FC_RUG at CheckThat! 2024: Few-Shot Learning Using GEITje for Check-Worthiness Detection in Dutch

Notebook for the CheckThat! Lab Task 1 at CLEF 2024

Sanne Weering[1], Tommaso Caselli[1,†]

[1]Rijksuniversiteit Groningen, Oude Kijk in't Jatstraat 26, 9712 EK Groningen, The Netherlands

Abstract

This contribution presents our approach for the CheckThat! 2024 Lab Task 1: Check-worthiness estimation. We focused on testing the abilities of `GEITje`, a large language model for Dutch based on `Mistral-7B`. We have experimented with different prompts varying the learning settings (zero-shot *vs.* few-shot) and the personas (helpful assistant *vs.* fact-checker). We selected our best model (helpful assistant with few-shot in-context learning) on the basis of the development data from the companion task of the CheckThat! 2022 Lab edition. We obtained a macro-F1 score of 0.657 and a F1-score on the positive class 0.594, ranking #6 out of 15 participants.

Keywords

Check-worthiness detection, Zero-shot learning, Few-shot learning, LLM, GEITje

1. Introduction

In our current digital age, where information and disinformation spread rapidly, it is essential to be able to assess the reliability of their content. Misinformation and fake news can have a serious impact on the public opinion and on decision-making processes [1]. Fact-checking is crucial in combating misinformation. Nowadays numerous fact-checking organizations exists[1] and yet fact-checking is mostly a manual activity conducted by a limited number of experts. These manual efforts are overwhelmed by the sheer volume of misinformation on online platforms. Tools based on machine learning techniques can help human fact-checkers by speeding up the fact-checking process. An area of potential useful impact is check-worthiness estimation.

The CLEF 2024 CheckThat! Task 1: Check-worthiness estimation [2] has been specifically designed to this end. The task is part of a battery of five additional tasks which target misinformation from different perspectives, including subjectivity of the message, persuasion techniques, rumor verification, and robustness to adversarial attacks [3].[2] Task 1 is framed as a binary classification, whose goal is to assign the check-worthiness status to a given message either extracted from Twitter/X or another source. The task is offered in Arabic, Dutch, and English. We have only focused on Dutch.

This paper presents our approach which aimed at investigating the capabilities of a recent monolingual large language model (LLM) for Dutch, `GEITje` [4]. We have conducted multiple experiments using the development data from the CheckThat! 2022 companion task on check-worthiness estimation [5]. While the annotation guidelines and the social media platform have remained the same [6], the Dutch data for the two editions differ for their topics: the 2022 edition is focused on COVID-19 while the 2024 data addresses climate change. The topic shift can pose an extra challenge for the LLM.

The remainder of the paper is organized as follows: Section 2 presents a short overview of the data used in our approach. In Section 3, we present a description of how `GEITje` has been developed, the specific model we have selected for the task, our prompts and their evaluation on the 2022 development data. We also present a detail description of the post-processing tasks required in order to extract the

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

† Corresponding author.

s.weering@student.rug.nl (S. Weering); t.caselli@rug.nl (T. Caselli)

0000-0003-2936-0256 (T. Caselli)

[1] https://edmo.eu/resources/repositories/fact-checking-organisations-in-the-eu/

[2] https://checkthat.gitlab.io/clef2024/.

answers for the official submission format, an often neglected component in recent work using LLMs for classification tasks. In Section 4, we report on the official results on the 2024 test set and to get a better understanding of what went wrong, we have conducted an error analysis. Lastly, Section 5, presents our conclusions and directions for future work.

2. Data

For Dutch, the lab organizers have provided training, development and test data from the CheckThat! 2022 edition for system development, and a new test set for this edition. A summary of the available data (and label distribution) is presented in Table 1. Across all datasets, the labels are not perfectly balanced, with a higher presence of the **not check-worthy** class (around 60% of the messages).

Table 1
Distribution of the CheckThat! 2024 Task 1 data for Dutch. Numbers in parentheses correspond to the percentages of the classes.

