



Universiteit
Leiden
The Netherlands

Opleiding Informatica

Comparing human opinions to GPT-3.5 generated opinions
on Dutch news articles

Rayden Tseng, s2641240

Supervisors:

Suzan Verberne (LIACS) & Peter van der Putten (LIACS)

Bachelor Thesis

Leiden Institute of Advanced Computer Science (LIACS)

www.liacs.leidenuniv.nl

09/07/2023

Abstract

ChatGPT (GPT-3.5) has drawn significant attention since its release. Its abilities have already been investigated for a wide variety of tasks in NLP. In this research, we investigate to what extent GPT-3.5 can generate human-like comments on Dutch news articles and what the best way is to do so. Our methods mainly analyze human likeness across multiple prompting techniques. In particular, we utilize zero-shot, few-shot and context prompts, each for two generated personas. We found that our fine-tuned BERT models easily distinguish human-written comments from GPT-3.5 generated comments, with none of the used methods performing noticeably better. Other findings were that providing the model with a persona did not result in different output and human comments consistently showed higher lexical diversity.

Contents

1	Introduction	1
1.1	Thesis overview	1
2	Background and related work	2
2.1	Transformers	2
2.2	Large Language Models and GPT-3.5	3
2.3	BERT	3
2.3.1	RobBERT	4
2.4	Previous research	4
3	Methods	5
3.1	Data collection	5
3.1.1	Human opinions	5
3.2	Generating opinions with GPT-3.5	6
3.2.1	Persona	7
3.2.2	Zero-shot	7
3.2.3	Few-shot	7
3.2.4	Context	8
3.2.5	Limitations	8
3.3	Evaluation through classification	9
3.3.1	Fine-tuning	9
4	Results	10
4.1	Output	10
4.2	Classification results	11
4.3	Similarity between comments	13
4.4	Lexical Diversity	16
4.5	Qualitative analysis	17
4.5.1	False positives	17
4.5.2	False negative	20
5	Discussion	21
5.1	Analysis of results	21
5.2	Limitations	22
6	Conclusions	23
	References	25
A	Confusion Matrices	25
B	Prompts and completions	27
C	Human output	29

1 Introduction

Natural Language Processing (NLP) is an interdisciplinary subfield of linguistics, computer science and artificial intelligence concerned with the interactions between computers and human language. In particular, it focuses on how to program computers to process and analyze large amounts of natural language data. Natural Language Processing has had remarkable advancements in recent years, especially with the development of Large Language Models (LLMs). For example, OpenAI’s GPT models have the ability to process and generate natural language texts, which are generally hard to distinguish from human written texts. In 2020, GPT-3 took the previous model GPT-2 to a whole new level. Due to its training data of 175 billion parameters, rather than the previous 1.5 billion, it could perform more advanced tasks such as writing code, summarizing text and translating languages [BMR+20].

ChatGPT, a new sibling of the GPT family, was finetuned for conversational interactions, allowing it to answer follow-up questions. It receives input via a prompt and generates a completion. With it being far more powerful and accessible than previous GPT models, it instantly drew great attention and raised questions concerning plagiarism, its impact on humanity, the automation of jobs and the ethics of AI. Since the public availability of GPT-3.5 (ChatGPT), its capabilities have been researched for a high range of tasks. It has for instance shown remarkable performance in tasks like text summarization [ZLZ23], machine translation [HAS+23] and classification tasks, such as hate speech [CCA21] or sentiment [WXD+23]. Only once has it been used to generate tweets, after which humans had to classify whether a tweet was human-written or generated by the LLM [SBAG23]. Until now, no one has investigated to what extent GPT-3.5 can simulate human-written opinions on news articles. Detecting whether opinionated texts are generated by GPT-3.5 or written by a human could be of importance to help mitigate potential misuse of the model, such as the spread of disinformation by bad actors.

In this thesis, the capability of GPT-3.5 to produce opinionated text is investigated. The research first consists of collecting opinionated comments generated by the model and by humans. The human comments are crawled from a newspaper website. We prompt in multiple ways, hopefully resulting in different levels of human likeness. Subsequently, we analyze to what extent the GPT-3.5 generated comments can be distinguished from human comments. We accomplish this by fine-tuning a Dutch-based BERT model. BERT has achieved amazing results in many language understanding tasks and can be fine-tuned to obtain new state-of-the-art results for a lot of tasks such as text-classification [SQXH19]. By fine-tuning it with labeled comments (either human or GPT-3.5), it can be used to classify whether unlabeled comments are human-written or machine-generated. Secondly, we utilize other metrics to investigate the differences between the outputs. Therefore, the research question of this thesis is: “To what extent can GPT-3.5 generate human-like opinions that are similar to human opinions on news articles and what is the best way to generate these?”

