Machine Learning and Natural Language Processing in Domain Classification of Scientific Knowledge Objects – Review

Samuel Machado and Jorge Oliveira e Sá[1]

[1] Algoritmi Center, University of Minho, Portugal

Abstract. The domain classification of scientific knowledge objects has been continuously improved over the years. Systems that can automatically classify a scientific knowledge object, through the use of artificial intelligence, machine learning algorithms, natural language processing, and others, have been adopted in most scientific knowledge databases to maintain internal classification consistency as well as to simplify the information arrangement. However, the amount of available data has grown exponentially in the last few years and now it can be found in multiple platforms under different classifications due to the implementation of different classification systems. Thus, the process of searching and selecting relevant data in research studies and projects has become more complex and the time needed to find the right information has continuously grown as well. Therefore, machine learning and natural language processing play an important role in the development and achievement of automatic and standardized classification systems that will aid researchers in their research work.

Keywords: Natural Language Processing, Machine Learning, Domain Classification, Scientific Knowledge Objects.

1 Introduction

The process of searching and selection of relevant data in research studies and projects have become more complex due to the huge amount of available data. In the data search process, researchers may or may not use various filters to restrain the amount of data that the used platform returns. These filters are used to classify the data and rearrange it under certain labels to simplify the search process.

During the search process, publication date, document type, and scientific domain are the most common filters applied to search processes and to the data itself. Regarding the scientific domain, this classification method is applied in almost every search, allowing the researcher to specify the scientific field that corresponds to the focus of the search.

However, multiple times data can be found under different search platforms with different domains associated. That happens because different platforms use different classification systems which cause data to be labeled differently, or even being possible

that the classification given to it could be wrong. Therefore, the inexistence of standardized and automatic systems for the classification of scientific knowledge objects (SKOs) has a major impact on the search process. Relevant information may not be found due to this inconsistency, and out of search focus information can cause several delays in the information search process.

Thus, this paper aims to provide an overview of the state-of-the-art in machine learning (ML) algorithms and natural language processing (NLP) techniques for SKOs domain classification through the review of the most recent studies and reviews that have been conducted in SKOs domain classification as well as identifying the most suitable algorithms to perform the domain classification.

The structure of this paper is composed as follows: Research Method, Machine Learning, and Natural Language Processing, Literature Review Results, and Conclusions.

2 Research Method

The research process and the write of this review were based on the guidelines provided by Webster & Watson [1].

2.1 Literature Search

To carry out the literature search that will support the research component of this project was used only one search platform: Scopus. Search criteria and terms were also defined to filter the quantity and quality of the information.

Thus, and since the project requires the study of the state of the art of ML and NLP techniques for the classification of SKOs, the search criteria YEAR> = 2014 was defined. With a temporal restriction of this amplitude, the results obtained from the queries introduced in Scopus are more current, that is, the solutions and investigations presented in the results adopt more recent techniques, thus allowing to verify the state of the art in this context. Table 1 synthesizes the obtained results of the search stage.

Table 1. Total number of results.

Search query	Number of results
1. "scientific articles classification" AND "NLP"	0
2. "scientific articles classification" AND "machine learning"	3
3. "scientific articles classification"	4
4. "classification of research articles"	5
5. "classification of research papers"	39
6. "domain classification" AND "scientific articles"	2
7. "domain classification" AND "research papers"	16
8. "domain classification" AND "research papers" AND "NLP"	2

Search query	Number of results
9. "domain classification" AND "NLP"	43
10. "topic classification" AND "research papers"	11
11. "topic classification" AND "NLP"	198
12. "topic classification" AND "machine learning"	872
13. "domain classification" AND "machine learning"	284
14. "semantic analysis" AND "NLP" AND "classification"	964
15. "subject classification" AND "NLP" AND "research paper"	2
16."subject classification" AND "machine learning" AND "research paper"	12

2.2 Literature Selection

Based on the results obtained in the previous section, a preliminary selection of the literature was made with a focus on the title, summary, keywords, and introduction to determine which results were within the context of the research focus. Therefore, this selection process resulted in 20 obtained SKOs. Table 2 shows the results of the literature selection.

