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 [1] and SBERT-based [2] 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

Permission to make digital or hard copies of all or part of this work for personal or  
classroom use is granted without fee provided that copies are not made or distributed  
for profit or commercial advantage and that copies bear this notice and the full citation  
on the first page. Copyrights for components of this work owned by others than the  
author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or  
republish, to post on servers or to redistribute to lists, requires prior specific permission  
and/or a fee. Request permissions from permissions@acm.org.

NLPIR 2022, December 16–18, 2022, Bangkok, Thailand
© 2022 Copyright held by the owner/authors(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9762-9/22/12 ... $15.00
https://doi.org/10.1145/3582768.3582795

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 [3–6] 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 [7–10]. 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 [11]. 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 [12–16]. 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 [1] and SBERT [2] embedding similarities. Previous work has mostly been evaluated against different weak baselines such as Word2Vec [17] 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 [10, 18], one of the most recent and well-performing similarity-based methods for unsupervised text classification, to be used with transformer-based language models.

2 RELATED WORK

Chang et al. [19] investigated unsupervised text classification under the umbrella name `Dataless Classification` in one of their earliest works. They used Explicit Semantic Analysis (ESA) [20] 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 [21–24]. 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 [7, 25–27].

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. [8] 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. [9] proposed to enrich label descriptions with expert keywords and subsequently conduct unsupervised classification based on Latent Semantic Analysis (LSA) [28] similarities. Stammbach and Ash [29] introduced DocSCAN, which produces semantic representations of text documents and uses Semantic Clustering by Adopting Nearest-Neighbors for unsupervised text classification. Schopf et al. [10] used Doc2Vec [30] to jointly embed word, document, and label vectors for subsequent similarity-based unsupervised text classification.

Similarly, Nam et al. [31] 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. [14] 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. [15] proposed to treat 0SHOT-TC as a textual entailment problem, while Ye et al. [32] 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 [28]. 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. [9] for unsupervised text classification.

Word2Vec: This produces semantic vector representations of words based on surrounding context words [17]. 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. [15] and Ye et al. [32] 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 [1]. 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 [33] that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings [2]. We use the same classification approach as described in the paragraph above, except that we now use SBERT embeddings instead of SimCSE embeddings.

3.2 Similarity-based Text Classification

As previously stated, numerous similarity-based approaches for unsupervised text classification exist. However, the recently introduced Lbl2Vec approach [10] 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 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
Algorithm 1 Split text document into paragraphs

Require:
\[ d = \text{text document} \]
\[ m_k = \text{max input sequence length of transformer-model } k \]
\[ \text{len}(x) = \text{numen of words in text } x \]

procedure SPLIT-DOCUMENT(d, \( m_k \))

\[ \text{sentences}_d \leftarrow \text{sentence_tokenize}(d) \]
\[ \text{paragraphs}_d \leftarrow \emptyset \]
\[ p \leftarrow \emptyset \]

for \ s \ in \ sentences \\
if \( \text{len}(p) + \text{len}(s) < \frac{m_k}{2} \) then \\
\[ p \leftarrow p + s \]
else \\
\[ \text{paragraphs}_d \leftarrow \text{paragraphs}_d + p \]
\[ p \leftarrow \emptyset \]

return \text{paragraphs}_d

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 [1] and SBERT [2] 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. [16] exist, preliminary experiments confirmed that the zero-shot entailment approach [15] 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 [34] 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 [35] 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 dataset is a common text classification benchmark dataset. It was introduced by Lang [36] 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 dataset is a collection of over 1 million news articles on different topics. The Zhang et al. [37] 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. [37] 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.

\[\text{groups.di.unipi.it/~gulli/AG_corpus_of_news_articles}\]

\[\text{qwone.com/jason/20Newsgroups}\]
4.4 Medical Abstracts

We obtained the raw Medical Abstracts dataset through Kaggle. 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.

| Class Name             | #training | #test  |
|------------------------|-----------|--------|
| Neoplasms              | 2530      | 633    |
| Digestive system diseases | 1195  | 299    |
| Nervous system diseases | 1540    | 385    |
| Cardiovascular diseases | 2441     | 610    |
| General pathological conditions | 3844 | 961    |
| **Σ**                  | **11550** | **2888** |

Table 1: Class distributions within the Medical Abstracts dataset.