1.1 Thesis overview

This thesis starts with some background information regarding this subject, GPT-3.5 and introduces our method briefly. Section 2 discusses some background information as well as related work with respect to this research. In Section 3, our methods are introduced and explained, giving the detailed implementation of our research. The results of the different experiments are shown in Section 4. Last, a conclusion, the limitations and potential further research is given in Section 6. This research

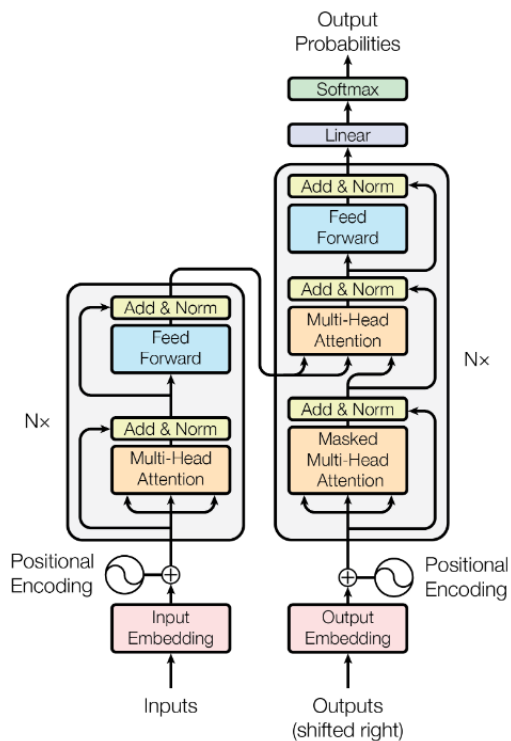


Figure 1: Model architecture [VSP+17]

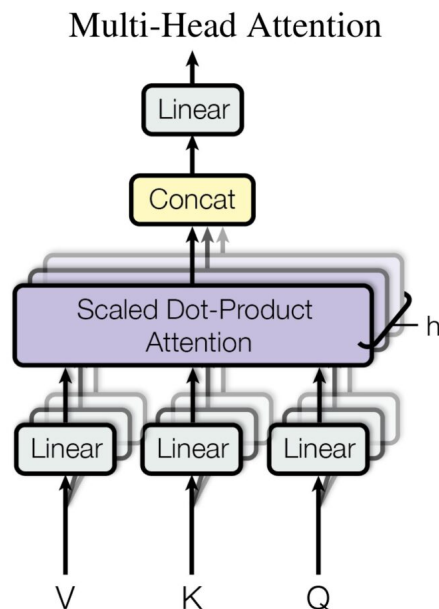


Figure 2: Multihead attention [VSP+17]

is a bachelor thesis at Leiden University, supervised by Dr. Suzan Verberne and Dr. Peter van der Putten from the Leiden Institute of Advanced Computer Science (LIACS).

2 Background and related work

In this section, background information as well as related work to this research is highlighted.

2.1 Transformers

Previous NLP tasks were performed by Recurrent Neural Networks (RNNs). RNNs have connections between their nodes which allow for cycles to occur in the network. It reads the input sequentially and as nodes are able to take information from prior inputs, the output of RNNs depends on prior elements in the sequence.¹ However, while processing larger texts, such models can become less effective. Since not all words are as relevant and with an increasing information chain, relevant information is more likely to get lost. Until June 2017, most NLP tasks were performed by RNNs. In the paper *Attention is all you need*, Vaswani et al. proposed a new network architecture, called the Transformer. This model was solely based on an attention mechanism and dispensed the recurrence and convolutions from earlier models, allowing for the Transformer to capture all contextual information in a sequence.

