Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches

Tim Schopf
Technical University of Munich, Department of Computer Science
Garching, Germany
tim.schopf@tum.de

Daniel Braun
University of Twente, Department of High-tech Business and Entrepreneurship
Enschede, Netherlands
d.braun@utwente.nl

Florian Matthes
Technical University of Munich, Department of Computer Science
Garching, Germany
matthes@tum.de

ABSTRACT

Text classification of unseen classes is a challenging Natural Language Processing task and is mainly attempted using two different types of approaches. Similarity-based approaches attempt to classify instances based on similarities between text document representations and class description representations. Zero-shot text classification approaches aim to generalize knowledge gained from a training task by assigning appropriate labels of unknown classes to text documents. Although existing studies have already investigated individual approaches to these categories, the experiments in literature do not provide a consistent comparison. This paper addresses this gap by conducting a systematic evaluation of different similarity-based and zero-shot approaches for text classification of unseen classes. Different state-of-the-art approaches are benchmarked on four text classification datasets, including a new dataset from the medical domain. Additionally, novel SimCSE [7] and SBERT-based [26] baselines are proposed, as other baselines used in existing work yield weak classification results and are easily outperformed. Finally, the novel similarity-based Lbl2TransformerVec approach is presented, which outperforms previous state-of-the-art approaches in unsupervised text classification. Our experiments show that similarity-based approaches significantly outperform zero-shot approaches in most cases. Additionally, using SimCSE or SBERT embeddings instead of simpler text representations increases similarity-based classification results even further.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Artificial intelligence**; **Machine learning**; **Unsupervised learning**; **Neural networks**; • **Information systems** → **Clustering and classification**.

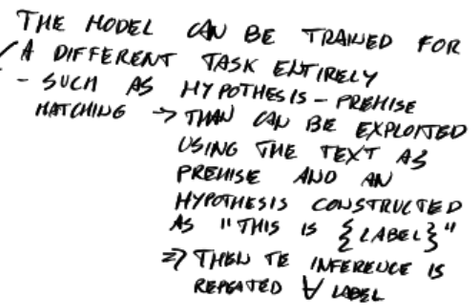
KEYWORDS

Natural Language Processing, Unsupervised Text Classification, Zero-shot Text Classification

1 INTRODUCTION

Unsupervised text classification approaches aim to perform categorization without using annotated data during training and therefore offer the potential to reduce annotation costs. Despite this possibility, unsupervised text classification approaches have attracted significantly less attention in contrast to supervised text classification approaches. As a result, extensive work is already being done to structure and evaluate the field of text classification with a focus on supervised approaches [11, 12, 19, 20] while little research has been conducted on evaluating unsupervised text classification approaches. This study bridges this gap by assessing the two most popular categories of unsupervised text classification approaches.

Generally, unsupervised text classification approaches aim to map text to labels based on their textual description, without using annotated training data. To accomplish this, there exist mainly two categories of approaches. The first category can be summarized under *similarity-based* approaches. Thereby, the approaches generate semantic embeddings of both the texts and the label descriptions, before attempting to match the texts to the labels using similarity measures such as cosine similarity [8, 29, 31, 34]. The second category uses *zero-shot learning (ZSL)* to classify texts of unseen classes. ZSL uses labeled training instances belonging to seen classes to learn a classifier that can predict testing instances belonging to different, unseen classes [38]. Although ZSL techniques employ annotated data for training, they do not use labels to provide information about the target classes and can use their knowledge of the previously seen classes to classify instances of unseen classes. Since pretrained zero-shot text classification (0SHOT-TC) models do not require training or fine-tuning on labeled data from the target classes, we classify them as an unsupervised text classification strategy. The highly successful deep learning performances of recent years have also stimulated research initiatives for 0SHOT-TC [17, 24, 27, 41, 43]. We argue, that one of the main differences between ZSL and similarity-based approaches is, that ZSL approaches use annotated data for seen classes to predict texts of unseen classes, whereas pure similarity-based approaches do not require seen classes at all.



We summarize the contributions of our work as follows::

- We evaluate the *similarity-based* and *zero-shot learning* categories for unsupervised text classification of topics. Thereby, we conduct experiments with representative approaches of each category on four different benchmark datasets, including a new text classification dataset from the medical domain.
- We propose simple but strong baselines for unsupervised text classification based on SimCSE [7] and SBERT [26] embedding similarities. Previous work has mostly been evaluated against different weak baselines such as Word2Vec [18] similarities which are easy to outperform and tend to overestimate the performance of new unsupervised text classification approaches.
- Since transformer-based text representations have been widely established as state-of-the-art for semantic text similarity in recent years, we further adapt Lbl2Vec [31], one of the most

recent and well-performing similarity-based methods for unsupervised text classification, to be used with transformer-based language models[1].

2 RELATED WORK

Chang et al. [2] investigated unsupervised text classification under the umbrella name "Dataless Classification" in one of their earliest works. They used Explicit Semantic Analysis (ESA) [6] to embed the text and label descriptions in a common semantic space and picked the label with the highest matching score for classification. Semantic embeddings are vector representations of texts that capture their semantic meaning and can be used as input for a variety of different Natural Language Processing (NLP) downstream tasks [1, 30, 32, 33]. Dataless classification is based on the idea that semantic representations of labels are equally relevant as learning semantic text representations and was subsequently further examined in [3, 16, 34, 35].

