KaramiTeam at IberAuTexTification: Soft Voting Ensemble for Distinguishing AI-Generated Texts

Mohammad Karami Sheykhlan[1,*], Saleh Kheiri Abdoljabbar[2] and Mona Nouri Mahmoudabad[1]

[1]University of Mohaghegh Ardabili, Daneshgah St., Ardabil, 5619911367, Iran
[2]University of Tabriz, Bahman Boulevard, Tabriz, 5166616471, Iran

Abstract

Large language models have revolutionized the field of natural language processing. As these models become more widespread, concerns about the spread of misinformation and potential misuse have grown. Consequently, distinguishing between texts written by humans and those generated by machines has become a significant challenge. In this paper, we describe our method for addressing the AuTexTification task at IberLEF 2024, which includes two main subtasks. The first subtask is a binary classification challenge that requires distinguishing between texts written by humans and those generated by AI. The second subtask is a multi-class problem involving six text generation models (A, B, C, D, E, and F). Both subtasks are conducted in multiple languages. We selected three BERT-like models as the baseline models and then used the soft voting technique to improve accuracy. The results of the test set showed that soft voting outperformed the individual models.

Keywords

AuTexTification challange, Ensemble learning, Machine-generated text detection, Transformers, Text classification

1. Introduction

The advancement of artificial intelligence and machine learning, particularly in the field of Natural Language Processing (NLP), has led to significant progress in automatic text generation. Models such as Generative Pre-trained Transformers (GPT) [1, 2, 3], Pathways Language Model (PaLM) [4], and BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) [5] can produce text that closely resembles human writing in terms of coherence, style, and grammar. These capabilities have broad applications, including conversational agents, code completion, machine translation, and generating radiology reports, impacting both economic and social spheres. Despite their benefits, these models also present challenges, such as the spread of misinformation, academic fraud [6, 7], and the creation of offensive or biased content [8, 9]. The AuTexTification task at IberLEF 2024 [10, 11] addresses these challenges by distinguishing between human-written and machine-generated text.

This research is focused on fine-tuning three advanced language models: Enhanced Representation through kNowledge Integration Multilingual (ErnieM) [12], BLOOM-560m and Multilingual Decoding-enhanced BERT with disentangled attention (mDeBERTaV3) [13]. Each model was individually adjusted to optimize its performance for our specific task. To further enhance the accuracy of our predictions, we employed an ensemble learning approach using soft voting. This technique combines the strengths of multiple models by averaging their predictions, thereby improving overall performance. Our final results demonstrated that the ensemble method significantly outperformed the standalone models. This finding underscores the effectiveness of integrating multiple models to achieve higher prediction accuracy and reliability in distinguishing between human-authored and AI-generated texts.

The paper is organized as follows: First, we provide an overview of related works, highlighting previous studies on AI text generation and detection. Next, the methodology section details our approach, including the fine-tuning of ErnieM, BLOOM-560m, and mDeBERTaV3 models and the use of soft voting for ensemble learning. In the Experiments section, we present our models' performance metrics

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
mohammadkaramisheykhlan@gmail.com (M. K. Sheykhlan); salehkheiri@gmail.com (S. K. Abdoljabbar); monanouri.m@gmail.com (M. N. Mahmoudabad)
0000-0003-2316-545X (M. K. Sheykhlan); 0009-0009-3328-7486 (S. K. Abdoljabbar)

and compare the ensemble method's effectiveness against individual models. Finally, the conclusion summarizes our findings, discusses their implications, and suggests directions for future research.

2. Related works

Text classification in NLP involves assigning texts to predefined categories using various models and algorithms, including traditional methods like Naive Bayes and advanced deep learning models like Bidirectional Encoder Representations from Transformers (BERT) [14] and GPT. These models process textual data to identify patterns that differentiate various categories, enabling applications such as author identification [15], author attribution [16], and detecting hate or offensive content [17].

The precision and effectiveness of text classification have greatly advanced due to improvements in machine learning techniques and the availability of extensive datasets. This progress facilitates more refined and accurate classifications, fostering innovation in fields such as customer service automation and content recommendation systems. Nonetheless, challenges persist, including managing ambiguous or context-dependent texts and ensuring the models operate without bias.

Due to the high sensitivity in this field, numerous studies have been conducted in recent years to distinguish between human-written and machine-generated texts. The PAN@CLEF 2024 shared task [18] provided a binary English training dataset for this purpose. However, the test set differed from the training data, and participants were asked to identify the human-authored text from two texts of a sample. Participants were required to submit a Docker file of their approach via the TIRA platform [19].

The SemEval 2024 task 8 [20, 21] consisted of three subtasks. In subtask A, participants needed to distinguish between human and machine-generated text using binary data. In subtask B, they had to predict which language model generated a given text if it was machine-written. In subtask C, the objective was to identify which part of the text had been altered.

In the first version of the AuTexTification shared task at IberLEF 2023 [22], various approaches were proposed to distinguish between human and machine-generated texts in both English and Spanish. Villegas-Trejo et al. [23] utilized traditional feature extraction algorithms and machine learning models to address both subtasks. Their findings indicated that the XGB model, when combined with the TF-IDF n-gram feature extraction method and enhanced with stylometric features, demonstrated the highest performance across their experiments. Scheibe and Mandl [24] have made significant contributions by leveraging transformer-based models for text classification tasks. Specifically, in subtask 1 of their study, they employed the DeBERTaV2 model to distinguish between human and machine-generated texts. Gritsay et al. [25] applied a fine-tuning approach to large pre-trained language encoder models, specifically XLM-RoBERTa, mDeBERTa, and MiniLM-V2. They passed the CLS token through three fully connected layers. Their observations showed that mDeBERTa achieved the best F1 score.

3. Methodology

This section will discuss the dataset and the proposed approach in detail. We will begin by describing the characteristics and composition of the dataset used in our experiments, including the data sources, and the data preparation. We will outline our proposed approach, including the models and techniques employed. This will encompass the fine-tuning of specific language models, the feature extraction methods applied, and the ensemble learning strategies implemented to enhance prediction accuracy. By providing a comprehensive overview of both the dataset and our methodology, we aim to offer a clear understanding of the foundations and innovations of our research.

3.1. Dataset

The AUtomated TEXt IdenTIFICATION on languages of the Iberian peninsula (IberAuTexTification) is an expanded version of the AuTexTification task at IberLEF 2023, focusing on more models, domains, and languages including Spanish, Catalan, Basque, Galician, Portuguese, and English. Participants in

this task develop models to differentiate between human-written and automatically generated texts (Subtask 1) and identify the specific model used for text generation (Subtask 2). The training dataset encompasses five diverse domains while the testing dataset includes two additional domains. Texts are generated using a variety of models such as GPT-3.5, GPT-4, LLaMA, Coral, Command, Falcon, and MPT, sourced from platforms like OpenAI, Amazon Bedrock, Anthropic, Cohere, AI21, Google Vertex AI, and Meta. Datasets are curated using TextMachina [26], incorporating texts from controlled domains like essays, news, social media, Wikipedia, WikiHow, and uncontrolled domains sourced from OSCAR [27] and Colossal Cleaned Multilingual Common Crawl. Enthusiasts in this field can visit the Zenodo website to access the training[1] and test datasets[2].

3.2. Data preparation

For our study, we utilized three advanced language models: mDeBERTaV3, ErnieM, and BLOOM-560m, each with its corresponding tokenizer to preprocess the text data. We determined that a token length of 170 tokens per sample was optimal based on the average length of our text samples, ensuring a balance between computational efficiency and preserving information. Any tokens beyond this limit were discarded to maintain consistency across all samples.

Given the resource constraints on Google Colaboratory, particularly the limited GPU availability, we implemented selective sampling for model fine-tuning. For both Subtasks 1 and 2, we selected 50,000 samples from the training dataset to fine-tune the BLOOM-560m model. This approach ensured that we could manage the computational load while still providing sufficient data for the model to learn effectively.

In Subtask 1, which involved distinguishing between human and machine-generated text, we selected a sample size of 60,000 for fine-tuning the mDeBERTaV3 model. For other scenarios and models, we utilized the entire training dataset, leveraging all available data to maximize the training effectiveness. This comprehensive approach aimed to ensure that our models were well-trained and capable of performing robust text classification tasks.