¹IBM - <https://www.ibm.com/topics/recurrent-neural-networks>

The transformer has an encoder-decoder structure. Its architecture is shown in Figure 1. It will, given a sequence, predict the next word (class) based on the highest calculated probability given the complete input context. The first step is feeding the inputs into a word embedding layer, mapping the words to a corresponding vector. Since the model does not have the ability to recur, positional embeddings must be added to the vectors, called positional encoding. Self-attention is the mechanism that relates tokens at different positions of a single sequence. It shows how relevant a word is to all other words in the same sequence. This is calculated by splitting the input into *Queries*, *Keys* and *Values*, through a process called Scaled Dot-Product Attention, as can be observed in Figure 2. By splitting the *Queries*, *Keys* and *Values*, this process can be performed in parallel and is called *Multi-Head Attention*. The scores afterwards are fed into an addition and normalization layer, which connects the original input to the scores and normalizes the values. After the Multi-Head Attention step, the input is fed into a Feed Forward layer and normalized again, such that the attention vectors are digestible by the new encoder or decoder block. The output of the encoder is a sequence of contextualized representations of all words from the input text.

During the training phase, the output sentence is fed to the decoder. It goes through the same embedding process, after which the attention is calculated. This is called Masked Self-Attention, meaning that the attention is only calculated for previous words in the sequence, rather than them all. The encoder-decoder attention stage uses both the outputs of the encoder and the decoder. This relates every token from both sequences. It is afterward sent to another Feed Forward network, after which a probability distribution can be calculated. The class with the highest probability will be the next token of the output sequence. Subsequently, the token is passed to the decoder input. This process repeats until the end of the sequence, resulting in a sequence of the most probable tokens. Recent work has shown that the Transformer is superior in quality while being more parallelizable and requiring less time to train compared to other models [VSP+17].

2.2 Large Language Models and GPT-3.5

Large Language Models are essentially the models that utilize the Transformer architecture, which enables them to capture contextual dependencies and generate human-like text. LLMs are general-purpose models, which are pre-trained on very large textual datasets to solve common language problems for a wide range of tasks, such as text classification, question answering, summarization and text generation. LLMs can be fine-tuned for specific tasks, by updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task [BMR+20].

The Generative Pre-trained Transformer 3.5 (GPT-3.5) is an example of an LLM. It was trained on 175 billion parameters, which is 10 times more than its predecessor GPT-2 [BMR+20]. Its architecture is decoder-only. This allows for open-ended generation. Unlike with encoder-decoder architectures, decoder-only models output is less scoped by the input which enlarges the space of acceptable output generations and therefore increases its potential for more diverse and creative responses [HBD+19]. Without any fine-tuning or gradient updates, it has shown strong performances in many NLP tasks on many datasets.

2.3 BERT

In 2019, Devlin et al. introduced a new language representation model called BERT, which stands for *Bidirectional Encoder Representations from Transformers*. In contrast to GPT-3.5, the BERT

model focuses on encoder representations, without a decoder component. Unlike traditional models that process text in a left-to-right or right-to-left manner, BERT takes into account both left and right context in all layers of the model [DCLT18]. BERT's pre-training makes use of a *Masked Language Model* (MLM). It randomly masks some tokens from the input, after which its objective is to predict the original token based on the context. MLM enables fusion between the left and right context, allowing to pre-train a deep bidirectional Transformer. On the other hand, it uses *Next Sentence Prediction* (NSP), which, given two sentences, determines whether the second sentence follows the first. MLM and NSP help BERT understand context across different sentences. It is relatively simple to fine-tune BERT for specific tasks, creating new state-of-the-art models. Since it is pre-trained, it already has a good language understanding and therefore fine-tuning only costs one additional output layer, allowing shorter training times.

2.3.1 RobBERT

RobBERT is the state-of-the-art Dutch BERT model. It is a large pre-trained general Dutch language model that can be fine-tuned on a given dataset to perform a wide range of NLP tasks [DWB20]. It uses a RoBERTa architecture and pre-training with a Dutch tokenizer. RoBERTa, a Robustly Optimized BERT Pre-training Approach modifies key hyperparameters, such as removing the NSP objective in pre-training, having much larger batches and training on longer sequences [LOG⁺19]. It has shown state-of-the-art results for various tasks, especially compared to other models when applied to smaller datasets [DWB20]. In this research, we use this model to fine-tune it for our specific use case.

