Team nlpln at PAN 2024: An Approach to Classifying Conspiratorial and Critical Public Health Narratives With Zero-shot and Sequence Labeling

Notebook for PAN at CLEF 2024

Biao Liu, Zhongyuan Han* and HaoJie Cao

Foshan University, Foshan, China

Abstract

The growing prevalence of conspiracy theories poses significant challenges to content moderation on digital platforms. In our team nlpln, we employ a two-pronged approach to tackle distinct classification tasks related to public health narratives. For the first binary classification task, we utilize a Zero-Shot Learning approach with prompt engineering and Large Language Model (LLM). This method enables the model to differentiate between critical and conspiratorial narratives without extensive labeled data. For the second token-level classification task, we fine-tune a pretrained BERT-based model to identify and classify key elements in the narratives with the sliding windows technique and sequence Labeling. Our experiments demonstrate results; in subtask1, DeepSeek V2 and Baseline models outperform others in classification tasks across both languages, with notably high MCC values; in subtask2, our model demonstrates higher precision in detecting oppositional narrative elements across both languages, but the baseline model achieves better overall performance with higher Span-F1 scores, indicating a superior balance between precision and recall.

Keywords

Conspiracy Theories, Fine-tune, LLM, Sequence Labeling, Prompt engineering

1. Introduction

Conspiracy theories present intricate narratives attributing significant events to secretive groups [1]. With the rise of public health narratives and other major societal events, there is an increasing need for robust classification methods to differentiate between critical and conspiratorial content. Existing research often need to effectively distinguish between critical and conspiratorial thinking, creating a gap that this study aims to address using advanced machine learning techniques. At PAN 2024, the track on conspiracy theories has introduced tasks that focus on analyzing texts reflecting oppositional thinking, distinguishing between conspiracy narratives and critical narratives [2, 3].

Our approach involves two main tasks. First, a binary classification using Zero-Shot Learning [4] with prompt engineering allows classification without extensive labeled data. Second, token-level classification with a fine-tuned BERT model enhances precision and context-aware classification. We preprocess the dataset to convert character-level comments into word-level tags, ensuring alignment with predefined categories. Our methods focus on distinguishing between critical and conspiratorial texts and detecting key elements of oppositional narratives, providing insights into the performance and reliability of the techniques applied. This paper introduces a robust approach combining Zero-Shot Learning [4] and BERT fine-tuning to classify public health narratives, addressing both binary and token-level tasks with promising results.

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

hyticen@gmail.com (B. Liu); hanzhongyuan@gmail.com (Z. Han*); caohaojie0322@163.com (H. Cao)

0009-0000-3031-9758 (B. Liu); 0000-0001-8960-9872 (Z. Han*); 0000-0002-8365-168X (H. Cao)

2. Related Work

Previous research has primarily focused on the binary classification of conspiratorial content, with notable advancements through domain-specific BERT models [5]. However, existing models often need to distinguish between critical and conspiratorial thinking, a gap this paper seeks to address. For evaluating the performance of our binary classification models, we use the Matthews Correlation Coefficient (MCC), which has been identified as a more reliable metric for imbalanced datasets [6]. Additionally, recent studies emphasize the role of intergroup conflict in conspiratorial narratives [7], highlighting the need for nuanced detection methods.

A significant contribution to this field comes from Liu et al., who introduced ConspEmoLLM, an open-source LLM specifically designed for conspiracy theory analysis. Their work is mainly relevant to our research as it integrates affective information (sentiment and emotions) into the model, enabling it to perform diverse tasks related to conspiracy theories [8].

The researchers explored the detection of COVID-19-related conspiracy theories using a combination of BERT ensembles, GPT-3 augmentation, and graph neural networks, demonstrating the advantages of ensemble methods in improving classification accuracy [9, 10, 11, 12]. Additionally, Phadke et al. examined the social factors that contribute to individuals joining conspiracy communities, highlighting the role of social interactions in the spread of conspiratorial thinking [13].

Recent research has also delved into the sociopsychological processes underlying engagement in conspiracist communities. Wagner-Egger [14] discussed the mechanisms that drive individuals towards conspiracy theories, including the dynamics of intergroup conflict and the reinforcement of in-group versus out-group narratives. This perspective aligns with the findings of Böhm, Rusch, and Baron [15], who reviewed the psychological theories and measures related to intergroup conflict, providing a comprehensive understanding of the factors that exacerbate conspiratorial beliefs.

