Enhancing Stance Detection in Low-Resource Brazilian Portuguese Using Corpus Expansion Generated by GPT-3.5

Dyonnatan F. Maia[1] **Nádia F. F. da Silva**[1]
Ellen P. R. Souza[2] **André P. de L. F. de Carvalho**[3]
[1] Institute of Informatics, Federal University of Goiás, Goiás, Brazil
[2]Centro de Informática, Federal University of Pernambuco, Pernambuco, Brazil
[3]Institute of Mathematics and Computer Science, University of São Paulo, São Paulo, Brazil
dyonnatan@discente.ufg.br, nadia.felix@ufg.br
ellen.ramos@ufrpe.br, andre@usp.br

Abstract

In Natural Language Processing, the Stance Detection task classifies the text standpoint towards a given target. Stance Detection for citizen political opinions is highly dynamic because bill trends can appear and disappear quickly, demanding Stance Detection to handle unseen topics. We investigate the potential of leveraging generative models as annotators to enrich the dataset and improve the classification models in the restricted Brazilian Portuguese language in a low-resource context. We propose to use the prompt to perform a zero-shot corpus expansion using a generative model as an annotator to enhance the specialist fine-tuned models. We tested the data augmentation method by training mBert and Bertimbau models on UlyssesSD, BrMoral, and MtTwitter datasets for unseen topics. The models using our proposed corpus expansion showed promising performance on unseen topics.

1 Introduction

In Natural Language Processing (NLP), the Stance Detection (SD) task aims to identify and classify the text standpoint towards a given target. The input can be composed of the tuple <target, text> and the commonly used output labels for classification are *Favor*, *Against*, or *Neither*. Scenarios like internet discussions, political bills, and the news are highly dynamic because trends can appear and disappear quickly, demanding Stance Detection to handle unseen topics. As shown in Example 1, it is possible to obtain various stances depending on the chosen topic.

The Transformer architecture has gained popularity in NLP fields with models that utilise either an encoder, a decoder, or both components of its architecture. The BERT-based language models (Devlin et al., 2019) are encoder-only and consist of millions of parameters, serving as a backbone for various downstream tasks. GPT-3.5 is a decoder-only model based on InstructGPT (Ouyang et al., 2022), which is a massive model with billions of parameters that achieve the tasks by predicting tokens that collectively generate textual responses.

Text	
"The Child's place is socializing in school."	
Topic	**Stance**
Homescholling	Against
School	Favour

Figure 1: Example of Stance Detection toward topics concerning school socialisation.

GPT-3.5 uses a prompting technique that aims to reduce the high fine-tuning cost. A prompt is a specific template to pad the task input, aiming to get valuable knowledge from these models and make them more adaptable to different tasks. To avoid fine-tuning, researchers have investigated the strength of prompting in NLP applications, including the SD task (Zhang et al., 2023b,a).

Despite promising results for several tasks, some empirical studies indicate that the zero-shot strategy with prompt-based models still did not surpass the specialised fine-tuned models (Bang et al., 2023) but showed better performance in data annotation compared to worker annotators (Gilardi et al., 2023a). Therefore, we investigate prompt-based capabilities to improve the performance of a specialist fine-tuned model for SD in the restricted context of Portuguese as low-resource language.

We propose the following contributions:

- We expanded by GPT-3.5 annotation the UlyssesSD (Maia et al., 2022) corpus that contains comments about bills discussed in the Brazilian Chamber of Deputies.
- We evaluate the capabilities of prompt-based models for Stance Detection compared to the BERT model.
- We evaluate the proposed use of the prompt-

based model as an annotator to improve a specialist BERT model.

This work is organised as follows: The second section provides an overview of related works. In the third section, we describe the approach used for the study and the related methods. Section 4 describes the data annotation process, experiments, and analysis of the results. Section 5 has our conclusions and the work limitations.

2 Related Works

The SemEval Task 6b (Mohammad et al., 2016) competition introduces the Stance Detection task. For BERT-like models, the prior baseline found that the BERT-joint (Allaway and McKeown, 2020), a type of sentence pair classification technique, showed better results than encoding topic and text separately. The domains in the Portuguese language are in the process of exploring (Pavan et al., 2020; Maia et al., 2022; Pavan and Paraboni, 2022) using techniques that were investigated for the English domains but with fewer data samples. Küçük and Can (2020); Küçük and Can (2022) identified at least 13 English datasets, but only one was identified by them in the Portuguese language, putting the Portuguese language at a drawback of low resources available compared to English.