2.4 Previous research

There has already been done some research on GPT-3.5's capabilities since the public availability. Many of these studies related to the capabilities of it. For instance, its capability to summarize has been investigated and compared with traditional fine-tuning methods on various benchmark datasets. This experimental analysis has shown that its extractive summarization performance is inferior compared to existing methods in terms of ROUGE score [ZLZ23]. Research by Hendy et al. has shown that GPT models have very competitive translation performance for high-resource languages, while still being limited for low-resource languages. It was also found that using a hybrid approach, which combines GPT models with other translators further enhances the quality of translations [HAS⁺23]. In addition to its capabilities to produce text, GPT has also been discovered for its ability to classify. Chiu et al. used zero-shot and few-shot prompting techniques to investigate whether GPT-3 can identify sexist or racist text. It has found that zero-shot averaged an accuracy between 55% and 67%, whereas few-shot learning achieved an accuracy that can be as high as 85% [CCA21]. An important remark here is that this was performed using the previous GPT-3 model, with GPT-3.5 potentially performing even better. Another preliminary study has investigated whether ChatGPT is a good sentiment analyzer, experimenting with different settings and methods. It has shown that ChatGPT exhibits impressive zero-shot performance in sentiment classification tasks and can rival a fine-tuned BERT model. Few-shot further enhances its performance, even surpassing fine-tuned BERT models in some cases [WXD⁺23]. Only once has GPT-3 been evaluated whether it could write human-like content in social media through the form of tweets. In this research, recruited individuals were asked to determine whether tweets were human-written or

Number	Title	Opinions	Date
1	Avondklok besproken als ‘serieuze optie’, maar invoering nog niet aan de orde	103	12/01/2021
2	Burgemeester Parijs: ‘Geen Russische atleten op Spelen zolang oorlog woedt’	143	03/02/2023
3	EU adviseert QR-code tot 9 maanden na laatste prik te laten gelden voor reizen	101	25/11/2021
4	Feyenoord-aanvoerder Kōkçü weigert vanwege religie regenboogband te dragen	194	16/10/2022
5	Jumbo stopt per direct met WK-reclamespot na storm van kritiek	144	02/11/2022
6	Minister rekent op 1.400 euro vergoeding voor studenten uit ‘pechgeneratie’	104	25/03/2022
7	Rusland valt Oekraïne aan, oorlog breekt uit	105	24/02/2022
8	Rutte biedt excuses aan voor slavernijverleden: ‘Aan alle nazaten tot hier en nu’	160	19/12/2022
9	Studentenorganisaties willen tijdelijke rem op komst internationale studenten	122	02/02/2023
10	Talpa wist volgens BOOS mogelijk al veel langer van misstanden bij The Voice	110	25/08/2022

Table 1: Titles of articles with number of comments and publication dates

machine-generated. It has shown that GPT-3, in comparison with humans, can produce accurate (dis)information that is easier to understand and that humans could not distinguish whether tweets were generated or written by humans [SBAG23].

3 Methods

In this section, we describe the methods that were used for this research. We provide a detailed, step-by-step description. This includes collecting the data through a parsing script. Secondly, we generate opinions with GPT-3.5 using the OpenAI API utilizing persona, zero-shot, few-shot and context methods. We investigated the difference in quality of the output by fine-tuning a Dutch-based BERT model. All code is written in Python, inside a virtual environment.

3.1 Data collection

In order to compare human opinions to artificial opinions, we chose specifically to collect opinions on Dutch news articles. The main requirement of an appropriate article was that it pertained a subject where the opinions generally differ, which would ideally lead to a high range of different opinions and perspectives on that topic. We made the choice to select the articles from *NU.nl*. This decision was based on the fact it features the integrated comment system *NUjij*. However, since an account was required to access the comments, each HTML page needed to be downloaded manually. For this research, the Chrome extension Save Page WE was used after having registered to the platform.² We collected a total of ten articles each containing at least a hundred comments. The articles and their corresponding amount of opinions are shown in Table 1.

