Benchmark for Research Theme Classification of Scholarly Documents

Óscar E. Mendoza[1] **Wojciech Kusa**[2] **Alaa El-Ebshihy**[2]
Ronin Wu[3] **David Pride**[4] **Petr Knoth**[4] **Drahomira Herrmannova**[5]
Florina Piroi[2] **Gabriella Pasi**[1] **Allan Hanbury**[2]

[1]University of Milano-Bicocca, Milan, Italy
oscar.espitiamendoza@unimib.it
[2]TU Wien, Vienna, Austria
[3]IRIS.ai, Stabekk, Norway
[4]Knowledge Media institute, The Open University, Milton Keynes, U.K.
[5]Elsevier, U.S.

Abstract

We present a new gold-standard dataset and a benchmark for the *Research Theme Identification* task, a sub-task of the Scholarly Knowledge Graph Generation shared task, at the 3rd Workshop on Scholarly Document Processing. The objective of the shared task was to label given research papers with research themes from a total of 36 themes. The benchmark was compiled using data drawn from the largest overall assessment of university research output ever undertaken globally (the Research Excellence Framework - 2014).

We provide a performance comparison of a transformer-based ensemble, which obtains multiple predictions for a research paper, given its multiple textual fields (e.g. title, abstract, reference), with traditional machine learning models. The ensemble involves enriching the initial data with additional information from open-access digital libraries and Argumentative Zoning techniques (Teufel et al., 1999b). It uses a weighted sum aggregation for the multiple predictions to obtain a final single prediction for the given research paper.

Both data and the ensemble are publicly available on https://www.kaggle.com/ and https://github.com/ProjectDoSSIER/sdp2022, respectively.

1 Introduction

With the recent demise of the widely used Microsoft Academic Graph (MAG) (Sinha et al., 2015), the scholarly document processing community is facing a pressing need to replace MAG with an open-source community-supported service. In order to create a comprehensive scholarly graph, it is challenging to correctly represent each paper as a node on the graph. This requires condensing meta-information, such as authorship, research organizations, research themes etc., of research papers to one node.

So far, the task of identifying research themes for a given scholarly document has been challenging due to the lack of large high-quality labelled data. This made it difficult both to train high-performance classification models as well as to compare models' performance across studies.

This paper provides a benchmark for research theme classification based on a large human-annotated corpus of scholarly papers across 36 themes defined by the UK Research Excellence Framework, the largest overall assessment of university research outputs ever undertaken globally (the Research Excellence Framework - 2014)[1] (Cressey and Gibney, 2014). The outcome of this paper is the product of the Scholarly Knowledge Graph Generation shared task which was part of the Scholarly Document Processing (SDP) workshop at COLING2022.

We started with a labelled dataset containing publications and subjects to which they belong (Section 3), which contains descriptions or abstracts, the first author, DOI, year of publication, and identifier to link the publication to the CORE (Knoth and Zdrahal, 2012) aggregator. We later enriched this dataset with further information including the full text, where available. This represents a new gold-standard dataset for theme classification of scholarly documents.

To establish a benchmark for research theme classification, we present experiments and evaluation results with traditional machine learning models and compare them to a more sophisticated transformer-based ensemble model.

Our transformer-based ensemble model exploits

[1]https://ref.ac.uk/2014/

Proceedings of the Third Workshop on Scholarly Document Processing, pages 253–262
October 12–17, 2022.

all textual fields for each scholarly document and maps these documents to CORE and Semantic Scholar (Fricke, 2018) to gather further external information. Thus, the ensemble consists of a transformer-based classifier used to produce multiple predictions for individual publications (split into multiple textual fields) that are aggregated to produce a single final prediction. We aggregate predictions from titles, abstracts, references, citations, and related titles for every publication, when available. Furthermore, we use abstracts, PDFs and full texts available to identify argumentative zones (Teufel et al., 1999b) to use them as additional fields. We report on the results of using aggregation for different combinations of these predictions.

The rest of the paper is organized as follows: Section 2 presents a discussion of the related work, focusing mainly on scientific document classification approaches and their evaluation. Section 3 describes details of building the new benchmark for theme classification. Section 4 discusses the ensemble we propose as a baseline and the system components in more detail. In Section 5, we describe the experimental settings. In Section 6, we discuss the evaluation results from a diverse set of experiments. Finally, we discuss the conclusion and the potential direction of future work in Sections 7 and 8.