Since the emergence of Large Language Models (LLMs), the number of generative models proposed and evaluated for tasks like question-answering has increased. These models are also being investigated for classification tasks, as presented in a study by Chae et al. (2023) on LLM classification. Brown et al. (2020) introduced the multilingual LLM GPT-3, which was later optimised for chat applications called ChatGPT and built under the proposed GPT-3.5 and GPT-4 models. For the Portuguese language specifically, Pires et al. (2023) introduces the *Sabiá* model with competitive results in the Portuguese language compared to multilingual models. Nonetheless, despite the competitive results compared to fine-tuned language models, the computational cost of using LLMs for massive data processing tasks for classification does not make them a viable option compared to lighter fine-tuned language models.

The advent of ChatGPT has increased the exploration of prompt techniques as far as LLMs capabilities; the prompt approach allows the users to guide the model to evaluate the task, a powerful tool that helps GPT-3.5 and GPT-4 to perform a wide range of tasks. Qiu and Jin (2024) compared ChatGPT to fine-tuned BERT; they used designed prompts and checked that ChatGPT outperforms the BERT model in a few-shot setting and can be helpful for data labelling. Gilardi et al. (2023b) found as results that the use of ChatGPT outperforms crowd workers for text-annotation tasks, while Wen and Fang (2023) investigated the use of prompt tuning in a low-resource language. Unlike the above works, we apply the strong zero-shot LLM models as data enhancement to analyse the improvement of BERT, a lower computational cost model, for SD classification.

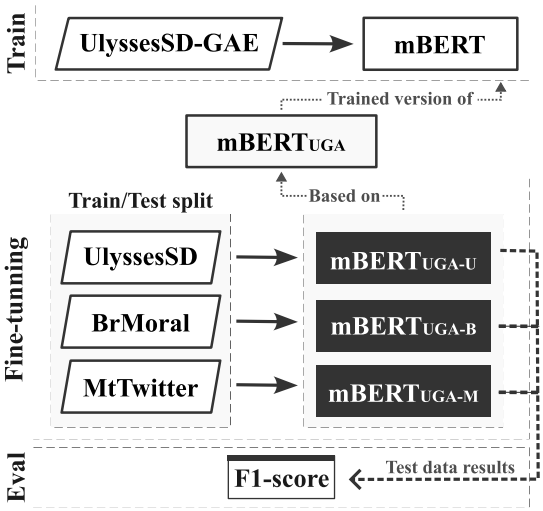

Figure 2: An overview of the proposed pipeline. It represents the mBERT classifier, whose Bertimbau classifier follows the same pipeline.

3 Our Approach

Our goal is to determine whether using a generative model as an annotator to enrich the dataset can enhance the performance of classification models for Stance Detection in unseen topics within the context of the low-resource Brazilian Portuguese language.First, we elect the generative model to be used as an annotator by comparing the GPT-3.5, GPT-4, and *Sabiá* (Pires et al., 2023) models using their respective API's on UlyssesSD test data. We collected and labelled the data according to Maia et al. (2022). Thus, we compiled the new dataset, referred to as "Ulysses Stance Detection with GPT-3.5 Annotation Expansion (UlyssesSD-GAE)".

We trained the BERT-like model with UlyssesSD-GAE, and then this model was fine-tuned with UlyssesSD train data and tested in UlyssesSD (Maia et al., 2022) test data; we also

Prompt
Você é um classificador de posicionamentos, Stance Detection. Diga o posicionamento das sentenças de acordo com o tópico: <topico>. Os rótulos são: neutro, favorável, contrário, nenhum, misto. Responda no formato csv: "<id>; <posicionamento>;" [You are a stance classifier, Stance Detection. State the stance of the sentences according to the topic: <topic>. The labels are: neutral, favour, contrary, none, mixed. Response in csv format: "<id>; <stance>;"]

Table 1: Prompt used to data annotation. The original text is in Portuguese, followed by the free translation.