Table 2. Number of selected, repeated and obtained results.

Search query	Selected results	Repeated results	Obtained results
1.	0	0	0
2.	0	0	0
3.	1	1	0
4.	0	0	0
5.	1	0	1
6.	1	1	0
7.	1	0	1
8.	0	0	0
9.	1	0	1
10.	2	1	1
11.	6	0	6
12.	5	2	3
13.	3	0	3
14.	5	3	2

Search query	Selected results	Repeated results	Obtained results
15.	1	0	1
16.	1	0	1
Total	**28**	**8**	**20**

2.3 Backward Tracking

This backward tracking process complements the literature search done previously as well as fix any informational gaps that might exist by obtaining additional relevant SKOs that the selected ones have referred to.

Thus, was developed a simple application that builds a cross-reference matrix between the references present in the selected literature to determine each one frequency, returning the number of occurrences of each reference from the complete given references list. To develop this application was made a preliminary analysis of the references list to detect which adjustments were needed. This process resulted in 14 more SKOs to complement the ones obtained previously, making a total of 34 initial SKOs.

3 Machine Learning and Natural Language Processing

NLP came to complement the existing approaches of information processing, enabling the understanding of texts and languages of human nature. These texts have different characteristics from structured texts since, in addition to often being in the form of free text, that is, unstructured text, they have intrinsic characteristics that may result in some additional meaning or purpose beyond the simple text. For a human being, these characteristics may be evident and will make sense, but for a computer, this is no longer the case.

Thus, the appearance of NLP has enabled the computer to identify these types of occurrences, for example, the expression of feelings, Sentimental Analysis, which has been increasingly explored in recent times. Thus, the connection between the techniques of NLP with ML has been providing increasingly better results. As the name implies, NLP is related to data processing and information extraction. It can be considered as the first step, after obtaining the data, to achieve a more rigorous and accurate classification system.

Regarding ML, one of the elements that have the greatest impact on the project is the selected approach(es). Typically, supervised and unsupervised approaches are the most common ones among ML implementations. Models based on supervised approaches need previously classified data sets to calibrate the model itself [2] whereas, in unsupervised approaches, they do not need the data to be classified, being able to make use of lexicons, to carry out the data classification [3].

3.1 Data Classification - ML

The implementation of ML algorithms is directly linked to the definition of the classification system, being the ML algorithms classification output restricted to the classification system defined. The identification and results of ML algorithms can be found in Section 4.

3.2 Data Processing - NLP

The most common approaches to the use of NLP techniques usually use a set of steps, in which the data obtained is processed. In the work of Romanov et al [4], in which a classification system for scientific texts in Russian was developed, an approach consisting of 5 steps was presented, namely: the removal of formulas that are frequent in scientific texts; the aggregation of metadata, which includes the title, keywords, and summary; transformation of data to lowercase; the removal of stop words that reduces the amount of existing information to just useful information; and the stemming of words, which consists of deflecting words to determine their lemma.

Nurfikri & Adiwijaya [5] presents an implementation based on four steps, namely case folding, tokenization, removal of stop words, and word lemmatization, the last two steps being common in both implementations presented.

Regarding the application of case folding, the transformation of the text by converting all existing letters to lower case letters as well as the possible removal of punctuation characters or even numbers [5] helps in removing existing noise in the data, and this removal makes it easier to perform the following steps to be performed.

As for the removal of stop words, it is a very common step in NLP implementations, since it significantly reduces the amount of abstract information [6] that will be processed by the algorithms, for just useful information. This removal allows the adopted algorithms to process this information more quickly, thus making a more efficient classification process. For the execution of this step, mention is made of two different possibilities. As a first possibility for the removal of stop words, we have the adoption of stop word dictionaries such as WordNet [5] to compare and remove these words from the data if there is a correspondence between them and the dictionary. Another option would be the adoption of a library that, using internal methods, performs this removal automatically. The SciKit Python [7] and NLTK [8][9] libraries are examples of libraries that can be used to perform this step.

