Exploring Stance Detection of Opinion Texts: Evaluating the Performance of a Large Language Model

Benchmarking the Performance of Stance Classification by GPT-3-Turbo

Niels Mateijsen

Supervisor(s): Catholijn Jonker, Morita Tarvirdians

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Niels Mateijsen
Final project course: CSE3000 Research Project
Thesis committee: Catholijn Jonker, Morita Tarvirdians, Mathijs Molenaar

Abstract

In April 2020, a Dutch research team swiftly analyzed public opinions on COVID-19 lockdown relaxations. However, due to time constraints, only a small amount of opinion data could be processed. With the surge of popularity in the field of Natural Language Processing (NLP) and the arrival of tools like ChatGPT, a number of tasks involving Large Language Models (LLMs) have become easier. This study aims to address the effectiveness of these LLMs on stance detection using this COVID-19 opinion corpus. The corpus is chunked and sampled to be used as input for OpenAI's GPT-3.5-Turbo LLM. The machine-generated stances are then evaluated against multiple binary classification metrics. It is shown that these models perform very well in the field of stance detection, with an average F-score of 0.895. However, a significant number of misclassifications are observed in one dataset. Therefore we conclude that while LLMs offer valuable guidelines, it is still crucial to verify their outputs when dealing with complex or important public matters.

1 Introduction

In recent months, the field of Natural Language Processing (NLP) has experienced a significant surge in popularity. This can be attributed to the emergence of powerful tools like ChatGPT, which makes use of the GPT-3 model, enabling a range of tasks to become easier or even feasible. Exciting advancements in various domains utilizing Large Language Models (LLMs) have been unveiled through recent studies, demonstrating promising results [3].

Among the tasks made easier by advancements in NLP is decision-making in public issues like the COVID-19 pandemic. In April 2020, a Dutch research team became aware that policymakers would come with big changes in the relaxation of the lockdown measures [14]. The team designed their study in 20 days and additionally collected and analyzed their data in 7 days. In this period, the public opinions of about 26,000 Dutch citizens about the proposed statements were collected. Because of the limited time, the researchers only had time to analyze 2,237 participants (8,5%).

This research has already served as a source of motivation for researchers to analyze this problem, leading to the development of innovative ideas and frameworks for more efficient analysis of such data. One such method is the HyEnA method [24]. This method uses human annotations in combination with automated methods to extract key arguments of a topic. HyEnA reduced the most human-intensive phase by 60%, allowing the analysis to be conducted very efficiently.

The HyEnA research makes use of an opinion corpus with annotated stance. But in some cases, the opinion data available is not annotated like in Twitter tweets. It can cost a lot of time and money to annotate each of these individual opinions. In that scenario, LLMs can replace humans to reduce these factors. This paper will answer the question of how effective OpenAI's GPT-3.5-Turbo LLM is at classifying stances from opinions. It is hypothesized that these new language models show promising results in this task as well as an easy implementation, indicating their potential for future utilization. This paper will make use of the same COVID-19 dataset as mentioned in [14]. However, this can also be used in other fields. For example, in differences of stances between richer and poorer countries using Twitter [20] or (political) meeting summarization [26] to create a summary of all the stances of the meeting participants. The contribution of this research lies in evaluating the effectiveness of GPT-3.5-Turbo LLM for stance classification, offering insights into its potential as a time-efficient and cost-effective alternative to human annotation in large-scale opinion analysis.

This paper will first give an overview of the dataset used in this study. Afterward, an outline of the methods used to gather results will be provided. Finally, we present the experiment's results and conclude the effectiveness of LLMs for the detection of opinion stances.

2 Related Work

Stance is defined as the viewpoint of a person concerning a target [8]. Usually, this can be denoted as either *pro* , *con* , or *neither*. Stance classification is the study of analyzing different sorts of text to extract the stance it has. Due to only having a small set of possible stances, this problem can also be defined as a multi-class classification problem. Stance classification can provide valuable insights into the opinions of groups in fields like social media discussions and political debates. Already many different approaches to the problem of stance detection and classification have been researched.