topic	comment	stance
Contratação [Contracting]	*A colocação de funcionários por contrato, o que pode trazer de volta a contratação de familiares dos políticos* [The placement of employees on a contract basis, which could bring back the hiring of politicians' family members]	against
Lei de Propriedade Industrial [Industrial Property Law]	*A concessão de direitos de propriedade industrial é um dever do Estado. Em todo o mundo funciona assim, exatamente pra evitar conflito de interesses. Como a iniciativa privada vai conceder direitos a ela mesma sem que esse tipo de conflito exista?* [The granting of industrial property rights is a duty of the State. It works like this all over the world, exactly to avoid conflict of interests. How will the private sector grant rights to itself without this type of conflict existing?]	favor
Reforma Administrativa [Administrative Reform]	*É necessário uma reforma administrativa, mas não dessa forma.* [Administrative reform is necessary, but not like this.]	neither

Table 2: Examples of UlyssesSD-GAE dataset. The original text is in Portuguese, followed by the free translation.

used BrMoral and MtTwitter (Santos and Paraboni, 2019) for cross-dataset evaluation, as shown in Figure 2.

The classifier models were built using the following baseline architecture: A selected BERT-like model served as the backbone, embedding the tuple <topic, text> jointly, as described in Allaway and McKeown (2020), and the model head consisted of a fully connected layer with a softmax output. We evaluated two models for the backbone: Bertimbau (Souza et al., 2020) and mBERT(Devlin et al., 2019). The multilingual BERT (mBERT) is a trained BERT model that includes the Portuguese language, which the trained mBERT classifier in UlyssesSD-GAE produced the fine-tuned model in Ulysses GPT-3.5 Annotation, named $\text{mBERT}_{\text{UGA}}$. We replicate the model in $\text{mBERT}_{\text{UGA-U}}$ for UlyssesSD fine-tuning and evaluation, and also in $\text{mBERT}_{\text{UGA-B}}$ and $\text{mBERT}_{\text{UGA-M}}$ for BrMoral and MtTwitter respectively (Fig. 2). Finally, we repeat the process above with the Bertimbau backbone, a trained BERT model for the Portuguese language that produced $\text{Bertimbau}_{\text{UGA}}$.

4 Results and Analysis

In this section, we report the results from the data annotation to the final evaluation of the models with our best approach. We then provide an analysis of the UlyssesSD-GAE performance.

4.1 Data annotation

We compared three large language generative models to choose the most suitable model for our experiments. We build a handcrafted prompt (Table 1) based on the GPT-4 outputs of some examples from each dataset, with the temperature of 0.2 for more deterministic results. We evaluated the prompt in the UlyssesSD test data on the GPT-4, GPT-3.5, and *Sabiá* models to identify the most accurate candidate for data annotation.

We define the range of classes according to the GPT-3.5 and GPT-4 capabilities. When we asked GPT-3.5 and GPT-4 using OpenAi's API to classify the text without defining the classes, the typical stance outcomes were *Favour*, *Against*, *Mixed*, *Neutral*, or *None*. We noticed that bounding the models to this broad range of labels yielded more accurate results than constraining them to the restricted standard classes *Favour*, *Against*, *Neither*. Therefore, we allowed those possibilities to improve the model's prediction; then we combined the outcomes *Neutral*, *None*, and *Mixed* into the category termed *Neither*. This final output simplification was made to fit the previously determined SD classes in the benchmark datasets.

Table 2 shows three comments with GPT-3.5 annotation as examples. The written comments are made in the context where the web page shows other poll answers with positive and negative points

	AC	CLT	LOAS	SP
GPT-3.5	**.836**	**.864**	.509	.465
GPT-4	.804	.763	**.550**	.437
Sabiá	.628	.725	.478	**.646**

Table 3: F1-macro zero-shot results applied to UlyssesSD test data.

about the proposed bill. Then, the citizens are asked to indicate their positive and negative findings separately.

Table 3 shows that GPT-3.5 achieved equivalent but smoothly better results than GPT-4, with the additional advantage of lower API costs. So, we elected GPT-3.5 to annotate the collected data from the Chambers of Debuties website and expand the UlyssesSD dataset to 5671 comments for 273 topics. For this study, we removed the topics presented on test data for our experiments with unseen topics, resulting in 4201 valid instances. The UlyssesSD-GAE has the labels distributed as shown in Table 4, in which 258 topics have less than 50 instances, and we can notice unbalanced data with 67% comments against the topic.