3.3. Transformer-based Models

BLOOM-560m, ErnieM, and mDeBERTaV3 are three state-of-the-art transformer-based models that have significantly advanced the field of NLP.

BLOOM-560m is a multilingual model developed as part of the BigScience initiative, featuring 560 million parameters. It is designed to provide open-access language processing capabilities across multiple languages, promoting inclusivity and transparency in NLP research. The model's extensive training on diverse datasets allows it to handle complex linguistic contexts, making it suitable for a wide range of large-scale NLP tasks.

ErnieM by Baidu integrates external knowledge sources, such as knowledge graphs, into its language representation learning. This integration enhances ErnieM's ability to generate contextually accurate and semantically rich text, especially in multilingual settings. Its ability to incorporate structured knowledge allows it to understand and process intricate linguistic patterns more effectively.

mDeBERTaV3 builds on the BERT architecture with a focus on improving performance through a disentangled attention mechanism. This mechanism separates content and positional information, enhancing the model's understanding of language nuances. Trained on a large and diverse dataset, mDeBERTaV3 excels in tasks requiring deep contextual understanding, such as text classification and machine translation, due to its improved training efficiency and comprehensive language processing capabilities.

[1] https://zenodo.org/records/10853560
[2] https://zenodo.org/records/11034382

3.4. Ensemble learning

In our approach, we utilized the Soft Voting technique for ensemble learning to enhance the performance of our text classification models. Soft Voting involves averaging the predicted probabilities of multiple models and selecting the class with the highest average probability as the final prediction. This method leverages the strengths of each individual model, compensating for their weaknesses and leading to improved overall accuracy and robustness.

We implemented Soft Voting with our three fine-tuned models: BLOOM-560m, ErnieM, and mDeBERTaV3. By combining the predictive power of these models, we were able to achieve superior performance compared to using any single model alone. The diversity in architecture and training methodologies of these models ensures a more comprehensive understanding of the text, thereby improving the reliability and accuracy of our predictions. Our experiments demonstrated that the Soft Voting ensemble approach significantly outperforms individual models in both Subtask 1 and Subtask 2, highlighting its effectiveness in distinguishing between human and machine-generated text across multiple languages and domains.

4. Experiments

This section outlines the experimental setup and procedures used to evaluate our models. First, we discuss the hyperparameter tuning process to optimize the performance of BLOOM-560m, ErnieM, and mDeBERTaV3 models. We then present the results of our experiments, highlighting the effectiveness of individual models and the benefits of using a Soft Voting ensemble approach. Through detailed analysis and comparative metrics, we demonstrate the superior performance of our ensemble method in distinguishing between human and machine-generated text and in accurately attributing text to its generative model.

4.1. Hyperparameter tuning and Evaluation

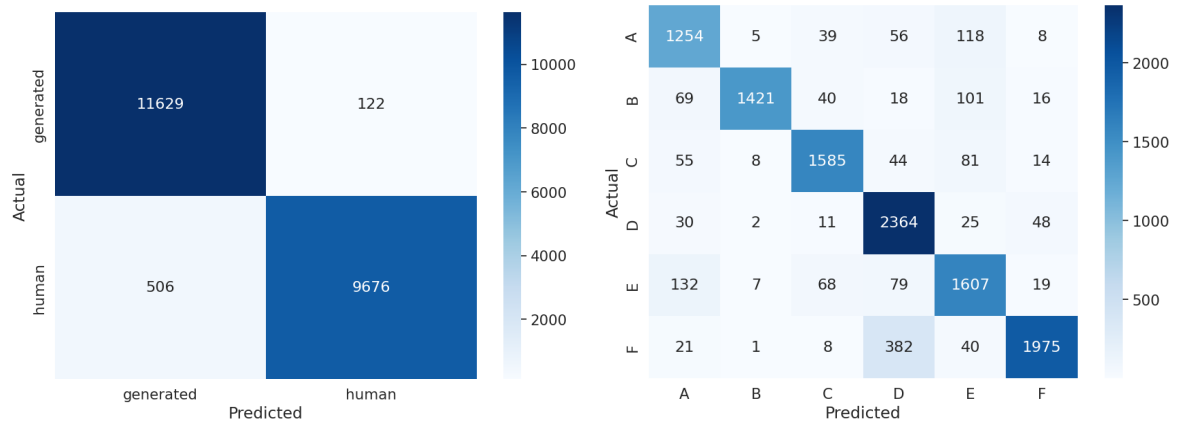
In this study, we used Google Collaboratory to fine-tune our models: BLOOM-560m, ErnieM, and mDeBERTaV3. The fine-tuning process was conducted using the Trainer API from the Hugging Face Transformers library [28], ensuring efficient training and evaluation. A learning rate of 5e-5 was consistently applied across all models to maintain stable and effective training. For ErnieM and mDeBERTaV3, we performed fine-tuning over 8 epochs, while BLOOM-560m was tuned for 3 epochs due to its larger size and complexity. Additional hyperparameters were uniformly set for all models to optimize performance: fp16 was enabled for mixed precision training, allowing faster computation and reduced memory usage; both per_device_train_batch_size and per_device_eval_batch_size were set to 8; a weight decay of 0.01 was applied to prevent overfitting by penalizing large weights; and gradient_accumulation_steps were set to 4 to simulate a larger batch size and stabilize training.