One of these approaches is based on probabilities [22]. This method makes use of the combination of a subjective word to determine the polarity (positive, negative, or neutral) and a target term. A probability distribution is generated from the polarity-target pairs to determine the stance of an opinion. Other approaches usually use some version of machine learning classifier models together with selected features like lexical, syntactic, domain-specific, and argumentation [1]. Some of these classifiers are Naïve Bayes [11, 12], Support Vector Machine [5, 11, 28], and random forest [11, 28].

Multiple different datasets can be used for stance classification. For example, SemEval is annual series of workshops dedicated to the advancements of NLP semantic analyses [19]. Each edition contains multiple tasks and datasets for different analysis like sarcasm detection and news similarity. A well-used task for stance classification is the SemEval-2016 Task 6. This contains 4,870 English annotated tweets about multiple targets like Climate Change and Abortion [13]. Another big dataset is P-Stance [9]. This is a large dataset with 21,574 annotated tweets which all relate to politics.

Stance classification can also be useful for meeting summarization. Together with discussion detection [21], the stance classification employed in this research offers valuable insights for discussions in meetings by summarizing the various stances of the participants. However, annotated meeting data is considered relatively scarce [23]. Classifying stances on meeting data can also be hard for humans, with a re-

ported 77% accuracy for political debates [27]. There has already been research done on using automated tools for these kinds of perspective detection for meetings in a political context [26].

3 Method

This study aims to analyze the stance annotation of public opinions using OpenAI's GPT-3-Turbo model. This is OpenAI's most capable and cost-effective GPT-3.5 model [15]. Firstly, an overview of the dataset used will be presented. Subsequently, the method for extracting stances will be described. Finally, performance metrics will be provided to evaluate the language model's effectiveness.

3.1 Data

For this research, we will use the opinion corpus from the public about the COVID-19 lockdown measures relaxation research. [14]. The dataset is publicly available as supplementary material from the HyEnA research [25]. Like in the HyEnA research, we will cover 3 of the 8 policy options that were collected. We will copy the same notation for the options by using a keyword as an identifier for the options. These identifiers are notated in uppercase. These options are:

1. *YOUNG people may come together in small groups.*
2. *All restrictions are lifted for persons who are IMMUNE.*
3. *REOPEN hospitality and entertainment industry.*

To analyze stance detection, the model will utilize the opinion corpus, which consists of three opinion lists corresponding to different policy options. The YOUNG list consists of 13,400 opinions, the IMMUNE list contains 10,567 opinions, and the REOPEN list comprises 12,814 opinions. Table 1 contains an overview of the entire dataset, separated by policy option. Each opinion in the corpus is annotated with its corresponding stance, which can be either *pro* (in favor) or *con* (against). These stances are annotated by the research participants associated with each opinion. In addition to the stance, there is a pre-computed quality score associated with each opinion, which serves as an indicator of the text's quality of the opinion.

3.2 Stance detection

To prepare the input for the LLM, the dataset is parsed, and each opinion is assigned a unique identifier to keep track of the opinion and its corresponding stance. Subsequently, the opinions of each option are parsed into a JSON format, forming a list that includes only the opinion's ID and text.

However, due to the OpenAI API's limitation of a maximum of 4,096 tokens for input, the data is chunked into segments of up to 4,096 tokens each. Considering the large volume of data, 20 chunks per option are uniformly sampled to be used as input. On average, each chunk contains 22 opinions. This value varies depending on the size of the opinion text.

For each chunk, we also add instructions for the LLM. These instructions contain the policy option for which the stance of the opinion needs to be evaluated. Additionally, the constraint is mentioned that the stance can only be either *pro* or *con*. Finally, the (JSON) output format is described with an example output to make sure the output can be easily parsed. The exact prompt used can be found in Appendix A.

3.3 Metrics

In the context of stance detection, the task can be approached as a binary classification problem, where opinions are classified as either for (*pro*) or against (*con*). As a result, we can employ the same set of evaluation metrics commonly used in binary classification to assess the performance of the Large Language Model. In this evaluation, we will focus on four key metrics: accuracy, precision, recall, and the F-score.