With the progress of text embeddings, the term "Dataless Classification" became less prevalent and was rather represented by the broad category of similarity-based approaches for unsupervised classification. Within this category, Sappadla et al. [29] embedded text documents and textual label descriptions with Word2Vec and used cosine similarity between text and label embeddings to predict instances of unseen classes. Haj-Yahia et al. [8] proposed to enrich label descriptions with expert keywords and subsequently conduct unsupervised classification based on Latent Semantic Analysis (LSA) [4] similarities. Stammbach and Ash [36] introduced DocSCAN, which produces semantic representations of text documents and uses Semantic Clustering by Adopting Nearest-Neighbors for unsupervised text classification. Schopf et al. [31] used Doc2Vec [14] to jointly embed word, document, and label vectors for subsequent similarity-based unsupervised text classification.

Similarly, Nam et al. [21] jointly embedded document, label, and word representations with Doc2Vec. However, they learned a ranking function for multi-label classification and attempted to predict instances of unseen classes in a zero-shot setting for classification. Zhang et al. [43] integrated four types of semantic knowledge (word embeddings, class descriptions, class hierarchy, and a general knowledge graph) in a two-phase framework for 0SHOT-TC. Yin et al. [41] proposed to treat 0SHOT-TC as a textual entailment problem, while Ye et al. [40] tackled 0SHOT-TC with a semi-supervised self-training approach.

3 TEXT CLASSIFICATION APPROACHES

3.1 Baselines

We compare the findings of current state-of-the-art unsupervised text classification approaches to some basic baselines to evaluate their performance.

LSA: Singular Value Decomposition (SVD) is used on term-document matrices to learn a set of concepts (or topics) related to the documents and terms [4]. For each dataset, we apply LSA to learn $n =$ *number of classes* concepts. Afterward, the text documents are classified according to the highest cosine similarity of resulting LSA vectors of documents and label keywords. A similar approach was used by Haj-Yahia et al. [8] for unsupervised text classification.

Word2Vec: This produces semantic vector representations of words based on surrounding context words [18]. A Skip-gram model with a vector size of 300 and a surrounding window of 5 is trained for each dataset. The average of word embeddings is then used to represent the text documents and label keywords. The text documents are predicted according to the highest cosine similarity of the resulting Word2Vec representations of documents and label keywords for classification. Similar approaches were used by Yin et al. [41] and Ye et al. [40] as baseline for 0SHOT-TC.

SimCSE: This is a contrastive learning framework that produces sentence embeddings which acieve state-of-the-art results in semantic similarity tasks [7]. Algorithm 1 is first used to separate the text documents into paragraphs because SimCSE models have a maximum input sequence length. Then, the average of SimCSE paragraph embeddings as text document representations and the average of SimCSE label keyword embeddings as class representations are employed. Finally, the text documents are classified according to the highest cosine similarity of the resulting SimCSE representations of document and label keywords.

SBERT: This is a modification of BERT [5] that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings [26]. We use the same classification approach as described in the paragraph above, except that we now use SBERT embeddings instead of SimCSE embeddings.

Algorithm 1 Split text document into paragraphs

Require:
d = text document
m_k = max input sequence length of transformer-model k
$len(x)$ = numer of words in text x
procedure SPLIT-DOCUMENT(d, m_k)
  sentences$_d$ ← sentence_tokenize(d)
  paragraphs$_d$ ← ∅
  p ← ∅
  for s **in** sentences **do**
    if $len(p) + len(s) < \frac{m_k}{2}$ **then**
      $p \leftarrow p + s$
    else
      paragraphs$_d$ ← paragraphs$_d$ + p
      p ← ∅
  return paragraphs$_d$

3.2 Similarity-based Text Classification

As previously stated, numerous similarity-based approaches for unsupervised text classification exist. However, the recently introduced Lbl2Vec approach [31] is focused on in this study. We chose Lbl2Vec to represent the similarity-based classification category since preliminary experiments confirmed improved performance compared with other similarity-based approaches. Lbl2Vec

[1] Code available: https://github.com/sebischair/Lbl2Vec

works by jointly embedding word, document, and label representations. First, word and documented representations are learned with Doc2Vec. Then, the average of label keyword representations for each class is used to find a set of most similar candidate document representations via cosine similarity. The average of candidate document representations, in turn, generates the label vector for each class. For classification, eventually, the documents are assigned to the class where the cosine similarity of the label vector and the document vector is the highest.

We adapt the Lbl2Vec approach, using transformer-based text representations instead of Doc2Vec to create jointly embedded word, document, and label representations. Since transformer-based text representations currently achieve state-of-the-art results in text-similarity tasks, we investigate the effect of the different resulting text representations on this similarity-based text classification strategy. In this paper, we use SimCSE [7] and SBERT [26] transformer-models to create text representations. We use the average paragraph embeddings per document as document representations. The paragraphs of documents are obtained by applying Algorithm 1. To find candidate documents for label vectors, the transformer-models create individual embeddings for each label keyword. Then, cosine similarity is used to find the documents that are most similar to the average of the label keyword embeddings for each class. After obtaining the candidate documents this way, the label vectors as an average of the candidate document representations for each class are computed. For classification, the documents are assigned to the class where the cosine similarity between the label vector and the document vector is the highest. In the following, the Lbl2Vec approach adapted with transformer-based text representations is referred to as Lbl2TransformerVec.

