Tonirodriguez at CheckThat!2024: Is it Possible to Use Zero-Shot Cross-Lingual Methods for Subjectivity Detection in Low-Resources Languages?

Notebook for the CheckThat! Lab Task2 at CLEF 2024

Antonio Rodríguez[1,*,†], Elisabet Golobardes[2,‡] and Jaume Suau[3,§]

[1] *La Salle Engineering, Universitat Ramon Llull, Barcelona, Spain*
[2] *La Salle Engineering, Universitat Ramon Llull, Barcelona, Spain*
[3] *Blanquerna, Universitat Ramon Llull, Barcelona, Spain*

Abstract

Subjectivity detection is a key task within natural language processing due to the challenges generated by new forms of journalism, the proliferation of misinformation and fake news, and existing concerns about the quality and integrity of journalism. Although subjectivity detection is an existing challenge in all languages, the amount of resources available to build these types of applications varies greatly among languages. In this paper, we present our participation in the CLEF2024 CheckThat! Lab Task2 [1], where we have attempted to apply Zero-Shot Cross-Lingual transfer techniques using the datasets for the five languages provided in Task2 (English, German, Italian, Bulgarian, and Arabic). For this, we have fine-tuned two multilingual models, mDeBERTa v3 and XLM-RoBERTa, on a subset of the dataset consisting of three of the languages provided in Task2, specifically English, German, and Italian, and we have applied Zero-Shot Cross-Lingual transfer to the other two languages available in Task2, Arabic and Bulgarian.

Keywords

Subjectivity Detection, Natural Language Processing, Fake News, Journalism, Misinformation, Transformers, Cross-lingual, Transfer Learning

1. Introduction

Currently, the proliferation of news sites and the widespread use of social networks have revolutionized the way news is consumed, giving rise to new forms of journalism [2]. However, these changes have introduced several challenges, including the proliferation of misinformation and fake news, the formation of "echo chambers" where news consumers limit their exposure to different points of view, and emerging concerns about the quality and integrity of journalism [3]. A common element in many of the identified challenges is the need to distinguish whether a news author is sharing objective information or expressing their own opinions, desires, or biases [4] [5]. The goal of Subjectivity Detection (SD) is to develop computational systems capable of implementing a binary classifier that can determine whether a text is objective or subjective.

CLEF2024 CheckThat! Lab Task2 [1] provides an opportunity to work on the challenges associated with subjectivity detection. This task aims to construct a binary classifier that can identify whether a text sequence, in the form of a sentence, is subjective or objective [6]. For the execution of Task2, the organizers have published five datasets in different languages (English, German, Italian, Bulgarian, and Arabic), plus an additional dataset that combines the previous five languages for the multilingual version of the task. The evaluation of the results presented will be carried out through the macro-averaged F1 between the two classes.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

arodriguez.delatorre@salle.url.edu (A. Rodríguez); elisabet.golobardes@salle.url.edu (E. Golobardes); jsuau@blanquerna.url.edu (J. Suau)

0000-0003-4654-0443 (A. Rodríguez); 0000-0001-8080-0769 (E. Golobardes); 0000-0003-4480-4441 (J. Suau)

This paper begins with the “Related Work” section, where a comprehensive review of previous research and studies relevant to the topic is conducted. This is followed by the “Data” section, which provides a detailed description of the structure and characteristics of the dataset provided for the Task2. The “Approach” section outlines the phases and techniques employed to conduct the research. In the “Results” section, the findings obtained from the implementation of the models used are presented and analysed using the metric macro F1. Finally, the “Conclusions” section provides a summary of the results, discusses the implications of the research, and suggests possible directions for future research.

2. Related Work

According to Liu[7], Subjectivity Detection (SD) is a field of study traditionally encompassed within a broader field known as Sentiment Analysis (SA) also referred to as opinion mining. Sentiment analysis is the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Sentiment Analysis is an area of research deeply studied in the last two decades.

Chaturvedi [8] categorizes methods for subjectivity detection into two main types: traditional syntax-centered NLP methods and semantics-based NLP approaches. Syntax-centered NLP can be broadly divided into three main categories: keyword spotting, lexical affinity, and statistical methods. The major issue with these methods is that they are highly language-specific and require the existence of databases and resources for each language in which they are to be applied. To address this issue, solutions such as translating content between languages lacking these resources and languages like English, which have a wealth of resources, have been adopted. However, the translation of sentences can lead to the loss of lexical information, such as word sense, resulting in low accuracy [8].