3. A Zero-Shot Learning Based Method for Distinguishing Between Critical and Conspiracy Texts

3.1. Method

3.1.1. Zero-Shot Learning

For the subtask1, we compare our zero-shot learning approach using LLMs with a baseline model. The baseline model is a fine-tuned BERT classifier, evaluated using the Matthews Correlation Coefficient (MCC). In a zero-shot setting, our approach utilizes various LLMs, including DeepSeek V2, Gemini 1.5 Pro, ZhiPu, Kimi, Claude3-Opus, and GPT-4o.

- **DeepSeek-V2**[1] is a powerful Mixture-of-Experts (MoE) language model characterized by economical training and efficient inference. It is comprised of 236 billion total parameters, with 21 billion activated for each token, and supports a context length of 128 thousand tokens.
- **Gemini 1.5 Pro**[2] is a large-scale foundational model developed by Google DeepMind that excels at performing reasoning tasks using text, images, audio, and video inputs. It boasts an impressive context window size of up to two million tokens, significantly outperforming its predecessors on various benchmarks.
- **ZhiPu AI**[3] is a Chinese company specializing in developing large-scale pre-training language models. Their flagship model, GLM-130B, was the only Asian model to be included in the Stanford Evaluation in 2022.
- **Kimi**[4] is an AI assistant developed by Moonshot, designed to provide users with a comprehensive suite of intelligent capabilities, including but not limited to information retrieval, data

[1] https://www.deepseek.com/
[2] https://deepmind.google/technologies/gemini/pro/
[3] https://chatglm.cn/
[4] https://kimi.moonshot.cn/

analysis, and more. While specific technical details regarding Kimi's architecture and capabilities are not widely available, it is positioned as a tool to help users "see a bigger world" through its intelligent assistance.

- **Claude3**[5] is a highly intelligent model offered by Anthropic, capable of handling complex analysis tasks involving multiple steps, higher-order mathematical computations, and coding challenges. As part of the Claude 3 family, Opus represents the most sophisticated model, providing superior performance for enterprise use cases at a competitive cost compared to other models in the market.
- **GPT-4o**[6] introduced by OpenAI, represents a significant advancement in multi-modal generative AI models. It supports a diverse range of input modalities, including text, audio, and images, enabling it to produce a corresponding variety of outputs. GPT-4o demonstrates exceptional performance across various benchmarks, particularly in visual and auditory understanding, setting new standards for real-time response times that approach human levels of interaction.

3.1.2. CoT Methodology

Chain-of-Thought (CoT) methodology is a prompting technique designed to improve the performance and accuracy of large language models (LLMs) on complex reasoning tasks by incorporating intermediate reasoning steps into the input prompts [16]. This approach guides the models to generate detailed reasoning processes, thus enhancing their problem-solving capabilities.

In our study, CoT prompts were specifically crafted to break down complex tasks into a series of coherent small steps, with each step building on the results of the previous one 3.1.2. The step-by-step reasoning methodology inherent in CoT enables LLMs to capture the essence of problems more accurately and demonstrate stronger performance and higher reliability on complex issues.

The CoT prompting process involves several key steps:

1. **Read the Text**: The model is instructed to carefully read the content in the provided field.
2. **Identify Key Themes**: It identifies key themes, phrases, or claims that may indicate the nature of the content.
3. **Evaluate Evidence**: The model assesses whether the claims made in the text are supported by credible evidence or are speculative.
4. **Determine Category**: Based on the evaluation, the model categorizes the text appropriately.
5. **Output the Result**: The final step involves formatting the response as a JSON object with relevant fields.