In this section, the terms TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) will be used to refer to specific data points. These values are specific to each class. For instance, if you wish to determine the precision of the model in predicting the *pro* stance, you would consider the *pro* stance as a positive data point.

Accuracy

Accuracy is the most simple metric. It measures how many times the model predicts a value correctly. Accuracy is described by dividing the total number of correct classifications by the total number of predictions, as shown by (1).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The accuracy metric is generally easy to comprehend and interpret. However, it can be misleading in cases where the dataset is unbalanced. For instance, if a dataset comprises 2% *con* opinions and a model consistently predicts *pro*, the accuracy will still be 98%. Therefore, it is recommended to utilize this metric cautiously and preferably in scenarios where the dataset is well-balanced.

Precision

Precision measures the proportion of correctly predicted cases that are positive. It is defined as the division of the number of correct positive classifications by the total number of positive classifications, as shown by (2).

$$precision = \frac{TP}{TP + FP} \tag{2}$$

The precision metric is useful for cases where False Positives are more costly than False Negatives.

Recall

Recall measures how many actual positive data points can be predicted correctly by a model. It shows the proportion of actual positive data points that were correctly identified. The metric is defined by dividing the number of correct positive classifications by itself and the number of incorrect negative classifications, as shown by (3).

$$recall = \frac{TP}{TP + FN} \tag{3}$$

The recall metric is useful for cases where False Negatives are more costly than False Positives.

Table 1: Stance Distribution per Policy Option

Policy Option	Number of Opinions	Percentage Pro	Percentage Con
YOUNG	13,400	66%	34%
IMMUNE	10,567	17%	83%
REOPEN	12,814	55%	45%

F-score

The F-score is a combination of the precision and recall metrics. It is defined as the harmonic mean of the precision and recall scores, as shown by (4).

$$\text{F-score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

Since the F-score is a combination of the precision and recall metrics, it considers both False Positives and False Negatives when evaluating a model. This is useful when the costs of both are of equal importance.

For our application, the costs of falsely classifying an opinion as *pro* or *con* are equally important. Therefore the F-score will be the best metric to evaluate the performance of the LLM.

4 Results

4.1 Quantitative results

The results of the analysis are presented in Table 2. The table provides information on the total number of opinions evaluated by GPT-3-Turbo, the total number of matches where the human-annotated stance aligns with the stance extracted by the LLM, and the total number of mismatches. Furthermore, two types of mismatches are described: instances where the LLM outputs *pro* while the expected output is *con* and instances where the output is *con* while the expected output is *pro*. Lastly, the table includes the total values across all policy options, as well as the average values.

An interesting result is the large number of *pro* opinions being falsely classified as *con* opinions in the YOUNG dataset. This is 10 times more than in the other data sets, and these 52 falsely classified opinions account for 60% of the total mismatched stances. Some examples are given in Table 3. Due to these findings, an examination of the opinion texts associated with these results was conducted. The analysis revealed that some opinions might be too ambiguous, to the point where even humans could have a hard time classifying the stance. Another reason for the false classifications might be that for this dataset, the opinion texts often contain negative expressions to describe a positive stance against the policy option. An example of such an opinion is: "*This is unfeasible and inhumane for these groups to sustain for a long time. Very harmful to their development*". Here the subject describes the current measures as unfeasible and inhumane. Therefore, allowing young people to come together in small groups is a good idea. However, this is a complex opinion for non-humans to understand and thus can be easily misclassified as a wrong stance. To address such errors, one approach that may prove beneficial is to ensure that the input provided to the model is completely unambiguous. This can help minimize any potential sources of confusion that might lead to inaccuracies in the model's outputs. However, this might not always be feasible depending on the context, like meeting data.

Opinion	Ann. Stance
This yields little profit and hardly enforceable for young people	Pro
They can't do that anyway and only frustrate everyone	Pro
To avoid difficulties such as domestic violence and stress.	Con
Young people do not become seriously ill under normal circumstances. They can provide group immunity, but they can't if you raise everyone.	Pro
This measure has a limited influence and is difficult to maintain"	Pro

Table 3: Examples of misclassifications of the YOUNG dataset.