Tokenization consists of analyzing and transforming data to generate terms [10], which means that complete sentences are transformed into arrays of words/tokens or, transform documents or paragraphs into arrays of simple phrases depending on the desired level. The NLTK library mentioned above also allows the execution of this step, resulting in more simplified information to be processed by the following steps.

Concerning the lemmatization of words and stemming of words, both approaches have similar characteristics having the same objective however, the result obtained may vary between them. Both approaches aim to transform words into their root word but using different methods. While the stemming process tries to identify the root word of each existing word in the data obtained by removing plurals of similar words [10] and removing existing prefixes and suffixes, lemmatization still uses the lexical context and semantics to identify similar words[10] and thus determine the root words. In this way, it can be said that the application of lemmatization should return better results

concerning the extraction of the root word of each existing word in the data since it takes into account the context in which they are inserted.

3.3 Data Semantic Analysis - NLP

Data semantic analysis is a component that plays an important role in obtaining a rigorous classification of SKOs. As the classification is directly related to their content, there is a need to understand it and classify it.

Examples of techniques that apply data semantic analysis are Part-of-Speech (PoS), Latent Dirichlet allocation (LDA), Bag of Words (BoW), Term Frequency – Inverse Document Frequency (TF-IDF), GloVe, and Word2vec. These techniques refer to the need to try to understand the data not only as a set of characters but as elements that have some value or meaning in a given context.

In the case of PoS tagging, the objective is to try to classify each word/token in a sentence using classes such as name, verb, adjective, preposition, among others [11] taking into account not only the basic characteristics defined by each language for each word but also the context in which the word is inserted as well as the relationship established between that word and the neighboring words in the sentence.

The LDA probabilistic model performs the representation of the data under a hierarchical scheme, defining that each data set is constituted by a certain set of topics and that each word/token belongs to a certain topic [12]. In this way, it is possible to carry out the classification of a given data set taking into account the topics observed in it, and that classification takes into account the words in the data as well as the probabilities of each word belonging to each topic. Thus, the same word can belong to multiple topics, but with a different probability of occurrence for each topic.

Another existing model is the BoW which, when complemented by other models such as the TF-IDF model, can become a powerful tool in classification tasks.

In the BoW model, two features can be verified. Firstly, this model creates a “bag” of words that will contain all the words in the total data set [13], functioning as a database of words that serve as the basis for the execution of the second functionality related to the frequency calculation algorithm inherent to this method. This algorithm, when exposed to a data set, counts the occurrences of each word, and associates each one with the same value, called weight [13]. Thus, it is possible to make a comparison between these weights assigned by the algorithm to that data set, with other data sets previously processed, thus allowing to perform some sort of classification.

The TF-IDF model, like the BoW method, also assigns a weight to each existing word to determine which words are most important in a given data set. However, this algorithm is not based solely on the frequency of occurrence of each word, but rather the total frequency of each word in the global data set [13]. Thus, this algorithm tends to devalue words with a high-frequency value in a global set [14], assuming the possibility that they are noise. In the context of classification, most of the words in the data do not contribute to the classification process, which can lead to deviations in the results obtained.

An alternative to the BoW model would be to use models such as GloVe [15] or Word2Vec [16]. These models differ from the BoW model in that they do not assign a weight to each word, but a vector. In this way, words can be represented in space

through their vectors, enabling the calculation of similarity between words [17] as well as the identification of relationships between them.

In this way, the use of semantic data analysis models will be indispensable in this project since the context in which the data is presented can determine the correctness or not regarding the classification of articles. For example, in the title "Co-training for topic classification of scholarly data" [18], "scholarly data", by itself, would obtain a classification in the area of education, which would conflict with the classification attributed to the first part of the title that is related to the training of algorithms for the classification of topics, which is framed in the area of computer science. With the absence of semantic analysis, the classification of this article would return practically the same weight for both parties, which is not correct since "scholarly data" is the object of classification of the algorithms, making this article an article with a weight higher education in the field of computer science and not in the field of education.

4 Literature Review Results

The present section aims to summarize the results obtained from the review of the obtained literature to determine the state-of-the-art in ML algorithms, and NLP techniques.