	Check-worthy	Not check-worthy	Total
CT22-Train	377 (40.85%)	546 (59.15%)	923
CT22-Development	28 (38.89%)	44 (61,11%)	72
CT22-Test	102 (40.48%)	150 (59.52%)	252
CT24-Test	397 (39.70%)	603 (60.30%)	1000

The 2022 data consist of tweets collected using keywords related to COVID-19 and the debate around COVID vaccines in a time period spanning between January 2020 till March 2021. The official 2024 test set has been built by extracting messages from Twitter/X but targeting a different topic (climate change and its associated debate) and from a different time period (January 2021 – December 2022), minimally overlapping with the data from the 2022 edition. The difference in time period and topic can pose an additional challenge for the model, although this can also offer insights on their robustness and portability.

3. Method and Prompts

In this section, we will provide additional information on how `GEITje` has been developed, the prompts, the post-processing steps needed to extract the answers, and the evaluation to select the best approach.

`GEITje` Model: `GEITje` is the first monolingual LLM for Dutch. The foundation model is based on `Mistral-7B` and has been obtained by training the model via full-parameter finetuning on 10 billion tokens from the Dutch Gigacorpus [3] and the Dutch portion of the MADLAD-400 webcrawling corpus [7]. The foundation model has been further aligned to follow instructions, answer questions, and hold dialogues resulting in `GEITje-7B-ultra` [8]. This model is an improved iteration of a previous version (`GEITje-7B-ultra-sft`) which was obtained using multiple datasets (Alpaca, Databrick-Dolly-15k, and Stack Overflow) automatically translated to Dutch using Open AI `gpt-3.5-turbo` and newly generated using `gpt-4-turbo` via Azure (no robots, ultrachat). `GEITje-7B-ultra` further finetunes on top of `GEITje-7B-ultra-sft` version using Direct Preference Optimization (DPO) with a 56 million tokens synthetic dataset obtained via `gpt-4-turbo` and `GEITje-7B-chat`. This version of `GEITje` has shown better performance in comparison to all other `GEITje`-based models when evaluated against the Dutch portion of the Open Multilingual LLM Evaluation Leaderboard [9].

[3] http://gigacorpus.nl

Prompts LLMs are known to present different behaviors according to the prompt they are subjected to [10]. Their performance is further influenced by two additional elements. The first is the persona variable. LLMs can be prompted to impersonate individuals with different expertise and socio-demographic backgrounds. Recent work has focused on assessing to what extent LLMs can simulate human behaviors when personas are included and how these variables contribute to the solution of a task [11, 12, 13]. The second is the exposure to examples, i.e., in-context learning. Although there is a relationship between models' size and performance, in-context learning has a major impact in the final performance of LLMs, especially when compared against zero-shot experiments [14].

Considering these factors, we have conducted a preliminary set of experiments to identify the most promising wording of the prompts. After this, we have devised four prompts by combining the following options: persona variables and in-context learning. All prompts contain a basic description of the task (i.e., whether a message is check-worthy or not) plus further specifications. Given that the check-worthiness of a message can depend on many factors, we decided to further explain the check-worthiness by introducing additional variables from the original annotation guidelines [15]. In particular, we specified that a message must be factual and potentially contains harmful content for the society. In this way, we combine two key questions for check-worthiness from the annotation guidelines.

For the persona variables, we experimented leaving it to its default value (i.e., a helpful assistant) or changing it to be a fact-checker assistant that has to decide whether a message must be checked or not. For the in-context learning, we have explored the use of zero-shot (i.e., no additional examples) and few-shot settings. For the few-shot setting, we have extracted six instances from the CheckThat! 2022 training data equally balanced between positive and negative classes. Finally, all prompts have been devised in such a way to force the model to return a structured output to make the extraction of the labels easier. In Table 2 we report the basic prompts. The original Dutch versions are in Appendix A for readability's sake. For the the few-shot experiments, the prompts presents the six examples before the instruction to classify a given message (i.e., "Classify the check-worthiness of the following tweet:")

Table 2
Basic prompts with the default and fact-checker persona variables.