4.2 Metrics

Table 4 presents the previously mentioned performance metrics. The metric values are categorized by different policy options, which are further divided into the two stances. Notably, for each metric, the policy-stance pair with the highest value is highlighted in bold. Additionally, the average values of each metric are mentioned at the end.

Overall the results of the metrics are quite good. Most of the values are above 0.9 which, is considered very good. However, as previously mentioned, the values for the YOUNG dataset concerning the *con* stance do contain an outlier. This high misclassification rate directly influenced the metrics and thus scored significantly lower than the other instances. Besides this outlier, the scores of the other datasets are similar.

As previously mentioned, in this specific use case, the F-score is considered the most appropriate metric for evaluation. With an average score of 0.895, it can be concluded that the LLM demonstrates a good ability to detect the stance of an opinion concerning a specific policy. If we compare the results to the highest rankings, as shown in an overview presented in [2], we can see that the LLM outperforms other methods on multiple different datasets. In addition to the findings of this study, it would be valuable to compare the performance of the LLM to other methods using well-established benchmark datasets, such as Semeval-16 Task 6 [13]. While the LLM showcased promising results in stance detection for COVID-19 policy opinions, evaluating its performance on a widely-used dataset like Semeval-16 Task 6 would provide a more comprehensive and objective assessment.

	Total	Match	Mismatch	Pro (Expected: Con)	Con (Expected: Pro)
YOUNG	479	423	56	4	52
IMMUNE	361	343	18	13	5
REOPEN	453	440	13	8	5
Total	1293	1206	87	25	62
Average	431	402	29	8.33	20.67

Table 2: Results of stance detection for the opinion corpus using GPT-3-Turbo.

It is important to note that the model's performance can vary significantly, as evident in the YOUNG dataset. Therefore, when employing Large Language Models for stance detection, it is crucial to exercise caution. While LLMs can serve as valuable guidelines, it is essential to verify their outputs to identify consistent errors. In critical scenarios such as the COVID-19 policy, relying solely on an LLM may not be deemed reliable enough, given that incorrectly annotated opinions can have significant consequences. In spite of the lower reliability, the usage of LLMs does reduce the time taken to annotate the opinions significantly when compared to manual annotation. For this sample size, the model took about 3 minutes without the need to recruit annotators.

5 Responsible Research

Given that we rely on data contributed by human participants and explore potential applications of LLMs, it becomes crucial to conduct a thorough evaluation of the ethical considerations associated with the research.

It is noteworthy that the research plan of the HyEnA method, responsible for generating the datasets employed in this study, received approval from an Ethics Committee [25]. Additionally, the research plan for the COVID opinion collection was also approved by the Ethics Board of the Delft University of Technology [14].

5.1 Human annotators

For the creation of the datasets, human annotators are used. However, it is important to note that this paper did not directly hire the used crowd workers for annotation. Instead, the dataset utilized in this paper was obtained from the HyEnA research [24], where 348 crowd workers were recruited through the crowd-sourcing platform Prolific [17]. Therefore, the dataset employed in this study is a result of the efforts undertaken by the HyEnA research.

The first concern is payment. A study has shown that in some crowd-sourcing platforms like Amazon Mechanical Turk, crowd workers get paid an extremely low amount of money, which can be as low as 30% of the minimum wage [6]. This is fortunately not possible when using Prolific, which has a minimum hourly income of £6,- (€6,96). For this research, each crowd worker was paid £7,50 per hour (€8,52 with respect to 2020). This is considered a fair price for crowd workers. It is lower than the minimum wage in countries in Europe like The Netherlands, where the wage was €9,70 in 2020 (for ages 21+) [18], but much higher in other countries like Bulgaria, where the wage was about €2.15 in 2020 [4].

An additional issue to consider is the potential fatigue experienced by crowd workers over time. Research indicates that workers on crowdsourcing platforms can experience significant fatigue after just one hour of work [30]. In the context of the HyEnA research, the task of Key Argument Annotation was designed to be feasible within a maximum time limit of one hour. By implementing this restriction, the researchers decreased the risk of excessive fatigue among the workers.