4.1 Data Classification Results - ML

To analyze the ML algorithms presented in the literature, the following Table 4 was built.

Table 3. ML algorithms analysis.

Literature	SVM	CRF	NN	NB	LR	LSTM	RF
[4]	X		X		X	X	X
[5]	X		X	X			
[6]							
[7]	X			X			
[8]	X		X			X	
[9]	X		X	X		X	
[18]	X			X			
[19]		X	X			X	
[20]			X				
[21]						X	
[22]	X	X					
[23]			X			X	X
[24]						X	

Literature	SVM	CRF	NN	NB	LR	LSTM	RF
[25]			X				
[26]	X			X			
[27]	X		X			X	
[28]			X				
[29]	X			X	X		
[30]			X				

Among the ML algorithms, there are the Support Vector Machine (SVM) and Naïve Bayes (NB) algorithms which, in addition to being the most traditional algorithms, continue to provide good results. In Romanov et al [4] 99% accuracy was obtained regarding the classification of scientific texts based on their abstracts. However, this high acuity value reveals low precision and recall values, 61% and 36% respectively, which is not ideal. The best set of analysis metrics for the methods was achieved through the use of the SVM algorithm, which was the one with the best results when compared to other algorithms such as Logistic Regression (LR), Random Forest (RF), Long Short Term Memory (LSTM) and a variant of the Neural Network (NN), called the Artificial Neural Network (ANN).

In the work by Kaplan et al. [22], a comparison was made between the SVM algorithm and the Conditional Random Field (CRF) algorithm using different processing characteristics. In this work, both algorithms obtained similar acuity values, being 72.4% and 72.7%, respectively SVM and CRF.

In another investigation conducted by Bhaskaran et al. [19], CRF and Bidirectional Long Short-Term Memory (Bi-LSTM) algorithms were compared in which, in the context of domain classification, the acuity values of 92.17% and 90.96% were achieved, respectively. These values were obtained using the GloVe model concerning the data processing component.

The Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) algorithms, the latter also a variant of NN, were compared by Semberecki & MacIejewski [9] given the classification of informative articles regarding their domain. The LSTM algorithm with 86.21% accuracy, proved to be more assertive compared to the CNN algorithm that obtained 82.07% accuracy. In this case, in the data processing component, the Word2vec model was used.

It should also be noted that many of the investigations carried out, in the context of text classification, propose their adaptations to the existing models and algorithms, as is the case of the work of Ghumade & Deshmukh [20], which proposes a system derived from Recurrent Neural Network (RNN). The proposed model, which obtained 97.5% of correct classifications, managed to overcome the results obtained by the NB, ANN, and RF algorithms of 92.3%, 91.1%, and 87.5%, respectively.

Regarding the results presented, it should be noted that the percentage values of the correctness of each algorithm, in the respective research project, are directly influenced by the level and quality of data processing before applying ML algorithms, as well as the quality of the data itself. Thus, comparative tests between different sets of steps and

NLP models with ML algorithms will have to be carried out, to try to achieve satisfactory results.

4.2 Data Processing and Data Semantic Analysis Results - NLP

In the same way as the previous section, the following Table 5 summarizes the information exposed in the literature regarding NLP techniques.

As for the NLP steps observed, there was a higher incidence in the data processing base steps, them being case folding, tokenization, and removal of stop words. In the application of semantic data analysis models, the vector representation models, in the context of this project, are those that have the greatest potential for contextualizing the data beyond the LDA model. This model is the most applied in multiple classification contexts since it is based on the definition of topics and the consequent classification of data based on a set of topics.

Table 4. NLP techniques analysis.