3.1.1 Human opinions

Once the articles had been downloaded and stored in the same directory, the parsing script was written. Prior to this, the HTML structure needed to be analyzed. Our implementation only included parent comments, since sub-comments might deviate from the initial topic. We found that the comments were contained in the *coral-comment-content* class and the text in the *textblock*

²Save Page WE - <https://chrome.google.com/webstore/detail/save-page-we/dhhpefjklgkmgfeafimnjhojgjamof>

paragraph class. Firstly, using the BeautifulSoup³ library, the text was extracted and secondly the comments. Subsequently, both the text and comments were individually written to files in newly created directories. It appeared that the comments inside the HTML pages had an inconsistent structure, resulting in varying outcomes during scraping. While some articles were scraped without any issues, others showed HTML tags or unusual punctuation. This includes unusual `
`, `
`, `<div>` or `<p>` tags for no apparent reason, alongside inconsistent use of ' or ". Since there was no other way to resolve this, all inconsistencies had to be corrected manually. The pseudo-code is shown in Algorithm 1 and Algorithm 2. Since GPT-3.5 its training data is up to September 2021, an important remark here is that all articles except for article 1 were published after this date. Example human comments are shown in Appendix C.

Algorithm 1 Parsing comments

```

1: Initialize article_dir
2: Initialize empty comment_list
3: for file in article_dir do
4:   soup = BeautifulSoup #create soup object
5:   comments = soup.findall('comments')
6:   for comment in comments do
7:     if comment has parent then
8:       Delete comment
9:   Add all comments to comment_list
10:  process_comments(comment_list)
11:  Write comment_list to new output file
12: return

```

Algorithm 2 Parsing text

```

1: Initialize article_dir
2: Initialize empty text_list
3: for file in article_dir do
4:   soup = BeautifulSoup #create soup object
5:   paragraphs = soup.findall('paragraph')
6:   for paragraph in paragraphs do
7:     Add paragraph to text_list
8:   Write text_list to new output file
9: return

```

3.2 Generating opinions with GPT-3.5

After correctly parsing the text and opinions of the articles, opinions were generated. Using the new *gpt-3.5-turbo* model, we were able to take advantage of a longer input context and incorporate human responses into the conversation as examples, whereas earlier models like *text-davinci-003* could not. Since each article contains approximately 100 human opinions, the goal was to generate the same number of artificial opinions.

Prompts

First of all, prompts had to be constructed. By providing the title of the article in the prompt, we first ran a set of test prompts to explore the quality of the responses of the model. It became evident that the model itself does not hold an opinion. Therefore it mostly gave formal, boring and factual perspectives on the subject. Together with my supervisor, we agreed to prompt GPT-3.5 to generate personas, such that subsequently we could generate opinions through a specific perspective, expecting more personalization in the opinions. Another observation after analyzing the test completions was that it consisted of opinions that were significantly shorter in length than human opinions. To address this, we provided an approximate length for each opinion in the prompt, calculated for each individual article. We used the average length of human comments on the article for determining the comment length. In total, we investigated four settings, described below. Examples of prompts are shown in Appendix B.

³BeautifulSoup Documentation - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

3.2.1 Persona

As mentioned earlier, two random personas were generated in advance and provided in this new contextual model. This was accomplished by appending the personas to the `system role`, such that the model knew how to ‘behave’. Every time the experiments were run, we prompted GPT-3.5 to generate these random personas. We ultimately came up with the following prompt:

```
Generate a persona. Use three sentences. Start with ‘You are’.
```

This prompt was constructed based on several criteria. Firstly, it was necessary to be concise and easily be easily understood. Secondly, since it would be used in other prompts to experiment with other settings, the output had to start with ‘You are’. Other options such as providing more information about what kind of persona it had to generate were also taken into consideration. In particular, we tried to provide the information that this persona likes to read an online news platform. However, after evaluating this, it became evident that it resulted in a more general persona based on the fact that it likes to read, rather than characteristics which would influence its opinion. It was nevertheless important to provide this information. We appended ‘You comment on an online newsplatform’ to the output manually. At last, we could pass this information to the other settings by appending the persona to the `role contents`. The implementation of generating personas is shown Algorithm 3. In Section 4, these personas are referred to as 1 and 2 or P1 and P2.

3.2.2 Zero-shot

The first setting was designed to utilize zero-shot learning. Recent work has shown that large language models exhibit the ability to perform reasonable zero-shot generalization to new tasks [SWR⁺21]. In this case of generating human-like perspectives, the model was solely prompted to generate opinions, with the title of the article as additional information. With this approach, we could examine GPT’s creativity to the fullest. However, depending on the topic of the article, the model generally has the least amount of context with this prompt. This led to the following prompt:

```
Give a list of 100 varied and critical opinions on the following news
article: ‘w’, where each opinion has an approximate length of ‘x’ words.
```

Here, `w` refers to the title of article and `x` to the approximate length of the comment. In Section 4, the zero-shot prompt is noted as ZS.