Default Persona	Fact-checker Persona
Assess whether tweets should be fact-checked. Tweets are verifiable only if they contain a verifiable factual assertion and if that assertion could be harmful. Choose one of the following labels [Yes, No]. Give the answer in the following format with the tweet between the tag [TWEET], the label between the tag [LABEL] and the explanation between the tag [EXPLANATION]: [Tweet: [TWEET], Label: [LABEL], Explanation: [EXPLANATION]] Classify the check-worthiness of the following tweet:	You are a fact checker assistant with the task of identifying messages that need to be fact-checked. Assess whether tweets should be fact-checked. Tweets are verifiable only if they contain a verifiable factual assertion and if that assertion could be harmful. Choose one of the following labels [Yes, No]. Give the answer in the following format with the tweet between the tag [TWEET], the label between the tag [LABEL] and the explanation between the tag [EXPLANATION]: [Tweet: [TWEET], Label: [LABEL], Explanation: [EXPLANATION]] Classify the check-worthiness of the following tweet:

Our prompts have been designed in such a way to minimize the impact of the potential explanations that the `GEITje` model could provide. As a matter of fact `GEITje-7B-ultra` is a chat-based models and this tends to results in the generation of verbose answers/explanations that may make it impossible to extract the required label. The presence of the tag [EXPLANATION] in our prompts serves this purpose, offering the models a dedicated "place" to provide the accompanying explanation(s).

Post-processing steps: Label extraction Once we have collected all the answers from the model, we have run a basic Python script that identify the [LABEL] tag and extract the associated answers. This approach worked for almost all cases, except two instances across all four prompts where the model failed to properly generate the answer in association with the [LABEL] tag. In one case, the model fails to use one of the required labels presenting its own variations. For instance, as illustrated by example 1, the model provided only the initial of the required label, forcing us to adjust the output. In the other case, the model generated the answer as the last token in the explanation section (see example 2 below). To avoid unnecessary penalties, we extracted the answers from these sentences as well.

1. TWEET: [...] Label: J Uitleg: De tweet bevat een verifieerbare bewering [...]. Daarom is het label "J".
   [TWEET: [...] Label: Y Explanation: This tweet contains a verifiable factual claim [...]. Therefore the label is "Y'
2. [Tweet: [...] Label: [LABEL], Uitleg: [UITLEG] Deze tweet bevat een verifieerbare feitelijke bewering [...]. Daarom is het label in dit geval "Ja"].
   [Tweet: [...] Label [LABEL], Explanation: This tweet contains a verifiable factual claim [...]. Therefore the label is "YES"

Results on CT22 Development To select the best prompt and setting, we have evaluated the four prompts against the CT22 development set. A summary of the results is presented in Table 3.

Table 3
Results for on CT22 development. We report Precision, Recall, Macro-F1 score and F1score for positive class (check-worthy). Best scores are in bold.

Prompt	Precision	Recall	Macro-F1	F1 check-worthy
Zero-shot, default persona	0.399	0.373	0.374	0.580
Zero-shot, fact-checker assistant	0.323	0.310	0.314	0.413
Few-shot, default persona	**0.656**	**0.662**	**0.657**	**0.600**
Few-shot, fact-checker assistant	0.556	0.558	0.550	0.500

The results indicate that, in general, the zero-shot setting is a much more challenging scenario than the few-shot one and that the persona variable does not have a positive impact. In both settings, the use of the fact-checker persona results in lower scores both at macro-level and on the positive class. It is remarkable that the F1-scores for the positive class in the zero-shot setting are consistently higher than the macro-F1 scores. It appears that this behavior may be due to an overgeneralization of the positive class by the model as an effect of the prompt instructions. This seems to be confirmed by the results for the few-shot settings. In this cases, we observed a generalized higher macro-F1 score, indicating a better performance on the negative class. On the basis of these results, we opted to run on the 2024 test data, a few-shot learning model with the default persona.

