Few-shot Learning Approach for Arabic Scholarly Paper Classification using SetFit Framework

ALZAMEL KHALED, ALAJMI MANAYER
Department of Computer Engineering,
Kuwait University,
Shadadiya,
KUWAIT

Abstract: Focus on the few-shot approach has increased recently for TC as it is competitive with fine-tuning models that need a large dataset [14]. In NLP, the process of using PTMs to classify new data is preferable to the expensive process of training a model from scratch. This can be considered a kind of TL, i.e., it focuses on reusing knowledge of PTMs to solve different problems, as long as the pre-training data is appropriately comparable. Transferring knowledge allows the model to circumvent the lack of data and enable FSL as a low-cost solution. To clarify, the term shot refers to a single example that is used for training, and the number of examples available for training is equal to N in N-shot learning. The focus of this study is on few-shot classification, which involves distinguishing between N classes using K examples of each. In this approach, N-way-K shot classification implies that each task involves N classes with K examples. In FSL, the model is able to predict a new class based on a few new examples [11] by transferring knowledge and contrasting examples. Such contrastive learning [5] has shown its effectiveness in different studies of various NLP tasks [20]. However, as far as we know, no previous studies have applied contrastive learning to standard Arabic for multi-class classification. This study aims to apply few-shot learning using a Siamese Network-based model(SN-XLM-RoBERTa [6]) to classify MSA texts in predefined classes labelled with the most common ministries' names. For this study, we extracted a new dataset from an AI-powered research tool. The model was fine-tuned by K examples per class. We experimented with various K values, including 10, 20, 50, 100, and 200. The results show that the accuracy in distinguishing between 6 classes using 200 examples of each is 91.076%. Moreover, the results indicated that employing few-shot learning, as in SN-XLM-RoBERTa, in classifying MSA texts can be a promising solution in case of an insufficient dataset or uncertain labelling. Few-Shot Learning (FSL) may contribute to the research domain by automating the classification process.

Keywords: BERT; contrastive learning; document classification; few-shot learning; sentence transformer; transfer learning

Received: April 16, 2024. Revised: October 17, 2024. Accepted: November 21, 2024. Published: December 27, 2024.

1. Introduction

In few-shot learning, the model is able to predict a new class based on a few new examples [11]. Traditionally, to add new classes to a dataset, we must update the model with sufficient data, in order to enhance performance. Recently, there has been an increased focus on the few-shot approach for text classification as it is competitive with fine-tuning models that need a large dataset [14]. Few-shot learning facilitates quick and effective generalization of new examples of the same class by learning similarities and differences between the input examples. Applying few-shot learning for multi-class text classification of standard Arabic may show promising results, as Arabic is not considered a high-level resource language. Multilingual models facilitate the concept of cross-lingual transfer learning, which permits the transfer of knowledge from higher-level resource languages to lower-level resource languages.

Transfer learning can be utilized in metric learning using a multilingual model. This will allow us to use cross-lingual transfer learning to classify texts. Metric learning can tackle the issue of few-shot classification by enabling a model, using a Siamese architecture (see Fig.1), to learn by comparing inputs [7]. A SNN only requires a few examples to predict with accuracy by comparing similarities between input feature vectors. SNNs are very popular in signature verification and face recognition or in the case of a low-resource dataset. This concept enables models to infer the meaning of a text in different languages, as cross-lingual learning can map distinct embeddings with a similar meaning to very close representations.

Sentence Transformer (ST), which is based on SNN, is a common approach to detecting semantic similarity. It produces a unique vector during contrastive learning by adapting a transformer model in a Siamese architecture to minimize the distance between similar sentences and maximize the distance between dissimilar sentences [12]. One popular transformer-based model that has represented a revolution in NLP is BERT. BERT has shown state-of-the-art results in different NLP tasks [4]. However, studies focusing on applying BERT to Arabic TC are still limited [1].

In our approach, fine-tuning was applied to the SN-XLM-RoBERTa to evaluate contrastive learning on classifying Arabic summaries of scholarly papers into ministries. This can be done by using the Siamese network-based model, SN-XLM-RoBERTa, to apply few-shot learning. In this study, we performed Arabic text classification with the following HuggingFace transformer model (see Table 1). We are there-