3.2.3 Few-shot

In the second setting, a few examples of human opinions were provided. The aim was to give the model more context and guidance to generate opinions similarly. Due to the limitations discussed in Section 3.2.5, only four examples are provided. Assuming the relevance is high, the comments with the most likes were selected. This method was most likely to perform best. It first of all learned directly from real examples and therefore might adapt its style and tone more accurately. This setting resulted in the following prompt:

```
Give a list of 100 varied and critical opinions on the following news
article: 'w', where each opinion has an approximate length of 'x' words.
Here are four examples: 'y'.
```

During the script, the variable `y` was replaced by the first four examples of the article. Later on, the few-shot approach is described as FS.

3.2.4 Context

In the third setting the experiments were run with is additional context. In addition to the zero-shot prompt, the introduction of the article was provided. The model, therefore, had more context to work with. This method may be beneficial since GPT-3.5 can produce more in-depth opinions, whereas the zero-shot prompt or few-shot can not. A potential downfall is that the introduction might not always contain any relevant or additional information. This resulted in the final prompt:

```
Give a list of 100 varied and critial opinions on the following news
article: 'w', where each opinion has an approximate length of 'x' words.
This is the introduction of the article: 'z'.
```

The `z` variable was substituted by the introduction of the corresponding article. The context prompt is later described as CL. We utilized ten different articles and three different prompting techniques, each with two different personas. Therefore, the number of total prompts is equal to 60 ($10 \times 3 \times 2$).

3.2.5 Limitations

In practice however, there were some limitations. Language models read text in chunks called tokens. Tokenization is the process of splitting text into smaller units called tokens, which is a fundamental preprocessing step for almost all NLP tasks [SSS⁺20]. Due to a maximum request of 4,096 tokens, which includes both the prompt and completion, the model would stop generating when that limit was reached. This, in particular for the few-shot and context prompt, resulted in unfinished completions. This limitation was also the reason we could not provide more examples or even the entire article in the prompt, which in theory would perform better. We came up with a solution by reducing the request of opinions to 25 (and 20 for the few-shot setting), such that the request was more likely to stay within the limit and increasing the number of requests per prompt to 4 (5 for few-shot).

Furthermore, the API had some limitations. Firstly, the server could overload with other requests, after which the code would terminate. Secondly, the API had a rate limit of only 60 requests per minute. As there were a total of 60 prompts, which all had to be executed at least 4 times, this was an issue. We resolved this by adding a `try` and `except` statement⁴, which would sleep for 60 seconds when an error occurred. As a result, we were able to generate nearly complete output for every prompt. In pseudo-code, the solution was implemented as shown in Algorithm 4.

⁴Python documentation - <https://docs.python.org/3/tutorial/errors.html>

Algorithm 3 Generating persona

```
1: prompt = "Generate a persona. Use
   three sentences. Start with 'You are'".
2: try:
3:   generate_completion(persona, prompt)
4:   persona = completion +
   "You comment on an online news platform")
5:   return persona
6: except Error:
7:   Sleep 20 seconds
```

Algorithm 4 Generating completions

```
1: persona_1 = generate_persona()
2: persona_2 = generate_persona()
3: for prompt in prompts do
4:   Initialize amount of prompt runs
5:   Write prompt to output file
6:   while j < prompt_runs do
7:     try:
8:       generate_completion(persona, prompt)
9:       Write completion to output file
10:    except Error:
11:      Sleep 60 seconds
12: return
```

3.3 Evaluation through classification

After we successfully prompted GPT-3.5, with 6 modes and ten articles we ideally would have 6000 generated opinions. To best evaluate the difference between the output of the prompts, different classification models were built. The data first needed to be correctly formatted such that the BERT model could process it. A small Python script was written, which iterates over all the outputs and then formats them to a readable database format. For this implementation, the .csv file format was chosen. An important remark here is that each record consists of a text *comment* and a *human* column, indicating the boolean value for Human- (1) or AI-generated (0). The data must be split into train and test sets. We split the data on article level, keeping all comments to the same article together in one partition. We processed the data in such a way, that we could examine an individual setting with every possible article as a test set. This in total, resulted in 120 files, where each file was either a train or test set, given a setting and article. Due to the lack of computational power available, the classification phase is performed using Google Colab.