3.3 Zero-shot Text Classification

0SHOT-TC is still relatively less researched, but nevertheless yields some promising approaches. Using pretrained 0SHOT-TC models can be considered an unsupervised text classification strategy, since no label information of target classes are required for training or fine-tuning. Although newer approaches such as the one of Liu et al. [17] exist, preliminary experiments confirmed that the zero-shot entailment approach [41] still produces state-of-the-art 0SHOT-TC results in predicting instances of unseen classes. As the name already implies, the zero-shot entailment approach deals with 0SHOT-TC as a textual entailment problem. The underlying idea is similar to that of similarity-based text classification approaches. Conventional 0SHOT-TC classifiers fail to understand the actual problem since the label names are usually converted into simple indices. Therefore, these classifiers can hardly generalize from seen to unseen classes. Considering 0SHOT-TC as an entailment problem, however, provides the classifier with a textual label description and therefore enables it to understand the meaning of labels.

Similarly, TARS [9] also uses the textual label description to classify text in a zero-shot setting. However, TARS approaches the task as a binary classification problem, where a text and a textual label description is given to the model, which makes a prediction about whether that label is true or not. The TARS authors state that this approach significantly outperforms GPT-2 [25] in 0SHOT-TC.

Since the zero-shot entailment approach currently produces state-of-the-art results in predicting instances of unseen classes and TARS also promises encouraging results, we select both approaches to represent the ZSL category for unsupervised text classification.

4 DATASETS

Our evaluation is based on four text classification datasets from different domains. As we use the semantic meaning of class descriptions for unsupervised text classification, we infer label keywords from each class name that serves the purpose of textual class descriptions. Thereby, the inference step simply consists of using the class names provided by the official documentation of the datasets as label keywords. In a few cases, we additionally substituted the class names with synonymous or semantically similar keywords, if we considered this to be a more appropriate description of a certain class.

4.1 20Newsgroups

The 20Newsgroups[2] dataset is a common text classification benchmark dataset. It was introduced by Lang [13] and comprised approximately 20,000 newsgroup posts, equally distributed across 20 different newsgroups classes. Appendix A.1 summarizes the classes and inferred label keywords.

4.2 AG's Corpus

The original AG's Corpus[3] dataset is a collection of over 1 million news articles on different topics. The Zhang et al. [44] version is used in this study, which was constructed by choosing the 4 largest classes from the original corpus. Each class contains 30,000 training samples and 1,900 testing samples. In total, the dataset consists of 127,600 samples. Appendix A.2 summarizes the classes and inferred label keywords.

4.3 Yahoo! Answers

The Yahoo! Answers dataset was constructed by Zhang et al. [44] and contains 10 different topic classes. Each class contains 140,000 training samples and 6,000 testing samples. In total, the dataset consists of 1,460,000 samples. Appendix A.3 summarizes the classes and inferred label keywords.

4.4 Medical Abstracts

We obtained the raw Medical Abstracts dataset through Kaggle[4]. The original corpus contains 28.880 medical abstracts describing 5 different classes of patient conditions, with only about half of the dataset being annotated. Furthermore, the original annotations consist of numerical labels only. A medical text classification dataset from this corpus by using only the labeled medical abstracts was created, adding descriptive labels to the respective classes, and splitting the data into a training set and a test set. Table 1 shows a summary of the processed Medical Abstracts dataset.

The inferred label keywords for each class are summarized in Appendix A.4. We make this corpus available under the Creative

[2] qwone.com/ jason/20Newsgroups
[3] groups.di.unipi.it/~gulli/AG_corpus_of_news_articles
[4] https://www.kaggle.com/datasets/chaitanyakck/medical-text

Class Name	#training	#test	∑
Neoplasms	2530	633	3163
Digestive system diseases	1195	299	1494
Nervous system diseases	1540	385	1925
Cardiovascular diseases	2441	610	3051
General pathological conditions	3844	961	4805
∑	11550	2888	14438

Table 1: Class distributions within the Medical Abstracts dataset.

Commons CC BY-SA 3.0 license[5] at https://github.com/sebischair/Medical-Abstracts-TC-Corpus.

5 EXPERIMENTAL DESIGN

For evaluation of different unsupervised text classification approaches, we use the datasets described in Section 4. Since we don't use label information to train the classifiers, we concatenate the training and test sets for each dataset and use the respective entire concatenated datasets for training and testing. After checking the Yahoo! Answers dataset for consistency, we observe that some answers we try to classify are empty or contain simple yes/no statements. Therefore, answers that are empty or consist of only one word are removed. We use the label keywords described in Appendix A for all text classification approaches to create class representations. Additionally, for the baselines and similarity-based approaches, we use the average of the respective label keyword embeddings as class representations. In contrast, for the zero-shot approaches, the respective label keywords of the 20Newsgroups, AG's Corpus, and Yahoo! Answers classes are concatenated with "and" and then used as hypotheses/label descriptions. For the Medical Abstracts dataset just the class names are used as hypotheses/label descriptions.

We use the approaches described in Section 3.1 as baselines for unsupervised text classification. For our SimCSE experiments, we use the *sup-simcse-roberta-large*[6] model. To create embeddings for the SBERT baseline approach, we use two different pretrained SBERT models. We choose the general purpose models *all-mpnet-base-v2*[7] and *all-MiniLM-L6-v2*[8], trained on more than one billion training pairs and expected to perform well on sentence similarity tasks. The *all-mpnet-base-v2* model is larger than the *all-MiniLM-L6-v2* model and guarantees slightly better quality sentence embeddings. The smaller *all-MiniLM-L6-v2* model, on the other hand, guarantees a five times faster encoding time while still providing sentence embeddings of high quality.

