Topic Classification of Publications

Identifying publication topics based on existing journals

Dayoung Lim

Supervisor(s): Diomidis Spinellis, Georgios Gousios

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 29, 2024

Name of the student: Dayoung Lim
Final project course: CSE3000 Research Project
Thesis committee: Diomidis Spinellis, Georgios Gousios, Koen Langendoen

Abstract

Accurate topic classification is crucial in the scientific community when it comes to finding relevant journals. However, the efficiency and accuracy of topic classification of publications do not seem to be at its best performance, especially with the fast-paced rise in the quantity of research papers. Our research aims to address this problem by utilizing state-of-the-art (SOTA) methods. We chose the ’April 2022 Crossref’ data set for the research, as Alexandria3k, the tool utilized for querying on the open data set, is tested on the same data. We stratified 50,000 data that have title, abstract, and work names, which are the roughly assigned topics. SOTA methods chosen for feature extraction and classification models are OpenAI Embeddings and XGBoost. Our research shows that this combination of SOTA methods has the potential to improve the performance of current topic classification of publications.

1 Introduction

Researchers are actively working on a solution to effectively query and select relevant literature [25], [10]. This is an important research topic, especially for the science community, as the pursuit of knowledge relies heavily on existing journals. Researchers depend on existing journals to identify emerging trends, recognize gaps in knowledge, and address them for scientific advancement. While the volume of published research grows exponentially, lack of transparency, repeatability, technical constraints in algorithms, along with poor management of journal database [24] led to increasing difficulty in searching relevant scientific journals efficiently.

An approach to addressing the inefficiencies in journal searching is the accurate identification of publication topics. Esha Datta [5], for instance, researches methods to automate journal subject classification using different combinations of methods. However, due to the limited data points available for the research, the study was not able to achieve the desired level of accuracy. This outcome highlights a common challenge in such research: the efficiency of multi-label classification relies on the data sets used. In addition to this, topic classification research is done for a specific subject area, like the medical field [8], [21]. In our research, we do not limit the data to a specific field of study except for the fact that they are scientific publications. This is more challenging work due to the wide range of topics, yet it is a more reasonable approach for the problem we are solving as common tools used for researchers, such as Google Scholar, also consist of journals from various fields.

This paper ultimately aims to answer the following question: ”*How can publication topics be identified and matched based on existing journal topic values?*” Our research focuses on advancing the method of topic classification for scientific publications using state-of-the-art (SOTA) feature extraction and supervised machine learning approaches. To achieve this, we utilize Alexandria3k (A3k), a command-line tool to perform relational queries on an open metadata set [23].

The report consists of five sections. In section 2, we begin by analyzing existing works on the topic in depth and identify a knowledge gap to be solved through this research. With the analysis, section 3 discusses the methods used for the research in detail and how this can lead to an improvement in the existing field. Using the mentioned methods, the research is conducted, and its results are evaluated in section 4. Section 5 elaborates on the overall implications and limitations of the study based on the results. Following the discussions, section 6 reflects on the ethical implications that the research might have. Lastly, section 7 gives suggestions of how the research can be developed further.

2 Related Works

Classification works have been researched from as early as the 1960s [14] due to their useful applications in various fields. Many researchers have examined the different classification models with feature extraction techniques to discover the optimal combination for specific applications. In order to apply appropriate methods for the topic classification task, acknowledging and understanding existing works is crucial.

2.1 Classification Models

Significant research has been dedicated to developing effective models for topic detection within given texts. This includes traditional methods such as Naive Bayes (NB) and Support Vector Machine (SVM) [26] as well as the ones from recent advancements in deep learning like Bidirectional Encoder Representations from Transformers (BERT) [14].

While traditional text classification models have shown promising results in most cases, they do not perform the best in multi-label classification on large data sets. SVM, for instance, is shown to be computationally expensive and less effective, especially on imbalanced data sets [2]. Meanwhile, deep learning models perform well on multi-label classification with large data sets, but they also encounter computational challenges. Furthermore, existing research on topic classification mainly uses full documents, which can be expensive, or a lot of short text-based classification works are sentence-based, which would be difficult to represent the content of the whole publication for instance.

