Advancing Topical Text Classification: A Novel Distance-Based Method with Contextual Embeddings

Andriy Kosar
Textgain &
University of Antwerp (CLiPS)
Antwerp, Belgium
andrew@textgain.com

Guy De Pauw
Textgain
Antwerp, Belgium
guy@textgain.com

Walter Daelemans
University of Antwerp (CLiPS)
Antwerp, Belgium
walter.daelemans
@uantwerpen.be

Abstract

This study introduces a new method for distance-based unsupervised topical text classification using contextual embeddings. The method applies and tailors sentence embeddings for distance-based topical text classification. This is achieved by leveraging the semantic similarity between topic labels and text content, and reinforcing the relationship between them in a shared semantic space. The proposed method outperforms a wide range of existing sentence embeddings on average by 35%. Presenting an alternative to the commonly used transformer-based zero-shot general-purpose classifiers for multiclass text classification, the method demonstrates significant advantages in terms of computational efficiency and flexibility, while maintaining comparable or improved classification results.

1 Introduction

Topical text classification remains an important task in text classification because it allows users to explore, analyze and organize large text collections. However, the nature of topical text classification is subjective as the content and context of the text are often perceived differently based on the intended audience. To address this, methods that dynamically explore topics are necessary, one of which is unsupervised text classification. This approach allows classifying text collections based on a predefined list of topics for further analysis.

Yin et al. (2019) outlined three primary techniques for unsupervised text classification: 1) evaluating the frequency of class labels in a text, 2) measuring the distance between class labels and text in a shared vector space, and 3) leveraging natural language inference with pre-trained classifiers to ascertain if a class label can be deduced from the text. With the advancement of transformer models, the latter method has gained increasing attention in the NLP community due to its successful outcomes (Yin et al., 2019; Ding et al., 2022). In this study we show that task-specific sentence embeddings trained on transformer models for distance-based topical text classification, can provide a flexible and efficient alternative to the aforementioned methods.

In this study, we undertake a comprehensive examination of unsupervised topical text classification utilizing contextual embeddings, and propose a methodology for generating sentence embeddings that are more appropriate for this task. To achieve this objective, we first evaluate a diverse array of existing contextual embeddings and their derived sentence embeddings on seven datasets across a broad spectrum of genres and topics. Subsequently, we explore the various options for training custom sentence embeddings, including the choice of training data, base models, and loss functions, with the aim of identifying the most suitable configuration for the given task. Finally, we assess the benefits and limitations of our proposed method.

The paper unfolds as follows: Section 2 outlines the previous research on unsupervised text classification; Section 3 presents the proposed method; Section 4 explains experiment setup; Section 5 presents evaluation results.

2 Related work

Unsupervised text classification, also referred to as dataless or zero-shot text classification, relies on semantic relatedness between class labels and documents for classification without requiring training data. Chang et al. (2008) pioneered this concept, employing Explicit Semantic Analysis (ESA) and Wikipedia as an external knowledge base to encode class labels and document texts within a single semantic space and classifying them based on proximity. This approach was further extended by Song and Roth (2014) for hierarchical text classification and by Song et al. (2016) for cross-lingual text classification.

With the introduction of neural word embed-

Proceedings of Recent Advances in Natural Language Processing, pages 586–597
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_064

dings by Mikolov et al. (2013a) and Mikolov et al. (2013b), these representations were also employed for unsupervised text classification. Sappadla et al. (2016) used word2vec for multi-label classification, while Haj-Yahia et al. (2019) leverages GloVe and word2vec to enrich class labels. Schopf et al. (2021) introduced Lbl2Vec, a method for retrieving documents with predefined topics, and Kosar et al. (2022) evaluated different neural word embeddings for topical text classification and proposed an improvement of class label representation with nearest words to a class label in one semantic space.

The emergence and success of large pre-trained language models (LLMs), initiated by Devlin et al. (2019), shifted unsupervised text classification towards natural language inference tasks. Yin et al. (2019) employed a textual entailment (TE) approach for unsupervised text classification by fine-tuning a pre-trained BERT model on multiple entailment corpora. Halder et al. (2020) presented the TARS method, a pre-trained BERT binary classifier for general text classification using various classification corpora. Ding et al. (2022) and Wang et al. (2022b) further advanced the entailment approach by fine-tuning models on Wikipedia categories (TE-Wiki) and enhancing model architecture (S-BERT-CAM), respectively. Laurer et al. (2022) showcased the exceptional performance of BERT NLI in zero-shot and few-shot scenarios across different text classification tasks. As LLMs continue to evolve, these methods have become increasingly dominant in unsupervised text classification.

Recently, the development of sentence embeddings introduced by Reimers and Gurevych (2019), with improved text representation, added additional push for improvement of various NLP tasks such as information retrieval and semantic search. Subsequent enhancements to sentence embeddings, like SGPT (Muennighoff, 2022), showcased their promising potential. Schopf et al. (2023) introduced Lbl2TransformerVec, an enhancement of the previously introduced Lbl2Vec for unsupervised text classification using sentence embeddings.

3 Proposed method

We formulate the problem of unsupervised topical text classification as follows: given a set of predefined topic categories, the objective is to classify texts based on the semantic relatedness between the topic name and the text content. Taking into account large amounts of data involved, and rapid changes in the data and topical categories, this classification should be done as efficiently as possible.

Of the two major methods of unsupervised text classification, the distance-based method with neural word embeddings is more computationally efficient but the transformer-based zero-shot classifiers has been shown to be more accurate due to its ability to better capture text semantics. To combine the advantages of these two methods we propose replacing the often used neural word embeddings with transformer-based embeddings tailored for this task.

For this purpose, we employ sentence embeddings introduced Reimers and Gurevych (2019) to embed both texts and topic names into a shared semantic space. However, instead of the typical training of sentence embeddings on text pairs that preserve the same level of abstraction and granularity, we propose training task-specific sentence embeddings on tag-text pairs, where tags serve as proxies for topics with a higher level of abstraction. To better demonstrate the distinctions between traditional and proposed methods, we provide examples of training data for both approaches.

SNLI and MS MARCO datasets typically are used for training sentence embeddings:

SNLI[1]. *Sentence 1*: A senior is waiting at the window of a restaurant that serves sandwiches, *Sentence 2*: A person waits to be served his food.

MS MARCO[2]. *Query*: when was the town of farragut tn incorporated, *Passage text*: In January of 1980, residents decided to incorporate by an overwhelming margin. The Town of Farragut was incorporated on January 16, 1980, with the first board of Mayor and Alderman elected on April 1, 1980.

Our approach suggests leveraging resources similar to Wikipedia categories and New York Times descriptors for training task-specific sentence embeddings:

Wikipedia. *Text*: Sojunghwa Sojunghwa is a century Korean concept that means Little China referring to the Joseon Dynasty After the Qing dynasty conquered the Han Ming dynasty Koreans thought that barbarians ruined the center of civilization of the world and so Confucianist Joseon Korea had become the new center of the world replacing Ming China hence the name Little China Tokugawa Japan and Vietnam also had a similar belief in themselves after the Qing Dynasty had taken over China Based on Sinocentrism the belief that China was the center of civilization in the world the Chinese

---

[1]Example obtained from: https://nlp.stanford.edu/projects/snli/snli_1.0.zip. Accessed March 15, 2023.

[2]Example obtained from: https://msmarco.blob.core.windows.net/msmarco/train_v2.1.json.gz. Accessed March 15, 2023.

believed that Korea then a tributary state was a highly civilized state Meanwhile the Koreans considered Japanese and Jurchen people to be barbarians or beasts under the distinction, *General category*: Philosophy by region.

NYT LDC. *Text*: No one here knew Diane O'Dell's secret. She was, said people who live in this wide spot in on a narrow rural road, a pleasant if somewhat standoffish neighbor and an affectionate mother. "Everybody in the area knows everybody," said John Karpauitzs, who lives a few doors down from the gray, tumble-down house that Ms. O'Dell shared with her common-law husband and their five children. "She was quiet. She kept mostly to herself. Not much else to say about her." There was nothing in her behavior, neighbors said, to indicate that she traveled with the corpses of three of her other children around the country for a decade. Ms. O'Dell, 49, was charged in Sullivan County, N.Y., on Tuesday with murdering three babies she bore in the early 1980's in Sullivan County... *General descriptor*: Murders and Attempted Murders.

Training sentence embeddings on texts that have been tagged with relevant topic labels or similar tags enhances the embeddings' ability to capture topic associations. As a result we obtain sentence embeddings that reinforce the association between topic labels and text content in shared semantic space. Subsequently, topical text classification is performed by assigning the topic label to the text with the closest proximity, as determined by cosine similarity.

4 Experiments

4.1 Experimental setup and evaluation

In our study, we evaluate the effectiveness of pre-trained contextual embeddings and custom-trained sentence embeddings on seven datasets. To obtain class label and text embeddings, we employed mean pooling as proposed by Reimers and Gurevych (2019) for the contextual embeddings. We employed a maximum sequence length of 128 and 256 tokens and did not perform any preprocessing on the texts. However, we report results only for the 128-token sequence length, as there was no significant difference observed in the performance on longer texts.

As a baseline, we utilized distance-based text classification with neural word embeddings, specifically word2vec (Mikolov et al., 2013a), as it has been reported by Kosar et al. (2022) to be more suitable for this task compared to other models. To obtain embeddings for compound class labels or texts, we computed the average of word embeddings of the constituent words present in the model's vocabulary.

Furthermore, we compared our results to TE-Wiki (Ding et al., 2022), an open-domain topic classification model that has been shown to outperform known zero-shot models and perform competitively with weakly-supervised methods.

To evaluate classification results, we employed accuracy as a metric to facilitate comparison with previous studies (Yin et al., 2019; Ding et al., 2022). Given the wide range of datasets and models utilized in our study, we based our conclusions on the general performance of the models (average accuracy). To provide a more comprehensive evaluation, the weighted average F1 score for each model has also been reported in Appendix A.3 Table 9.

4.2 Datasets

We tested our proposed method on seven English datasets that covered a variety of genres, including Wikipedia extracts (DBPedia, Lehmann et al., 2015), news headlines and articles (AGNews - Zhang et al., 2015, RCV1-v2 - Lewis et al., 2004 and New York Times[3]), academic articles (S2ORC - Lo et al., 2020), Q&A (Yahoo - Zhang et al., 2015), social media posts (Twitter - Antypas et al., 2022) and e-commerce product descriptions (Amazon - Ni et al., 2019). These datasets offer a diverse array of class labels, including both simple topics like business and complex ones like the environment and natural world, and cover a wide range of subjects from science and technology to pet supplies.

For the DBPedia, Yahoo, and AGNews datasets, we used texts and class labels provided by Ding et al. (2022) to compare our results with theirs. For the remaining datasets, apart from Twitter, we randomly picked 380-500 texts per class from the sources mentioned above. The objective behind sampling these datasets is to facilitate a larger number of experiments while simultaneously reducing the environmental impact typically associated with the research process. All datasets exhibit an equal distribution of examples across classes, with the exception of Twitter.

The statistics of the datasets are shown in Table 1. A list of class labels for all datasets is included in Appendix A.1.

4.3 Pre-trained contextual embeddings

We conducted a comparison of two types of pre-trained contextual embeddings: the standard transformer-based version, and a modified version called "sentence embeddings" which are designed

[3]The dataset was built using full text articles and metadata collected from the New York Times newspaper over the past 20 years.

Dataset	Size	Classes	Mean tokens	Std tokens
DBPedia	70000	14	46	21
Yahoo	100000	10	81	88
AGNews	7600	4	36	10
RCV	8100	18	286	191
S2ORC	8550	19	166	88
NYT	8500	17	889	557
Twitter	3399	6	26	12
Amazon	5700	15	91	68

Table 1: Corpora statistics.

to produce improved text representation. Our aim was to determine whether these pre-trained models could be used for unsupervised topical text classification.

To evaluate the standard pre-trained contextual embeddings, we used several widely-known models including GPT, BERT, RoBERTa, XLNet, GPT-2, BART, and T5, as described in the works of Liu et al. (2020) and Min et al. (2021). Additionally, we included MPNet in our study since it was used as the basis for training high-performing sentence embeddings (Reimers and Gurevych, 2019).

For the pre-trained sentence embeddings, we tested a number of models including "all MPNet Base v2", GTR-T5, Sentence T5, and E-5, which are among the top performers on the Massive Text Embedding Benchmark (MTEB) Leaderboard[4]. In addition to these models, we also evaluated commercially available text embeddings: OpenAI[5] and Cohere[6].

We provide the list of the tested models in Table 2.

Model	Attribution
Plain models	
GPT	Radford and Narasimhan (2018)
BERT base uncased	Devlin et al. (2019)
RoBERTa base	Liu et al. (2019)
XLNet base cased	Yang et al. (2019)
GPT-2	Radford et al. (2019)
BART base	Lewis et al. (2019)
T5 base	Raffel et al. (2020)
MPNet base	Song et al. (2020)
Sentence embeddings	
all MPNet base v2	Reimers and Gurevych (2019)
GTR-T5 base	Ni et al. (2021)
Sentence T5 base	Ni et al. (2022)
E-5 base	Wang et al. (2022a)
SGPT (125M)	Muennighoff (2022)

Table 2: Evaluated models.

[4] https://huggingface.co/spaces/mteb/leaderboard. Accessed March 15, 2023.

[5] Model: text-embedding-ada-002. Accessed October, 2022.

[6] Model: large. Accessed October, 2022.

4.4 Trained task-specific sentence embeddings

In order to train task-specific sentence embeddings, we experimented with two datasets: the Wikipedia dataset, as presented by Ding et al. (2022), and the NYT LDC dataset, as presented by Sandhaus (2008). The Wikipedia dataset comprises of articles from Wikipedia, along with their corresponding high-level categories (e.g., Politicians, Musical Groups, Civil Engineering, etc.), with a total of 674 unique categories. The NYT LDC dataset, on the other hand, includes full-text news articles from The New York Times newspaper, as well as additional metadata, including article headlines, sections, general descriptors, etc. From the NYT LDC dataset, we utilized the text of the articles and the general descriptors (e.g. Politics and Government, Medicine and Health, Baseball, etc.). After preprocessing, we obtained a total of 1,622 unique high-level descriptors. A list of the top 20 tags for each dataset can be found in Appendix A.2 Table 7 and 8.

As the base models we used plain contextual embeddings BERT, BART, T5 and MPNet. Additionally, we experimented with existing sentence embeddings such as "all MPNet base v2", GTR-T5 and Sentence T5 as base models in order to evaluate the possibility of leveraging fine-tuned sentence embeddings on related tasks (e.g. semantic textual similarity and semantic search), to enhance the training process and achieve enhanced performance.

As a part of our study we also evaluated three types of loss functions, mainly Cosine Similarity Loss, Contrastive Loss (Hadsell et al., 2006) and Multiple Negatives Ranking Loss (Henderson et al., 2017). Additionally we tested an enhanced version of Contrastive Loss - Online Contrastive Loss.

We replicated the training setup used by Ding et al. (2022) in their TE-Wiki model to compare our results. This included using a maximum sequence length of 128, batch size of 64, learning rate of 5e-5, and training for one epoch with 1500 training steps. We used a text from a dataset and an assigned tag (high-level category or general descriptor) as a positive pair and a randomly selected tag from the remaining tags for a negative pair. We also preprocessed the text by truncating it to 200 tokens for Wikipedia and 600 characters for the NYT LDC dataset. We conducted each training experiment five times with different seeds and report the average accuracy.

5 Results and analysis

5.1 Comparing pre-trained contextual embeddings

The results of our experiments (Table 3) reveal that pre-trained transformer-based contextual embeddings exhibit poor performance in distance-based text classification in comparison to neural word embeddings, and are less suitable for this task. This finding is consistent with the findings of Reimers and Gurevych (2019), who demonstrate that averaged GloVe embeddings show superior performance compared to BERT averaged embeddings on the Semantic Textual Similarity task. Additionally, we observed that the T5 model achieved the highest performance among the models evaluated.

5.2 Comparing pre-trained sentence embeddings

Our results (Table 3) show that using modified sentence embeddings improves distance-based text classification compared to plain contextual embeddings and often performs better than neural word embeddings for the same task. However, none of the current models surpass the performance of the TE-Wiki model for zero-shot open domain topic classification. It is worth noting that the OpenAI embeddings (text-embedding-ada-002) have exceptional overall performance and outperform the TE-Wiki model on four datasets (RCV, NYT, Tweets, Amazon).

5.3 Effect of training task-specific sentence embeddings

Our experiments (Table 3) with training task-specific sentence embeddings on four base transformer models - BERT, BART, MPNet, and T5 - on the Wikipedia dataset demonstrate that all the models outperform existing sentence embeddings on unsupervised distance-based text classification tasks. Additionally, these models also exhibit superior overall performance compared to TE-Wiki, despite similar training setups and training data. We also observe similar or better performance when BERT and BART models are trained on different data, particularly the NYT LDC datasets with the Multiple Negatives Ranking Loss (Table 4 and 5). This leads us to the conclusion that the proposed method is not limited to the specific base models or training data.

The results of our experiments, training task-specific sentence embeddings on the Wikipedia dataset with pre-trained sentence embeddings, show improvement in classification accuracy (Table 3). Additional analysis during training reveals that training custom sentence embeddings based on pre-trained sentence embeddings can boost performance, even with a limited amount of training data (Figure 1).

5.4 Loss selection

The experiments on BERT and BART models trained individually on Wikipedia and NYT LDC datasets (Tables 4 and 5 indicate that the Multiple Negatives Ranking Loss is the preferred option for training loss, especially in cases where there are no negative training examples. It in general outperforms all other evaluated losses by a large margin. The Online Contrastive Loss, commonly used for training sentence embeddings, performs second best.

5.5 Number of training steps

The examination of the models' progression during the training phase, conducted retrospectively after every 100 steps, reveals (as illustrated in Figure 1) that the majority of the models attain greater than 90% of their optimal capacity within the first 100 steps, with the exception of the T5 model. Furthermore, it was noted that the pre-trained sentence embeddings displayed superior initial performance, yet with the incorporation of additional training data, the discrepancy in performance between plain transformers and pre-trained sentence embeddings on relevant tasks becomes narrower.

5.6 Effect of removing known labels

To evaluate the model's generalization to unseen labels, we removed labels that appear in the evaluation datasets from the training data. To do this, we lemmatized all words, filtered out determiners and conjunctions, and removed punctuation. If a label in the training data overlapped with or was a subset of a label in the evaluation data, the corresponding example was removed. For instance, if "computers and the internet" appeared in the training data, but "computers internet" was in the evaluation data, the former would be removed. Similarly, if a single-word label in the training data, such as "toys," was a subset of the label "toys and games" in the evaluation data, the former would be removed. As a result, 60 unique tags were removed from the Wikipedia data and 35 from the NYT LDC dataset.

Model	Year	DBPedia	Yahoo	AGNews	RCV	S2ORC	NYT	Tweets	Amazon	AVG
baseline										
word2vec	2013	70.6	37.4	72.1	36.6	26.1	30.3	51.1	31.8	44.5
TE-Wiki	2022	90.2	57.3	79.6	56.5	44.2	59.5	61.5	51.5	64.1
pre-trained contextual embeddings										
GPT	2018	25.6	32.5	25.7	24.7	8.7	9.0	19.7	31.4	22.1
BERT base uncased	2019	23.1	13.4	35.7	13.2	10.1	4.4	5.5	12.8	14.8
RoBERTa base	2019	8.0	9.0	29.8	6.3	6.0	5.4	11.4	12.6	11.1
XLNet base cased	2019	7.2	10.0	25.0	7.0	5.9	5.9	3.0	6.7	8.8
GPT-2	2019	13.3	9.4	26.2	8.5	8.3	5.9	11.3	6.4	11.1
BART base	2020	29.5	16.2	47.8	15.8	7.8	12.9	27.8	12.4	21.3
MPNet base	2020	7.3	10.1	24.8	6.3	7.6	8.3	30.5	7.6	12.8
T5 base	2020	27.5	31.3	51.5	18.3	12.9	9.1	38.3	18.4	25.9
pre-trained sentence embeddings										
all MPNet base v2	2021	74.8	50.0	73.8	50.8	43.4	58.4	57.0	58.3	58.3
GTR T5 base	2021	70.8	42.6	62.3	38.8	31.1	41.4	30.6	53.6	46.4
Sentence T5 base	2022	79.6	48.4	70.9	48.9	39.2	55.5	75.1	65.4	60.4
E5 base	2022	74.3	40.4	71.5	58.8	46.0	52.6	62.4	53.2	57.4
SGPT (125M)	2022	44.3	38.8	51.3	37.4	29.0	25.6	59.4	31.9	39.7
pre-trained commercial embeddings										
OpenAI	2022	76.6	52.1	70.8	58.9	43.2	63.7	63.0	66.0	61.8
Cohere	2022	47.9	39.9	44.2	47.4	35.5	28.4	48.2	54.1	43.2
sentence embeddings for topical text classification										
trained on Wikipedia										
BERT base uncased	2019	86.8	57.6	80.3	63.2	51.0	62.9	65.8	59.2	65.8
BART base	2020	87.3	**59.2**	79.6	58.6	48.3	60.5	72.7	55.9	65.3
MPNet base	2020	89.2	54.3	81.6	**66.9**	51.6	**66.0**	72.2	59.8	67.7
T5 base	2020	84.4	57.1	**82.5**	65.6	50.6	60.8	73.0	56.8	66.3
trained with pre-trained sentence embeddings on Wikipedia										
all MPNet base v2	2021	89.5	58.2	80.9	65.0	**52.5**	62.9	74.2	64.9	68.5
GTR T5 base	2021	**90.9**	56.9	81.5	65.0	48.1	62.7	70.6	67.6	67.9
Sentence T5 base	2022	88.4	57.7	82.3	64.7	48.6	64.0	**75.7**	**68.8**	**68.8**

Table 3: Comparison of the results (accuracy) obtained from distance-based text classification with pre-trained contextual embeddings, pre-trained sentence embeddings, custom trained sentence embeddings on the Wikipedia dataset with Multiple Negatives Ranking Loss.

Our experiments, as shown in Table 6, indicate that there has been a slight decline in the performance of the model when known labels are removed from the training data. However, despite this decline, the model still performs well when compared to the TE-Wiki model. This highlights the model's ability to generalize and apply to unseen labels.

5.7 Error analysis

Our examination of incorrect topic label predictions for associated texts revealed three main issues: 1) sentence embeddings sometimes fail to capture the actual meaning of a text when language from a different topical domain is used; 2) the predicted label accurately represents the text's true meaning, but may differ from the annotated label, as both topics can be relevant to the text; and 3) the text may have an inaccurately annotated label.

To better illustrate these problems, below we provide an example for each.

AGNews. *Text*: The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com) . SPACE.com - TORONTO, Canada – A second team of rocketeers competing for the #36;10 million Ansari X Prize, a contest for privately funded suborbital space flight, has officially announced the first launch date for its manned rocket. *Annotated label*: technology; *predicted label*: sports.

AGNews. *Text*: Dutch Retailer Beats Apple to Local Download Market. AMSTERDAM (Reuters) - Free Record Shop, a Dutch music retail chain, beat Apple Computer Inc. to market on Tuesday with the launch of a new download service in Europe's latest battleground for digital song services. *Annotated label*: technology, *predicted label*: business.

Tweets. *Text*: I m trying to access GenBank and other URL sites, but all come back as not available. Anybody else having this problem? Is the server down? @National Library of Medicine@ @NCBI@ *Annotated label*: business & entrepreneurs, *predicted label*: science & technology.

Moreover, we noticed that categories with overlapping or similar meanings can be misclassified. In our experiments with the S2ORC dataset, abstracts from subjects such as biology, chemistry, geography, and geology were inaccurately classified

Loss	DBPedia	Yahoo	AGNews	RCV	S2ORC	NYT	Tweets	Amazon	AVG
Wikipedia									
Multiple Negatives Ranking Loss	**86.8**	**57.6**	**80.3**	**63.2**	**51.0**	**62.9**	65.8	**59.2**	**65.8**
Cosine Similarity Loss	82.2	57.0	80.1	51.8	49.9	53.3	71.8	49.2	61.9
Contrastive Loss	82.0	57.3	80.1	53.8	50.8	53.6	**72.8**	49.0	62.4
Online Contrastive Loss	85.8	56.0	78.9	54.5	50.0	58.2	71.7	55.8	63.9
NYT LDC									
Multiple Negatives Ranking Loss	**76.4**	**55.9**	**85.4**	**64.0**	**47.3**	**65.4**	**76.4**	**62.2**	**66.6**
Cosine Similarity Loss	65.3	50.0	84.1	56.9	37.7	57.9	60.0	42.8	56.8
Contrastive Loss	55.6	46.7	82.8	56.2	39.1	58.2	62.3	42.1	55.4
Online Contrastive Loss	60.2	49.5	80.0	58.9	44.5	61.4	62.7	51.7	58.6

Table 4: Comparison of the results (accuracy) obtained from distance-based text classification after applying four different losses for training custom sentence embeddings based on BERT base model on Wikipedia and NYT LDC datasets.

Loss	DBPedia	Yahoo	AGNews	RCV	S2ORC	NYT	Tweets	Amazon	AVG
Wikipedia									
Multiple Negatives Ranking Loss	**87.3**	59.2	79.6	**58.6**	**48.3**	**60.5**	72.7	**55.9**	**65.3**
Cosine Similarity Loss	78.7	61.0	**81.1**	45.0	45.9	55.1	74.0	45.7	60.8
Contrastive Loss	79.4	**61.7**	80.2	46.1	46.1	55.9	**75.9**	45.7	61.4
Online Contrastive Loss	84.4	59.9	78.2	45.9	46.4	56.3	69.1	50.8	61.4
NYT LDC									
Multiple Negatives Ranking Loss	**76.6**	**57.6**	83.7	**59.4**	36.9	**67.8**	**64.8**	**58.1**	**63.1**
Cosine Similarity Loss	56.6	48.1	**84.9**	51.9	32.2	60.7	51.2	41.4	53.4
Contrastive Loss	59.6	48.8	84.6	52.4	35.5	59.7	49.5	41.8	54.0
Online Contrastive Loss	65.4	48.3	80.6	53.0	**37.4**	59.9	47.7	49.9	55.3

Table 5: Comparison of the results (accuracy) obtained from distance-based text classification after applying four different losses for training custom sentence embeddings based on BART base model on Wikipedia and NYT LDC datasets.

as environmental science (Appendix A.4 Figure 2). This could be due to the interdisciplinary nature of environmental science, which encompasses several of these subjects and may result in similar semantic representations for the texts and topic labels.

5.8 Computational efficiency and flexibility

The proposed method exhibits a greater degree of computational efficiency in comparison to the TE-Wiki model and similar NLI/TE classifiers. This is due to the fact that the proposed method only requires inference to be performed on the total number of classes and text examples (n class labels + n texts), as opposed to the former methods which require inference for each class label and text pair (n class labels * n texts). Our experiments with measuring time performance of two methods in the same set up (BERT base model, sequence length 128 and batch size 256) on DBPedia (14 classes), Yahoo (10 classes), and AGNews (4 classes) datasets demonstrate a significant reduction in computational time with the proposed method. Specifically, the proposed method was found to reduce computational time by a factor of 15, 11, and 4 times on the respective datasets. Notably, the benefit of our method increases substantially when dealing with a larger number of classes.

The proposed method not only increases computational efficiency, but also offers greater flexibility. By pre-computing text representations, text classification can be updated to a new schema by simply re-computing the representation for topic labels. In contrast, any changes to the topical schema or labels in zero-shot classifiers require reclassifying all results. This is often necessary when the text distribution is unknown and multiple classification iterations are required.

6 Conclusion & Future work

In this study, we examine the performance of contextual embeddings and neural word embeddings in distance-based topical multiclass text classification tasks. Our findings indicate that plain contextual embeddings are suboptimal for such tasks compared to neural word embeddings. Additionally, sentence embeddings, which have been shown to have improved representation capabilities for semantic similarity and search tasks, still do not surpass the performance of transformer-based zero-shot general-purpose classifier proposed by Ding et al. (2022). A plausible explanation for this under-

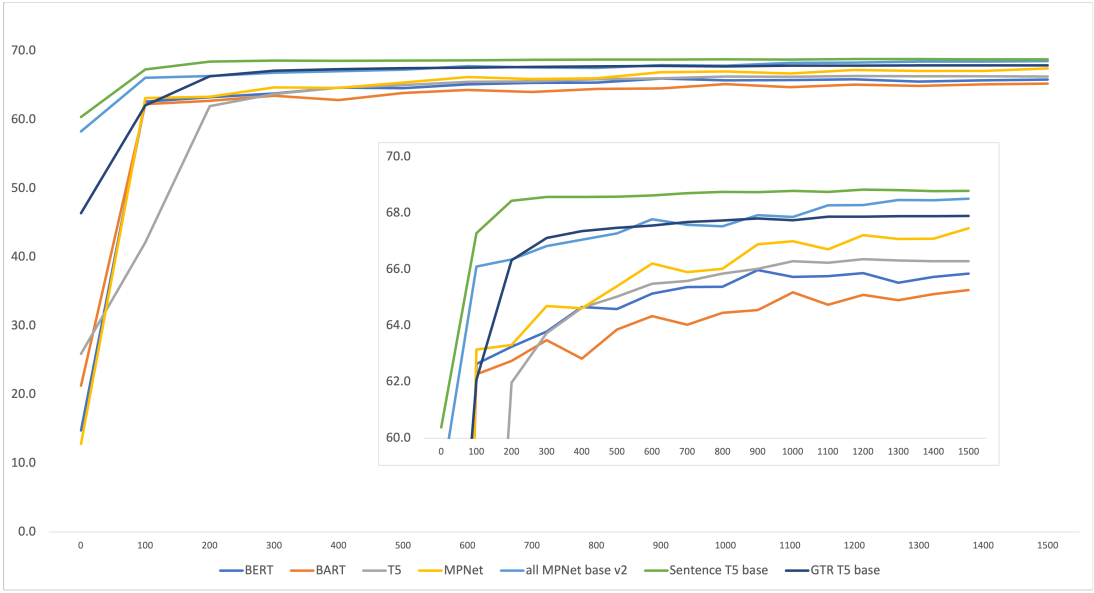

Figure 1: Comparison of classification results (average accuracy) on seven datasets after incremental training of custom sentence embeddings based on pre-trained contextual embeddings and pre-trained sentence embeddings on the Wikipedia dataset.

	DBPedia	Yahoo	AGNews	RCV	s2orc	NYT	Tweets	Amazon	AVG
					Wikipedia				
All	**86.8**	**57.6**	80.3	**63.2**	**51.0**	**62.9**	**65.8**	**59.2**	**65.8**
Unseen	85.0	55.7	**81.2**	62.4	48.8	62.3	61.5	58.8	64.5
Difference %	-2.1	-3.4	1.2	-1.1	-4.2	-1.0	-6.5	-0.7	-2.2
					NYT LDC				
All	**76.4**	55.9	**85.4**	**64.0**	47.3	**65.4**	**76.4**	**62.2**	**66.6**
Unseen	73.9	**57.3**	85.1	**64.0**	**48.8**	62.8	74.0	59.2	65.6
Difference %	-3.2	2.6	-0.4	0.0	3.1	-4.0	-3.1	-4.8	-1.2

Table 6: Comparison of the results (accuracy) obtained from distance-based text classification after removing same or similar labels from training data. Trained BERT base model on Wikipedia and NYT LDC datasets.

performance is that sentence embeddings primarily focus on both lexical and semantic overlap, potentially overlooking the abstract aspects of topical relationships.

To address these limitations, we introduce the concept of task-specific sentence embeddings that enforce the relationship between topic labels and text in a shared semantic space. This enhances their suitability for distance-based topical multiclass text classification. Our method is model and training data agnostic and can be applied with various transformer-based models and trained on plain texts tagged with relevant topic labels. The results demonstrate comparable or improved performance compared to state-of-the-art transformer-based zero-shot general-purpose classifiers and offer additional benefits such as increased computational efficiency and greater flexibility in topical text classification.

The promising avenues for future research involve addressing the limitations of shallow semantic representation of texts using sentence embeddings and extending the proposed method to enable multilabel topical text classification.

Acknowledgments

This research was funded by Flanders Innovation & Entrepreneurship (VLAIO), grant HBC.2021.0222.

References

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the*

23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08, page 830–835. AAAI Press.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Hantian Ding, Jinrui Yang, Yuqian Deng, Hongming Zhang, and Dan Roth. 2022. Towards open-domain topic classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 90–98, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Zied Haj-Yahia, Adrien Sieg, and Léa A. Deleris. 2019. Towards unsupervised text classification leveraging experts and word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 371–379, Florence, Italy. Association for Computational Linguistics.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2022. Unsupervised text classification with neural word embeddings. *Computational Linguistics in the Netherlands Journal*, 12:165–181.

Moritz Laurer, W v Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying–addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *ArXiv*, abs/2003.07278.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Bonan Min, Hayley H. Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ArXiv*, abs/2111.01243.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus LDC2008T19. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*.

Tim Schopf, Daniel Braun, and Florian Matthes. 2021. Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies - WEBIST,*, pages 124–132. INSTICC, SciTePress.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR)*, NLPIR 2022, New York, NY, USA. Association for Computing Machinery.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1579–1585. AAAI Press.

Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. 2016. Cross-lingual dataless classification for many languages. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2901–2907. AAAI Press.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. 2022b. Generalised zero-shot learning for entailment-based text classification with external knowledge. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 19–25.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation, and Entailment Approach. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Appendix

A.1 Corpora topic labels

1. **DBPedia**: album; animal; artist; athlete; building; company; film; novel publication book; plant tree; politics; river mountain lake; school university; transportation; village.
2. **Yahoo Answers**: business finance; computers Internet; education reference; entertainment music; family relationships; health; politics government; science mathematics; society culture; sports.
3. **AGNews**: business; politics; sports; technology.
4. **RCV**: arts, culture, entertainment; biographies, personalities, people; crime, law enforcement; defence; disasters and accidents; domestic politics; environment and natural world; health; human interest; international relations; labour issues; religion; science and

technology; sports; travel and tourism; war, civil war; weather; welfare, social services.

5. **S2ORC**: art; biology; business; chemistry; computer science; economics; engineering; environmental science; geography; geology; history; materials science; mathematics; medicine; philosophy; physics; political science; psychology; sociology.
6. **NYT**: arts; automobiles; books; business; education; fashion & style; food; health; home & garden; movies; politics; real estate; science; sports; technology; theater; travel.
7. **Tweets**: arts & culture; business & entrepreneurs; daily life; pop culture; science & technology; sports & gaming.
8. **Amazon**: automotive; books; cell phones and accessories; gift cards; industrial and scientific; magazine subscriptions; movies and tv; musical instruments; office products; pet supplies; software; sports and outdoors; tools and home improvement; toys and games; video games.

A.2 Training data

Wikipedia	
Surnames	54284
Musical groups	45153
Writers	44117
Musicians	28991
Books	28689
Video games	21970
Ethnic groups	21939
Politicians	18403
Vehicles	18139
Women	17303
Rivers	17268
Composers	16764
Plants	15990
Government	15463
Chemistry	14766
Astronomy	14554
Music	14286
Civil engineering	14234
Generals	13561
Film	13549

Table 7: Top 20 high-level categories of Wikipedia dataset.

NYT LDC	
Politics and Government	200798
Finances	151958
United States International Relations	113384
United States Politics and Government	102084
Corporations	87340
Company Reports	79580
International Relations	68493
Elections	68479
Medicine and Health	68081
Armament, Defense and Military Forces	65514
Music	55645
Presidential Elections (US)	55466
Books and Literature	54083
Law and Legislation	50823
Baseball	47334
Crime and Criminals	47274
Education and Schools	45192
Weddings and Engagements	44595
United States Armament and Defense	44488
Terrorism	43201

Table 8: Top 20 general descriptors of NYT LDC dataset.

A.3 Classification results

Model	Year	DBPedia	Yahoo	AGNews	RCV	S2ORC	NYT	Tweets	Amazon	AVG
baseline										
word2vec	2013	67.6	35.0	71.4	34.6	24.1	32.4	52.9	30.7	43.6
TE-Wiki	2022	90.1	55.5	79.8	53.4	41.7	57.7	65.3	49.8	61.6
pre-trained contextual embeddings										
GPT	2018	13.1	26.0	11.8	18.6	2.8	4.6	20.0	27.7	15.6
BERT base uncased	2019	16.2	8.2	26.8	6.4	3.0	1.3	1.7	5.7	8.7
RoBERTa base	2019	2.5	2.9	18.6	1.5	1.0	1.1	12.9	5.5	5.8
XLNet base cased	2019	1.1	1.8	10.0	2.7	1.2	0.7	0.4	0.8	2.3
GPT-2	2019	6.7	3.7	17.6	3.3	2.9	1.9	10.0	3.3	6.2
BART base	2020	22.6	11.7	45.2	7.2	4.6	8.3	28.0	7.9	16.9
MPNet base	2020	1.3	2.4	12.6	1.7	2.1	2.7	21.3	1.7	5.7
T5 base	2020	17.9	29.7	51.4	11.8	6.9	2.9	25.1	12.9	19.7
pre-trained sentence embedding										
all MPNet base v2	2021	73.6	49.2	73.5	48.5	42.7	58.3	62.5	56.2	58.1
GTR T5 base	2021	70.2	39.9	60.8	37.2	29.9	41.0	33.9	53.2	45.8
Sentence T5 base	2022	78.8	46.9	70.3	48.1	37.3	55.1	75.4	64.1	59.5
E5 base	2022	72.9	37.1	70.6	57.4	44.8	53.2	65.0	52.1	56.6
SGPT (125M)	2022	35.0	35.2	51.3	30.5	24.3	20.4	63.1	29.8	36.2
OpenAI	2022	75.2	47.8	70.3	54.8	42.8	61.9	66.6	63.6	60.4
Cohere	2022	37.6	36.5	35.5	41.7	30.2	17.6	53.2	49.7	37.8
sentence embeddings for topical text classification										
BERT base	2019	86.4	56.3	80.1	60.7	50.5	62.3	69.5	58.8	65.6
BART base	2020	86.9	**57.7**	79.1	55.5	48.2	59.3	74.9	55.1	64.6
MPNet base	2020	87.7	53.8	80.3	**64.2**	50.8	**64.1**	74.1	60.3	66.9
T5 base	2020	83.3	55.9	82.7	63.4	49.3	61.1	74.3	55.8	65.7
all MPNet base v2	2021	89.1	57.1	80.6	62.0	**52.1**	62.6	76.7	63.6	68.0
GTR T5 base	2021	**90.7**	55.5	81.4	62.0	47.9	61.9	73.6	66.7	67.5
Sentence T5 base	2022	87.7	56.7	**82.1**	61.7	48.5	63.2	**77.8**	**67.6**	**68.1**

Table 9: Comparison of the results (weighted average F1) obtained from distance-based text classification with pre-trained contextual embeddings, pre-trained sentence embeddings, custom trained sentence embeddings on the Wikipedia dataset with Multiple Negatives Ranking Loss.

A.4 Error analysis

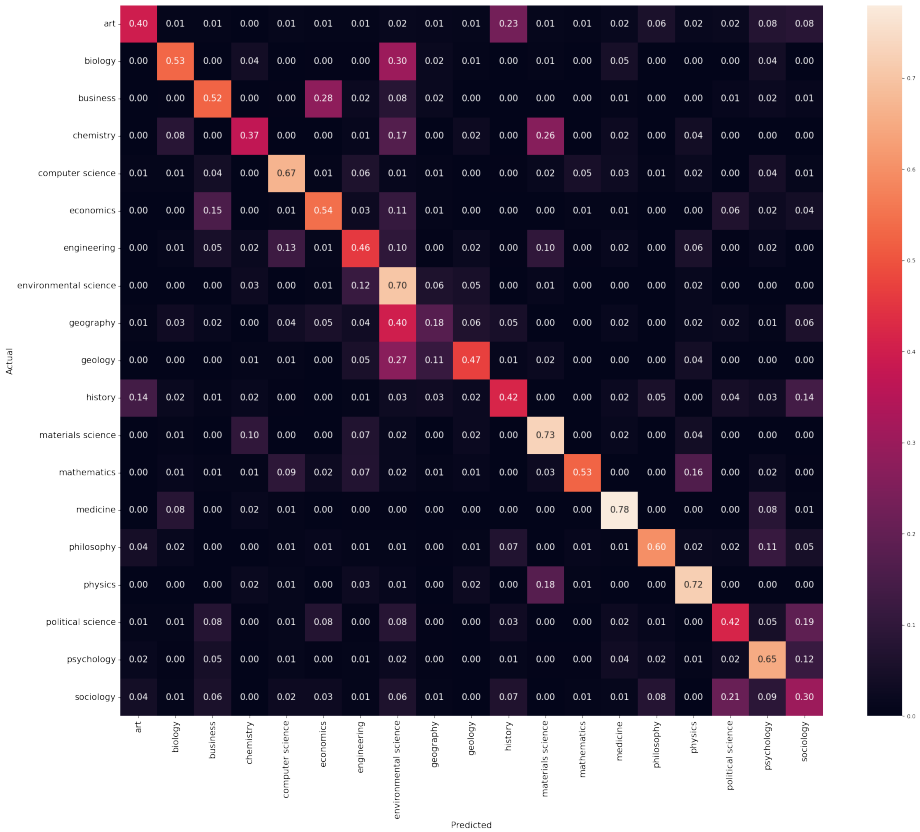

Figure 2: Confusion matrix for classification results of "all MPNet base v2" model trained on the Wikipedia high-level categories with Multiple Negatives Ranking Loss for S2ORC dataset.