For evaluation of similarity-based text classification, we apply the approaches described in Section 3.2. Similar to the SimCSE and SBERT baseline approaches, we generate text embeddings for the Lbl2TransformerVec approach using the *sup-simcse-roberta-large*, *all-mpnet-base-v2*, and *all-MiniLM-L6-v2* models.

For evaluation of 0SHOT-TC, we use the zero-shot approaches described in Section 3.3. We conduct experiments with three different pretrained zero-shot entailment models: a DeBERTa [10] model [9] trained on the MultiNLI [39], Fever-NLI [37], LingNLI [23], and DocNLI [42] datasets, a large BART [15] model[10] trained on the MultiNLI dataset, and a smaller DistilBERT [28] model[11] trained on the MultiNLI dataset. For TARS experiments, we use the BERT-based pretrained *tars-base-v8*[12] model. Since *tars-base-v8* pretraining is partly done on AG's Corpus, we don't conduct TARS experiments on this dataset.

5.1 Hypotheses

We had four main hypotheses prior to conducting the experiments.

(1) **0SHOT-TC models yield better text classification results than similarity-based approaches:** The 0SHOT-TC models investigated in this paper use a cross-encoder architecture which allows them to compare the input text and the textual label description simultaneously, while performing self-attention over both. In contrast, the similarity-based approaches encode the input text and label description separately. For semantic text similarity tasks, cross-encoders have proven to perform better than calculating cosine similarities for separately encoded texts. Hence we expect a similar outcome for unsupervised text classification.

(2) **Using larger Pretrained Language Models (PLMs) results in better classification performances:** Although this may seem obvious, we nevertheless want to examine whether the outcomes of using larger PLMs justify their drawbacks during training and inference.

(3) **Classification results of PLM-based approaches are highly domain dependent:** We assume that, PLM-based approaches lose some of their classification performance when dealing with very domain-specific corpora, since this specific domain may be underrepresented in the training data. Therefore, we anticipate that for certain domains, approaches like Lbl2Vec that trains unsupervised models on the classification data from scratch might perform comparably better.

(4) **With increasing length of text documents, the performance of SimCSE and SBERT-based approaches decreases:** SimCSE and SBERT representations are most effective if the texts are embedded as a whole and no truncation strategy is used. Since we compute the document representations as the average of their respective paragraph embeddings, we assume that the quality of SimCSE and SBERT document embeddings decreases with increasing text length, resulting in worse classification performance accordingly.

TO HELP THE MODEL THEY DIDN'T USE ONLY THE LABEL - BECAUSE IT'S NOT RAPRESENTATIVE ENOUGH OF THE CLASS THEY WANT
=> INSTEAD THEY USED THE LABEL THEY WANTED + 2/3 SYNONYMOUS
=> THAN DID THE MEAN OF THE EMBEDDINGS.
SEE APPENDIX A FOR THE LABEL KEYWORD THEY USED

X
LIB2TRANSFORMERVEC OTPERFORMED ALL THE OTHER APPROACHES

✓
THE PERFORMANCE INCREASE BUT THEY ARE ALSO SLOWER.
THERE IS A TRADE-OFF TO TAKE INTO ACCOUNT

X
THE DATASET PROPOSE VARY IN TOPIC YET LIB2TRANSFORMERVEC ACHIEVED CONSISTENT PERFORMANCE.

X
THEY COULDN'T CONFIRM THE HYPOTHESIS BECAUSE OF STATISTICAL INSIGNIFIANCE.

[5] https://creativecommons.org/licenses/by-sa/3.0
[6] princeton-nlp/sup-simcse-roberta-large
[7] sentence-transformers/all-mpnet-base-v2
[8] sentence-transformers/all-MiniLM-L6-v2
[9] MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c
[10] facebook/bart-large-mnli
[11] typeform/distilbert-base-uncased-mnli
[12] https://flair.informatik.hu-berlin.de/resources/models/tars-base/

6 EVALUATION

Table 2 shows the performances of unsupervised text classification approaches for each dataset, measured in F_1-scores. We can observe that none of the baselines achieves the highest F_1-score on any dataset based on these data. This indicates that the use of advanced unsupervised text classification approaches usually yields better results than simple baseline approaches. However, we observe that the LSA and Word2Vec approaches generally yield the worst results and are easy to outperform. In contrast, the SimCSE and SBERT baselines produce strong F_1-scores that even some of the advanced approaches could not surpass in certain cases. Furthermore, the SimCSE and SBERT baseline approaches may produce better results than the Lbl2Vec similarity-based approach on three datasets. We nevertheless can deduce that the use of advanced similarity-based approaches generally produces better unsupervised text classification results than the use of simple baseline approaches. Specifically, the Lbl2TransformerVec approaches using SBERT embeddings appear to be promising, as they consistently perform well across all datasets and outperform the baseline results. In contrast, the 0SHOT-TC approaches perform consistently weak and in the majority of cases did not even manage to outperform the baseline results. However, the DeBERTa zero-shot entailment model could classify the domain-specific medical abstracts surprisingly well and achieved the best F_1-score of all classifiers on this dataset. Nevertheless, considering that all 0SHOT-TC models yielded disappointing results in all remaining experiments and also failed to outperform the baselines, our first hypothesis can be rejected.

Concerning our second hypothesis, the results are less obvious. On the one hand, the large DeBERTa zero-shot entailment model always significantly outperforms the smaller BART-large and DistilBERT zero-shot entailment models. Additionally, the BERT-based TARS model performs slightly better than the smaller DistilBERT zero-shot entailment model, except in case of the domain-specific Medical Abstracts dataset. Conversely, *all-mpnet-base-v2* and *all-MiniLM-L6-v2*-based approaches tend to produce unsupervised classification results that are fairly close to each other. Although these results are quite similar and sometimes even approaches based on the smaller *all-MiniLM-L6-v2* model perform better, we nevertheless see that approaches based on the larger *all-mpnet-base-v2* produce slightly better results in most cases. Therefore, we find sufficient support for our second hypothesis in the case of similarity-based unsupervised text classification approaches, with even stronger support in case of 0SHOT-TC.

		20Newsgroups	AG's Corpus	Yahoo! Answers	Medical Corpus
Baselines	**LSA**	17.89	41.17	15.82	31.61
	Word2Vec	12.87	28.22	12.55	25.00
	SimCSE	42.84	80.10	49.90	34.94
	SBERT (all-MiniLM-L6-v2)	57.89	68.57	43.77	46.53
	SBERT (all-mpnet-base-v2)	59.75	70.84	51.25	46.34
Similarity-based TC	**Lbl2Vec**	**65.71**	74.63	44.26	43.03
	Lbl2TransformerVec (SimCSE)	58.79	**83.79**	53.32	39.60
	Lbl2TransformerVec (all-MiniLM-L6-v2)	63.01	80.88	52.87	54.57
	Lbl2TransformerVec (all-mpnet-base-v2)	64.69	80.05	**55.84**	56.46
0SHOT-TC	**TARS**	17.65	-	34.60	10.92
	Zero-shot Entailment (DistilBERT)	16.27	59.48	31.81	25.74
	Zero-shot Entailment (BART-large)	38.54	68.24	40.21	56.86
	Zero-shot Entailment (DeBERTa)	47.19	72.57	43.09	**57.28**

BEST OF ALL

Table 2: F_1-scores (micro) of examined text classification approaches on different datasets. The best results on the respective dataset are displayed in bold. Since we use micro-averaging to calculate our classification metrics, we realize equal F_1, Precision, and Recall scores respectively.

Figure 1 shows a more detailed view of the classification results by visualizing the F_1-scores of classification models for the individual classes of all datasets. Here we observe that the overall performance of classifiers is class-dependent. While all classifiers generally yield good results for some classes (e.g. the sports classes of the 20Newsgroups and AG's Corpus datasets), all classifiers performed considerably worse for other classes (e.g. "talk.religion.misc [20Newsgroups]" or "Education & Reference [Yahoo! Answers"]). When we compare the performance of the Lbl2Vec model, which was trained from scratch, to that of PLM-based approaches, we discover that all approaches produce similar results for many classes. In some classes, however, Lbl2Vec clearly outperforms F_1-scores of all other PLM-based approaches (e.g. in the "comp.sys.mac.hardware", "misc.forsale", or "alt.atheism" classes of the 20Newsgroups dataset). Unfortunately, this fact can't be generalized from individual classes to the entire domains. For example, Lbl2Vec scores relatively well in "comp.sys.mac.hardware (20Newsgroups)" and "comp.windows.x (20Newsgroups)" classes, but performs significantly worse than PLM-based models in "comp.os.ms-windows.misc (20Newsgroups)", despite all classes belonging to the same domain. We conclude that although a model trained from scratch can yield better results than PLM-based approaches in some cases, as demonstrated by the Lbl2Vec results on the 20Newsgroups dataset, we do not find sufficient support for our third hypothesis.