Literature	Case Folding	Tokenization	Stop Words	Lematization	Stemming	LDA	POS	BOW	TF-IDF	Glove	Word2Vec
[4]		X	X						X		
[5]	X	X					X	X	X		
[6]		X						X			X
[7]		X					X				
[8]		X									X
[9]	X			X	X	X					
[10]	X	X	X	X	X	X			X		
[15]	X	X	X	X	X		X		X		X
[19]	X	X					X			X	X
[20]		X	X		X				X		
[25]	X	X	X		X	X		X			
[29]	X	X								X	X
[30]	X		X		X						
[31]			X			X			X		
[32]	X	X		X	X		X				
[33]	X	X	X			X	X	X	X	X	X
[34]	X	X									
[35]	X	X	X	X							
[36]		X									X

5 Conclusions

The amount of available information, on diverse scientific platforms and databases that exist nowadays, can make it difficult to find the right one. That is why the existence of a standardized platform that could provide the most acuity possible in information-seeking would have a major impact on research projects.

The existence of a standardized classification system that could provide the most acuity possible SKOs classification, enables the information-seeking process with a major impact on researchers and in their research projects.

To accomplish the implementation of the desired classification system, NLP and ML play an important role in data analysis and classification. As for the NLP steps, there was a higher incidence case folding, tokenization, and removal of stop words for data processing. In the application of semantic analysis, we highlight the vector representation models GloVe and Word2Vec, since they seemed to have the greatest potential for contextualizing the data.

The results from the NLP application should feed the ML algorithms so that they can perform the correct classification of SKOs. Thus, the importance of this set of steps regarding the processing of data, that has a direct influence on the final result to be obtained, is visible. It should also be noted that the processing of the data, and the steps that constitute it, are variable, that is, the techniques exposed may not provide the best possible results because the used data has a direct influence on these same results. Thus, it will be necessary to carry out comparative tests between different sets of processing steps and models, to determine which is the best set of steps and models that provides the best results for the classification of SKOs.

From the identified ML algorithms, the traditional SVM and NB algorithms continue to be the most used, and with satisfactory results in terms of accuracy. However, the LSTM algorithm proved to be a viable alternative to traditional algorithms, having also obtained good results in terms of the accuracy of the classifications.

Future work will be related to the development of the classification system hopping to reach a standardized classification system that solves the information-seeking problems present on the current and future research projects

Acknowledgments

This work has been supported by IViSSEM: POCI-01-0145-FEDER-28284.

References

[1] J. Webster and R. T. Watson, “Analyzing the Past To Prepare for the Future : Writing a Literature Review,” *MIS Q.*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[2] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, “The State-of-the-Art in Twitter Sentiment Analysis,” *ACM Trans. Manag. Inf. Syst.*, vol. 9, no. 2, pp. 1–29, Aug. 2018.

[3] N. Boudad, R. Faizi, R. Oulad Haj Thami, and R. Chiheb, “Sentiment analysis in Arabic: A review of the literature,” *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 2479–2490, Dec. 2018.

[4] A. Romanov, K. Lomotin, and E. Kozlova, “Application of natural language processing

algorithms to the task of automatic classification of Russian scientific texts," *Data Sci. J.*, vol. 18, no. 1, pp. 1–17, Aug. 2019.

[5] F. S. Nurfikri and Adiwijaya, "A comparison of Neural Network and SVM on the multi-label classification of Quran verses topic in English translation," in *Journal of Physics: Conference Series*, 2019, vol. 1192, no. 1.

[6] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, Dec. 2019.

[7] S. Saini, S. P. Singh, and R. Agarwal, "Augmented machine learning ensemble extension model for social media health trends predictions," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 7, pp. 482–486, Jul. 2019.

[8] Y. Cheng, Z. Ye, M. Wang, and Q. Zhang, "Document classification based on convolutional neural network and hierarchical attention network," *Neural Netw. World*, vol. 29, no. 2, pp. 83–98, 2019.

[9] P. Semberecki and H. MacIejewski, "Deep learning methods for subject text classification of articles," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, 2017, pp. 357–360.

[10] F. L. Hernández, E. V. Pérez, J. Javier Rainer Granados, and R. G. Crespo, "A Nondisturbing Service to Automatically Customize Notification Sending Using Implicit-Feedback," *Sci. Program.*, vol. 2019, 2019.

[11] J. wei Fan *et al.*, "Part-of-speech tagging for clinical text: wall or bridge between institutions?," *AMIA Annu. Symp. Proc.*, vol. 2011, no. May 2014, pp. 382–391, 2011.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.