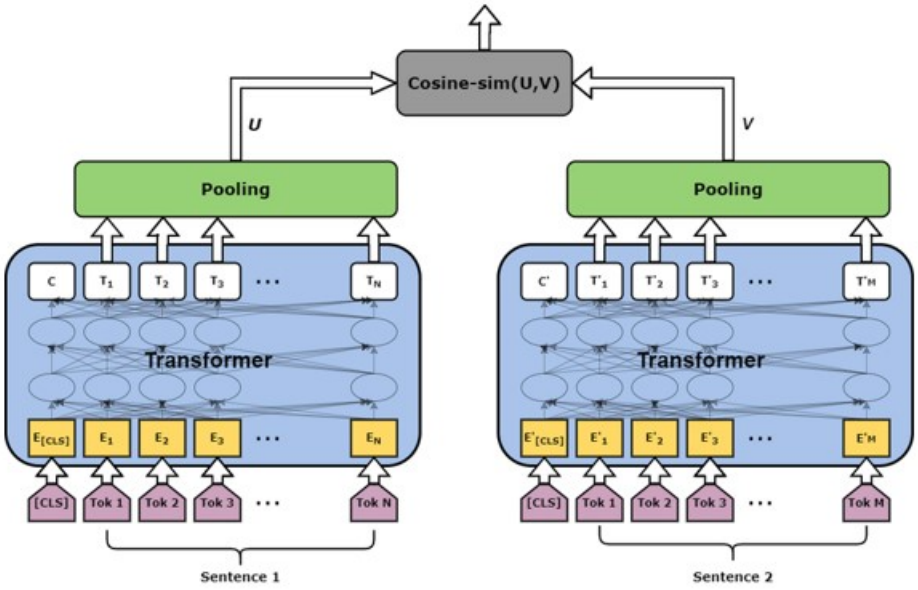

Figure 1. The Architecture of a Siamese Sentence Transformer [17].

Table 1. HuggingFace transformer models used in this study.

Model as referred to in this study	Model name on Hugging Face
SN-XLM-RoBERTa	"symanto/sn-xlm-roberta-base-snli-mnli-anli-xnli"

fore confronted with proving the suitability of exploiting FSL in classifying Arabic summaries of scholarly papers.

For this study, we extracted a new dataset from an AI - powered research tool. The new dataset has a total number of 5996 MSA summaries of scholarly papers and is categorized into six different groups of approximately the same size, which are: Islamic Affairs, Commerce and Industry, Health, Justice, Electricity and Water, and Education. Each label represents a ministry name, as each text's content relates to one ministry's affairs. This study aims to provide scientific assistance to the ministries in Kuwait to develop a work plan derived from different Arabic scholarly papers. In this study, we focus on classifying different scholarly papers, in order to forward them to the target label.

The structure of the work is as follows: Section 2 describes literature reviews, which include different research areas related to this study. Section 3 introduces the dataset-gathering process and a description of the distribution of Arabic texts per class, along with the methods used to fine-tune the selected models. Section 4 shows the obtained results. Further, it provides an analysis and discussion of the results. Finally, Section 5 describes conclusions drawn from the findings, as well as the scope for future work.

2. Background

NLP researchers are focused on improving the ability of machines to understand and solve problems, as well as to learn to do complex tasks. Transformers represent a revolution in NLP by succeeding in replacing the existing idea of using recurrency in sequence-to-sequence models [16]. The transformer family utilizes either the encoder, decoder, or both to understand language or generate text. Encoder-based models aim to encode input sequences into contextualized representations that can be utilized for various tasks. These tasks may require a comprehensive understanding of the entire sentence, such as text classification, named entity recognition, sentiment analysis and extractive question answering (AKA machine reading comprehension). Some examples of models that utilize this architecture include BERT, DistilBert [13], and RoBERTa [8]. In this respect, the advent of BERT, a model that represented a revolution in NLP, has shown outstanding performance in various NLP tasks [4].

Reimers and his colleagues presented Sentence-BERT [12], which uses contrastive learning to derive meaningful sentence embeddings that can be compared using cosine similarity (see Fig.2).

A new approach, called SETFIT, was introduced by Intel Labs, UKP lab, and Hugging Face [14]. SETFIT stands for Sentence Transformer Fine-tuning. It is a prompt-free framework for fine-tuning sentence transformers for few-shot learning. It permits the use of a pre-trained sentence transformer model to support NLP tasks. Using a Sentence Transformer (ST), which is based on SNN is a powerful way to facilitate quick and effective generalization of new examples.