2 Related Work

Classifying scholarly documents is an important task, whether for understanding the dynamics of scientific fields or simply for organizing scientific literature more effectively. In previous literature, it typically relies on textual features such as titles, author keywords, and abstracts, as well as the interrelationships between the documents (i.e., citations and co-authorship). Full texts are frequently not available and processing a large amount of text can be computationally expensive.

A wide variety of classification features have been proposed at different levels of granularity, e.g., themes, topics, and subjects. A large proportion of classification methods rely on semantic similarity (Wang and Koopman, 2017; Semberecki and Maciejewski, 2017; Salatino et al., 2022; Hande et al., 2021; Boyack and Klavans, 2018). Others include approaches for clustering documents based on keyword co-occurrence (Van Eck and Waltman, 2017; Kim and Gil, 2019). Further approaches leverage the relationship graph representation built from citations and co-authorship (Taheriyan, 2011; Shen et al., 2018; Hoppe et al., 2021).

One promising but unexplored approach to theme classification is using information about argumentative zoning (AZ) (Teufel et al., 1999b). AZ refers to the examination of the argumentative status of sentences in scientific articles and their assignment to specific argumentative zones. Its main goal is to collect sentences that belong to predefined zones, such as "claim" or "method". Annotated AZ corpora has been created by (Teufel et al., 1999a,b; Teufel and Moens, 2002; Teufel et al., 2009) with approaches to AZ identification reported in (Liu, 2017). In this work, we aim to test to what extent can the AZ signal support classification of scholarly documents into research themes.

Classification models previously appplied to this task include traditional machine learning models, such as k-Nearest Neighbours (Waltman and Van Eck, 2012; Łukasik et al., 2013), K-means (Kim and Gil, 2019) and Naïve Bayes (Eykens et al., 2021). It has been reported that these models encounter performance challenges related to overly coarse classifications and low accuracy (Daradkeh et al., 2022). There are applications of deep neural networks (NN) models as well, such as convolutional NN (Rivest et al., 2021; Daradkeh et al., 2022) and recurrent NN (Semberecki and Maciejewski, 2017; Hoppe et al., 2021). More recent deep learning approaches take advantage of pretrained language models (Kandimalla et al., 2021; Hande et al., 2021).

One of the common practices to evaluate approaches for classifying scientific text is to use classification systems from digital libraries (Kandimalla et al., 2021; Gialitsis et al., 2022; Taheriyan, 2011; Gündoğan and Kaya, 2020), such as the ACM Computing Classification System[2], the Web of Science Categories[3] and Science-Metrix[4]. Other practices involve generating automatic annotations for scientific collections that can be completely synthetic (Waltman and Van Eck, 2012) or curated by experts (Salatino et al., 2022; Eykens et al., 2021; Daradkeh et al., 2022; Hande et al., 2021; Pech et al., 2022). However, to date, there has been no established benchmark to evaluate these approaches.

We present a new high-quality benchmark for evaluating research theme classification, used for

[2] ACM Computing Classification System
[3] Web of Science Categories
[4] Science-Metrix

the first time in the Scholarly Knowledge Graph Generation Shared Task.

3 Inital Dataset Creation

As previously discussed, one of the significant challenges faced in the domain is the lack of large-scale labelled data for research theme classification. For the shared task, a completely new gold-standard dataset was compiled using data drawn from the U.K.'s Research Excellence Framework (REF) 2014 exercise (Cressey and Gibney, 2014). In total, 191,000 research outputs were submitted by 154 higher education and research institutions, and these were then peer-reviewed by experts from each domain. The REF divided research outputs into 36 'Units of Assessment' (UoA) or domain areas. The institutions themselves selected to which Unit of Assessment each output was submitted.

