Ensemble Pre-trained Transformer Models for Writing Style Change Detection

Notebook for PAN at CLEF 2022

Tzu-Mi Lin, Chao-Yi Chen, Yu-Wen Tzeng and Lung-Hao Lee

Department of Electrical Engineering, National Central University, Taiwan

Abstract

This paper describes a proposed system design for Style Change Detection (SCD) tasks for PAN at CLEF 2022. We propose a unified architecture of ensemble neural networks to solve three SCD-2022 edition tasks. We fine-tune the BERT, RoBERTa and ALBERT transformers and their connecting classifiers to measure the similarity of two given paragraphs or sentences for authorship analysis. Each transformer model is regarded as a standalone method to detect differences in the writing styles of each testing pair. The final output prediction is then combined using the majority voting ensemble mechanism. For SCD-2022 Task 1, which requires finding the only one position of a single style at the paragraph level, our approach achieves a macro F1-score of 0.7540. For SCD-2022 Task 2 to detect the actual authors of each written paragraph, our method achieves a macro F1-score of 0.5097, a Diarization error rate of 0.1941 and a Jaccard error rate of 0.3095. For SCD-2022 Task 3 to find located writing style changes at the sentence level, our model achieves a macro F1-score of 0.7156. In summary, our method is the winning approach in the list of all intrinsic approaches.

Keywords

Ensemble Learning, Pre-trained Models, Authorship Analysis, Plagiarism Detection

1. Introduction

PAN hosts a series of shared tasks for digital text forensics [1]. PAN-2018 introduced a Style Change Detection (SCD) to differentiate between multiple authors. The goal of the SCD-2022 shared task seeks to identify author switches within multi-author documents [2]. Given a document combined from the StackExchange questions and answers, participants are asked to solve three tasks illustrated in Figure 1: 1) Style Change Basic: given a text written by two authors that contains a single style change only, the developed system should find the position of such change; 2) Style Change Advanced: the developed system should be able to assign each paragraph of a multi-author text to a particular author; 3) Style Change Real-Word: for a text written by two or more authors, the developed system should find all positions of writing style change at the sentence level. Tasks 1 and 3 are basically similar but at different levels of granularity, i.e. paragraph-level or sentence-level. The output for Task 1 is represented as a list, supporting a binary classification for each pair of consecutive paragraphs within the given

document, where '0' means no style change and '1' denotes the only one style change in the task setting. Task 3 has a similar output format, presenting comparisons between each pair of consecutive sentences which may contain multiple changes. Task 2 is relatively difficult because the actual number of authors is unknown (though limited to five). Thus, the first author is identified as '1', the second author appearing in the document is referred to '2', and so on.

Figure 1: Possible scenarios for SCD-2022 tasks

Task 1	Task 2	Task 3
output: [0, 0, 0, … , 1]	output: [1, 2, …, 3, 2, 4, 5]	output: [1, 1, 0, 1, 1]
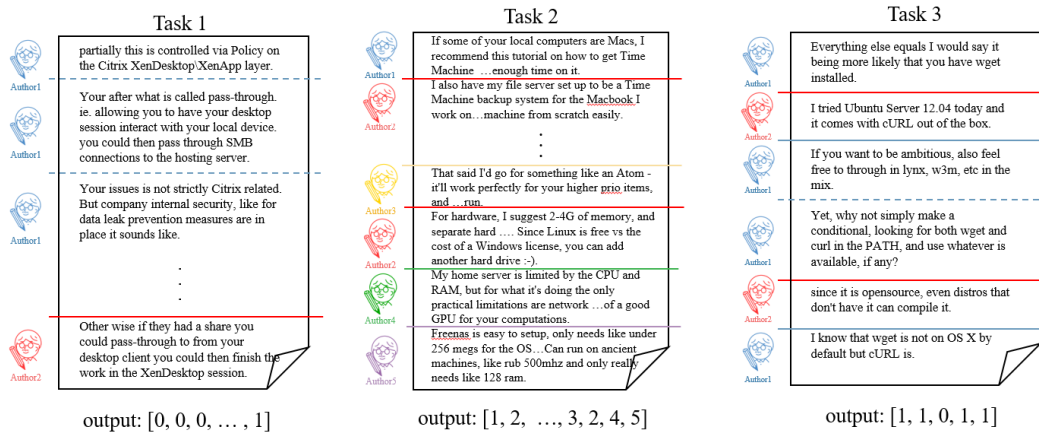
This paper describes our developed **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) system for the style change detection task of PAN-2022 evaluation. Our solution explores the use of pre-trained transformer models including BERT, RoBERTa and AlBERT, and fine-tuning of the downstream classification task for writing style change detection. In addition, system performance is enhanced using a majority voting ensemble mechanism. For Task 1, our ensemble solution had a macro F1-score of 0.7540. For Task 2, our system achieved a macro F1-score of 0.5097, a Diarization error rate of 0.1941 and a Jaccard error rate of 0.3095. For Task 3, our model resulted in a macro F1-score of 0.7156.