4. Results on CT24 Test

In Table 4, we report the overview of the results on the CT24 test data. We report also the results of the baseline - provided by organizers and based on a random classifier - and the scores of the best system for comparison.

With an F1 score of 0.594 our submission, FC_RUG, has ranked #6 out of 15 participants for the Dutch data. Our few-shot learning approach has easily outperformed the baseline of the organizers but it struggles against the best system. The difference in performance is 0.138 points, clearly indicating a margin of improvement. If we compare this score to the one obtained on the development data, we do not observe a huge drop, suggesting a better portability of in-context learning models when compared to fine-tuned encoders or other supervised methods.

Table 4
Results on the CT24 test data. We report the official scores based on F1 on the positive class (check-worthiness) and the rank.

Model	F1 check-worthy	Rank
Best system	0.732	1
FC_RUG	0.594	6
Baseline	0.438	14

To gain a better understanding of the model's behavior we have conducted an error analysis on 100 misclassified messages from the CT24 test data. Our analysis did not limit itself to identify classes of errors but further took into consideration the explanations offered by the LLM.

At macro level, we have distinguished between False Positive (FP) errors (i.e., messages wrongly predicted as check-worthy) and False Negative (FN) errors (i.e., messages wrongly predicted as not-check-worthy). In general the distribution of the errors between these two broad categories is quite balanced, with 48 FN instances and 52 FP cases. In addition to this, we have identified three fine-grained classes of errors in common, namely: (*i*) assessment of verifiable and harmful claims; (*ii*) labels not aligned with the explanation; (*iii*) classification of a message from the prompt. A summary of the distribution of these errors in presented in Table 5.

Table 5
Errors that are made by the model divided into different categories with their corresponding counts.

Error class	False Positive	False Negatives
Assessment of verifiable and harmful claim	36	38
Label not aligned with explanation	12	1
Classification of example from prompt	4	8

The majority of errors concerns the assessment of the check-worthiness of the message. For the FN instances, a deeper inspection has identified that in 18 cases, the model fails completely to identify the presence of a verifiable claim, and in 20 other instances it considers the message as not check-worthy because the content is not considered to be harmful, clearly an error induced by our formulation of the prompt. This is an error that is also present in the FP instances - where 18 cases are considered not check-worthy because deemed not harmful. On the contrary, the other 18 instances are wrongly assessed as containing verifiable and harmful content when they are not.

For the other errors, it seems that the FPs tend to give rise to a dis-alignment between the proposed labels and the explanations - which correctly assess the non check-worthiness of the messages. Lastly, both classes of errors present classification of messages that are in the prompts rather than the one that is proposed. Notably, the classified example is consistently either the first or last instance in the prompt.

5. Conclusions & Future work

This paper presents our approach at using a monolingual LLM for Dutch chat- and instruction-tuned, `GEITje-7B-ultra`, to assess the check-worthiness of tweets. We explored different prompt designs and settings, including zero *vs.* few-shot learning, and modification of the persona variable (default helpful assistant *vs.* fact-checking assistant). We used the data from the CheckThat! 2022 companion task to identify the instances in the few-shot scenario and to evaluate our best prompt settings, which resulted in a few-shot with default persona. Our submission obtained a macro-F1 score of 0.657 and a F1 score over the positive class of 0.594, ranking #6 over 15 participants. Moving forward, continued experimentation with different prompt structures and formulations can help identify more effective prompts for check-worthiness assessment. For example, it could be interesting to see the effect on the

results when using others examples in the prompts. Additionally, it would be very useful to improve the process of label extraction.

Acknowledgments

All experiments have been conducted on the Hábrók HPC Cluster of the University of Groningen.