5.2 LLM usage

As previously stated, the initial problem addressed decision-making on public matters such as COVID-19. The incorporation of Language Models (LLMs) can significantly enhance the effectiveness of these decision-making processes. Nevertheless, these models are not flawless and may occasionally yield incorrect outputs. Additionally, ensuring accountability for the use of LLMs can be complex. Therefore, it is crucial to evaluate the challenges associated with employing this LLM-based approach for significant and consequential decisions. It is recommended to have an expert review the outputs generated by the LLM in order to guarantee fairness and ethical decision-making.

Another potential problem with the usage of LLMs is the potential bias. Experiments using ChatGPT show that outputs of a language model can have biases towards specific targets, as shown in [29]. They show that a significant number of neutral tweets with stances towards a target like "legalization of abortion" would be classified as *pro*. Even though the COVID-19 dataset does not contain any neutral opinions, it is hard to detect whether misclassified stances are the result of such biases.

Finally, it is essential to review the privacy statement associated with the utilized model/API. The data utilized in this research was sourced from publicly available information and did not include any personal or re-identifiable data. However, it is crucial to recognize that this may not hold for other datasets. It is always important to consider that the information provided to the model may be utilized for alternative purposes. For instance, according to the OpenAI Data Usage Policy: "OpenAI will not use data submitted by customers via our API to train or improve our models, unless you explicitly decide to share your data with us for this purpose" [16]. This demonstrates that the usage of the OpenAI API is relatively secure. Nonetheless, the data provided is retained for a period of 30 days and can be accessed by authorized OpenAI employees, which still entails the risk of potential data exposure, such as through a data breach.

	Stance	Metrics			
		Accuracy	Precision	Recall	F-Score
YOUNG	*pro*	0.883	**0.989**	0.876	0.929
	con		0.514	0.932	0.663
IMMUNE	*pro*	0.950	0.814	0.920	0.864
	con		0.983	0.957	0.969
REOPEN	*pro*	**0.971**	0.969	**0.981**	**0.975**
	con		0.974	0.959	0.967
Average		0.934	0.874	0.938	0.895

Table 4: Performance metrics of stance detection using GPT-3-Turbo.

6 Future Work and Limitations

In this research, the GPT3-Turbo model was utilized. However, the newer model GPT-4 by OpenAI has shown promising results. It has been demonstrated that GPT-4 outperforms any GPT-3 model in multiple fields [7, 10]. Access to the API of the GPT-4 model was not available at the time of this study. Therefore, it is suggested to investigate this newer model, along with other promising Large Language Models, to analyze their performance.

Another possibility for future work is the utilization of multiple datasets. This research focuses solely on COVID-19 related opinions but can be applied to any referendum-like dataset. One could analyze the performance of LLMs on different topics.

Lastly, the involvement of human annotators could be considered in research. In the dataset used for this study, the individuals providing their opinions also annotated their stances. By recruiting human annotators to manually annotate each opinion, one could examine the differences in stance detection between human-generated and LLM-generated annotations. When using human annotators, it is also important to calculate the inter-annotator agreement. If the results are similar, opting to use an LLM in situations where no annotations are available could be a potential approach for future research.

7 Conclusion

The purpose of this study is to examine the efficiency of Large Language Models (LLMs) in determining the stance of an opinion regarding a particular policy option. We specifically focused on the application of LLMs to large datasets containing COVID-19-related opinions concerning the relaxation of regulations. Our findings indicate that the LLM shows a strong performance, achieving an average F-score of 0.895 in a short time span. Nevertheless, there were instances where the LLMs did not perform as effectively, mainly due to ambiguity in the opinion text. Thus it is important to consider these factors when utilizing LLMs.

References

[1] Aseel Addawood, Jodi Schneider, and Masooda Bashir. Stance classification of twitter debates: The encryption debate as a use case. pages 1–10, 07 2017.

[2] Rong Cao, Xiangyang Luo, Yaoyi Xi, and Yaqiong Qiao. Stance detection for online public opinion awareness: An overview. *International Journal of Intelligent Systems*, 37(12):11944–11965, 2022.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[4] CountryEconomy.com. National minimum wage - bulgaria, 2020.

[5] Marcelo Dias and Karin Becker. INF-UFRGS-OPINION-MINING at SemEval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 378–383, San Diego, California, June 2016. Association for Computational Linguistics.

[6] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI

'18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[7] Anis Koubaa. GPT-4 vs. GPT-3.5: A Concise Showdown. 4 2023.

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

[9] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online, August 2021. Association for Computational Linguistics.

[10] John C. Lin, David N. Younessi, Sai S. Kurapati, Oliver Y. Tang, and Ingrid U. Scott. Comparison of gpt-3.5, gpt-4, and human user performance on a practice ophthalmology written examination. *Eye*, 2023.

[11] Liran Liu, Shi Feng, Daling Wang, and Yifei Zhang. *An Empirical Study on Chinese Microblog Stance Detection Using Supervised and Semi-supervised Machine Learning Methods*, page 753–765. Lecture Notes in Computer Science, 2016.

[12] Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker. NLDS-UCSC at SemEval-2016 task 6: A semi-supervised approach to detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, 2016.

[13] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 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, June 2016. Association for Computational Linguistics.

[14] Niek Mouter, Jose Ignacio Hernandez, and Anatol Valerian Itten. Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures. *PLOS ONE*, 16(5):e0250614, 2021.

[15] OpenAI. Openai models. https://platform.openai.com/docs/models/gpt-3-5, 2023.

[16] OpenAI. Openai policies - api data usage policies. https://https://openai.com/policies/api-data-usage-policies, 2023.

[17] Prolific Academic Ltd. Prolific, 2023.

[18] Rijksoverheid. Bedragen minimumloon 2020, 2020.

[19] SemEval. Semeval. https://semeval.github.io/, 2023.

[20] Chongtham Rajen Singh and R. Gobinath. *Hypothesis Testing of Tweet Text Using NLP*, page 95–108. Lecture Notes on Data Engineering and Communications Technologies, 2023.

[21] Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34, Antwerp, Belgium, September 2007. Association for Computational Linguistics.

[22] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[23] Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic, November 2021. Association for Computational Linguistics.

[24] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. Hyena: A hybrid method for extracting arguments from opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pages 17–31, Amsterdam, the Netherlands, 2022. IOS Press.

[25] Michiel van der Meer, Enrico Liscio, Catholijn M Jonker, Aske Plaat, Piek Vossen, and Pradeep K Murukannaiah. Hyena: A hybrid method for extracting arguments from opinions: Supplementary material. https://liacs.leidenuniv.nl/ meermtvander/publications/hyena/, 3 2022.

[26] David Vilares and Yulan He. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[27] Marilyn Walker, Pranav Anand, Rob Abbott, Jean Fox Tree, Craig Martell, and Joseph Eichenbaum. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53:719–729, 11 2012.

[28] Jiaming Xu, Suncong Zheng, Jing Shi, Yiqun Yao, and Bo Xu. Ensemble of feature sets and classification methods for stance detection. volume 10102, pages 679–688, 12 2016.

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

[30] Ying Zhang, Xianghua Ding, and Ning Gu. Understanding fatigue and its impact in crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Sup-*

ported Cooperative Work in Design ((CSCWD)), pages 57–62, 2018.

Appendix A Input prompt

```
You will be provided with a list of opinions. The input will be a JSON formatted list.
Each element (opinion) contains an id and text. For each element (opinion) in the JSON
list, evaluate the stance of that opinion against this statement: Young people may come
together in small groups. The stance can only be 'pro' or 'con'. Give the output in another
JSON formatted list where each element of the list should contain the id of the input element
and a stance field which contains the stance of the opinion.

    Constraint: Even if uncertain about the stance you must pick either 'con' or 'pro'

    Output example: {
        "output": [
            {
            "id": 947,
            "stance": 'con'
            },
            {
            "id": 936,
            "stance": "pro"
            }
        ]
    }

    Input:

    {'opinions': [{'id': 12821, 'text': 'This is not already ...the virus.'}]}
```