The rest of this paper is organized as follows. Section 2 investigates related studies for the previous editions of the SCD task. Section 3 describes the NCUEE-NLP system for the SCD-2022 tasks. Section 4 presents results evaluation and performance comparisons. Conclusions are finally drawn in Section 5.

2. Related Work

The 2018 edition of the SCD task focused on detecting whether a document is single-authored or multi-authored using intrinsic analyses [3]. This task belonged to a binary classification problem, in which the document contained style changes, indicating at least two authors. An ensemble of supervised learning models including SVM, Random Forest, AdaBoost, MLP and LightGBM was proposed to make predictions of style change [4]. Parallel attention networks were used to focus on the hierarchical structure of the language and the parse tree features of a sentence for style change detection [5]. Three feature types, including text statistics, hashing,

and high dimensionality were extracted to independently train the classifiers using an ensemble learning strategy to detect writing style changes [6]. Basic stylometry analysis based on word frequencies was used for efficient detection of style change [7]. The Bidirectional Echo State Network was also applied to detect writing style changes [8].

The SCD-2019 edition added a subtask for detecting the actual number of authors within a document [9]. Two clustering algorithms based on threshold and window merge were proposed to yield a number of clusters corresponding to the number of authors [10]. If multiple authors are detected in a document, the K-means and hierarchical clustering algorithms are used to group the documents written by the same author [11].

The SCD-2020 edition added a subtask to detect style changes between two consecutive paragraphs [12]. This past year, a paragraph representation based on character, lexical and syntactic features was proposed, using a clustering algorithm for style change detection [13]. The BERT pre-trained bidirectional model was used to generate the embedding representation and train a random forest classifier to detect style changes between two paragraphs [14].

The SCD-2021 edition also added determining the number of authors and locating specific author changeovers within the text [15]. Style change and author identification were regarded as binary classifications based on writing similarity measurements [16]. A stacking ensemble was developed and trained separately on previous text embeddings and features to enhance style change detection performance [17]. The logistic regression classifier was used to determine whether two successive paragraphs were written by the same author or not [18]. An LSTM-powered attribution algorithm was proposed for the style change detection task [19]. An end-to-end Siamese neural network was presented to detect stylistic similarities between two texts to determine the location of authorship changes [20].

In summary, deep learning-based neural computing approaches achieved promising results in the previous editions of the SCD shared task. This motivates us to explore the use of pre-trained transformer models to address the authorship analysis issue for writing style changes.

3. Ensemble Pre-trained Transformer Models

Figure 2: Our proposed NCUEE-NLP system architecture for the SCD-2022 tasks
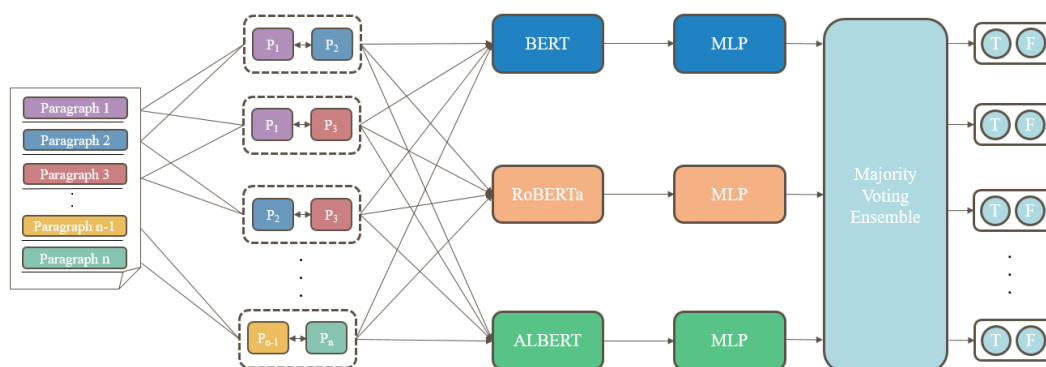
We propose a unified architecture of ensemble neural networks to solve three SCD-2022 tasks. Figure 2 shows our system architecture for style change detection, comprised of two main parts: 1) pre-trained transformer models; and 2) an ensemble mechanism.

The pre-trained model is a kind of saved neural network that was previously trained on a large-scale data set. Instead of building a model from scratch, pre-trained models can be used to solve similar problems, using transfer learning techniques to fine-tune the pre-trained model to fit the downstream task.