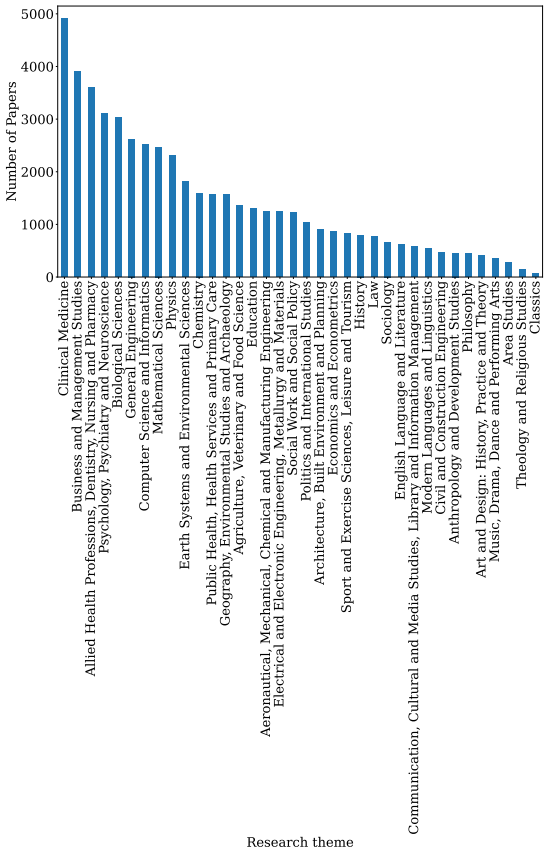

Figure 1: Breakdown of the dataset by theme.

The data from the REF exercise, therefore, provides a near-perfect starting point for the task of automatically identifying research themes as the UoA labels were manually assigned to each output by the expert academics responsible for its production.

For each output, the following were available from the REF data; publication title, publication year, publication venue, name of institution, and Unit of Assessment. These fields were fully populated for 190,628 out of 190,963 submissions to the outputs category of the REF process. We further enriched each record with the DOI, CORE id, and abstract (where available). The CORE id is used to identify the actual research article held by the CORE service[5]. Not all papers in the dataset are open access, therefore the full-text content of all papers is not available. For non-open access papers, CORE often still has the metadata for these articles.

For the data used in this shared task, separate test and train datasets were generated. From the full REF dataset, 51,560 randomly selected records were used for the train set, and a separate 10,000 were selected for the test set. The datasets were then verified to ensure that there was no overlap between the two sets. Figure 1 shows the cross-domain (theme) breakdown of all records used for this task.

4 Classification Ensemble

This section depicts the approach we used to estimate probabilities of academic publications belonging to specific theme and the heuristics we follow for classification. In general, we want to exploit all the information available for the scholarly documents that need to be classified. Academic publications are typically well-structured documents with multiple textual fields and metadata. We rely on open-access platforms to enrich the data with additional information (Section 4.2).

Currently, Transformer-based contextual language models like ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019) outperform most feature-based representation methods. We use a classifier based on contextual word embeddings to evaluate the utility of individual textual fields in the classification of Academic publications.

4.1 Transformer-based Classifier

We rely on the pre-trained general language model BERT (Devlin et al., 2019), which achieves outstanding performance on different NLP tasks through fine-tuning for the downstream tasks (Acheampong et al., 2021), in this case, multiclass classification.

[5]https://core.ac.uk

We allow all layers of BERT to be updated as we are learning the relevant context from the training data. A custom operation is added on top of the model, which takes the last hidden state tensor from the encoder and then passes it to a linear layer. At the end of the linear layer, we have a vector with a size equal to the number of classes, and each element corresponds to a category of the provided labels. Specifically, we use the following setting to build the model base:

Input layer. It builds the model's input sequence. The input sequence is segmented according to the WordPiece embeddings and the token vocabulary. The final input representations are then produced by adding the position embeddings, word embeddings, and segmentation embeddings for each token.

BERT encoder. It consists of multiple Transformer blocks and multiple self-attention heads that take an input of a sequence of a limited number of tokens and output the representations of the sequence. The representation can be a specific hidden state vector or a time-step sequence of hidden state vectors.

Output layer. It consists of a simple linear layer with a Softmax classifier on top of the encoder for computing the conditional probability distributions over predefined categorical labels.

The cross-entropy loss is used to optimize the model with the Adam optimizer.

4.2 Data Enrichment

Taking advantage of the open access libraries available for scientific publications, we search for complementary data for each example provided for the task. Specifically, we use the CORE (Knoth and Zdrahal, 2012) and the Semantic Scholar (Ammar et al., 2018) APIs to map publication titles to the various fields available for each publication.

The original task dataset includes mainly titles with metadata. Our goal with the enrichment is to collect more information related to the publication to better match the themes. After mapping the papers to results from the search using the APIs, we add a list of references and citations, full papers, abstracts, and PDFs, for the cases when they are available. Moreover, we search for five recommended papers using the title for every publication using the CORE API.

We believe that regardless of the performance of the classification model, if there is enough evidence for a publication to belong to a specific theme, we should be able to classify it with enough certainty. For instance, given a publication title, which can be ambiguous, we hypothesize that considering the multiple references or citations leads to disambiguation and deciding effectively to which topic this publication should belong. The list of references or citations can be classified the same way as single inputs, and the classification result can consider the multiple corresponding outputs for the final decision.

Since there is no guarantee that this data is available for all the original samples, we exploit all available sections, including the full text and PDFs. However, since processing such an amount of text is expensive, we use AZ (Teufel et al., 1999b). Here, we define four zones that cover the main components of scientific articles, namely: *Claim*, *Method*, *Result* and *Conclusion*.

In order to extract sentences that cover the four zones from the available PDF scientific articles, we follow an approach similar to a previously proposed approach by El-Ebshihy et al. (2020), which generates an article summary by expanding the article abstract. To sum up, the sentence selection and labeling with zones process goes as follows: (1) we convert the PDF papers to an XML format using the GROBID PDF parser (Lopez, 2009), which identifies the paragraphs of the article, (2) the paragraphs are fed into a Solr[6] index, (3) the sentences in the article's abstract are passed as queries to the Solr index in order to find the top most similar paragraphs to the abstract sentences, (4) sentences of the retrieved paragraphs, as well as the sentences of the abstract, are labeled to zones using a pre-trained BERT model based on the approach proposed by Accuosto et al. (2021), and (5) we use the labeled sentences to extend our training data with four extra text fields that represent the *Claim*, the *Method*, the *Result* and the *Conclusion* — we refer to these extra fields as Argumentative Zones. In case we cannot find the PDF source of the article, we use the article abstract, if found, to generate these fields.

4.3 Extending Labels to Enriched Data

During training, the model takes text examples together with the labels associated with them. Since examples for this task are academic publications,

[6] https://lucene.apache.org/solr/

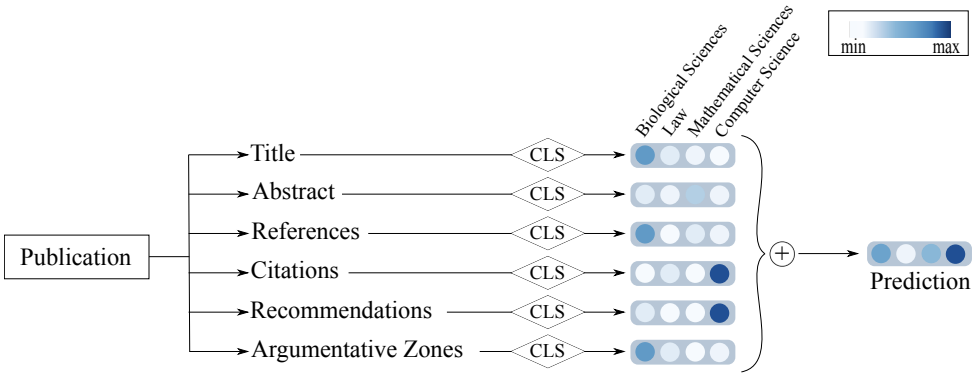

Figure 2: Ensemble for research theme classification. CLS stands for classifier.

and we want to use different sections independently, we rebuild the dataset considering each section as a single sample but associated with the same publication, and we use the same label for all samples of the same publication.

In this way, we end up with an extended version of the initial dataset, in which new samples are created for titles, abstracts, citations, references, and recommendations.

4.4 Aggregating Predictions from Enriched Data

During inference time, we compute multiple predictions associated with the same publication. These predictions can either agree or disagree, so we formulate the final prediction as the aggregation of the different predictions. Figure 2 illustrates the prediction procedure used to obtain the final theme prediction for a publication in which various sections are evaluated as independent samples with the classifier. Section 5 describes how this aggregation is parameterized for the experiments.

5 Experimental setup

5.1 Dataset

Statistics for the initial dataset are provided in Table 1. Most of this dataset's publications do not contain abstracts, additional metadata, or PDFs. Theme identification algorithms should be robust to these missing features and work well when only titles are available.

	Train	Test
Size	51,560	10,000
% of Publications		
– available via CORE API	91.6%	92.4%
– with abstract	31.8%	31.7%
– with PDF	24.6%	25.6%
– with full text	6.3%	6.4%
– with references	8.4%	7.6%

Table 1: Dataset statistics.

5.2 Training Settings

Given the labelled training samples, we train the model using two different sets. The first training set consists of the list of titles, while the second takes both titles and available abstracts. We argue that although more information can be available per publication, the labels provided match only titles and abstracts, and further assumptions can hurt the model's performance. However, we define an additional training set under our data enrichment procedure. We refer to the first model as BERT_T and to the second one as BERT_{T+A}.