As an approach to balancing computational cost and performance, this study has chosen to focus on using abstracts and titles. Furthermore, considering the size of the data set of Crossref (160GB) and that the task is multi-label classification, it is important to consider the scalability and efficiency of a model. As a result, SOTA eXtreme Gradient Boosting (XGBoost) has been investigated in [22].

2.2 Feature Extraction

Traditional models often utilize feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec for model performance improvement [20]:

- TF-IDF [4]: term *t*, document *d*, total number of documents *N*, number of documents containing term *t df*

$$TF(t, d) = \frac{\text{number of times t appears in d}}{\text{total number of terms in d}}$$
$$IDF(t) = \log \frac{N}{1+df}$$
$$TF - IDF = TF(t, d) * IDF(t)$$

- Word2Vec: Represents words as dense vectors, where semantically similar words are mapped to nearby points in the vector space.

Despite the frequent use of these methods in many research [1], [17], there are crucial limitations. For instance, TF-IDF cannot perform semantic analysis, as it only considers the frequency of a term. On the other hand, Word2Vec can conduct semantic analysis but is not optimal for multilingual datasets and struggles with newly encountered words [3].

The SOTA in this domain is represented by OpenAI Embeddings, offering advanced semantic analysis capabilities [12], [15], along with high efficiency, speed, and relatively low cost. However, as it is not a free tool, there is limited research on using it for topic classification.

2.3 Knowledge Gap

Similar works for topic classifications have been researched using various methods, including the ones mentioned in previous section, on different applications. This includes topic identification for news articles [6] and social media [13].

For multi-label classification on publications, the works have been done on full documents, which can be costly, or on bibliometric features [11]. One of the most relevant research is done specifically in the biomedical field [21]. For instance, A. Deepika and N. Radha conducted research on an abstract-based classification, where they achieved 94% accuracy [8]. The main difference that is important to note, however, is that we use a multilingual dataset of publications across any scientific domain as well as we aim to perform a multi-label classification, whereas the Deepika and Radha's work primarily focuses on English-based journals that are specific to the biomedical and life sciences fields.

Esha Dotta [5], as mentioned before, performed topic classification research in an automated way. The study mainly compares the combination of different methods: TF-IDF + Linear Support Vector Classification, Embeddings (SciBERT) + Linear Support Vector Classification, and OpenAI LLM + sentence completion. This study provides an overview of classifying publications based on the titles of journals and articles. A title-based approach can be cost-efficient; however, as the research conclusion also suggests, using more data, such as a full document, can lead to a different result. Our research utilizes more information than just the title of the data set: abstract and title.

3 Methodology

Based on existing works and in-depth research, we use OpenAI Embeddings and the XGBoost combination to classify publications. Crossref data set, specifically the April 2022 version, is used as A3k has been tested against it [23]. This data set contains information about a varying corpus of journals in multiple languages across different fields of study, and a number of journals already have topics assigned that are based on Scorpus topic values, making it an ideal data set for topic classification. The overall workflow is shown in figure 1, and each step is discussed in this section. Technical implementation of each steps can be found in Github [7].

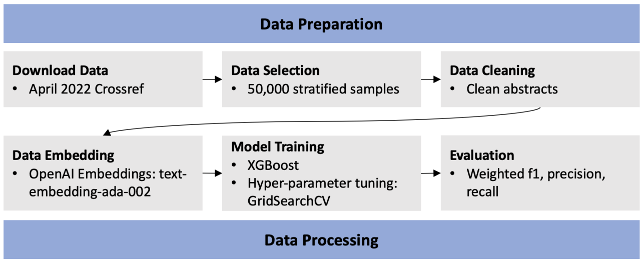

Figure 1: Workflow of the research

3.1 Data Preparation

As the Crossref data set contains various information, including those that are not related to the research, as can be seen in Appendix A, separate data selection step is performed. Among the information existing in Crossref, we use the title, abstract, and work names of the journals for the research. The reason why these fields were chosen can be explained as follows:

- The title provides a brief overview of what the publication is about. There are works done around topic classification based on the title, which has been shown to be one of the most effective ways to do a classification [18].
- The abstract contains a summary of major findings from the research; hence, it contains keywords and terms that are relevant to topic finding.
- Work names contain the topic(s) for some journals. This data is the ground truth to be trained and tested against throughout the research.