3.3.1 Fine-tuning

We analyze the human likeness of the generated comments by classifying them. As BERT can be fine-tuned to obtain state-of-the-art results for tasks like text-classification, we fine-tuned Delobelle’s *robbert-v2-dutch-base* BERT model, which is the state-of-the-art Dutch BERT model [DWB20], in several ways. The goal of the models was to predict whether a given opinion is human-generated or GPT-3.5-generated. To ensure the models were properly trained, as mentioned earlier, we fine-tuned them via 10-fold cross-validation. Here for each fold, we trained the model on the other 9 of the 10 article sets and evaluated its performance on the remaining set. In this research, 10 models were trained for each setting, rather than a single model. We made this choice to prevent the model from overfitting the training data and to ensure it could generalize well to unseen data. Besides, training more models enhanced the overall reliability of the results.

To fine-tune the BERT model, several steps had to be taken. First, the generated train and test sets needed to be imported. In order for the model to read the input, the input text needed to be tokenized using the tokenizer from the pre-trained model. Since we did not use an existing dataset from the HuggingFace library, our tokenized data had to be converted to a suitable *Dataset* object, in order for the trainer to run without errors.

Before training the model, the *TrainingArguments* needed to be initialized. *TrainingArguments* are a subset of arguments that relate to the training process. For this analysis, we used the default values provided by the HuggingFace tutorial.⁵ Specifically, the learning rate at which the model adapts its parameters while training was set to 2×10^{-5} and the batch size per training core to 8. The weight decay was set to 0.01, while the evaluation and save strategies were configured to epoch. The number of training epochs was set to 3.

To evaluate the model, we used the F1 score, recall and precision as metrics. These methods can be imported via the Python *evaluate* library. In addition to these metrics, we also created a confusion matrix.

4 Results

The results of the research are discussed in this section. It contains an overview of the output of the prompts and the metrics of the fine-tuned models. The effect of the persona in the prompts is analyzed by utilizing the ROUGE metric for text similarity between the comments, the lexical diversity is measured and example outputs are shortly highlighted.

4.1 Output

Before utilizing all prompts, the two personas were generated, which resulted in the following two completions:

```
Je bent een man van middelbare leeftijd, getrouwd en vader van twee kinderen. Je werkt al twintig jaar als accountant bij een groot bedrijf en je houdt van tennis in je vrije tijd. Je reageert op een online nieuwsplatform.
```

Translated to English:

```
You are a middle-aged man, married and father of two children. You have worked as an accountant at a large company for 20 years and you enjoy playing tennis in your spare time. You comment on an online news platform.
```

and

```
Je bent een 32-jarige marketingmanager die houdt van gezelligheid en reizen. Je hebt een drukke baan, maar vindt het belangrijk om tijd te spenderen met familie en vrienden. In je vrije tijd ben je graag buiten en doe je aan hardlopen en yoga. Je hebt een passie voor koken en probeert altijd nieuwe recepten uit. Je reageert op een online nieuwsplatform.
```

Translated to English:

```
You are a 32-year-old marketing executive who loves socialising and travelling. You have a busy job, but find it important to spend time with family and friends.
```

⁵HuggingFace tutorial - <https://huggingface.co/docs/transformers/training>

In your spare time, you like to be outdoors and do running and yoga. You have a passion for cooking and are always trying out new recipes. You comment on an on-line news platform.

Ideally, the methods described in Section 3.2 would result in 100 comments per setting per article. In practice, a total of 5855 generated comments were generated, where a few prompts did not result in the full completion due to the limitations. As a consequence, balanced datasets could not be generated in some cases. We wanted to achieve this anyway, so in some cases, we had to generate the remaining responses manually. In Table 2, the article, setting and number of generated comments are shown. A setting consists of a prompt and a generated persona. ZS refers to the zero-shot prompt, FS to few-shot and CL to the context prompt. For instance, FS-2 indicates few-shot with persona 2.