2.1 Sn-xlm-Roberta-base-SNLI-MNLI-ANLI-XNLI

SN-XLM-RoBERTa is a model that uses the Siamese network and is based on the XLM-RoBERTa Base developed by HuggingFace [6]. Its function is to convert sentences and paragraphs into a 768-length vector. This mapping makes it a

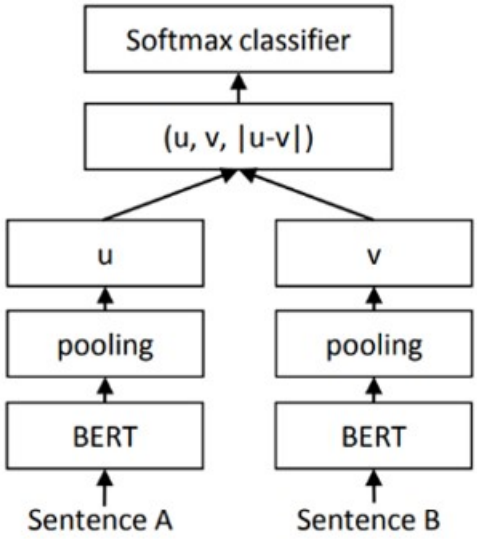

Figure 2. Sentence-BERT is a very common architecture used for Semantic similarity search tasks based on a Siamese architecture [12]

.

sentence transformer model that can be used in the SETFIT environment. SN-XLM-RoBERTa supports the Arabic language as it is based on XLM-RoBERTa, which is a cross-lingual language model. It was trained on SNLI, MNLI, ANLI, and XNLI.

The acronym SNLI stands for the Stanford-Natural-Language-Inference corpus. It includes 570,000 pairs of English sentences written by humans. Each pair is manually classified as either entailment, contradiction, or neutral [2]. MNLI is the Multi-Genre Natural Language Inference corpus, containing 433K sentence pairs with textual entailment information. It offers different genres of written and spoken English, making it more challenging and extensive than the SNLI corpus it's derived from. These genres include fiction, letters, and telephone speech. This variation of sources in the corpus makes it suitable to evaluate models on the full breadth of the highly complex English language [19]. ANLI refers to the Adversarial Natural Language Inference, a corpus of texts collected via an iterative procedure where humans and the model check predictions of sentence relationships (as entailment, contradiction, or neutral) against each other. ANLI has over 103K examples [10]. The XNLI corpus is an extension of the English-only MNLI and ANLI corpora, including the translation of sentence pairs into 14 additional languages, resulting in 112.5 annotated pairs in 15 languages [3]. The purpose behind its creation was to solve the problem of evaluating a model that was trained on a high-level resource language to predict accurate labels of entailment, contradiction, or neutral in other languages at test time.

3. Methodolgy

In this section, we will discuss the fine-tuned transformer-based encoder model on our dataset. Fine-tuning was applied to a sentence transformer called SN-XLM-RoBERTa. In this study, we performed Arabic text classification with the following HuggingFace transformer models (see Table 1).

3.1 Dataset

In this study, we want to implement a classification task that can be used to classify MSA summaries from scholarly papers into predefined categories. We will use six categories labeled with the most common ministries' names: Islamic Affairs, Commerce and Industry, Health, Justice, Electricity and Water, and Education. The purpose is to provide ministries with relevant research papers. The first step in data discovery is understanding the problem and finding a relevant dataset. In this study, we scraped textual data from Semantic Scholar, a research tool that utilizes AI to analyze scientific literature. Every summary was labeled manually, considering the responsibilities and visions of each respective ministry. Table 2 shows the number of scraped Arabic summaries per label.

3.2 Fine-tuning Methodology

In our approach, we involved SETFIT to take advantage of employing sentence transformers for Arabic text classification. The following subsection explains the training process of the SETFIT method.

3.3 Setfit Approach

SETFIT is based on sentence transformers [12]. It was made available by a group of researchers led by Intel Lab, UKP Lab, and Hugging Face [14]. A ST outputs a dense vector that represents textual data. However, Reimers and his colleagues proposed sentence transformers that apply contrastive learning to create sentence embeddings with semantic meaning [12]. Each unique sentence embedding is a contextualized representation of an input sequence of data. In essence, they work by minimizing the distance between pairs of semantically similar sentences and maximizing the distance between dissimilar ones. SETFIT is a new few-shot text classification approach.

The training process of the SETFIT library includes two stages. First, a pre-trained ST is fine-tuned on a small number of text pairs using a contrastive Siamese architecture. The trained model then generates text embeddings, and these embeddings are used to train a classification head (see Fig.3). A noticeable advance in accuracy was achieved using SETFIT. By using only eight labeled examples of the customer review dataset, the accuracy is competitive with a three-times-larger model, which was fine-tuned on 3000 examples [14].

Table 2. The number of scraped summaries per label.