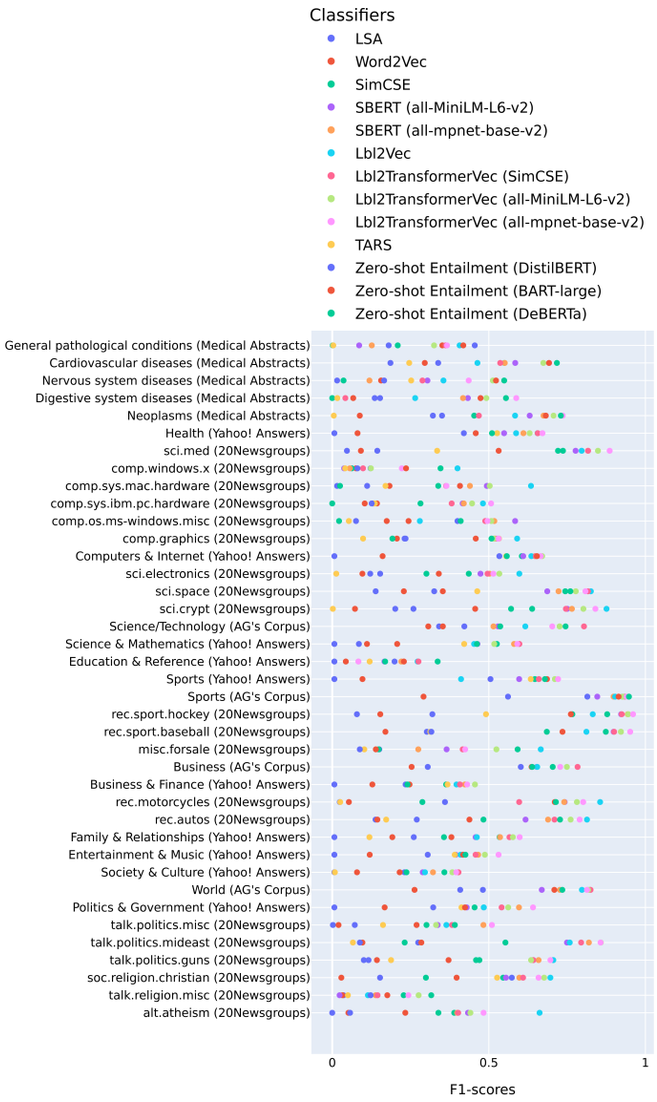


Figure 1: F_1-scores of classification models for the individual classes of all four benchmark datasets.

Model	Kendall's τ	p-value
SimCSE	-0.16	0.16
SBERT (all-MiniLM-L6-v2)	0.07	0.52
SBERT (all-mpnet-base-v2)	0.04	0.73
Lbl2TransformerVec (SimCSE)	-0.08	0.46
Lbl2TransformerVec (all-MiniLM-L6-v2)	-0.03	0.82
Lbl2TransformerVec (all-mpnet-base-v2)	0.03	0.80

Table 3: Results of the correlation analysis to measure the relationship between X = average number of document words of each class in all four benchmark datasets and Y = F_1-scores of each class in all four benchmark datasets.

To test our fourth hypothesis, we perform a correlation analysis measuring monotonic relationships between the F_1-scores of the transformer-based classification approaches per class and the average number of document words per class. We choose Kendall's τ as correlation coefficient, because of its robustness against outliers and the small dataset. Further, we determine a significance level of 0.05. Table 3 shows the results of this correlation analysis. We can observe that all correlation coefficients are close to zero. Therefore, we can't identify a correlation trend. Moreover, all p-values exceed our defined significance level of 0.05 by far, indicating our test results are statistically insignificant. As a result, we find no support for our fourth hypothesis and reject it.

7 LIMITATIONS

One significant limitation of this evaluation is that only unsupervised text classification results for the *topic* aspect are considered. This means that we consider classification results based on topics that describe what a text document is about only. However, text classification can be seen in a broader context where aspects such as *emotion* or *situation* are predicted as well [41]. We only focus on unsupervised similarity-based approaches and 0SHOT-TC approaches that can classify the entire datasets without requiring training or fine-tuning on parts of the datasets. Self-training approaches which address the problem as a semi-supervised task or ZSL approaches that use parts of the datasets for training or fine-tuning, may lead to different results. Although we try to generalize from the datasets

and approaches examined in the experiments, our evaluation is limited to those datasets and approaches nonetheless.

8 CONCLUSION

The evaluation of unsupervised text classification approaches in Section 6, has shown that similarity-based approaches generally outperform 0SHOT-TC approaches in a variety of different domains. 0SHOT-TC approaches tend to produce relatively bad results and are therefore hardly eligible for unsupervised text classification problems. In comparison, similarity-based approaches appear to predict instances of unknown classes more accurately. The characteristics of text embeddings enable representations of similar topics or classes to be located close to each other in embedding space. This implies that text representation approaches which are able to cluster topics in embedding space coherently also perform well in unsupervised text classification. This characteristic is also evident in our work. DensMap [22] visualizations of document representations in embedding space used for classification in this work are shown in Appendix A.5. To improve similarity-based text classification results even further, we can use additional, different, or more descriptive label keywords than the ones we used for evaluation [8, 31].

We showed that using larger PLMs yield better results for 0SHOT-TC, but this is not always the case for similarity-based approaches. Therefore, unsupervised text classification using smaller PLMs can be conducted in order to benefit from faster inference without necessarily sacrificing much performance in terms of F_1-score.

Our evaluation shows that simple approaches such as LSA or Word2Vec are easy to outperform and therefore are not recommended to be used as baselines for text classification of unseen classes. However, our proposed SimCSE and SBERT baseline approaches generate strong unsupervised text classification results, outperforming even some more advanced classifiers. Therefore, we propose to use SimCSE and SBERT baselines for evaluating unsupervised text classification approaches and 0SHOT-TC performance on unseen classes in future work.

Lbl2TransformerVec, our proposed similarity-based text classification approach yields best F_1-scores for almost all datasets. This is largely due to the great text-similarity characteristics of SimCSE and SBERT representations. Therefore, we believe that future unsupervised text classification work will benefit considerably from enhanced text embedding representations.

REFERENCES

[1] Daniel Braun, Oleksandra Klymenko, Tim Schopf, Yusuf Kaan Akan, and Florian Matthes. 2021. The Language of Engineering: Training a Domain-Specific Word Embedding Model for Engineering. In *2021 3rd International Conference on Management Science and Industrial Engineering* (Osaka, Japan) *(MSIE 2021)*. Association for Computing Machinery, New York, NY, USA, 8–12. https://doi.org/10.1145/3460824.3460826

[2] Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of Semantic Representation: Dataless Classification. In *AAAI*. 830–835. https://www.aaai.org/Library/AAAI/2008/aaai08-132.php

[3] Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. 2015. Dataless Text Classification with Descriptive LDA. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 2224–2231. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9524

[4] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* 41 (1990), 391–407. https://cis.temple.edu/~vasilis/Courses/CIS750/Papers/deerwester90indexing_9.pdf