The inferred label keywords for each class are summarized in Appendix A.4. We make this corpus available under the Creative Commons CC BY-SA 3.0 license 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 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 and all-MiniLM-L6-v2, 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 [38] model, trained on the MultiNLI [39], Fever-NLI [40], LingNLI [41], and DocNLI [42] datasets, a large BART [43] model trained on the MultiNLI dataset, and a smaller DistilBERT [44] model trained on the MultiNLI dataset. For TARS experiments, we use the BERT-based pretrained tars-base-v8 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 under-represented 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.

6 EVALUATION

Table 2 shows the performances of unsupervised text classification approaches for each dataset, measured in F₁-scores. We can observe that none of the baselines achieves the highest F₁-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 F₁-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 F₁-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

| Baselines | 20Newsgroups | AG’s Corpus | Yahoo! Answers | Medical Corpus |
|-----------|--------------|-------------|----------------|----------------|
| 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          |
| 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          |
| 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          |

Table 2: F₁-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 F₁, Precision, and Recall scores respectively.
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.

Figure 1 shows a more detailed view of the classification results by visualizing the F1-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 cases, however, Lbl2Vec clearly outperforms F1-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)" and "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.

| 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 $Y =$ F1-scores of each class in all four benchmark datasets.

To test our fourth hypothesis, we perform a correlation analysis measuring monotonic relationships between the F1-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. However, text classification can be seen in a broader context where aspects such as emotion or situation are predicted as well [15]. 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. DenseMap [45] visualizations of document representations in embedding space used for classification in this work are shown in Appendix A.5 in Figure 2. 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 [9, 10].

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 F1-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.

Lb2TransformerVec, our proposed similarity-based text classification approach yields best F1-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] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552.
[2] Wilco Beimers and Iryna Gurevych. 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), pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.
[3] Marcin Michal Mirończuk and Jaroslav Protasiewicz. A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106:36–54, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.03.058. URL https://www.sciencedirect.com/science/article/pii/S095741741830215X.
[4] Ammar Ismael Kadhim. Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1):273–292, January 2019. doi: 10.1007/s10462-018-9677-1. URL https://doi.org/10.1007/s10462-018-9677-1.
[5] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. Information, 10(4), 2019. ISSN 2078-2489. doi: 10.3390/info10040150. URL https://www.mdpi.com/2078-2489/10/4/150.
[6] Shervin Minaei, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlou, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. ACM Comput. Surv. 54(3), apr 2021. ISSN 0360-0300. doi: 10.1145/3439726. URL https://doi.org/10.1145/3439726.
[7] Yangguo Song and Dan Roth. 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/9956.
[8] Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Mencia, and Johannes Furnkranz. Using semantic similarity for multi-label zero-shot classification of text documents. In Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2016. URL https://www.esann.org/sites/default/files/proceedings/legacy/es2016-174.pdf.
[9] Zied Haj-Yahia, Adrien Sieg, and Léa A. Deleris. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1036. URL https://aclanthology.org/P19-1036.
[10] Tom Schopf, Daniel Braun, and Florian Matthes. 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. INSTEC, SciTePress, 2021. ISBN 978-989-758-536-4. doi: 10.5220/0010300000000148.
[11] Wei Wang, Vincent W. Zheng, Han Yu, and Chunya Miao. A survey of zero-shot learning: Settings, methods, and applications. ACM Trans. Intell. Syst. Technol., 10(2), Jan 2019. ISSN 2157-6904. doi: 10.1145/3293938. URL https://doi.org/10.1145/3293938.
[12] Pushpakanth Kumar Pushp and Muktabh Mayank Srivastava. Train once, test anywhere: Zero-shot learning for text classification. ArXiv, abs/1712.05972, 2017. URL https://arxiv.org/abs/1712.05972.
[13] Anthony Biao and Ramakant Kavuluru. 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, pages 3132–3142, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1352. URL https://aclanthology.org/D18-1352.
[14] Jingqing Zhang, Piayaw Lertvittayakumjorn, and Yike Guo. 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), pages 1031–1040, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1108. URL https://aclanthology.org/N19-1108.
[15] Wengpeng Yin, Jamaa Hay, and Dan Roth. 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), pages 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404. URL https://aclanthology.org/D19-1404.
[16] Tengfri Liu, Yongli Hu, Junbin Gao, Yanfeng Sun, and Baocai Yin. Zero-shot text classification with semantically extended graph convolutional network. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 8352–8359, 2021. doi: 10.1109/ICPR48806.2021.9419114.
[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In
## 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       | general, pathological, |
| conditions                 | 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).