Topic	Total	Favor (%)	Against (%)	Neither (%)
Con	501	79.2	15.2	5.6
RA	404	6.2	87.9	5.9
ED	192	21.9	70.3	7.8
LSN	191	35.1	51.3	13.6
(265 others)	2913	26.2	62.8	11.0
All	4201	23.2	67.0	9.8

Table 4: Distribution of instances in UlyssesSD-GAE according to topics. The whole topic's real names are Con: "Contratação"/"Hiring"; RA: "Reforma Administrativa"/"Administrative Reform"; "Estatuto do ED: "Desarmamento"/"Disarmament Statute"; LSN: "Lei de Segurança Nacional"/"National Security Law".

4.2 Experiments

We employed the experiments in the UlyssesSD dataset and cross-dataset validation in the BrMoral and MtTwitter datasets. The BrMoral is the elicited data from Santos and Paraboni (2019) and comprises 4.080 comments across eight topics extracted from a poll on morality questions. The MtTwitter dataset has 13.771 comments about politics, distributed across five topics. The UllysesSD dataset was collected from the website of the Chamber of Deputies of Brazil[1] in a poll section about political bills. The UlyssesSD dataset has 20 topics related to political bills and 1.935 comments with citizen opinions manually annotated. We performed over the same test data sample from Maia et al. (2022) for our results.

The test data comprises the topics "*Ajuda de custo/Subsistence allowance*" (AC), "*Consolidação das Leis do Trabalho/Consolidation of labour laws*" (CLT), "*Lei Orgânica de Assistência Social/Organic Social Assistance Law*" (LOAS), and "*Servidores Públicos/Public Servants*" (SP) from UlyssesSD. The BrMoral test data topics were "Same-sex marriage" (SSM) and "Church tax exemptions" (CTE), while the MtTwitter dataset included "Racial quotas" (RQ) and "Drug legalisation" (DL).

Training settings: The implementation was based on PyTorch (Paszke et al., 2019), transformers (Wolf et al., 2020), and it ran on hardware NVIDIA® V100 GPU. We use the AdamW optimiser with a learning rate of 2e-5 and no bias correction, implementing smooth weight decay rates with two groups of parameters, 0.01 and after 0.001 for bias, gamma, and beta. Appling a mini-batch size 16 and training in 10 epochs takes an average of 5 minutes for each model to be fine-tuned.

Applying the proposed approach, we trained two models with the UlyssesSD-GAE in the first step. Next, we replicated and fine-tuned the $\text{mBERT}_{\text{UGA}}$ models with UlyssesSD, BrMoral, and MtTwitter train data samples and evaluated the model on the test data. Then, we repeat the process and get the mean of 5 runs for the results shown in Table 5.

Table 5 shows that the models fitted by UlyssesSD-GAE (i.e., UGA versions) outperform the baseline models on most topics. $\text{Bertimbau}_{\text{UGA}}$ achieved the best performance in most topics and the best result in the simple average of all topics and significantly outperforms the other models in weighted averages considering the sample size of each topic; this means that the model also obtains better predictions in more instances than other models. This result strengthens the hypothesis that the new annotated examples enrich the understanding of the text, improving performance.

We conclude that the results show improvement in the model, especially in the cross-dataset evaluation, which shows the best results on most topics and the averaged scores despite the relatively low

[1] https://www.camara.leg.br/enquetes/

	UlyssesSD				BrMoral		MtTwitter			
	AC	CLT	LOAS	SP	SSM	CTE	RQ	DL	s.avg	w.avg
mBERT	.778 ±.003	.908 ±.05	**.883 ±.0**	.899 ±.006	.472 ±.003	**.517 ±.01**	.437 ±.005	.364 ±.01	.657 ±.007	.453 ±.009
$\text{mBERT}_{\text{UGA}}$	.889 ± .002	.991 ± .01	.880 ± .008	.991 ± .004	.502 ± .02	.509 ± .009	.545 ± .03	.366 ± .01	.709 ±.0014	.513 ±.022
Bertimbau	**.901 ± .002**	.989 ± .01	**.883 ± .0**	.993 ± .002	.485 ± .002	.478 ± .01	.512 ± .009	.389 ± .009	.704 ±.007	.500 ±.008
$\text{Bertimbau}_{\text{UGA}}$	.891 ± .004	**.993 ± .004**	.880 ± .002	**.994 ± .003**	**.633 ± .017**	.202 ± .04	**.633 ± .003**	**.635 ± .003**	**.733 ±.016**	**.623 ±.012**

Table 5: Comparison of stance classifications using the F1_{macro} score. Results are averaged over five runs with their respective standard deviations. The **s.avg** represents the simple average for F1_{macro} of each topic, and the **w.avg** is the average weighted by the sample size of the topics with the pooled standard deviation.

number of new instances and less accurate data than tuned models, as shown in Table 3 compared to Table 5.

5 Conclusions

We proposed using an LLM with prompt instructions to perform zero-shot data labelling to improve the Language Model fine-tuning applied to Stance Detection in Brazilian Portuguese domains. Our study demonstrates the potential of leveraging generative models as annotators to enrich SD datasets, particularly in a low-resource language. We utilised a BERT-like model trained on an augmented UlyssesSD dataset, annotated by large generative models, including GPT-3.5 and GPT-4. Our model showed promising performance across different topics related to political bills, as benchmarked against the UlyssesSD, BrMoral, and MtTwitter datasets.

For future work, there are many methods we can explore to improve the results and address unsolved problems. It could be to generate synthetic data using the LLMs to balance the dataset and compare whether the bias of the imbalanced model will be reduced in base models like BERT. Additionally, there are prompts for optimising the responses, like chain-of-thoughts reasoning (Wei et al., 2022) to improve the data annotation.

Limitations

GPT-3.5 and GPT-4 use and analyses come from a not fully disclosed architecture, and the models are only available via API. Additionally, because we executed our LLMs with a nonzero temperature, we gained interesting outcome variety for the annotation but also randomness that did not allow full reproducibility, which is not an impactful problem for annotation once we compile the final dataset. We consider the annotator model that reached the values closest to human annotation, that is, the annotators' bias, where there is no guarantee that it is the best possible labelling.

Acknowledgements

This work has been supported by the AI Center of Excellence (Centro de Excelência em Inteligência Artificial - CEIA) of the Institute of Informatics at the Federal University of Goiás (INF-UFG). Ellen Souza and Nadia Félix are supported by FAPESP with an agreement between USP and the Brazilian Chamber of Deputies. To the CEIA, to the Institute of Artificial Intelligence (IAIA), and to research funding agencies, to which we express our gratitude for supporting the research.

References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023a. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023b. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Dilek Küçük and Fazli Can. 2022. Stance detection and open research avenues. *arXiv preprint arXiv:2210.12383*.

Dyonnatan Ferreira Maia, Nádia Felix Felipe da Silva, Ellen Polliana Ramos Souza, Augusto Sousa Nunes, Lucas Caetano Procópio, Guthemberg da S Sampaio, Márcio de Souza Dias, Adrio Oliveira Alves, Dyéssica F Maia, Ingrid Alves Ribeiro, Fabíola Souza Fernandes Pereira, and Andre Carlos Ponce de Leon Ferreira de Carvalho. 2022. Ulyssessd-br: Stance detection in brazilian political polls. In *EPIA Conference on Artificial Intelligence*, pages 85–95. Springer.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Matheus Camasmie Pavan, Wesley Ramos dos Santos, and Ivandré Paraboni. 2020. Twitter moral stance classification using long short-term memory networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12319 LNAI:636–647.

Matheus Camasmie Pavan and Ivandré Paraboni. 2022. Cross-target stance classification as domain adaptation. In *Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part I*, page 15–25, Berlin, Heidelberg. Springer-Verlag.

Ramon Pires, Hugo Queiroz Abonizio, Thales Sales Almeida, and Rodrigo Frassetto Nogueira. 2023. [inline-graphic not available: see fulltext] sabiá: Portuguese large language models. In *Intelligent Systems - 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25-29, 2023, Proceedings, Part III*, volume 14197 of *Lecture Notes in Computer Science*, pages 226–240. Springer.

Yunjian Qiu and Yan Jin. 2024. Chatgpt and finetuned bert: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, 21:200308.

Wesley Santos and Ivandré Paraboni. 2019. Moral stance recognition and polarity classification from Twitter and elicited text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1069–1075, Varna, Bulgaria. INCOMA Ltd.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhihao Wen and Yuan Fang. 2023. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 506–516, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2023a. How would stance detection techniques evolve after the launch of chatgpt?

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023b. Investigating chain-of-thought with chatgpt for stance detection on social media.