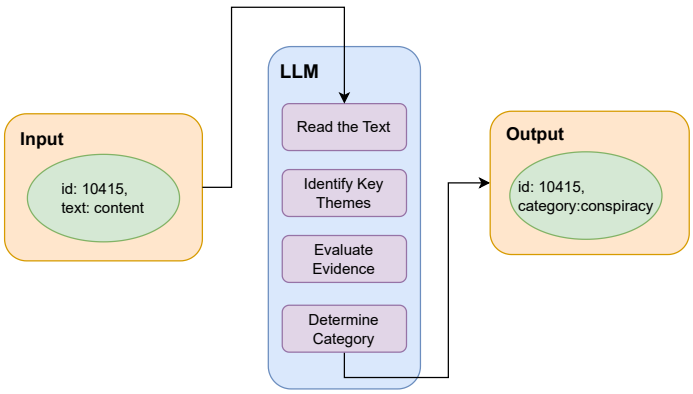

Figure 1: LLM processing flow

[5]https://www.anthropic.com/claude
[6]https://openai.com/index/hello-gpt-4o/

CoT Prompt

You are a professional information reviewer tasked with a binary classification task. You need to differentiate between:

Detail: I will provide you with a file in JSON format. You need to analyze the content in the text field and determine whether it is CRITICAL or CONSPIRACY.

Criteria:

- **CRITICAL:** Messages that question public health decisions, policies, or actions based on critique, factual information, or legitimate concerns without promoting a conspiracist mentality.
- **CONSPIRACY:** Messages that suggest the pandemic or public health decisions are the result of a malevolent conspiracy by secret, influential groups, often without credible evidence.

Step-by-Step Instructions:

1. **Read the Text:** Carefully read the content in the text field of each JSON object.
2. **Identify Key Themes:** Identify key themes, phrases, or claims that may indicate whether the text is questioning public health decisions or promoting a conspiracy theory.
3. **Evaluate Evidence:** Assess whether the claims made in the text are backed by credible evidence or whether they are speculative and lack factual support.
4. **Determine Category:** Based on your evaluation:
    - If the text provides reasoned critique, factual information, or raises legitimate concerns without promoting a conspiracist mentality, categorize it as CRITICAL.
    - If the text suggests the pandemic or public health decisions are the result of a malevolent conspiracy by secret, influential groups, categorize it as CONSPIRACY.
5. **Output the Result:** Format your response as a JSON object with the id and category fields.

Our LLM-based approach instructs the models to output their classifications in JSON format. The prompt includes specific instructions for formatting the output as a JSON object with 'id' and 'category' fields. We then use a Python JSON parser to extract the classification results. In cases where the LLM output does not conform to the expected JSON structure, we implement error handling to extract the classification based on keyword matching, defaulting to 'CRITICAL' if the classification cannot be reliably determined.

3.2. Results

For subtask1, distinguishing between critical and conspiracy texts, we evaluated the performance of several prominent large language models (LLMs) using the Matthews Correlation Coefficient (MCC) scores, based on the official train dataset provided by the PAN 2024 competition organizers consisting of 4000 gold-standard labels. MCC is a statistical metric ranging from -1 to 1, where 1 indicates perfect classification, 0 indicates performance no better than random guessing, and -1 indicates total misclassification. It is particularly effective for assessing the accuracy of classification models on imbalanced datasets.Each model was assessed using the same dataset to ensure comparability and reliability of the results. Table 1 presents the MCC scores for the different LLMs.

From Table 1, presents the models evaluated, including DeepSeek V2, Gemini 1.5 Pro, Baseline, ZhiPu, Kimi, Claude3-Opus, and GPT-4o. Among these, DeepSeek V2 and Baseline stand out with

[6]The baseline model was evaluated on the test set, while the results presented here for the other models are based solely on their performance on the training data.

Table 1
MCC Values for Various Models in ES and EN on Train Dataset

Model	MCC-ES	MCC-EN
DeepSeek V2	0.668	0.784
Gemini 1.5 Pro	0.509	0.601
Baseline	**0.668**	**0.796**
ZhiPu	0.584	0.651
Kimi	0.657	0.721
Claude3-Opus	0.579	0.681
GPT-4o	0.568	0.631

notably high MCC values in both languages: 0.668 for MCC-ES and 0.784 for MCC-EN in DeepSeek V2, and 0.668 for MCC-ES and 0.796 for MCC-EN in Baseline.While competitive, it lags behind the leading models with MCC values of 0.509 (MCC-ES) and 0.601 (MCC-EN). Models like ZhiPu, Kimi, Claude3-Opus, and GPT-4o show moderate MCC values ranging between 0.568 to 0.657 (MCC-ES) and 0.631 to 0.721 (MCC-EN), indicating varied levels of effectiveness in classification tasks compared to DeepSeek V2 and Baseline.

The MCC values highlight the varying degrees of effectiveness of different models incorporating zero-shot learning to distinguish between critical and conspiracy texts in Spanish and English. The baseline model establishes a high benchmark, particularly in English, indicating strong generalization without additional training. DeepSeek V2 stands out due to its high performance in both languages, suggesting it could be a preferred model for applications requiring robust zero-shot learning.

4. A Fine-Tuning Based Method for Detecting Elements of the Oppositional Narratives

4.1. Dataset

The dataset for this task consists of two text corpora, one in English and one in Spanish, sourced from the Telegram platform and related to the COVID-19 pandemic [2] and contains 4000 JSON entries. Each JSON object includes an id, a text field with a paragraph of text, a category label, and an annotations array. The text field contains content that can vary widely, but in this case, it includes statements related to conspiracy theories. The category field classifies the overall nature of the text, while the annotations array identifies specific segments within the text, marking them with different categories (e.g., `NEGATIVE_EFFECT`, `VICTIM`, `AGENT`). Each annotation specifies the exact span of the text it refers to, with starting and ending character positions.

4.2. Method

On the one hand, we use a sliding window method to handle the problem of long texts. Due to the input length limitation of the BERT model, when the text length exceeds the model's maximum input length, we split the text into multiple overlapping windows, each not exceeding the maximum input length. This method ensures that the model can process all textual information while preserving contextual information across windows. On the other hand, we fine-tune the BERT model to adapt it to specific tasks. We fine-tune a pre-trained BERT model for token-level classification tasks to identify and classify key elements within the narratives. We will introduce these two parts separately.

4.3. Sliding Window Method

Due to the limitation of the maximum input length of the BERT model, we employed a sliding window method to prevent information loss. When the length of the text exceeds the model's maximum input

length, we split the text into multiple overlapping windows, each not exceeding the maximum input length. This method ensures the model can process all textual information while preserving contextual information across windows. In our implementation of the sliding window method, we used a window size of 512 tokens with a stride of 256 tokens. This configuration allows for a 50% overlap between adjacent windows, ensuring that no information is lost at the boundaries. For each window, we perform token-level classification and then aggregate the results, resolving any conflicts at the overlapping regions by selecting the prediction with the highest confidence score.

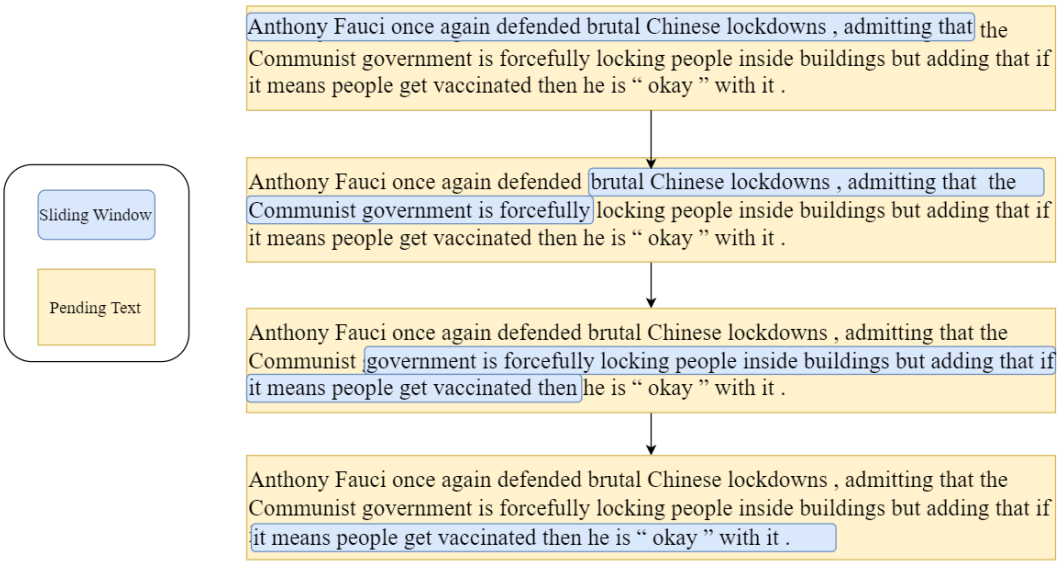

Figure 2: Sliding Windows

4.3.1. Fine-Tuning

For subtask2, detecting elements of the oppositional narratives, we fine-tune a pretrained BERT-based model to identify and classify key elements within the narratives. The approach we use involves three main stages:

- Dataset Construction
- BERT Model Fine-tuning
- Hyperparameter Tuning

We define a preprocessing function that converts character-level comments into word-level tags. This function iterates over each text and its annotations, assigning the appropriate `"B-"` (beginning) or `"I-"` (inside) tags to the words based on the character position of the annotations, ensuring that these tags conform to predefined categories. Any tags that do not fit the predefined categories will be skipped. The processed word segments and tags are converted to numerical format, where tags are indexed based on their position in the category list.

4.4. Experiment

4.4.1. Data Processing

To balance the dataset, we split it into training and test sets, using 70% of the data for training and 30% for testing.

The dataset used for this study was loaded from a JSON file, which contained textual data along with corresponding annotations. Each entry in the dataset includes a text field and an annotations

field. The annotations field comprises a list of entities, each defined by a start character, end character, and category label.

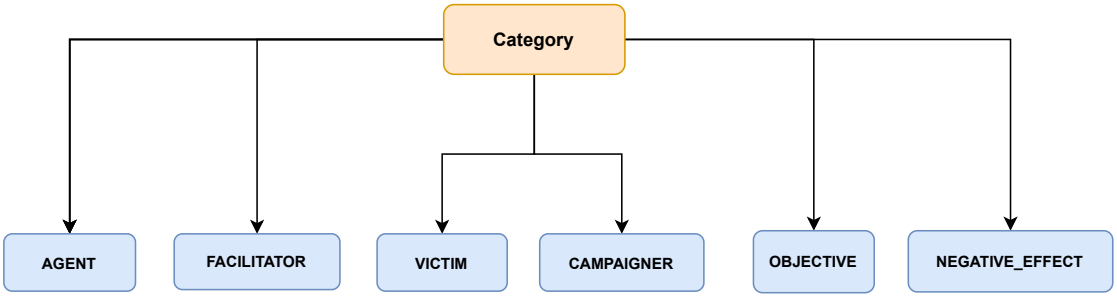

Figure 3: Category

4.4.2. Results

In the task of detecting elements of oppositional narratives, we evaluated the models' performance on English (EN) and Spanish (ES) corpora using Span-P, Span-R and Span-F1 metrics.

From Table 2 presents there performance metrics for detecting elements of oppositional narratives using fine-tuned BERT models on English (EN) and Spanish (ES) corpora. It compares our model against the baseline model, Baseline-BETO. For the English corpus, our model achieves a Span-P of 0.527, surpassing the baseline's 0.468, indicating higher precision. However, the baseline model performs better in Span-F1 with a score of 0.532 compared to our model's 0.334, suggesting it balances precision and recall more effectively.

In the Spanish corpus, our model also shows superior precision with a Span-P of 0.517 against the baseline's 0.453. Nevertheless, the baseline model achieves a slightly higher Span-F1 score (0.49 versus 0.467), demonstrating better overall performance in balancing precision and recall.

The detailed analysis reveals several key points about the models' performance. First, our model demonstrates higher precision across both languages, indicating it is more accurate in identifying elements of oppositional narratives. However, the considerably lower Span-F1 scores highlight a substantial issue with recall. The baseline model, on the other hand, maintains a better balance between precision and recall, as reflected in its higher Span-F1 scores. This suggests that while our model can make precise predictions, it fails to capture as many relevant instances, reducing its overall effectiveness.

Table 2
Model Score in EN and ES on Test Dataset

Language	Model	Span-P	Span-F1	Span-R
EN	nlpln	**0.527**	0.334	0.3303
EN	Baseline-BETO	0.468	**0.532**	**0.6334**
ES	nlpln	**0.517**	0.467	0.4426
ES	Baseline-BETO	0.453	**0.493**	**0.5621**

Overall, Our Model is more precise in detecting elements of oppositional narratives, with a lower false positive rate. On the other hand, the Baseline-BETO model maintains a better balance between precision and recall, especially in the English corpus. This suggests that different models may have distinct advantages depending on the application context. If the task prioritizes precision, Our Model is more suitable; however, for a balanced performance, the Baseline-BETO model might be the better choice.