On the other hand, semantic methods based on embeddings, RNNs, Convolutional Networks, and Transformers have gained significant relevance recently. They offer more accurate results than methods based on syntactic features, but they present their own challenges, as they require large datasets for each language in which we want to work. The creation of these datasets is complex and can generate problems such as ambiguity when classifying sentences [8] or annotator bias [9]. To avoid these problems, a recent paper published by F. Antici et al. [10] proposes annotation guidelines with the aim of unifying criteria and avoiding previous problems while experimenting with monolingual, multilingual, and cross-lingual Transformers scenarios between English and Italian languages.

Schumacher [11], starting with a multilingual BERT model, achieves good results for cross-language entity linking. From there, he explores Zero-Shot Cross-Lingual transfer between different languages and obtains robust results with a slight degradation when the model is applied to a language for which fine-tuning has not been performed. He concludes that although multilingual Transformer models make a good transfer between languages, issues remain in disambiguating similar entities unseen in training.

The objective of this paper is to address the question of the viability of using Zero-Shot Cross-Lingual transfer for subjectivity detection. To this end, we will fine-tune two multilingual Transformer models and analyze the results obtained within the framework of the CLEF2024 CheckThat! Lab Task2 [1]. To achive this goal, we will employ DeBERTa [12, 13] and RoBERTa [14] for the monolingual approach and their multilingual versions, MDeBERTa [12, 13] and XLM-RoBERTa [14], respectively for the multilingual approach. These models are evolutions built upon BERT that significantly enhance the results achieved by multilingual BERT, particularly in low-resource languages [15].

3. Data

The six datasets provided for the execution of Task2 exhibit varying characteristics in terms of size and distribution of objective and subjective sentences. In all datasets, objective sentences are labeled with the tag “OBJ”, while subjective sentences are labeled as “SUBJ”. As shown in Table 1, the Bulgarian dataset, which is the smallest, comprises a total of 1043 texts, 729 of which are included in the training dataset. In contrast, the Italian dataset contains a total of 2280 sentences, 1613 of which are in the training dataset. Furthermore, an examination of the datasets reveals a distribution bias in favour of the "OBJ" class across all datasets, although the extent of this bias varies depending on the language. For instance, while the bias is only 55.69% in favour of "OBJ" sentences in Bulgarian, this bias increases to 76.32% and 76.37% for Italian and Arabic, respectively.

Table 1
Datasets and Distribution of classes

English:	**Objective**	**Subjective**	**Total**
Train	532 (64.10%)	298 (35.90%)	830
Dev	106 (48.40%)	113 (51.60%)	219
Dev Test	116 (47.74%)	127 (52.26%)	243

Italian:	**Objective**	**Subjective**	**Total**
Train	1231 (76.32%)	382 (23.68%)	1613
Dev	167 (73.57%)	60 (26.43%)	227
Dev Test	323 (73.41%)	117 (26.59%)	440

German:	**Objective**	**Subjective**	**Total**
Train	492 (61.50%)	308 (38.50%)	800
Dev	123 (61.50%)	77 (38.50%)	200
Dev Test	194 (66.67%)	97 (33.33%)	291

Bulgarian:	**Objective**	**Subjective**	**Total**
Train	406 (55.69%)	323 (44.31%)	729
Dev	59 (55.66%)	47 (44.34%)	106
Dev Test	116 (55.77%)	92 (44.23%)	208

Arabic:	**Objective**	**Subjective**	**Total**
Train	905 (76.37%)	280 (23.63%)	1185
Dev	227 (76.43%)	70 (23.57%)	297
Dev Test	363 (81.57%)	82 (18.43%)	445

Multilingual:	**Objective**	**Subjective**	**Total**
Train	3568 (69.16%)	1591 (30.84%)	5159
Dev	250 (50.00%)	250 (50.00%)	500
Dev Test	250 (50.00%)	250 (50.00%)	500

The multilingual dataset, the largest among all, is composed of a subset of sentences provided in each of the other datasets across all subsets (training, validation and test). However, due to its composition, it also exhibits a bias in favour of the "OBJ" class, accounting for 69.16% of the dataset.

4. Approach

In our research, we adopted a dual approach. Initially, we employed a monolingual approach that leveraged Transformers, placing the focus on the English language. Subsequently, we implemented a second phase, utilizing multilingual Transformers with a dual purpose: to enhance the results obtained in the first phase with the monolingual Transformers by increasing the size of the training set, and to verify the Zero-Shot Cross-Lingual transfer capabilities of the model. This means that a model that is fine-tuned in certain languages can be applied to other languages without any specific training.