References

[1] C. López-Marcos, P. Vicente-Fernández, Fact Checkers Facing Fake News and Disinformation in the Digital Age: A comparative analysis between Spain and United Kingdom, Publications 9 (2021). URL: https://www.mdpi.com/2304-6775/9/3/36. doi:10.3390/publications9030036.

[2] M. Hasanain, R. Suwaileh, S. Weering, C. Li, T. Caselli, W. Zaghouani, A. Barrón-Cedeño, P. Nakov, F. Alam, Overview of the CLEF-2024 CheckThat! Lab Task 1 on Check-Worthiness Estimation of Multigenre Content, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[3] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[4] E. Rijgersberg, B. Lucassen, GEITje: een groot open Nederlands taalmodel, 2023. URL: https://github.com/Rijgersberg/GEITje.

[5] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. Kartal, Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets, in: CLEF 2022: Conference and Labs of the Evaluation Forum, volume 3180 of *CEUR Workshop Proceedings*, CEUR Workshop Proceedings (CEUR-WS.org), 2022, pp. 368–392.

[6] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 611–649. URL: https://aclanthology.org/2021.findings-emnlp.56. doi:10.18653/v1/2021.findings-emnlp.56.

[7] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, O. Firat, MADLAD-400: A Multilingual and Document-level Large Audited Dataset, Advances in Neural Information Processing Systems 36 (2024).

[8] B. Vanroy, GEITje-ultra-sft, 2024. URL: https://huggingface.co/BramVanroy/GEITje-7B-ultra-sft.

[9] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2021. URL: https://doi.org/10.5281/zenodo.5371628. doi:10.5281/zenodo.5371628.

[10] A. Salinas, F. Morstatter, The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance, 2024. arXiv:2401.03729.

[11] G. V. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies, in: International Conference on Machine Learning, PMLR, 2023, pp. 337–371.

[12] T. Hu, N. Collier, Quantifying the Persona Effect in LLM Simulations, 2024. arXiv:2402.10811.
[13] C. Sun, K. Yang, R. G. Reddy, Y. R. Fung, H. P. Chan, C. Zhai, H. Ji, Persona-DB: Efficient Large Language Model Personalization for Response Prediction with Collaborative Data Refinement, arXiv preprint arXiv:2402.11060 (2024).
[14] S. Lu, I. Bigoulaeva, R. Sachdeva, H. T. Madabushi, I. Gurevych, Are Emergent Abilities in Large Language Models just In-Context Learning?, 2023. arXiv:2309.01809.
[15] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. D. S. Martino, A. Abdelali, H. Sajjad, K. Darwish, P. Nakov, Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms, 2021. URL: https://europepmc.org/article/PPR/PPR310594.

Appendix A - Prompts

Zero-shot, default persona

> Beoordeel of tweets gecontroleerd moeten worden op de feiten. Tweets zijn alleen controleerbaar als ze een verifieerbare feitelijke bewering bevatten en als die bewering schadelijk kan zijn. Kies een van de volgende labels [Ja, Nee]. Geef het antwoord in het volgende formaat waarbij de tweet tussen de tag [TWEET], het label tussen de tag [LABEL] en de uitleg tussen de tag [UITLEG] staat: [Tweet: [TWEET], Label: [LABEL], Uitleg: [UITLEG]] Classificeer de controleerbaarheid van de volgende tweet:

Zero-shot, fact-checker assistant

> Je bent een factchecker-assistent met de taak om zinnen te identificeren die op feiten gecontroleerd moeten worden. Tweets zijn alleen controleerbaar als ze een verifieerbare feitelijke bewering bevatten en als die bewering schadelijk kan zijn. Kies een van de volgende labels [Ja, Nee]. Geef het antwoord in het volgende formaat waarbij de tweet tussen de tag [TWEET], het label tussen de tag [LABEL] en de uitleg tussen de tag [UITLEG] staat: [Tweet: [TWEET], Label: [LABEL], Uitleg: [UITLEG]] Classificeer de controleerbaarheid van de volgende tweet:

Few-shot, default persona

> Beoordeel of tweets gecontroleerd moeten worden op de feiten. Tweets zijn alleen controleerbaar als ze een verifieerbare feitelijke bewering bevatten en als die bewering schadelijk kan zijn. Kies een van de volgende labels [Ja, Nee]. Geef het antwoord in het volgende formaat waarbij de tweet tussen de tag [TWEET], het label tussen de tag [LABEL] en de uitleg tussen de tag [UITLEG] staat: [Tweet: [TWEET], Label: [LABEL], Uitleg: [UITLEG]] Hier zijn enkele voorbeelden:
>
> Tweet: "RTLnieuws Het #RIVM en het kabinet MinPres hebben via nalatig en gebrekkig beleid #Nederland gebracht in de wereldwijde top van meeste doden per inwoner en nu draait de #propaganda machine op volle toeren zodat de #VVD nog harder kan stijgen in de peilingen. Lijkt #NoordKorea wel.", Label: Ja, Uitleg:
> Tweet: "Epidemioloog: "Coronavirus kan 60% van de wereld besmetten" https://t.co/fwOozlC9QV", Label: Ja, Uitleg:
> Tweet: "Een andere aanpak om #Corona in te dammen: effectiever, minder schadelijk voor de economie, maar wel de privacy in het geding: 'Testen, opsporen, isoleren' https://t.co/dWTulQZ6n2", Label: Ja, Uitleg:
> Tweet: "Maatregelen coronavirus: zorgen bij Gooiland nemen verder toe - https://t.co/ao4jSo5Ze5", Label: Nee, Uitleg:
> Tweet: "RIVM heel laat vandaag met publicatie cijfers. Om 14:30 nog niets en website overbelast", Label: Nee, Uitleg:
> Tweet: "Je kan in Amsterdam beter een vliegtuig hebben dan een auto. #gratisparkeren

#COVID19NL", Label: Nee, Uitleg:
Classificeer de controleerbaarheid van de volgende tweet:

Few-shot, fact-checker assistant

Je bent een factchecker-assistent met de taak om zinnen te identificeren die op feiten gecontroleerd moeten worden. Tweets zijn alleen controleerbaar als ze een verifieerbare feitelijke bewering bevatten en als die bewering schadelijk kan zijn. Kies een van de volgende labels [Ja, Nee]. Geef het antwoord in het volgende formaat waarbij de tweet tussen de tag [TWEET], het label tussen de tag [LABEL] en de uitleg tussen de tag [UITLEG] staat: [Tweet: [TWEET], Label: [LABEL], Uitleg: [UITLEG]] Hier zijn enkele voorbeelden:
Tweet: "RTLnieuws Het #RIVM en het kabinet MinPres hebben via nalatig en gebrekkig beleid #Nederland gebracht in de wereldwijde top van meeste doden per inwoner en nu draait de #propaganda machine op volle toeren zodat de #VVD nog harder kan stijgen in de peilingen. Lijkt #NoordKorea wel.", Label: Ja, Uitleg:
Tweet: "Epidemioloog: "Coronavirus kan 60% van de wereld besmetten" https://t.co/fwOozlC9QV", Label: Ja, Uitleg:
Tweet: "Een andere aanpak om #Corona in te dammen: effectiever, minder schadelijk voor de economie, maar wel de privacy in het geding: 'Testen, opsporen, isoleren' https://t.co/dWTulQZ6n2", Label: Ja, Uitleg:
Tweet: "Maatregelen coronavirus: zorgen bij Gooiland nemen verder toe - https://t.co/ao4jSo5Ze5", Label: Nee, Uitleg:
Tweet: "RIVM heel laat vandaag met publicatie cijfers. Om 14:30 nog niets en website overbelast", Label: Nee, Uitleg:
Tweet: "Je kan in Amsterdam beter een vliegtuig hebben dan een auto. #gratisparkeren #COVID19NL", Label: Nee, Uitleg:
Classificeer de controleerbaarheid van de volgende tweet: