 3. Retrieving dense document regions using spatial clustering of applications based on hierarchical density with noise

csv file referred to as a corpus, in which each document represents one record (line).

Each row has several basic keys required for data organization:

1. Index: a unique sequential number for each document.
2. Identifier: a unique string for each document that represents a regulatory or legal document, serving as the actual identifier as it is referenced in the original source document.
3. Title: string name of the regulatory or legal document.
4. Summary: string brief description of the content of the regulation or document.
5. Details: string full description and content of the document as it appears in the original source.

For the purpose of supervised learning, the model is only trained on the data that are in the Details column from the corpus. These are then assembled into a single list, which is preprocessed to remove stop words, punctuation marks and diacritics, as well as to perform root detection and normalization.

Preprocessing stage includes the following sub-stages:

1. Tokenization: this step involves breaking the text into individual words or tokens. This is an important first step in any text processing pipeline since most algorithms work with individual words or tokens rather than whole sentences or documents. Tokenization can be performed using various methods: space-based tokenization, regular expressions, and rule-based tokenization.
2. Stop word removal. Stop words are common words that are often removed from a text because they do not make much sense and their presence may reduce the accuracy of text analysis. Examples of stop words are prepositions, conjunctions, and articles (and, the, of, in, etc.). By removing stop words, it is possible to reduce the dimensionality of the data and increase the efficiency of subsequent text processing steps.
3. Grounding (stemming). This step involves bringing words back to their root form, also known as the base form. The goal of stemming is to group similar words together, even if they are nonidentical. Stemming can be performed using Porter's algorithm, Snowball algorithm and Lancaster's algorithm.
4. Normalization. This step involves converting words to a standardized form to ensure consistency and reduce redundancy in the text. Examples of normalization include converting all words to lower case, converting all numbers to digits, and removing punctuation marks. By helping reduce noise in the data, normalization improves the accuracy of subsequent text processing steps.

Following this step, the necessary dataset of preprocessed documents is obtained for training the model.

2.4. Model training

Training the Top2Vec model involves several parameters that need to be set based on the characteristics of the dataset and the desired result. In this work, the Top2Vec model was trained using the following parameters:

- Documents: corpus of preprocessed documents for clustering.
- min count: the minimum number of times a word must appear in order to be included in the model dictionary.
- Embedding model: a type of embedding model used to create document attachments. In this case, the Document-to-Vector (Doc2Vec) model was used. This represents an unsupervised deep learning model that can learn vector representations of documents to enable efficient similarity computation and topic modeling.
- To split the documents: a Boolean value indicating whether the documents should be split for processing. In this work, the value "split documents" is true.
- Document blocker: a method used to separate documents. In this case, a sequential fragmenter was used. This is a method of dividing documents into smaller sequential fragments of fixed length to facilitate training and processing of machine learning models.
- Fragment length: the maximum length of each document fragment. In this case, the block length is set as 5, i.e., each document is split into blocks of 5 sentences. This parameter controls the level of detail of the splitting process: smaller values result in smaller fragments and potentially more detailed topics, while larger values result in larger fragments and potentially more general topics.
- Maximum number of fragments for each document. In this case, the maximum number of fragments is set as 2, i.e., each document is split into no more than 2 fragments of 5 sentences in length. Thus, if the document is longer than 10 words, it will be divided into 2 fragments of 5 words each, while, if the document is shorter than 6 words, it will be considered as one fragment.
- HDBSCAN arguments: hyperparameters of HDBSCAN clustering algorithm, minimum cluster size, distance measure and cluster selection method.
- Speed: the speed for model training. The paper establishes deep learning, which is the most sophisticated and efficient training method.

Once these parameters are set, the Top2Vec model is ready to be trained.

Optimum values are obtained after several experiments with different parameters and comparison of results. The model is trained on a large number of legal documents. Because of this and the deep learning capabilities, the training process takes several hours. The result of the training process is a model that has access to a set of clusters, each containing a representative document together with a list of similar documents. These clusters can be used to explore the corpus and identify patterns or themes in legal documents.

2.5. Queries processing

At this step, the user query is processed to match the preprocessed documents that were used to train the model. The same preprocessing methods are applied to the user query as for the documents. The preprocessed query is then used to create a set of candidate documents for the model.

2.6. Documents retrieval

In order to obtain the documents associated with a query using Top2Vec, the first step is to embed the query in the same vector space as the documents. This can be done by passing the query through the same neural network that was used to embed the documents. The result is a vector representing the query in the same multidimensional vector space.

Top2Vec then computes the cosine similarity between the query vector and all document vectors in the vector space. The cosine similarity score ranges from −1 to 1, where 1 indicates complete similarity, 0 indicates no similarity, while −1 indicates complete dissimilarity.

The cosine similarity between vectors **q** and **d** can be calculated as follows:

$$\cos(\mathbf{q},\mathbf{d}) = \frac{\mathbf{q}\cdot\mathbf{d}}{(\|\mathbf{q}\|\cdot\|\mathbf{d}\|)},$$

where $\|\mathbf{q}\|$ and $\|\mathbf{d}\|$ are the Euclidean norms of vectors **q** and **d**, and $\mathbf{q}\cdot\mathbf{d}$ is their scalar product [14].

3. ASSESSMENT AND DISCUSSION

The proposed document retrieval approach was evaluated using a set of real user queries related to felonies, including criminal offenses, in Arabic legal documents. First, a set of 100 user queries from lawyers working with such documents was generated. To evaluate the effectiveness of the approach, the 10 most popular documents for each query were compared with a set of relevant documents identified by legal experts. A relevant document is one that contains information relevant to the query topic even if it does not explicitly answer the query. For example, a document about a murder investigation will be considered relevant to a murder query. Two standard evaluation metrics are used to measure the performance of the proposed approach: precision and completeness. Accuracy measures the share of retrieved relevant documents, while completeness measures the share of retrieved relevant documents. Specifically, accuracy and completeness are defined as follows [15]:

$$\text{Precision} = \frac{\text{Relevant Documents Retrieved (RDR)}}{\text{Total Documents Retrieved (TDR)}},$$

$$\text{Recall} = \frac{\text{Relevant Documents Retrieved (RDR)}}{\text{Total Relevant Documents (TRD)}},$$

where RDR is the number of extracted relevant documents, TDR is the total number of extracted documents, and TRD is the total number of relevant documents.

In order to ensure a single summary measure of overall performance, an F1 score representing a harmonic mean of accuracy and completeness was evaluated. To compare the performance of the proposed approach with the baseline, a simple keyword-based approach retrieves documents containing at least one keyword from a user query. The same set of queries and evaluation metrics are used in the comparison. Experiments show that the proposed document retrieval approach outperforms the keyword-based approach in terms of both accuracy and completeness, achieving retrieval accuracy of 87% and retrieval completeness of 80% as compared to an accuracy of 66% and completeness of 62% for the keyword-based approach. The F1 score for the proposed approach is 0.83, while the F1 score for the keyword-based approach is 0.63.

These results show that utilizing natural language processing and machine learning techniques for document retrieval can significantly improve the performance of legal information retrieval systems. By processing user queries and documents using advanced algorithms, the proposed approach retrieves relevant legal documents with higher accuracy, reducing the burden on legal practitioners to manually search large document repositories. Nevertheless, it should be noted that this approach can still be improved. For example, it may not be effective in retrieving relevant documents for highly specialized legal domains that require a more nuanced understanding of legal language. Nevertheless, it is crucial to recognize that the current approach has room for improvement. In addition, potential distortions inherent in training data may lead to misrepresentations in the results obtained.

All in all, the experimental results demonstrate the promising potential of the proposed document retrieval approach in developing more efficient legal information retrieval systems. Further research and development in this area could lead to even more efficient algorithms that can better support legal practitioners in their work.

CONCLUSIONS

The proposed approach to legal document retrieval uses natural language processing techniques, including the Top2Vec model and the HDBSCAN clustering algorithm. A comparison of this approach with a keyword-based approach demonstrates its superiority in terms of accuracy and completeness. Specifically, the proposed approach achieves a retrieval accuracy of 87% and a retrieval completeness of 80%. This can significantly improve legal document retrieval by making it faster and more accurate. However, it is important to note its limitations and room for future research and development. In conclusion, the work underscores the relevance of using advanced natural language processing techniques in the legal domain, as well as successfully demonstrating their potential to improve the efficiency and accuracy of legal document retrieval.

Authors' contributions

K.S. Jafar—conducted the majority of the research from conception to execution, including the design of the methodology, implementation of the Topic-to-Vector algorithm, and evaluation of results.

A.A. Mohammad—contributed to the literature review, identifying relevant studies and background information related to legal information retrieval, and assisted in drafting the introduction.

A.H. Issa—made significant contributions to writing the discussion section, interpreting the findings within the context of existing literature, and contributed to the conclusion of the article.

A.V. Panov—supervised the research process, provided guidance on research direction, and assisted in refining the methodology and results.

REFERENCES

1. Sleimi A., Sannier N., Sabetzadeh M., Briand L., Dann J. Automated extraction of semantic legal metadata using natural language processing. In: *2018 IEEE 26th International Requirements Engineering Conference* (*RE*). IEEE; 2018. P. 124–135. https://doi.org/10.1109/RE.2018.00022
2. Rogers A., Gardner M., Augenstein I. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surveys*. 2023;55(10):1–45. https://doi.org/10.1145/3560260
3. Alanazi S.S., Elfadil N., Jarajreh M., Algarni S. Question Answering Systems: A Systematic Literature Review. *International Journal of Advanced Computer Science and Applications* (*IJACSA*). 2021;12(3):359. https://doi.org/10.14569/IJACSA.2021.0120359
4. Sansone C., Sperl´ı G. Legal Information Retrieval systems: State-of-the-art and open issues. *Inform. Syst.* 2022;106:101967. https://doi.org/10.1016/j.is.2021.101967
5. Sartor G., Araszkiewicz M., Atkinson K., et al. Thirty years of Artificial Intelligence and Law: the second decade. *Artif. Intell. Law*. 2022;30(4):521–557. https://doi.org/10.1007/s10506-022-09326-7
6. Zhong H., Xiao C., Tu C., et al. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. 2020. *arXiv:2004.12158* [cs.CL]. https://arxiv.org/abs/2004.12158v5
7. Abu Shamma S., Ayasa A., Yahya A., et al. Information extraction from Arabic law documents. In: *2020 IEEE 14th International Conference on Application of Information and Communication Technologies* (*AICT*). IEEE; 2020;1–6. https://doi.org/10.1109/AICT50176.2020.9368577
8. Hammami E., Faiz R. Topic Modelling of Legal Texts Using Bidirectional Encoder Representations from Sentence Transformers. In: *Advances in Information Systems, Artificial Intelligence and Knowledge Management. Conference paper. International Conference on Information and Knowledge Systems.* Cham: Springer Nature Switzerland; 2023. V. 486. P. 333–343. https://doi.org/10.1007/978-3-031-51664-1_24
9. Angelov D. Top2Vec: Distributed Representations of Topics. 2020. *arXiv:2008.09470* [cs.CL]. https://arxiv.org/abs/2008.09470v1
10. Karas B., Qu S., Xu Y., Zhu Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis. *Front. Artif. Intell*. 2022;5:948313. https://doi.org/10.3389/frai.2022.948313
11. Vianna D., de Moura E.S., da Silva A.S. A topic discovery approach for unsupervised organization of legal document collections. *Artif. Intell. Law*. 2023;Online First. https://doi.org/10.1007/s10506-023-09371-w
12. McInnes L., Healy J., Astels S. hdbscan: Hierarchical density-based clustering. *J. Open Source Softw*. 2017;2(11):205. https://doi.org/10.21105/joss.00205
13. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. *arXiv*, *preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805v2
14. Salton G., McGill M.J. *Introduction to Modern Information Retrieval*. N.Y.: McGraw-Hill; 1983. 472 p.
15. Manning C.D., Raghavan P., Schütze H. *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press; 2008. 492 p.

About the authors

Kamel S. Jafar, Postgraduate Student, Department of Corporate Information Systems, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: zhafar.k@edu.mirea.ru. Scopus Author ID 57552322300, https://orcid.org/0009-0004-1791-1406

Ali A. Mohammad, Master Student, Faculty of Computer Science, HSE University (11, Pokrovsky bulv., Moscow, 109028 Russia). E-mail: amokhammad_1@edu.hse.ru. https://orcid.org/0009-0002-3533-568X

Ali H. Issa, Postgraduate Student, Department of Automated Control Systems for Biotechnological Processes, BIOTECH University (11, Volokolamskoye sh., Moscow, 125080 Russia). E-mail: ali.issa.rus@gmail.com. https://orcid.org/0009-0001-3699-222X

Alexander V. Panov, Cand. Sci. (Eng.), Associate Professor, Department of Corporate Information Systems, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: panov_a@mirea.ru. RSCI SPIN-code 8571-9729, https://orcid.org/0009-0003-0310-3638

Об авторах

Жафар Камел С., аспирант, кафедра корпоративных информационных систем, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: zhafar.k@edu.mirea.ru. Scopus Author ID 57552322300, https://orcid.org/0009-0004-1791-1406

Мохаммад Али А., магистрант, Факультет компьютерных наук, ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ) (109028, Россия, Москва, Покровский бульвар, д. 11). E-mail: amokhammad_1@edu.hse.ru. https://orcid.org/0009-0002-3533-568X

Исса Али Х., аспирант, кафедра автоматизированных систем управления биотехнологическими процессами, ФГБОУ ВО «Российский биотехнологический университет» (РОСБИОТЕХ) (125080, Россия, Москва, Волоколамское шоссе, д. 11). E-mail: ali.issa.rus@gmail.com. https://orcid.org/0009-0001-3699-222X

Панов Александр Владимирович, к.т.н. доцент, профессор кафедры корпоративных информационных систем, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: panov_a@mirea.ru. SPIN-код РИНЦ 8571-9729, https://orcid.org/0009-0003-0310-3638

Translated from Russian into English by Lyudmila O. Bychkova
Edited for English language and spelling by Thomas A. Beavitt