4.1. Monolingual Models

The primary objective of the monolingual phase was to enhance the results provided by Task2 as a baseline. The baseline is based on a two-step approach. First, Sentence-BERT [16] is used to transform

Table 2
Sample of sentences in English and their classification

sentence	label
The fiscal swing was more like 4% of gdp in the two years from 2008 to 2010.	OBJ
As a result, monetary easing did not translate into a big expansion of credit.	OBJ
A key element in the fraud was crying racism.	SUBJ
Demands upon the public credit for social service are most difficult to resist.	SUBJ

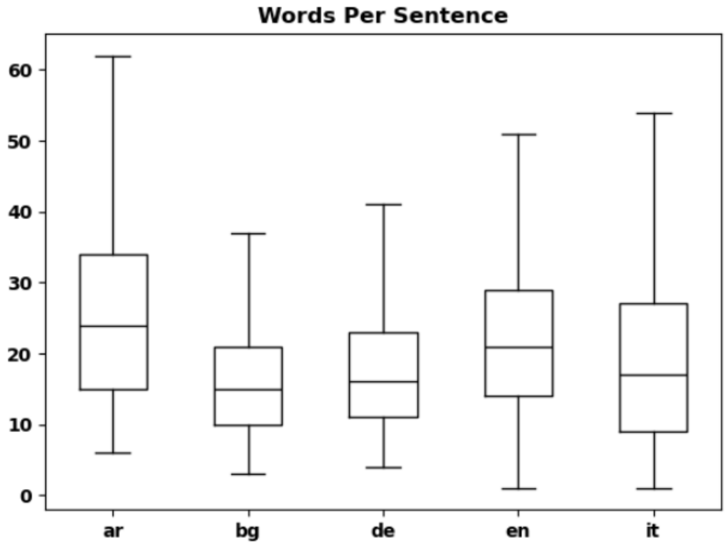

Figure 1: Distribution of the number of words per sentence for each of the languages considered in the Task2.

each sentence into a high-dimensional vector representation capable of capturing its semantic meaning. In the second step, a classifier is constructed by training a Logistic Regression model on the vectors generated in the previous step. To improve the results provided by the baseline, we utilized various Transformers such as DeBERTa v3 Large [12, 13], RoBERTa Large [14] and BART [17] Large MNLI [17] that uses the entailment approach [18].

BART Large MNLI [17] is a Transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder pretrained on English. BART is pre-trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. BART is particularly effective when fine-tuned for text generation tasks (e.g., summarization, translation) but also performs well for comprehension tasks (e.g., text classification, question answering). In this study, we selected the checkpoint for bart-large after it had been trained on the MultiNLI (MNLI) dataset. Yin et al. [18] proposed a method for using pre-trained NLI models as ready-made Zero-Shot sequence classifiers. The method works by posing the sequence to be classified as the NLI premise and constructing a hypothesis from each candidate label.

4.2. Multilingual Models

In the second phase of our study, we utilized multilingual Transformers. Although these models have architectures and training procedures similar to their monolingual counterparts, they differ in that the corpus used for their pretraining consists of documents in many languages. The multilingual transformer models used in this study were MDeBERTa Base and XLM-RoBERTa Base. These models use masked language modeling as a pretraining objective and are trained jointly on texts in over one hundred languages. By pretraining on vast corpora across numerous languages, these multilingual Transformers enable Zero-Shot Cross-Lingual transfer. This implies that a model fine-tuned on one language can be applied to others without any additional training. The characteristics of these models

Table 3
Results of the Monolingual Models trained in EN and applied to the Final Test dataset for EN.

	F1 Macro	SUBJ F1
Baseline EN	0.63	0.45
Deberta V3 Large	0.73	0.60
Roberta Large	**0.74**	0.59
BART Large MNLI	0.69	0.51

are as follows:

MDeBERTa V3 Base: [12, 13] mDeBERTa is multilingual version of DeBERTa which use the same structure as DeBERTa and was trained with CC100 multilingual data. The mDeBERTa V3 base model comes with 12 layers and a hidden size of 768. It has 86M backbone parameters with a vocabulary containing 250K tokens which introduces 190M parameters in the Embedding layer. This model was trained using the 2.5T CC100 data as XLM-R.

XLM-RoBERTa: XLM-RoBERTa is a multilingual version of RoBERTa. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. Following the work of XLM and RoBERTa, the XLM-RoBERTa or XLM-R model takes multilingual pretraining one step further by massively upscaling the training data [19]. Using the Common Crawl corpus, its developers created a dataset with 2.5 terabytes of text; they then trained an encoder with MLM on this dataset. Since the dataset only contains data without parallel texts (i.e., translations), the TLM objective of XLM was dropped. This approach beats XLM and multilingual BERT variants by a large margin, especially on low-resource languages [15].

The objective pursued through this cross-lingual approach is to utilize the same model across different languages, as the resulting linguistic representations can be well generalized across languages for various subsequent tasks, such as classification in our case. To this end, we have fine-tuned the multilingual models in English, German, and Italian, and applied them to the rest of the languages available in Task2, Arabic and Bulgarian.

5. Results

In the initial phase of this research, we focused on the English language, applying fine-tuning to various monolingual models with the aim of achieving optimal results as measured by the macro F1 metric, as outlined in the guidelines for Task2. We selected three distinct Transformer-based models for this purpose: DeBERTa Large, RoBERTa Large, and BART Large MNLI. We used Kaggle as the platform for training these models. The results of this process are presented in Table 3.

The models DeBERTa v3 Large and RoBERTa Large yield very similar results for the English language, with the best result being achieved by RoBERTa Large, scoring 0.74 on the test dataset. A much larger model, BART Large MNLI, which in principle should be capable of a greater understanding of language, performs worse, likely due to the dataset size not allowing it to generalize the characteristics of subjective language. As this model does not have an equivalent multilingual model, we have discarded it for the subsequent phases of the research. In any case, all trained models significantly outperform the baseline result provided for Task2 in English.

In the second phase of the research, we fine-tuned the multilingual models equivalent to the models selected in Phase 1 on a training dataset composed of the union of the data provided in Task2 for English, Italian, and German languages. Given the increased size of the training dataset, we used the base models, which are smaller in size, instead of the large models. Therefore, we replaced DeBERTa v3 Large with MDeBERTa v3 Base, and instead of RoBERTa Large, we used XLM-RoBERTa Base. As we can observe in Table 4, in all cases, the MDeBERTa v3 Base model outperforms the XLM-RoBERTa Base by a wide margin. In the case of the English language, we narrowly missed surpassing the result obtained by RoBERTa Large in the previous phase, but we matched the result obtained by DeBERTa v3 Large with a base model. The results obtained in the German and Italian languages are noteworthy,

Table 4
Results of the Multilingual Models trained in EN+IT+DE and applied to the Final Test datasets for EN,IT,DE.

	Baseline		**MDeBERTa V3 Base**		**XLM-RoBERTa Base**	
	F1 Macro	**SUBJ F1**	**F1 Macro**	**SUBJ F1**	**F1 Macro**	**SUBJ F1**
English	0.63	0.45	**0.73**	0.58	0.69	0.50
German	0.69	0.63	**0.85**	0.80	0.82	0.75
Italian	0.63	0.50	**0.83**	0.74	0.65	0.43

Table 5
Results of the Multilingual Models trained in EN+IT+DE and applied to the Final Test datasets for AR, BG.

	Baseline		**MDeBERTa V3 Base**		**XLM-RoBERTa Base**	
	F1 Macro	**SUBJ F1**	**F1 Macro**	**SUBJ F1**	**F1 Macro**	**SUBJ F1**
Arabic	**0.49**	0.40	0.48	0.29	0.45	0.23
Bulgarian	**0.75**	0.72	0.69	0.61	0.64	0.53

Table 6
Hyperparameters for the best performing models

Hyperparameter	**Best Monolingual Model**	**Best Multilingual Model**
Model	FacebookAI/roberta-large	microsoft/mdeberta-v3-base
Training Dataset	Train_EN	Train_EN+Train_IT+Train_DE
Num Train Epochs	5	3
Train Batch Size	8	16
Eval Batch Size	8	8
Learning Rate	5e-5	2e-5
Weight Decay	0.01	0.01
Warmup Steps	500	200

where we achieved scores of 0.85 and 0.83 respectively, significantly surpassing the baseline provided by Task2 for these languages.

In order to ensure the reproducibility of the results obtained with both the monolingual and multilingual approaches, Table 6 displays the models, training dataset, and hyperparameters used to train the models that achieved the best results when applied to the Final Test Dataset.

Finally, we sought to verify the Zero-Shot Cross-Lingual properties of both models by applying the models trained with the English, Italian, and German language datasets to the test datasets for the Bulgarian and Arabic languages without any specific fine-tuning for them.