[13] S. George K and S. Joseph, "Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature," *IOSR J. Comput. Eng.*, vol. 16, no. 1, pp. 34–38, 2014.

[14] C. Nicholson, "A Beginner's Guide to Bag of Words & TF-IDF," 2019. [Online]. Available: https://pathmind.com/wiki/bagofwords-tf-idf. [Accessed: 18-Jan-2020].

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, pp. 1532–1543.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

[17] A. I. Wiki, "A Beginner's Guide to Word2Vec and Neural Word Embeddings | Skymind," 2019. [Online]. Available: https://pathmind.com/wiki/word2vec. [Accessed: 18-Jan-2020].

[18] C. Caragea, F. Bulgarov, and R. Mihalcea, "Co-training for topic classification of scholarly data," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2357–2366.

[19] S. K. Bhaskaran, C. Sreejith, and P. C. Rafeeque, "Neural networks and conditional random fields based approach for effective question processing," in *Procedia Computer Science*, 2018, vol. 143, pp. 211–218.

[20] T. G. Ghumade and R. A. Deshmukh, "A document classification using NLP and recurrent neural network," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 632–636, Aug. 2019.

[21] R. Jing, "A Self-attention Based LSTM Network for Text Classification," in *Journal of Physics: Conference Series*, 2019, vol. 1207, no. 1.

[22] D. Kaplan, T. Tokunaga, and S. Teufel, "Citation block determination using textual coherence," *J. Inf. Process.*, vol. 24, no. 3, pp. 540–553, 2016.

[23] E. Khabiri, W. M. Gifford, B. Vinzamuri, D. Patel, and P. Mazzoleni, "Industry Specific Word Embedding and its Application in Log Classification," 2019, pp. 2713–2721.

[24] C. Montenegro, R. Santana, and J. A. Lozano, “Data generation approaches for topic classification in multilingual spoken dialog systems,” in *ACM International Conference Proceeding Series*, 2019, pp. 211–217.

[25] A. Y. Romanov, K. E. Lomotin, E. S. Kozlova, and A. L. Kolesnichenko, “Research of neural networks application efficiency in automatic scientific articles classification according to UDC,” in *2016 International Siberian Conference on Control and Communications, SIBCON 2016 - Proceedings*, 2016.

[26] D. Tang, B. Qin, and T. Liu, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classificatio,” in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1422–1432.

[27] D. Tang, B. Qin, and T. Liu, “Document modeling with gated recurrent neural network for sentiment classification,” in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1422–1432.

[28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations ofwords and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013.

[29] G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification,” 2003.

[30] Y. Kim, “Convolutional neural networks for sentence classification,” in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, pp. 1746–1751.

[31] K. M. O. Nahar, A. F. Al Eroud, M. Barahoush, and A. M. Al-Akhras, “SAP: Standard Arabic profiling toolset for textual analysis,” *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 222–229, Apr. 2019.

[32] Z. Ratkovic, W. Golik, and P. Warnier, “Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach.,” *BMC Bioinformatics*, vol. 13 Suppl 1, 2012.

[33] X. Yue, G. Di, Y. Yu, W. Wang, and H. Shi, “Analysis of the combination of natural language processing and search engine technology,” in *Procedia Engineering*, 2012, vol. 29, pp. 1636–1639.

[34] J. Zhou, B. cheng Li, and G. Chen, “Automatically building large-scale named entity recognition corpora from Chinese Wikipedia,” *Front. Inf. Technol. Electron. Eng.*, vol. 16, no. 11, pp. 940–956, Nov. 2015.

[35] K. M. O. Nahar, N. Alhindawi, O. M. Al-Hazaimeh, R. M. Alkhatib, and A. M. Al-Akhras, “NLP and IR based solution for confirming classification of research papers,” *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 16, pp. 5269–5279, 2018.

[36] G. I. Ulumudin, A. Adiwijaya, and M. S. Mubarok, “A multilabel classification on topics of qur’anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF,” in *Journal of Physics: Conference Series*, 2019, vol. 1192, no. 1, pp. 1–4.