English Label	of summaries
Islamic Affairs	999
Commerce and Industry	1000
Health	1003
Justice	1003
Electricity and Water	989
Education	1002

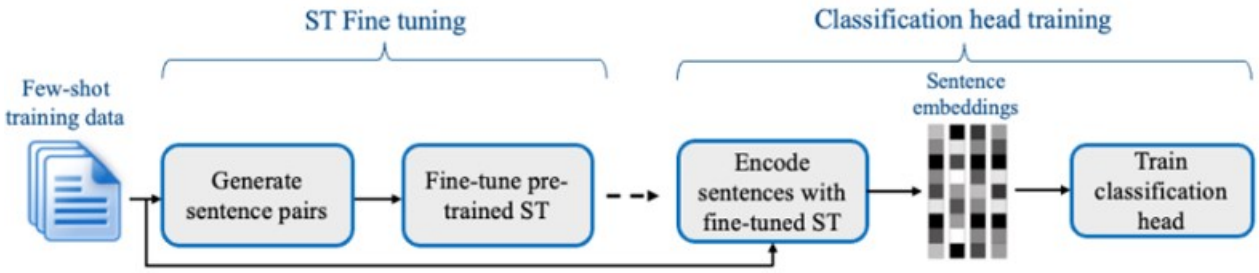

Figure 3. SETFIT's process of two-stage training [14].

Table 3. Accuracy results for few-shot learning using SN-XLM-RoBERTa.

Classification Task	Metric	# of examples per class	Sn-XLM-R
Multi-class Arabic texts	Acc.	10	85.905%
		20	85.488%
		50	87.573%
		100	89.992%
		200	91.076%

The SETFIT method's objective is to create data in order to augment labeled data in few-shot settings. This can be done by generating $2 \times N$ training pairs in each sentence-pair iteration, where N is the total number of training samples per task. The pseudo code of sentence pairs generation is depicted in Figure 4. In other words, we want to model a small set of K lablled examples $D = \{(x_i, y_j)\}$, in which x_i and y_j are sentences and their class labels, respectively. For each class label $c \in C$, the authors generate a set of R positive triplets; $T_p^c = \{(x_i, x_j, 1)\}$ (pairs of sentences randomly chosen from the same class c). A set of R negative triplets $T_n^c = \{(x_i, x_j, 0)\}$ (pairs of sentences randomly chosen from different classes) are generated as well. The final objective is to produce T by concatenating the positive and negative triplets across all class labels; $T = \left\{ \left(T_p^0, T_n^0\right), \left(T_p^1, T_n^1\right), \ldots, \left(T_p^{|C|}, T_n^{|C|}\right) \right\}$, where $|C|$ represents the number of class labels, and $|\mathrm{T}| = 2R|C|$ represents the number of pairs, where R is a hyperparameter. To create the augmented data, R is used to indicate the number of times the sampling process should be conducted [14]. After finetuning the sentence transformer in a contrastive way, each training example in the original labeled training data is encoded into a vector, AKA sentence embedding. The dataset for training the classification head includes sentence embeddings and their corresponding class labels. The default model used by the authors as a classification head is a logistic regression model. SETFIT is a publicly available open-source environment that has shown promising results in classification for German, Japanese, Mandarin, French, and Spanish, both in-language and across different languages.[15].

4. Results and Discussion

4.1 Results

This study explores the use of N-way, K-shot classification in few-shot learning. N refers to the number of classes while K represents the number of examples per class. In our approach, the dataset used is categorized with six labels, and the model was fine-tuned by K examples per class. We experimented with various K values, including 10, 20, 50, 100, and 200. The results of the few-shot learning experiment using SN-XLM-RoBERTa are summarized in Table 3.

The results show that using SN-XLM-RoBERTa led to promising results in classifying Arabic texts, as shown in Table 3. This can be effective when dealing with datasets that lack sufficient data or when there is uncertainty in labeling. Using a Sentence Transformer (ST), which is based on SNN, as done in SN-XLM-RoBERTa is a powerful way to learn general features of unlabeled datasets. It performs very well at data-scarce tasks in a specific domain. It can be said that few-shot learning has significant results in classifying MSA texts. Undoubtedly, few-shot learning faces some limitations that can be improved. In this study, SN-XLM-RoBERTa, sentence embeddings are created by using the mean-pooling method. Thus, we can't be sure whether applying other techniques, such as CLS-pooling, would enhance the accuracy of our task or not. Moreover, we suggest using monolingual and bilingual (ST) models to study the impact of them on Arabic as a non-high-level-resource language and compare the results to ours. It might be interesting to fine-tune the ST model using output Arabic texts from ChatGPT **[9]**.

5. Conclusion

This study aims to fine-tune a (ST) model to classify Arabic research papers into pre-defined classes and investigate the accuracy of SN-XLM-RoBERTa to automate and facilitate a TC task for Arabic, using a new dataset. It is categorized into six different ministry names, as each summary's content relates to one ministry's affairs. Specifically, we worked on fine-tuning the pre-trained model SN-XLM-RoBERTa on Arabic text classification. The results indicated that employing few-shot learning, as in SN-XLM-RoBERTa, in classifying MSA texts can be a promising solution in case of an insufficient dataset or uncertain labelling. Few-Shot Learning (FSL)

```
Train.SentencesPairs = []
Train.SentencePairsLabel = []
Repeat #iterations
 for idxA in TrainSentences
   currentSentence = sentences[idxA]
   posSentence = select random sentence from same category of currentSentence
   posPair = [currentSentence, posSentence]
   Train.SentencesPairs. append.(posPair), Train.SentencePairsLabel. append.(True)
   negSentence = select random sentence from distant category of currentSentence
   negPair = [currentSentence, negSentence]
   Train.SentencesPairs. append.(negPair), Train.SentencePairsLabel. append.(False)
```

Figure 4. The pseudo-code of Sentence-pairs generation [18].

may contribute to the research domain by automating the classification process.

References

[1]. Ali Saleh Alammary. 2022. BERT Models for Arabic Text Classification: A Systematic Review. Applied Sciences 12, 11 (2022). DOI:http://dx.doi.org/10.3390/app12115720

[2]. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 632–642. DOI:http://dx.doi.org/10.18653/v1/D15 -1075

[3]. Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2475–2485. DOI:http://dx.doi.org/10.18653/v1 /D18-1269

[4]. 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), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI:http://dx.doi.org/10.18653/v1/N19-1423

[5]. R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2. 1735–1742. DOI: http://dx.doi.org/10.1109/CVPR.2006.100

huggingface.co. 2021. symanto/sn-xlm-roberta-base-snlimnli- anli-xnli. (2021). https://huggingface.co/symanto/sn-x lm-roberta-base-snli-mnli-anli-xnli?doi=true

[6]. Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, and others. 2015. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, Vol. 2. Lille, 1– 30. https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

[7]. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (07 2019). DOI: http://dx.doi.org/10.48550/arXiv.1907.11692

[8]. Brady Lund and Ting Wang. 2023. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? Library Hi Tech News 40 (02 2023). DOI:http://dx.doi.org/10.1108/LHTN-01-2023-0009

[9]. Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A New Benchmark for Natural Language Understanding. (10 2019). https://aclanthology.org/2020.acl-main.441.pdf

[10]. Archit Parnami and Minwoo Lee. 2022. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. (03 2022). DOI:http://dx.doi.org/10.48550/arXiv.2203.04 2 91

[11]. 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 (EMNLPIJCNLP), Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. DOI:http://dx.doi.org/10.18653/v1/D19-1410

[12]. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (2020). https://arxiv.org/abs/19 10.01108

[13]. Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. (2022). https://arxiv.org/abs/2209.11055

[14]. Luke Bates Daniel Korat Oren Pereg MosheWasserblat Unso Eun Seo Jo, Lewis Tunstall. September 26, 2022. SetFit: Efficient Few-Shot Learning Without Prompts — huggingface. co. https://huggingface.co/blog/setfit. (September 26, 2022).

[15]. A Vaswani. 2017. Attention is all you need. Advances in Neural Information Processing Systems (2017). https: //proceedings.neurips.cc/paper_files/paper/2017/file/3f5 ee243547dee91fbd053c1c4a845aa-Paper.pdf

[16]. Xin Wang and Huimin Yang. 2022. MGMSN: Multi- Granularity Matching Model Based on Siamese Neural Network. Frontiers in bioengineering and biotechnology 10, 3 (March 2022), 839586. DOI:http://dx.doi.org/10.3389/fbioe.2022.839586

[17]. Moshe Wasserblat. Dec 14, 2021. Sentence Transformer Fine-Tuning (SetFit): Outperforms GPT-3 on few-shot Text-Classification while. . . — towardsdatascience.com. https: //towardsdatascience.com/sentence-transformer-fine-tunin g-setfit-outperforms-gpt-3-on-few-shot-text-classificat ion-while-d9a3788f0b4e. (Dec 14, 2021).

[18]. Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. (2018). https://arxiv.org/abs/1704.05426

[19]. Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022. Contrastive Data and Learning for Natural Language Processing. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, Miguel Ballesteros, Yulia Tsvetkov, and Cecilia O. Alm (Eds.). Association for Computational Linguistics, Seattle, United States, 39–47. DOI: http://dx.doi.org/10.18653/v1/2022.naacl-tutorials.6

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.