We can observe in Table 5 that for both Arabic and Bulgarian languages, the results obtained in each case are worse than the baseline provided for both languages by Task2. Therefore, we must conclude that for subjectivity detection, there is no significant transfer of learning from one language to others without having seen examples in the second language during training. Consequently, we cannot rely on this feature of multilingual models for subjectivity detection in low-resource languages.

We believe that there could be several reasons why cross-lingual transfer has not worked, which should be analyzed in greater depth in subsequent studies. Lauscher [20] highlights the pretraining corpora size of the target language and the structural language similarity between languages as the main factors for the success of cross-lingual transfer.

In the final ranking for Task2, we achieved the second position out of a total of 15 participating teams in English language, with a final result for the Macro F1 score of 0.7372 and a SUBJ F1 score of 0.58. In Arabic, we obtained the fifth position out of a total of 7 participating teams, with a Macro F1 score of 0.4551 and a SUBJ F1 score of 0.25.

6. Conclusion

Our contribution to Task2 of CheckLab!2024 Subjectivity [1] aimed to determine, based on the provided datasets, whether it is possible to use the Zero-Shot Cross-Lingual feature of multilingual models to detect subjectivity in low-resource languages. The conclusion we reached is that it is not possible. However, given that this is a widespread problem that applies to all languages, we believe it would be interesting to continue investigating other non-multilingual Transformer-based approaches to help detect subjectivity in low-resource languages. Although the answer to our research question was negative, during the research process, we fine-tuned an MDeBERTa v3 Base model that achieved second place for English in Task2, with a score of 0.7372. It also achieved excellent results for German and Italian, with scores of 0.85 and 0.83 respectively, although we did not actively participate in the competition for these languages. As future lines of work, we propose adding Bulgarian and Arabic datasets, which we have not used to train the MDeBERTa v3 Base model, to see if adding more languages improves the model. It would also be relevant to analyze the use of Downsampling and Oversampling techniques to mitigate the bias present in the available datasets between objective and subjective sentences.

Acknowledgments

This research was made possible through the funding of Project TED2021-130810B-C22 by the Ministry of Science and Innovation of the Government of Spain.

The authors express their gratitude to the Smart Society Research Group at La Salle Engineering, Universitat Ramon Llull, the Digilab Research Group at Blanquerna, Universitat Ramon Llull and the anonymous reviewers for their insightful comments. They also extend their thanks to the organizers of the CheckThat!2024 Lab Task2 [1] for making this lab possible.

References

[1] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[2] G. Cardoso, Networked Communication. People are the Message, Editora Mundos Sociais, Lisboa, 2023.

[3] R. Nielsen, S. Ganter, The Power of Platforms: Shaping Media and Society, 2022. doi:10.1093/oso/9780190908850.001.0001.

[4] G. Canella, Journalistic power: Constructing the "truth" and the economics of objectivity, Journalism Practice 17 (2023) 209–225. URL: https://doi.org/10.1080/17512786.2021.1914708. doi:10.1080/17512786.2021.1914708. arXiv:https://doi.org/10.1080/17512786.2021.1914708.

[5] J. Birks, Evolving journalism norms: objective, interpretive and fact-checking journalism, in: The Routledge companion to political journalism, Routledge, London, 2021, pp. 62–71.

[6] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[7] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, May 2012.

[8] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Inf. Fusion 44 (2018) 65–77. URL: https://api.semanticscholar.org/CorpusID:46764901.

[9] M. Geva, Y. Goldberg, J. Berant, Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, 2019, pp. 1161–1166. URL: https://aclanthology.org/D19-1107. doi:10.18653/v1/D19-1107.

[10] F. Antici, F. Ruggeri, A. Galassi, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on English news articles, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 273–285. URL: https://aclanthology.org/2024.lrec-main.25.

[11] E. Schumacher, J. Mayfield, M. Dredze, Cross-lingual transfer in zero-shot cross-language entity linking, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 583–595. URL: https://aclanthology.org/2021.findings-acl.52. doi:10.18653/v1/2021.findings-acl.52.

[12] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[13] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[15] T. W. Lewis Tunstell, Leandro von Werra, Natural Language Processing with Transformers, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2022.

[16] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[18] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, ArXiv abs/1909.00161 (2019). URL: https://api.semanticscholar.org/CorpusID:202540839.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[20] A. Lauscher, V. Ravishankar, I. Vulic, G. Glavas, From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020. URL: http://dx.doi.org/10.18653/v1/2020.emnlp-main.363. doi:10.18653/v1/2020.emnlp-main.363.