Selecting all data that has these three elements mentioned above results in 10,663,104 data points. We performed several initial runs to compare the performance of OpenAI Embeddings and XGBoost with low cost as well as using different features to select for the final run. For this, we selected 10,000 data within 50 topic values. For the final run, we used stratification technique to sample 50,000 data points as conducting the research on 10M data is costly. Stratification technique is where a random amount of sample is taken for each stratum, which would work names in our research, while preserving the proportion of the stratum equal to the original data set [16].

With the selected data, a cleaning process is done for the abstract in order to reduce the noise of the data. As it can be seen from table 1, there are elements that are not relevant for understanding the abstract content.

Abstract
<jats:p>The subgenus includes P. exanthematica exanthematica (Scopoli, 1763) ... <jats:p>
<jats:title>Abstract</jats:title><jats:p>Routine silvicultural practices continue ... <jats:p>
<jats:p>El presente trabajo de investigación intenta poner de manifiesto... </jats:p>

Table 1: Example abstracts before data cleaning

As OpenAI Embeddings can perform semantic analysis, data cleaning is relatively simpler compared to existing work on other feature extraction models. Other feature extraction models require lowercase, stemming, and the removal of all punctuation. However, these steps for our research can result in different contextual meanings resulting in worse semantic understanding. It is also important to note that, as Crossref is a multilingual data set, performing certain data cleaning steps, such as removing punctuation, can differ the meaning of the text, it is not performed. Instead, we proceed with other standard data cleaning processes, such as removing HTML tags, LaTex artefacts, and URLs, normalizing white space, and trimming leading and trailing white space. In addition to this, Crossref contains abstracts with the term 'abstract' in the beginning to further clarify that the upcoming sentences are part of an abstract. This, however, is unnecessary as it does not add meaning to the content of the abstract. Therefore, this word has been removed. Table 2 shows how the abstract has been cleaned following the mentioned steps compared to table 1.

Abstract
The subgenus includes P. exanthematica exanthematica (Scopoli, 1763) ...
Routine silvicultural practices continue ...
El presente trabajo de investigación intenta poner de manifiesto ...

Table 2: Example abstracts after data cleaning

3.2 Data Embedding

To convert the data into machine-understandable language, an embedding step is needed. In our research, the OpenAI Embeddings, specifically the text-embedding-ada-002 model, is used. The main reason for this choice is because it provides semantic analysis, which we expect to perform better than previous feature extraction techniques, which are based on bags or words [27]. Furthermore, this embedding model provides support for multilingual data sets, unlike the previous versions, and is transformer-based [19].

The input for OpenAI Embeddings needs to be one string. As the study uses both the title and the abstract of publications, a step to combine these fields for each data point into one string is needed. As this embedding model understands texts based on context, a simple step of adding 'title:' and 'abstract:' and combining them would be sufficient. An example of this can be seen in table 3. The labels, in the meantime, are created by splitting the work names data into comma-symbols.

Title	Abstract
Species of the subgenus Psacasta s. ...	The subgenus includes P. exanthematica exanthematica (Scopoli, 1763) ...
Effects of Thinning and Herbicide Treatments on ...	Routine silvicultural practices continue ...
El cambio de nivel: todo un desafío	El presente trabajo de investigación intenta poner de manifiesto...

Title + Abstract
title: Species of the subgenus Psacasta s. ... abstract: The subgenus includes P. exanthematica exanthematica (Scopoli, 1763) ...
title: Effects of Thinning and Herbicide Treatments on ... abstract: Routine silvicultural practices continue ...
title: El cambio de nivel: todo un desafío abstract: El presente trabajo de investigación intenta poner de manifiesto ...