From the wide range of available pre-trained models, we select the following transformer models to tackle this task:

- Bidirectional Encoder Representations for Transformers (BERT) [21]
  BERT uses an encoder architecture with an attention mechanism to construct a transformer-based neural network architecture, providing state-of-the-art results in a wide variety of natural language processing tasks. BERT was pre-trained on two tasks: 1) masked language models: 15% of tokens were masked to train the BERT and the model then predicts the original value of the masked words based on the context; 2) next sentence prediction: BERT was trained to predict whether a following sentence was probable or not based on the previous sentence. Through the pre-training phase, BERT learns contextual embedding for representations from large-scale data sets. After pre-training, BERT can be fine-tuned on smaller data sets to optimize its performance on specific tasks.
- a Robust optimized BERT pre-training approach (RoBERTa) [22]
  RoBERTa is a replication study of BERT pre-training that carefully measures the impact of key parameters and training data size. The model modifications include removing the next sentence predictions, dynamically changing the masking pattern applied to the training data, and training with large batches.
- A Light BERT (ALBERT) [23]
  ALBERT was proposed to improve the training and results of the BERT architecture, using three different techniques: factorization of the embedding matrix, cross-layer parameter sharing, and inter-sentence coherence prediction.

Each pre-trained transformer model is regarded as a standalone model. The training datasets are used to fine-tune the language model of the individual pre-trained transformer and its connected Multi-Layer Perceptron (MLP) as a classifier. For SCD-2022 Task 1, each pair of two consecutive paragraphs are used for fine-tuning, along with their labeled classes ('1' means change and otherwise '0'). To solve the SCD-2022 tasks under a unified framework, we convert the Task 2 label to the binary classes of style changes [16]. The label of the first paragraph (denoting as P_1) is initialized as '1', indicating the first author. The second paragraph (P_2) is compared with the P_1. If this pair obtains the class label '1' (a style change), reflecting that the author is different from P_1 (the author label is '2'). Otherwise, if the class label is '0', the author of P_2 should be the same as that for P_1 (the author label is '1'). Similarly, every subsequent paragraph is compared against all preceding ones (e.g., P3 is compared with P1 and P2) to produce an ordered author list (limited to at most five distinct authors). For SCD-2022 Task 3, instead of paragraph pairs, we used every sentence pair with their style change labels to train their corresponding transformer models.

The ensemble mechanism uses multiple learning models to obtain better classification performance. We then use a majority voting ensemble [24], in which each transformer model makes a independent classification (i.e., a vote 0 or 1) for each testing instance. The final system prediction output is the one that receives a majority of votes.

4. Evaluation

4.1. Data

The experimental datasets were mainly provided by task organizers [2]. Table 1 presents a statistical summary of three datasets, which were based on user posts from the StackExchange sites on different topics. A total of 2,000 documents in Dataset 1 were provided for Task 1, while 10,000 documents each from Datasets 2 and 3 were provided for Tasks 2 and 3. Each dataset was mutually exclusive and split into three parts. The training set, including ground truth labels, was used to develop and train the model, and accounted for 70% of the whole dataset. We also used the 11,2000 documents in the 2021 edition dataset of this shared task for training data augmentation. The validation set, accounting for 15% of the whole dataset, was used to evaluate and optimize our model, while the remaining 15% was used for system performance evaluation.

Table 1
Data statistics for SCD-2022 tasks

Dataset	Dataset 1		Dataset 2		Dataset 3	
	#authors	#para.	#authors	#para.	#authors	#sent.
Training set	2,800	10,989	21,000	52,723	21,000	111,992
Validation set	600	2,441	4,500	11037	4,500	23,605
Test set	-	2,432	-	11,393	-	24,055

4.2. Settings

The pre-trained BERT[1], RoBERTa[2] and AlBERT[3] models were downloaded from HuggingFace [25]. On an Nvidia DGX-1 server using a V100 GPU we optimized the hyper-parameter values for our model implementation with the following settings: maximum sequence length 256; learning rate 0.00005; dropout 0.25; epoch 10 and batch size 100 for all three models.

System implementation was deployed on the TIRA platform for performance evaluation [26]. The three tasks were evaluated independently using the macro-averaged F1-score. Two additional measures – the Diarization Error Rate (DER) and Jaccard Error Rate (JER) – were used to provide additional perspectives on the results obtained for Task 2. A higher F1-score and a lower DER and JER indicate more accurate detection performance.

---

[1]https://huggingface.co/bert-base-uncased
[2]https://huggingface.co/roberta-base
[3]https://huggingface.co/albert-base-v2

4.3. Results

Table 2 shows the results on the SCD-2022 validation set. Among individual transformer models, RoBERTa significantly outperformed BERT and ALBERT for all evaluation metrics in all three tasks. The three ensemble transformer models enhanced task performance with some improvements, though slightly underperforming RoBERTa in Task 1.