[6]The data for the evaluation results comes from the test data

5. Conclusion

This study aimed to develop NLP models that distinguish between critical and conspiratorial texts and identify key elements of oppositional narratives. Our approach combined Zero-Shot Learning with BERT fine-tuning to address the binary and token-level classification tasks, respectively.

Several limitations still need to be addressed in our current work. First, the reliance on Zero-Shot Learning may introduce variability in performance depending on the quality of prompt engineering. Second, while our fine-tuning process has shown promising results, there is a need to investigate the impact of different preprocessing techniques and model configurations on the final outcomes. Another open question is how well our models handle evolving narratives, especially as new conspiracy theories and public health critiques emerge. Addressing these limitations and questions will be crucial for advancing the effectiveness and reliability of our classification approach.

Acknowledgments

This work is supported by the Social Science Foundation of Guangdong Province, China (No.GD24CZY02)

References

[1] K. M. Douglas, R. M. Sutton, What are conspiracy theories? a definitional approach to their correlates, consequences, and communication, Annual Review of Psychology 74 (2023).

[2] D. Korenčić, B. Chulvi, X. Bonet Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the oppositional thinking analysis pan task at clef 2024, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024–Conference and Labs of the Evaluation Forum, 2024. URL: https://doi.org/10.5281/zenodo.10680586.

[3] Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification - Condensed Lab Overview, 2024.

[4] Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, Q. M. J. Wu, Zero-shot learning - the good, the bad and the ugly, arXiv preprint arXiv:2011.08641 (2020).

[5] A. Giachanou, B. Ghanem, P. Rosso, Detection of conspiracy propagators using psycho-linguistic characteristics, Journal of Information Science 49 (2023) 3–17.

[6] D. Chicco, N. Tötsch, G. Jurman, The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Mining 14 (2021) 13.

[7] R. Böhm, H. Rusch, J. Baron, The psychology of intergroup conflict: A review of theories and measures, Journal of Economic Behavior & Organization 178 (2020) 947–962.

[8] Zhiwei Liu and Boyang Liu and Kailai Yang, Paul Thompson and Sophia Ananiadou, Computer science the university of manchester, manchester, united kingdom, mailto:zhiwei.liu-2@postgrad.manchester.ac.uk,boyang.liu-2@postgrad.manchester.ac.uk,kailai.yang@postgrad.manchester.ac.uk,paul.thompson@manchester.ac.uk,sophia.ananiadou@manchester.ac.uk, ???? Accessed: 2023-11-09.

[9] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, P. Rosso, Definitions matter: Guiding GPT for multi-label classification, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 4054–4063. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.267. doi:10.18653/v1/2023.findings-emnlp.267.

[10] Y. Peskine, G. Alfarano, I. Harrando, P. Papotti, R. Troncy, Detecting covid-19-related conspiracy theories in tweets, 2023.

[11] Y. Peskine, P. Papotti, R. Troncy, Detection of covid-19-related conpiracy theories in tweets using transformer-based models and node embedding techniques, 2023.
[12] Y. Peskine, D. Korenčić, I. Grubišić, P. Papotti, R. Troncy, P. Rosso, Definitions matter: Guiding gpt for multi-label classification, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 4054–4063.
[13] S. Phadke, M. Samory, T. Mitra, What makes people join conspiracy communities? role of social factors in conspiracy engagement, Proceedings of the ACM on Human-Computer Interaction 4 (2021) 223:1–223:30.
[14] P. Wagner-Egger, A. Bangerter, S. Delouvée, S. Dieguez, Awake together: Sociopsychological processes of engagement in conspiracist communities, Current Opinion in Psychology 47 (2022) 101417.
[15] R. Böhm, H. Rusch, J. Baron, The psychology of intergroup conflict: A review of theories and measures, Journal of Economic Behavior Organization 178 (2020) 947–962.
[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models (2022). arXiv:2201.11903.