For both subtasks in our study, we employed the macro F1 score as the primary evaluation metric. The macro F1 score is particularly well-suited for imbalanced datasets as it calculates the F1 score for each class independently and then averages them, giving equal weight to each class regardless of its frequency. This approach ensures that the performance of our models is evaluated comprehensively across all classes, providing a balanced measure of precision and recall.

4.2. Results

We began by partitioning the training data into 80% for training and 20% for validation. After fine-tuning our models on the training set, we evaluated their performance individually and using a Soft voting ensemble method, which combined the outputs of ErnieM, mDebertaV3, and BLOOM-560m.

For this study, we submitted two runs for each subtask. In Subtask 1, we submitted the output of mDebertaV3 as run1 and the Soft voting ensemble as run2. In Subtask 2, we submitted the output of BLOOM-560m as run1 and the Soft voting ensemble as run2. The test set results indicated that the Soft



(a) Confusion matrix for Subtask 1.

(b) Confusion matrix for Subtask 2.

Figure 1: Confusion matrix of the soft voting model.

Table 1
Evaluation measures on the validation set. The best result is given in bold.

	mDebertaV3	ErnieM	BLOOM	Soft voting	Hard voting
Subtask 1	95.55	95.54	92.1	**97.11**	**97.11**
Subtask 2	78.63	80.17	**92.56**	86.84	86.36

Table 2
Evaluation measures on the test set. The best result is given in bold

	Run1	Run2
Subtask 1	62.33	**63.15**
Subtask 2	48.06	**49.30**

voting approach consistently outperformed the individual models in all subtasks. The results for the validation set are detailed in Table 1, while the results for the test set are provided in Table 2.

We examine the confusion matrix of the best model on the test data (soft voting). For Subtask 1 (Figure 1a), the confusion matrix shows that out of 11,751 generated texts, 11,629 were correctly identified, and only 122 were misclassified as human-written. Conversely, out of 10,182 human-written texts, 9,676 were correctly identified, and 506 were misclassified as machine-generated. This high level of accuracy demonstrates the effectiveness of our approach in distinguishing between human and machine-generated texts.

In Subtask 2 (Figure 1b), the confusion matrix reveals the performance of the soft voting approach in a multi-class classification scenario. The model achieved high accuracy across different categories, with notable performance in classes D and F, where 2,364 and 1,975 instances were correctly classified, respectively. However, there were some misclassifications, such as in class E, where 132 instances were incorrectly labeled as class A. Despite these challenges, the overall results underscore the robustness and efficiency of the soft voting ensemble method in handling diverse and complex text classification tasks.