We train the model for 10 epochs, with early stopping based on the performance measured using the evaluation metric (see Section 6.1) and patience of 3 epochs. The training samples are picked randomly, searching for a uniform distribution over the classes per batch. To prevent overfitting in case of unbalanced batches, we use the weighted cross-entropy loss, and assign the weights dynamically, according to the result of the random selection of samples in the batch. We use 16384 samples from the training set per epoch divided into batches of 64 samples, and train the models on an Nvidia Quadro RTX 8000 GPU.

5.3 Prediction Settings

As well as the training strategy, we evaluate the utility of having multiple predictions per publication in the test set compared to a single prediction. To do so, we prepare different evaluation sets, following the same training set schema. Thus, we evaluate the model using only titles, then using titles and abstracts, and finally, using the set created under

our data enrichment procedure.

Since we have to produce a single prediction per publication, and the sets are not uniform, in the sense that certain publications may not have extra fields (see Table 1, for instance, abstracts are available for only 32% of publications), we parameterise the prediction aggregation based on the different sets of fields. We consider the aggregation to be a weighted sum. The motivation for selecting a weighted sum, instead of just summing up the outputs is that we can introduce offsetting through the weights. Thus, we give an advantage to the labelled fields in the original dataset over the extended data.

For our experiments, in the case of the set with titles and abstracts, we use uniform weighting. In the case of the extended set, we assign weights such that: 0.5 is distributed uniformly between title and abstract, and 0.5 is uniformly distributed between all the additional fields available per publication. This setting is compared experimentally to a uniform weighting across all the fields.

6 Results

6.1 Evaluation metrics

The evaluation metric used for evaluating classification results is micro F1-Score. The F1 score, commonly used in machine learning, measures accuracy using the statistics precision and recall.

The F1 metric weighs recall and precision equally, and a good classification algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

6.2 Baseline Models

We implement several baseline models for comparison to the ensemble described in Section 4:

K-nearest neighbours classifier with Tf-idf representation

Logistic Regression classifier with Tf-idf representation

Naïve Bayes classifier with Tf-idf representation

Support Vector Machine classifier with Tf-idf representation

fastText classifier (Joulin et al., 2016) with word vectors pretrained on wikipedia[7]

[7] https://dl.fbaipublicfiles.com/fasttext

We also present scores using two dummy classifiers: selecting the most frequent category and sampling from a multinomial distribution parameterised by prior probabilities. All classifiers except for fastText are implemented using scikit-learn (Pedregosa et al., 2011).

6.3 Validation Results

Given the provided training data, we create balanced splits such that 60% is used for train, 10% for early stopping and 30% for validation. All the sets are enriched following the process described earlier. Table 2 shows some preliminary results for experiments we perform to select the model and the training setup. We compare the two different BERT models with traditional models. The performance of the model trained using titles and abstracts is slightly better, and we use it for further experiments.

Model name	Titles	Titles and abstracts
Dummy: most frequent	— 0.095 —	
Dummy: stratified random	— 0.048 —	
K-nearest Neighbours	0.132	0.468
Logistic Regression	0.457	0.498
Naïve Bayes	0.460	0.493
Support Vector Machine	0.474	0.506
fastText	0.454	0.473
BERT$_T$	0.498	–
BERT$_{T+A}$	**0.500**	**0.512**

Table 2: Micro F1-score results comparison using different input features for prediction. BERT$_T$ stands for BERT model trained on titles only, BERT$_{T+A}$ means model trained on both titles and abstracts.

Furthermore, we evaluated the utility of enriching the dataset by comparing predictions from titles only with aggregated predictions using titles and additional available fields. Table 3 shows that adding information improves the classification for all three experiments. Notice that the experiments are not comparable to each other because the dataset samples are different. Subsamples are selected such that corresponding sections are available for all documents.

Table 4 shows the results obtained for the validation set using different variants of ensemble. In general, we are able to improve the performance of the classification while adding more data, although the difference between the experiments is small. The best score reached is 0.526, using titles, abstracts, citations, references and the argumentative

Sections	Sample size	F1-score (title)	F1-score (all sections)
Title + Abs.	31.3%	0.503	0.539
Title + Cit. + Refs	25.4%	0.492	0.541
Title + AZ	1.6%	0.548	0.552

Table 3: Three experiments testing the utility of individual sections on BERT$_{T+A}$. The augmentation is evaluated by independent sections combined with titles. Samples are selected such that corresponding sections are available for all documents.

Title	Abs.	Cit.	Refs	AZ	Recs.	F1
×	–	–	–	–	–	0.500
×	×	–	–	–	–	0.512
×	×	×	×	–	–	0.523
×	×	×	×	×	–	**0.526**
×	×	×	×	×	×	0.525

Table 4: Validation results using different fields for BERT$_{T+A}$. The experiments vary in the prediction and aggregation settings. The aggregations we use are simply weighted sums with uniform weights and assigned arbitrarily according to Section 5.3.

zones.

For the best configuration, we also show the confusion matrix (see Figure 3). For convenience, we show the results for only the 25 most frequent classes and we group the rest of them in a single class. It should be noted that for Clinical Medicine, most of the examples where the model's prediction is incorrect are classified as Allied Health Professions, Dentistry, Nursing and Pharmacy, and Biological Sciences. Similar behaviour can be observed with related fields of study. Further analysis must be done to evaluate overlapping between disciplines.

6.4 Test Results

In this section, we show the results for the test set (see Table 5). In general, we see a positive impact with our approach considering that we could not get additional information for all the items in the original dataset.

In this set of experiments, we evaluate a different aggregation setting, uniform weighting through all the fields (run 4), and the result is the best score for the set of runs. Furthermore, we also evaluate an additional model trained with all the fields available (run 5), and we see no improvements.

7 Discussion

In this work, we first released a new gold-standard human-annotated dataset of over 60k papers complete with paper metadata, research themes and additional textual information including the papers' abstract and full-text where available. In future, it would be possible to further extend the size of the presented dataset to include all REF2014 and now the recently finalised REF2021 papers, which both used the same research themes classifications. This would result in an annotated dataset of over quarter of a million papers. To our knowledge, our work was the first to utilise REF research evaluation for the purposes of building machine learning models for themes classification and highlighted the significant potential of this dataset for developing state-of-the-art models.

Run	Title	Abs.	Cit.	Refs	AZ	Recs.	Agg.	F1
1	T+P	T+P	–	–	–	–	U	0.569
2	T+P	T+P	P	P	P	–	C	0.575
3	T+P	T+P	P	P	P	P	C	0.571
4	T+P	T+P	P	P	P	P	U	**0.577**
5	T+P	T+P	T+P	T+P	–	T+P	C	0.556

Table 5: Test results with different experimental (Run) settings. The experiments vary in the training (T), prediction (P) and aggregation (Agg.) settings. The aggregations we use are simply weighted sum with uniform weights (U) and compensation weights (C) assigned according to section 5.3.

Second, we use this dataset to establish a new benchmark for research theme classification, testing a range of classic machine learning models under the same laboratory conditions. Unsurprisingly, our results confirm that models trained with both titles and abstracts as input features consistently achieve higher results than when using titles alone. These results hold both for baseline models and our newly introduced ensemble BERT model. While the results confirm that the BERT-based ensemble model outperforms traditional models, the performance of SVMs is only marginally worse.

It is interesting to note that using all available features for training (run 5) decreases the score compared to the model trained on titles and abstracts only. We hypothesise that a large proportion of false negatives can be attributed to noise introduced by reference sections within the full texts, especially for closely aligned domains. The confusion matrix (Figure 3) shows that many of the incorrect classifications happened in closely related domains (Clinical Medicine / Biological Science for example).

This is indicative of the difficulty of this task, particularly when presented with closely matched

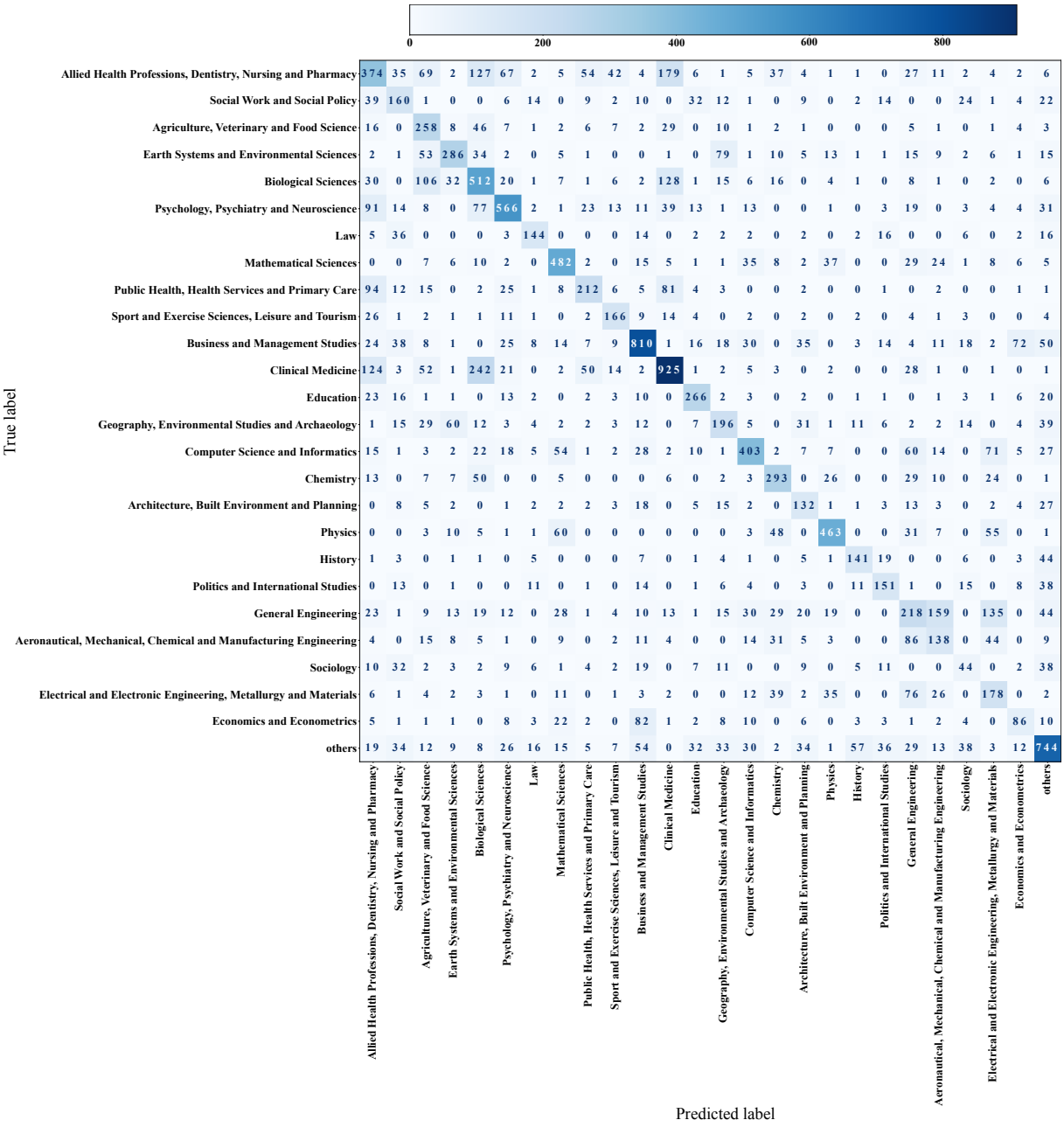

Figure 3: Confusion Matrix for validation results for 25 most frequent classes. The remaining 11 classes are grouped in the 'others' category.

or overlapping domains. Indeed, one limitation of our approach may be the classification of each paper into a single research field. In real-world examples, a paper could often be classified into multiple domains. Another limitation is that our ensemble model requires the availability of both title and abstract, which are necessary for the AZ approach, which we have seen contributes to the performance.

Assigning research themes to scholarly documents has wide-ranging applications. These include enhanced domain-specific search, for instance search in Chemistry is a complex task due to the need to index chemical compounds, and identifying emerging research trends. Further, a significant problem with current bibliometric methodologies is accounting for cross-disciplinary differences in both publishing and citation practices. Identifying the research theme enables accounting for disciplinary differences by, for instance, calculating normalised citation counts.

In future work, we would like to measure the importance of weight assignments for augmented predictions and consider the overlap between disci-

plines to evaluate ways of disambiguating predictions falling into related themes.

8 Conclusion

We have introduced a new large human annotated gold-standard dataset and a benchmark for research theme classification of scholarly documents. The work was conducted in the context of the *Extracting Research Themes* task from the 2022 edition of the Scholarly Knowledge Graph Generation shared task. The task was to identify the main research theme from a taxonomy of 36 classes, introduced by the UK Research Excellence Framework.

Our experiments addressed the effect of using a variety of textual fields on the prediction performance. Enriching the supplied training and testing data with external textual information (e.g., PDF source, full-text article, references) using open-access sources improved the results of our models. However, we have demonstrated that this enrichment might also introduce additional noise.

We presented a new transformer-based classifier model based on BERT and used it to obtain multiple predictions for a given research article for each textual field. We experimented with a variety of aggregation functions to produce the final prediction. Despite incomplete and noisy data, the results show that our ensemble model has a small positive impact on the classification performance.

Acknowledgements

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

References

Pablo Accuosto, Mariana Neves, and Horacio Saggion. 2021. Argumentation mining in scientific literature: from computational linguistics to biomedicine. In *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36.* CEUR Workshop Proceedings.

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8).

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

Kevin W Boyack and Richard Klavans. 2018. Accurately identifying topics using text: Mapping pubmed. In *STI 2018 Conference Proceedings*, pages 107–115. Centre for Science and Technology Studies (CWTS).

Daniel Cressey and Elizabeth Gibney. 2014. Uk releases world's largest university assessment. *Nature*.

Mohammad Daradkeh, Laith Abualigah, Shadi Atalla, and Wathiq Mansoor. 2022. Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics. *Electronics*, 11(13):2066.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alaa El-Ebshihy, Annisa Maulida Ningtyas, Linda Andersson, Florina Piroi, and Andreas Rauber. 2020. ARTU / TU Wien and artificial researcher@ LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

Joshua Eykens, Raf Guns, and Tim CE Engels. 2021. Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. *Quantitative Science Studies*, 2(1):89–110.

Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145.

Nikolaos Gialitsis, Sotiris Kotitsas, and Haris Papageorgiou. 2022. Scinobo: A hierarchical multi-label classifier of scientific publications. *arXiv preprint arXiv:2204.00880*.

Esra Gündoğan and Mehmet Kaya. 2020. Research paper classification based on word2vec and community discovery. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 1032–1036. IEEE.

Adeep Hande, Karthik Puranik, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Domain identification of scientific articles using transfer learning and ensembles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 88–97. Springer.

Fabian Hoppe, Danilo Dessì, and Harald Sack. 2021. Deep learning meets knowledge graphs for scholarly data classification. In *Companion proceedings of the web conference 2021*, pages 417–421.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Bharath Kandimalla, Shaurya Rohatgi, Jian Wu, and C Lee Giles. 2021. Large scale subject category classification of scholarly papers with deep attentive neural networks. *Frontiers in research metrics and analytics*, 5:600382.

Sang-Woon Kim and Joon-Min Gil. 2019. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9(1):1–21.

Petr Knoth and Zdenek Zdrahal. 2012. Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12).

Haixia Liu. 2017. Automatic argumentative-zoning using word2vec. *CoRR*, abs/1703.10152.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

Michał Łukasik, Tomasz Kuśmierczyk, Łukasz Bolikowski, and Hung Son Nguyen. 2013. Hierarchical, multi-label classification of scholarly publications: modifications of ml-knn algorithm. In *Intelligent tools for building a scientific information platform*, pages 343–363. Springer.

Gerson Pech, Catarina Delgado, and Silvio Paolo Sorella. 2022. Classifying papers into subfields using abstracts, titles, keywords and keywords plus through pattern detection and optimization procedures: An application in physics. *Journal of the Association for Information Science and Technology*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Maxime Rivest, Etienne Vignola-Gagné, and Éric Archambault. 2021. level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. *PloS one*, 16(5):e0251493.

Angelo Salatino, Francesco Osborne, and Enrico Motta. 2022. Cso classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics. *International Journal on Digital Libraries*, 23(1):91–110.

Piotr Semberecki and Henryk Maciejewski. 2017. Deep learning methods for subject text classification of articles. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 357–360. IEEE.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. *arXiv preprint arXiv:1805.12216*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Mohsen Taheriyan. 2011. Subject classification of research papers based on interrelationships analysis. In *Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation*, pages 39–44.

Simone Teufel, Jean Carletta, and Marc Moens. 1999a. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502.

Simone Teufel et al. 1999b. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Nees Jan Van Eck and Ludo Waltman. 2017. Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics*, 111(2):1053–1070.

Ludo Waltman and Nees Jan Van Eck. 2012. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392.

Shenghui Wang and Rob Koopman. 2017. Clustering articles based on semantic similarity. *Scientometrics*, 111(2):1017–1031.