[5] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (Hyderabad, India) *(IJCAI'07)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1606–1611. https://www.ijcai.org/Proceedings/07/Papers/259.pdf

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[8] 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*. Association for Computational Linguistics, Florence, Italy, 371–379. https://doi.org/10.18653/v1/P19-1036

[9] 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*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3202–3213. https://doi.org/10.18653/v1/2020.coling-main.285

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *CoRR* abs/2006.03654 (2020). arXiv:2006.03654 https://arxiv.org/abs/2006.03654

[11] Ammar Ismael Kadhim. 2019. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review* 52, 1 (Jan. 2019), 273–292. https://doi.org/10.1007/s10462-018-09677-1

[12] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text Classification Algorithms: A Survey. *Information* 10, 4 (2019). https://doi.org/10.3390/info10040150

[13] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*. 331–339. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.6286

[14] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 1188–1196. https://proceedings.mlr.press/v32/le14.html

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[16] Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia Sycara. 2016. Joint Embedding of Hierarchical Categories and Entities for Concept Categorization and Dataless Classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 2678–2688. https://aclanthology.org/C16-1252

[17] Tengfei Liu, Yongli Hu, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2021. Zero-Shot Text Classification with Semantically Extended Graph Convolutional Network. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 8352–8359. https://doi.org/10.1109/ICPR48806.2021.9411914

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[19] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning–Based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* 54, 3, Article 62 (apr 2021), 40 pages. https://doi.org/10.1145/3439726

[20] Marcin Michał Mirończuk and Jarosław Protasiewicz. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications* 106 (2018), 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

[21] Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. All-in Text: Learning Document, Label, and Word Representations Jointly. *AAAI Conference on Artificial Intelligence* (2016). https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12058

[22] Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. 2020. Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability. *bioRxiv* (2020). https://doi.org/10.1101/2020.05.12.077776

[23] Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does Putting a Linguist in the Loop Improve NLU Data Collection?. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 4886–4901. https://doi.org/10.18653/v1/2021.findings-emnlp.421

[24] Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train Once, Test Anywhere: Zero-Shot Learning for Text Classification. *ArXiv* abs/1712.05972 (2017). https://arxiv.org/abs/1712.05972

[25] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019). https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[26] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[27] Anthony Rios and Ramakanth Kavuluru. 2018. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3132–3142. https://doi.org/10.18653/v1/D18-1352

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019). https://arxiv.org/abs/1910.01108

[29] Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Mencia, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *Proc. European Sympo- sium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. https://www.esann.org/sites/default/files/proceedings/legacy/es2016-174.pdf

[30] Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A Decade of Knowledge Graphs in Natural Language Processing: A Survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Online only, 601–614. https://aclanthology.org/2022.aacl-main.46

[31] 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,*. INSTICC, SciTePress, 124–132. https://doi.org/10.5220/0010710300003058

[32] Tim Schopf, Simon Klimek, and Florian Matthes. 2022. PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*. INSTICC, SciTePress, 243–248. https://doi.org/10.5220/0011546600003335

[33] Tim Schopf, Peter Weinberger, Thomas Kinkeldei, and Florian Matthes. 2022. Towards Bilingual Word Embedding Models for Engineering: Evaluating Semantic Linking Capabilities of Engineering-Specific Word Embeddings Across Languages. In *2022 4th International Conference on Management Science and Industrial Engineering (MSIE)* (Chiang Mai, Thailand) *(MSIE 2022)*. Association for Computing Machinery, New York, NY, USA, 407–413. https://doi.org/10.1145/3535782.3535835

[34] Yangqiu Song and Dan Roth. 2014. On Dataless Hierarchical Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 28, 1 (Jun. 2014). https://ojs.aaai.org/index.php/AAAI/article/view/8938

[35] 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* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2901–2907. https://www.ijcai.org/Proceedings/16/Papers/412.pdf

[36] Dominik Stammbach and Elliott Ash. 2021. DocSCAN: Unsupervised Text Classification via Learning from Neighbors. *ArXiv* abs/2105.04024 (2021). https://arxiv.org/abs/2105.04024

[37] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[38] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2, Article 13 (jan 2019), 37 pages. https://doi.org/10.1145/3293318

[39] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[40] Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot Text Classification via Reinforced Self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3014–3024. https://doi.org/10.18653/v1/2020.acl-main.272

[41] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. 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)*. Association for Computational Linguistics, Hong Kong, China, 3914–3923. https://doi.org/10.18653/v1/D19-1404

[42] Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 4913–4922. https://doi.org/10.18653/v1/2021.findings-acl.435

[43] Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating Semantic Knowledge to Tackle Zero-shot Text Classification. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1031–1040. https://doi.org/10.18653/v1/N19-1108

[44] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-Level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 649–657. https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

A APPENDIX

A.1 20Newsgroups Class Summary

Class Name	Label Keywords
alt.atheism	atheism
comp.graphics	computer, graphics
comp.os.ms-windows.misc	computer, os, microsoft, windows
comp.sys.ibm.pc.hardware	computer, system, ibm, pc, hardware
comp.sys.mac.hardware	computer, system, mac, hardware
comp.windows.x	computer, windows
misc.forsale	forsale
rec.autos	cars
rec.motorcycles	motorcycles
rec.sport.baseball	sport, baseball
rec.sport.hockey	sport, hockey
sci.crypt	encryption
sci.electronics	electronics
sci.med	medical
sci.space	space
soc.religion.christian	religion, christianity
talk.politics.guns	politics, guns
talk.politics.mideast	politics, arab
talk.politics.misc	politics
talk.religion.misc	religion

Table 4: 20Newsgroups class names and inferred label keywords.

A.2 AG's Corpus Class Summary

Class Name	Label Keywords
World	government
Sports	sports
Business	business
Science/Technology	science, technology

Table 5: AG's Corpus class names and inferred label keywords.

A.3 Yahoo! Answers Class Summary

Class Name	Label Keywords
Society & Culture	society, culture
Science & Mathematics	science, mathematics
Health	health
Education & Reference	education, reference
Computers & Internet	computers, internet
Sports	sports
Business & Finance	business, finance
Entertainment & Music	entertainment, music
Family & Relationships	family, relationships
Politics & Government	politics, government

Table 6: Yahoo! Answers class names and inferred label keywords.

A.4 Medical Abstracts Class Summary

Class Name	Label Keywords
Neoplasms	neoplasms
Digestive system diseases	intestine, system, diseases
Nervous system diseases	nervous, system, diseases
Cardiovascular diseases	cardiovascular, diseases
General pathological conditions	general, pathological, conditions

Table 7: Medical Abstracts class names and inferred label keywords.

A.5 DensMAP Dataset Visualizations

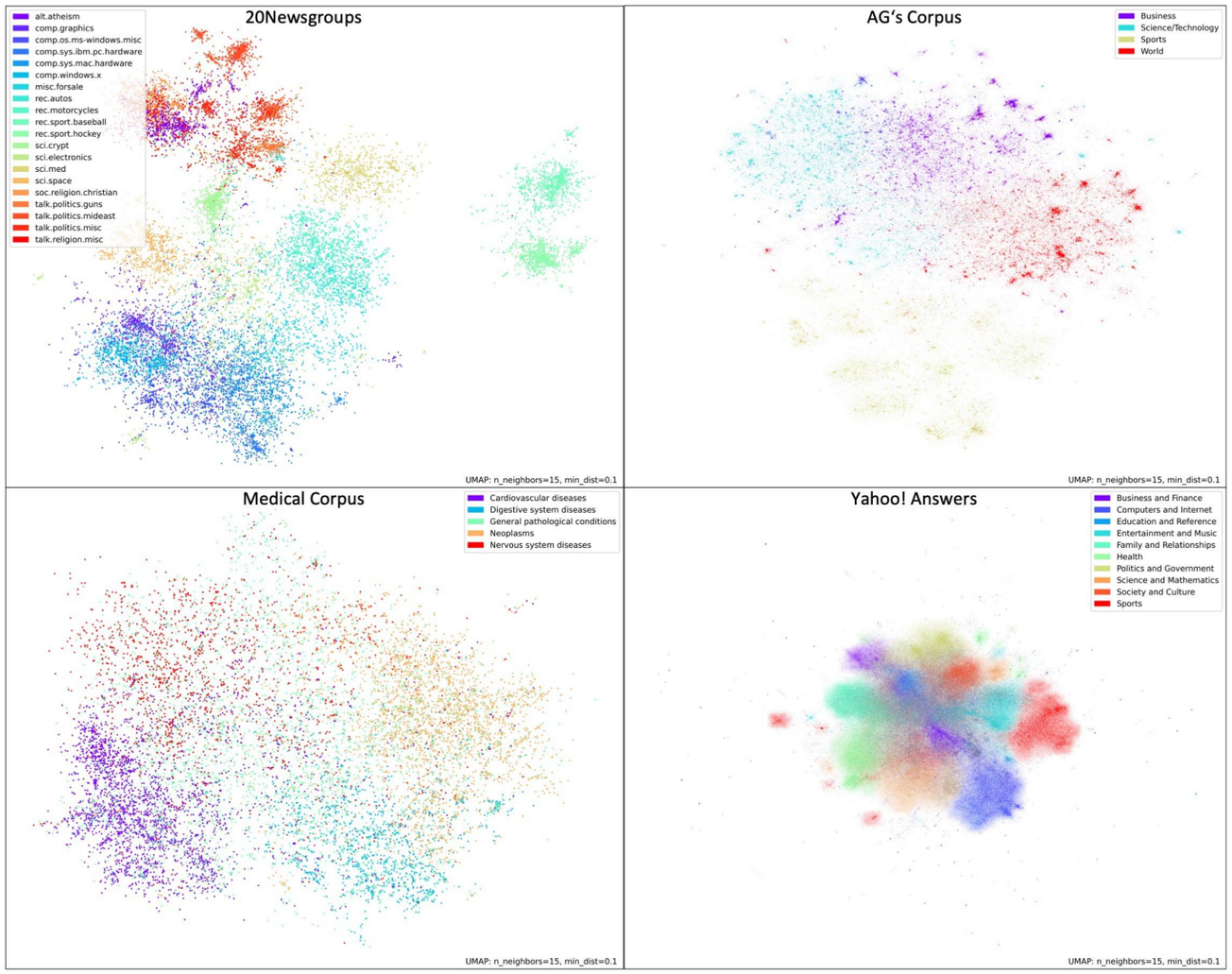


Figure 2: DensMAP visualizations of the document representations for each dataset described in Section 4. The document representations were created by applying the average paragraph embedding strategy described in Section 3.1 using SBERT (*all-mpnet-base-v2*).