Table 3: Example combined input for OpenAI Embeddings

For the initial run, it costed $0.2 while it for stratified sample, it costed roughly $2.8. This cost can be predicted based on the token estimation method provided by OpenAI [19].

3.3 Model Training

The results from the OpenAI Embeddings step are used by XGBoost to train and test its performance for topic classification. The data set is split in such a way that 80% is used for training and the remaining for testing. With the split data set, the training data utilizes scikit-learn's OneVSRestClassifier for multi-label classification along with XGBClassifier provided by XGBoost. XGBClassifier consists of several key hyper-parameters:

- Objective specifies the learning objective. For this research, it is set as 'binary:logistic' which implies logistic regression for binary classification and outputs probability [9].
- The eval metric is an evaluation metric for the validation data and is set based on the objective. For the 'binary:logistic' objective, the 'logloss' eval metric is commonly used, and we also use this specific one.
- eta is a learning rate that ranges between 0 and 1. The default is 0.3.
- n_estimators is the number of boosting rounds for the gradient boosting model, and the default is set to 100.
- max_depth is the maximum depth of the tree. The value of max_depth starts at 0, and there is no limit. However, the higher the number, the more complex the model gets and the higher the probability of overfitting [9].

GridSearchCV is used for tuning these hyper-parameters. The parameter grid for tuning the final run is 'eta: 0.0001, 0.3, 0.5, 0.8', 'n_estimators: 50, 100, 500, 1000', and 'max_depth: 3, 6, 16, 20'. Running the grid search on this parameter grid resulted in 0.8, 1000, and 20 as optimal values for eta, n_estimators, and max_depth for the stratified data.

4 Results

For the initial runs, we had two rounds of comparison. Firstly, we compared the performance of the combination of OpenAI Embeddings and XGBoost to a baseline method, BM25 and XGBoost. The result was that our method performed slightly better than BM25, where the weighted f1 score was 0.56 and 0.51, respectively. After this, a comparison of different fields chosen for the classification was mainly researched into, for instance, selecting only abstract or title + author + abstract or title + abstract. The results were evaluated using weighted f1, and for further understanding, the micro-average precision and recall have been taken into account.

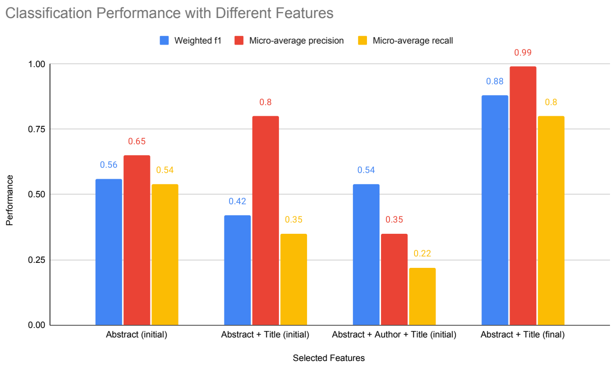

Figure 2: Performance comparison of initial runs with different features and the final run

As it can be seen from figure 2's initial runs, overall the performance of choosing only the abstract performed better in terms of weighted f1. However, for the abstract and title combination, despite weighted f1 showing slightly lower performance than the other two, its micro-average precision is significantly high. This implies that when the model predicts, then the outcome is highly likely to be correct. One of the reasons why precision is high but recall is low could have been caused by the label imbalance as well as the inaccurate representation of the data set. As a result, we decided to use abstract and title features for the larger set of data, along with stratification sampling technique. These decisions were made to ensure accurate representation of the original data set and to increase the possibility of enhanced recall performance.

Figure 2's last result shows the final results for running our chosen methods. Each of the evaluation metric scores of weighted f1, micro-average precision, and micro-average recall showed 0.88, 0.99, and 0.8. These are significantly higher scores than the initial run on an unstratified 10,000 data set with 50 topic values. The detailed analysis and discussion are written in section 5.

5 Discussion

The results of the research are important to discuss in order to critically reflect on what they convey in terms of answering the research question. Furthermore, by acknowledging the limitations of the work, future researchers will be able to take into consideration how the research can be improved.

5.1 Result Analysis

As it can be seen from figure 2, the performance difference between the initial setup and the final has shown a drastic difference. This could be due to several factors. The following paragraphs discuss these possible reasons and what the results imply in relation to the research question we have.

One of the factors that could have caused a significant difference is the quantity and quality of the data. As mentioned earlier, initial data limits the topic values to be represented at 50, whereas in the actual run on samples, we made sure that all topic values are included. Furthermore, the amount of data we used increased from 10,000 to 50,000, meaning there is more data for the model to be trained on for better performance. Along with these, the main important factor is the usage of the sampling technique. Instead of choosing random works within the 50 topic values, we were able to accurately represent the original data by stratifying. This is another factor why the sample size increased, such that there is more data to reflect on the increased topic values.

The results reflect the importance of quality and quantity of data. Moreover, the combination of OpenAI Embeddings and XGBoost showed promising results. Considering A. Deepika and N. Radha using abstract achieved around 0.95 weighted f1 score [8] on English only and non-multi-label classification, our result of 0.88 weighted f1, 0.99 precision, and 0.8 recall is a comparable figure. Especially with 0.99 precision, it is hopeful for a better result when using the method with the whole Crossref data set. With the consideration of limitations discussed in the following subsection and related future works suggested, our proposed method has room for even better results than it is at the moment.

5.2 Limitations

The main limitation of this research is its high computational cost. This includes the physical cost that is incurred by using OpenAI Embeddings and the time it takes to work through the whole process. Due to this, the full Crossref data set was not used, but only a stratified sample was. As mentioned earlier, the quantity of data also affects the outcome. Furthermore, despite each topic being represented in the actual experiment, there is no consideration of representing different languages in the stratification method. The larger the data set, the higher the chance of including more data for different languages.

Continuing with the cost limitation, it is important to acknowledge that the research is only performed on the Crossref data set. This implies that the method is not tested against a random collection of data. Even if this method performed well in this research, that does not imply that this particular topic classification method would work on other sets of data, and vice versa.

Another limitation to notice is that the correctness of the original data set has not been checked. Even if it is a public data set, it is possible that there are a few where the topics are misclassified. The verification process was not performed due to the time limitation. There are 10M relevant pieces of data, and it is difficult to go through them one by one. There are different ways to solve this problem and one of them is discussed in section 7.

6 Responsible Research

The research contains several ethical aspects, particularly biases, that could have affected the results. It is important to critically reflect on them and be aware of them in order to thoroughly understand the steps we have taken for topic classification. In addition to this, the reproducibility of our method is discussed in order to ensure the research is done responsibly.

6.1 Ethical Aspects

A number of biases arise from the nature of the Crossref data set. The data set is predominantly in English, implying that there is significantly less data for the model to learn and train to predict correctly for non-English publications. This can lead to inaccurate topic classification for the languages that are not represented enough. Similarly, the Crossref data set contains journals from varying fields, and the representation of each topic is imbalanced. As a result, a selection bias can exist for overrepresented topics. Along with the characteristics of the data set, it is important to note that the research has only been tested on Crossref; hence, the performance of the procedures done might not show similar results with different data sets.

Along with the characteristics of the Crossref data set, OpenAI embedded systems have their own limitations regarding biases. According to OpenAI, it has been detected that ”the models encode social biases such as stereotypes or negative sentiment towards certain groups” [19]. This could have affected the embedding procedure where it interprets natural languages, where social biases might come into play, and encodes into embedding. Furthermore, the different conventions of titles and abstracts could have affected the extraction of topics, making the accuracy of the classification lower.

6.2 Reproducibility

The data set used, the ’April 2022 Public Data File’ from Crossref, is publicly available. However, due to the size of the data set, it can be difficult to reproduce. Our GitHub repository [7] contains instructions on how to get the data and prepare them for consecutive steps of the research. In order to reproduce the research steps, it is recommended to have around 160GB of storage available for the data set itself.

As mentioned earlier, the source code is available on the GitHub repository, which is accessible to everyone. The repository includes files of codes for steps described in section 3, evaluation methods, and instructions on how to run them to obtain the results. However, the intermediate results are not included due to size limit of files from GitHub.

7 Future Work

To improve the research results further based on the limitations discussed earlier, verification of the methods on different conditions is needed. Some of the ways to perform this can be testing the methods on different data sets other than Crossref is advised, performing verification after classifying works without work names, or manually comparing the outcome that does not match with topics already in Crossref after classifying. This would require two raters and a referee, and these roles can be selected within the group. A way of determining the sample size is by using Cochran’s formula.

Furthermore, when working on a similar research, take into consideration that working with the full Crossref data has the potential to give more insight to the research. However, it can be costly, which is why we used the stratification sampling technique instead. Hence, if the resources allow, using the full data set can be an option.

Lastly, as of January 25, 2024, a new embedding models ’text-embedding-3-small’ and ’text-embedding-3-large’ have been announced for release, which have stronger performance than the model that we currently use [28]. This can be used for further improvement in the performance of topic classification.

8 Conclusion

With the difficulties that researchers are facing with regards to searching relevant journals due to poor classification of publications, we have conducted research on the question: *How can publication topics be identified and matched based on existing journal topic values?* The research started with identifying the problem, existing related works, and knowledge gaps, preparing data in a way to ensure the best outcome, choosing appropriate methods, evaluating the results and discussing the limitations.

Existing works regarding topic classification have mainly two reasons why they are not suitable for our research: traditional models like Support Vector Machine do give promising results but are not optimal for multi-label classification, and recent models like Bidirectional Encoder Representations from Transformers are computationally expensive. Considering that Crossref is multilingual metadata, state-of-the art feature extraction tool and classification model like OpenAI Embeddings and XGBoost are chosen for the research.

After identifying the knowledge gap and deciding on the tools to be used for the research, the data preparation step was performed. We started with 10,000 data with 50 work names to compare OpenAI and XGBoost combination to the BM25 and XGBoost combination. OpenAI has shown a slightly better weighted f1. Based on the result, we compared the results of selecting different features and decided to proceed the research with abstract and title as the features. Based on the results, we used stratification sampling technique to get 50,000 data that included title, abstract, and work names from the whole data set. After sampling, data cleaning process including removing HTML tags, LaTex artefacts, and handling white spaces was performed. Prepared data was then processed via OpenAI Embeddings, with title and abstract represented in one string. The results were used by XGBoost to train and test the labels. By tuning the hyper-parameters using GridSearchCV, we ensured that the best result was obtained.

The results of our methods are promising, giving 0.88, 0.99, and 0.8 for each weighted f1, micro-averaged precision, and micro-averaged recall. The performance improved significantly with the increase in the amount of data used for the research and the usage of the stratification sampling method to accurately represent the original data set.

Despite the methods showing promising results, it is important to consider that the research has not done a verification of the methods on different data sets other than Crossref and works without topic values. Future works can include solving these limitations as well as making use of the newer model for OpenAI Embeddings: text-embedding-3-small and text-embedding-3-large.

Overall, our research on using SOTA tools, like OpenAI and the XGBoost combination, has shown to be an assuring answer to the research question. Meanwhile, by considering the limitations and future works mentioned earlier, our research also has the potential to improve for better performance.

References

[1] CAHYANI, D. E., AND PATASIK, I. Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics 10*, 5 (2021), 2780–2788.

[2] CERVANTES, J., GARCIA-LAMONT, F., RODRÍGUEZ-MAZAHUA, L., AND LOPEZ, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing 408* (2020), 189–215.

[3] CHANDRAN, S. Introduction to text representations for language processing-part 2, Nov 2021.

[4] CHEN, K. Introduction to natural language processing-tf-idf, May 2021.

[5] DATTA, E. *Attempts at automating journal Subject classification* (2023).

[6] DAUD, S., ULLAH, M., REHMAN, A., SABA, T., DAMAŠEVIČIUS, R., AND SATTAR, A. Topic classification of online news articles using optimized machine learning models. *Computers 12*, 1 (2023).

[7] DAYOUNG, L. Topic classification of publications. https://github.com/DayoungLim/topic-classification-publications, 2024.

[8] DEEPIKA, A., AND RADHA, N. Performance analysis of abstract-based classification of medical journals using machine learning techniques. In *Computer Networks and Inventive Communication Technologies* (Singapore, 2022), S. Smys, R. Bestak, R. Palanisamy, and I. Kotuliak, Eds., Springer Singapore, pp. 613–626.

[9] DEVELOPERS, X. Xgboost parameters. https://xgboost.readthedocs.io/en/stable/parameter.html.

[10] GREWAL, A., KATARIA, H., AND DHAWAN, I. Literature search for research planning and identification of research problem. *Indian Journal of Anaesthesia 60*, 9 (Sep 2016), 635.

[11] HU, Y.-H., TAI, C.-T., LIU, K. E., AND CAI, C.-F. Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. *Journal of Informetrics 14*, 1 (2020), 101004.

[12] KHEIRI, K., AND KARIMI, H. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning, 2023.

[13] KUSUMAWARDANI, R., AND BASRI, M. Topic identification and categorization of public information in community-based social media. *Journal of Physics: Conference Series 801*, 1 (jan 2017), 012075.

[14] LI, Q., PENG, H., LI, J., XIA, C., YANG, R., SUN, L., YU, P. S., AND HE, L. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol. 13*, 2 (apr 2022).

[15] MARS, M. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences 12*, 17 (2022).

[16] MENON, S. Stratified sampling in machine learning. https://medium.com/analytics-vidhya/stratified-sampling-in-machine-learning-f5112b5b9cfe, Dec 2020.

[17] MOHAMMED, M., AND OMAR, N. Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PLOS ONE 15*, 3 (2020).

[18] NOVA, K. Machine learning approaches for automated mental disorder classification based on social media textual data. *Contemporary Issues in Behavioral and Social Sciences 7*, 1 (Apr 2023), 70–83.

[19] OPENAI. Openai embeddings. https://platform.openai.com/docs/guides/embeddings/limitations-risks, 2023.

[20] SEETHALAKSHMI, Y. M. M., ANDAVAR, S., AND RAJ, R. S. P. A survey on feature extraction techniques, classification methods and applications of sentiment analysis. *Brazilian Archives of Biology and Technology 66* (2023), e23220654.

[21] SING, D. C., METZ, L. N., AND DUDLI, S. Machine learning-based classification of 38 years of spine-related literature into 100 research topics. *Spine 42*, 11 (2017), 863–870.

[22] SONG, R., LI, T., AND WANG, Y. Mammographic classification based on xgboost and dcnn with multi features. *IEEE Access 8* (2020), 75011–75021.

[23] SPINELLIS, D. Alexandria3k documentation. https://dspinellis.github.io/alexandria3k/index.html, 2022.

[24] SPINELLIS, D. Open reproducible scientometric research with Alexandria3k. *PLoS ONE 18*, 11 (Nov. 2023), e0294946.

[25] VAN WEE, B., AND BANISTER, D. Literature review papers: The search and selection process. *Journal of Decision Systems* (2023), 1–7.

[26] WANG, S., AND MANNING, C. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Jeju Island, Korea, July 2012), H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, and J. C. Park, Eds., Association for Computational Linguistics, pp. 90–94.

[27] WINASTWAN, R. Text classification with transformer encoders, Aug 2023.

[28] ZHUANG, J., BRAUNSTEIN, A., NEELAKANTAN, A., JIAO, J., BALTESCU, P., BOYD, M., YATBAZ, M., KIVLICHAN, I., PENG, A., AGRAWAL, A., AND ET AL. New embedding

models and api updates. https://openai.com/blog/new-embedding-models-and-api-updates, Jan 2024.

Appendix

A Crossref Schema

This is the data schema of Crossref data set. It consists of the tables and columns of the Crossref.

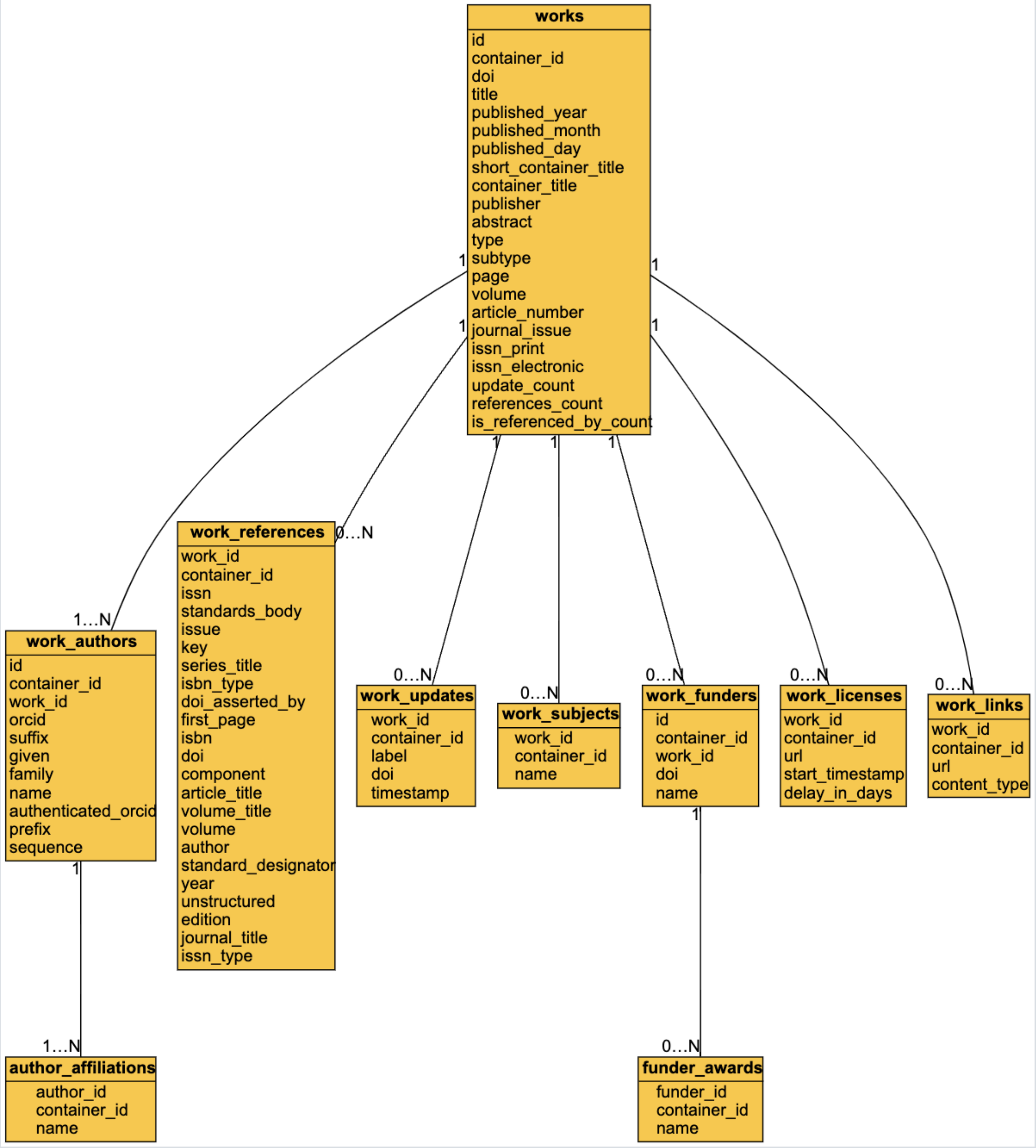

Figure 3: Crossref Schema [23]