5. Conclusion

In this study, we explored the efficacy of three language models—ErnieM, mDebertaV3, and BLOOM-560m—for the tasks of distinguishing between human-written and machine-generated texts and attributing generated texts to specific models. We fine-tuned these models using a carefully partitioned training

dataset and applied a Soft voting ensemble method to enhance prediction accuracy. Our experiments demonstrated that the Soft voting ensemble approach significantly outperformed individual models in both subtasks. Specifically, for Subtask 1, the combined model yielded better results compared to mDebertaV3 alone, and for Subtask 2, it outperformed BLOOM-560m. These findings underscore the potential of ensemble learning to improve the robustness and accuracy of AI text detection systems. The results from the validation and test sets clearly indicated that our ensemble method could generalize well across different domains and languages, reflecting the broader applicability of our approach. This work contributes to the ongoing effort to develop reliable detectors for distinguishing between human and AI-generated texts, highlighting the importance of ensemble methods in achieving higher accuracy and robustness. Future research could explore the integration of additional models and the application of advanced ensemble techniques to further enhance performance. Additionally, addressing the challenges of model biases and ensuring fairness in AI-generated text detection remain critical areas for further investigation.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, Journal of Machine Learning Research 24 (2023) 1–113.

[5] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model (2023).

[6] D. R. Cotton, P. A. Cotton, J. R. Shipway, Chatting and cheating: Ensuring academic integrity in the era of chatgpt, Innovations in education and teaching international 61 (2024) 228–239.

[7] J. P. Wahle, T. Ruas, F. Kirstein, B. Gipp, How large language models are transforming machine-paraphrased plagiarism, arXiv preprint arXiv:2210.03568 (2022).

[8] K. C. McLean, M. A. Fournier, The content and processes of autobiographical reasoning in narrative identity, Journal of research in personality 42 (2008) 527–545.

[9] R. Gagiano, H. Fayek, M. M.-H. Kim, J. Biggs, X. Zhang, Iberlef 2023 autextification: Automated text identification shared task–team od-21 (2023).

[10] A. M. Sarvazyan, J. Á. González, F. Rangel, P. Rosso, M. Franco-Salvador, Overview of iberautextification at iberlef 2024: Detection and attribution of machine-generated text on languages of the iberian peninsula, Procesamiento del Lenguaje Natural 73 (2024).

[11] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[12] X. Ouyang, S. Wang, C. Pang, Y. Sun, H. Tian, H. Wu, H. Wang, Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora, arXiv preprint arXiv:2012.15674 (2020).

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

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[15] H. B. Giglou, M. Rahgouy, T. Rahgooy, M. K. Sheykhlan, E. Mohammadzadeh, Author profiling: Bot and gender prediction using a multi-aspect ensemble approach., in: CLEF (Working Notes), 2019.
[16] M. Rahgouy, H. B. Giglou, T. Rahgooy, M. K. Sheykhlan, E. Mohammadzadeh, Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach., in: CLEF (Working Notes), 2019.
[17] M. K. Sheykhlan, J. Shafi, S. Kosari, Pars-hao: Hate speech and offensive language detection on persian social media using ensemble learning, Authorea Preprints (2023).
[18] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.
[19] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.
[20] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. Mohammed Afzal, T. Mahmoud, T. Sasaki, T. Arnold, A. Aji, N. Habash, I. Gurevych, P. Nakov, M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1369–1407. URL: https://aclanthology.org/2024.eacl-long.83.
[21] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, et al., Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection, arXiv preprint arXiv:2404.14183 (2024).
[22] A. M. Sarvazyan, J. Á. González, M. Franco-Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains, arXiv preprint arXiv:2309.11285 (2023).
[23] Z. Villegas-Trejo, H. Gómez-Adorno, S.-L. Ojeda-Trueba, Exploring text representations for detecting automatically generated text (2023).
[24] T. Scheibe, T. Mandl, Univ. of hildesheim at autextification 2023: Detection of automatically generated texts (2023).
[25] G. Gritsay, A. Grabovoy, A. Kildyakov, Y. Chekhovich, Automated text identification: Multilingual transformer-based models approach (2023).
[26] A. M. Sarvazyan, J. Á. González, M. Franco-Salvador, Textmachina: Seamless generation of machine-generated text datasets, arXiv preprint arXiv:2401.03946 (2024).
[27] J. Abadji, P. O. Suarez, L. Romary, B. Sagot, Towards a cleaner document-oriented multilingual crawled corpus, arXiv preprint arXiv:2201.06642 (2022).
[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).