Table 3 shows the results on the SCD-2022 test set are similar when compared with the validation results. Our NCUEE-NLP system achieved a macro F1-score of 0.7540 for Task 1; a macro F1-score of 0.5097, a Diarization error rate of 0.1941 and Jaccard error rate of 0.3095 for Task 2; and a macro F1-score of 0.0.7156 for Task 3.

Table 2
Results of transformer models on the SCD-2022 validation set

	Task 1	Task 2			Task 3
	macro F1-score	macro F1-score	DER	JER	macro F1-score
BERT	0.7144	0.5032	0.2078	0.3370	0.7059
RoBERTa	0.7561	0.5059	0.1989	0.3415	0.7164
ALBERT	0.7046	0.4664	0.2345	0.3807	0.6876
Ensemble	0.7424	0.5384	0.1907	0.3107	0.7276

Table 3
Results of transformer models on the SCD-2022 test set

	Task 1	Task 2			Task 3
	macro F1-score	macro F1-score	DER	JER	macro F1-score
BERT	0.7095	0.4919	0.2138	0.3337	0.6962
RoBERTa	0.7659	0.5026	0.2006	0.3237	0.7045
ALBERT	0.6917	0.4598	0.2681	0.3726	0.6736
Ensemble	0.7540	0.5097	0.1941	0.3095	0.7156

5. Conclusion

This study describes the model design, system implementation and performance of the NCUEE-NLP system in the PAN 2022 style change detection tasks. We selected pre-trained transformer models as the staring points and fine-tuned the corresponding downstream classification tasks. Our unified architecture used a majority voting ensemble mechanism to determine final system detection. Based on evaluation results, our ensemble transformer-based neural networks can achieve promising results with different task settings, although considerably more work is needed for Task 2, especially in terms of both error rate measures.

Acknowledgments

References

[1] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: A. Barron-Cedeno, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022.

[2] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings, 2022.

[3] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), Working Notes Papers of the CLEF 2018 Evaluation Labs, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[4] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection—Notebook for PAN at CLEF 2018, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[5] M. Hosseinia, A. Mukherjee, A Parallel Hierarchical Attention Network for Style Change Detection—Notebook for PAN at CLEF 2018, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[6] K. Safin, A. Ogaltsov, Detecting a Change of Style Using Text Statistics—Notebook for PAN at CLEF 2018, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[7] J. Khan, A Model for Style Change Detection at a Glance—Notebook for PAN at CLEF 2018, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[8] N. Schaetti, Character-based Convolutional Neural Network for Style Change Detection—Notebook for PAN at CLEF 2018, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[9] E. Zangerle, M. Tschuggnall, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2019, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[10] S. Nath, Style Change Detection by Threshold Based and Window Merge Clustering

Methods, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[11] C. Zuo, Y. Zhao, R. Banerjee, Style Change Detection with Feed-forward Neural Networks, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[12] E. Zangerle, M. Mayerl, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[13] D. Castro-Castro, C. Rodríguez-Losada, R. Muñoz, Mixed Style Feature Representation and B0-maximal Clustering for Style Change Detection—Notebook for PAN at CLEF 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[14] A. Iyer, S. Vosoughi, Style Change Detection Using BERT—Notebook for PAN at CLEF 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[15] E. Zangerle, M. Mayerl, , M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[16] Z. Zhang, Z. Han, L. Kong, X. Miao, Z. Peng, J. Zeng, H. Cao, J. Zhang, Z. Xiao, X. Peng, Style Change Detection Based On Writing Style Similarity—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-198.pdf.

[17] E. Strøm, Multi-label Style Change Detection by Solving a Binary Classification Problem—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-191.pdf.

[18] R. Singh, J. Weerasinghe, R. Greenstadt, Writing Style Change Detection on Multi-Author Documents—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-190.pdf.

[19] R. Deibel, D. Löfflad, Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-163.pdf.

[20] S. Nath, Style change detection using Siamese neural networks—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-183.pdf.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of NAACL-HLT 2019 (2019) 4171–4186. doi:https://doi.org/10.48550/arXiv.1810.04805.

[22] 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 (2019). doi:https:

//doi.org/10.48550/arXiv.1907.11692.

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, International Conference on Learning Representations 2020 (2020) 4171–4186. doi:https://doi.org/10.48550/arXiv.1909.11942.

[24] L.-H. Lee, Y.-S. Wang, C.-Y. Chen, L.-C. Yu, Ensemble multi-channel neural networks for scientific language editing evaluation, Institute of Electrical and Electronics Engineers Access (2021) 158540 − 158547. doi:10.1109/ACCESS.2021.3130042.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing (2019). doi:https://doi.org/10.48550/arXiv.1910.03771.

[26] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.