Article	ZS-1	FS-1	CL-1	ZS-2	FS-2	CL-2	Total
1	100	92	100	100	73	100	565
2	100	95	100	100	100	100	595
3	100	71	100	100	75	100	546
4	100	73	100	100	100	100	573
5	100	88	100	100	100	100	588
6	100	100	100	100	100	100	600
7	100	100	100	100	88	100	588
8	100	100	100	100	100	100	600
9	100	100	100	100	100	100	600
10	100	100	100	100	100	100	600
Total	1000	919	1000	1000	936	1000	5855

Table 2: Amount of generated comments per setting

On preliminary manual analysis of the completions, a few things immediately stood out. Firstly, it became evident that each prompt did not complete in a single run. This was evident as there was less variation within one batch than between different batches. Batches showed clear differences in overall writing style. It appears that the output relies on earlier produced tokens in the same request. An example is shown in Figure 3. Since a batch contained 25 comments, there is a clear difference in writing style between the first batch (15 to 25) and the second batch (26 to 37).

A second preliminary observation we made was the nature of the generated comments. While these could be classified as opinions, they exhibited mostly factual text or reasoned arguments rather than expressing emotional viewpoints, even having utilized the few-shot prompt. Besides, the comments looked rather formal, as opposed to the human-written comments which often contained more informal language and slang.

4.2 Classification results

Our main results are presented in Table 3. Each row contains a single setting. The columns represent the average and standard deviation of metrics of all the fine-tuned models. It is noticeable that all values in Table 3 are relatively high. Initially, this seems positive. However, we are most interested in the lowest scores, since it reveals the areas where the model encountered the most challenges in

15 Het lijkt erop dat Talpa alleen reageert als ze worden geconfronteerd met beschuldigingen, in plaats van proactief te handelen.
16 Het is hypocriet van John de Mol om BOOS te beschuldigen van valse beschuldigingen, terwijl er duidelijk sprake is van serieuze klachten.
17 Het is tijd dat er binnen de muziekindustrie meer aandacht wordt besteed aan veiligheid en respect voor medewerkers.
18 Het is triest dat Jeroen Rietbergen meerdere vrouwen heeft misbruikt en dat Talpa hier blijkbaar zo weinig tegen heeft gedaan.
19 Het feit dat Talpa mogelijk al langer op de hoogte was, maakt hen medeplichtig aan dit wangedrag.
20 Het is tijd voor Talpa om te stoppen met het negeren van ernstige problemen en hun verantwoordelijkheid te nemen.
21 Het is belangrijk dat iedereen binnen Talpa die op de hoogte was van de misstanden, ter verantwoording wordt geroepen.
22 Het is verontrustend dat Talpa blijkbaar meer bezorgd is over de kritiek die ze krijgen, dan over het welzijn van de slachtoffers.
23 Het is schandalig dat de leidinggevenden van The Voice mogelijk hebben weggekeken bij zulk wangedrag. Dit is onacceptabel en moet worden aangepakt.
24 Het is tijd voor een cultuuromslag binnen de hele muziekindustrie, zodat dit soort wangedrag niet langer wordt getolereerd.
25 Het is onbegrijpelijk dat Talpa niet proactief heeft gehandeld toen ze op de hoogte waren van de misstanden. Dit toont aan dat er binnen het bedrijf een gebrek
26 Dit is ontzettend teleurstellend om te horen over Talpa en The Voice of Holland. Het is betreurenswaardig dat er dergelijke misstanden kunnen voorkomen in zo'n
27 Het feit dat leidinggevenden mogelijk langer afwisten van seksueel wangedrag is onacceptabel en verontrustend.
28 Ik hoop dat Talpa dit serieus neemt en gepaste maatregelen neemt om deze zaken aan te pakken en te voorkomen in de toekomst.
29 Het is voorlopig nog onduidelijk wat er precies is gebeurd, maar de beschuldigingen moeten serieus genomen worden en grondig onderzocht worden.
30 Het is belangrijk dat er een veilige werkomgeving is voor alle medewerkers, ongeacht hun functie binnen The Voice of Holland.
31 Dit toont opnieuw aan hoe belangrijk het is dat slachtoffers van seksueel wangedrag de moed hebben om hun verhaal te vertellen en dat het serieus genomen wordt.
32 Seksueel wangedrag moet in welke vorm dan ook altijd keihard worden veroordeeld en de dader moet altijd worden gestraft.
33 Het is triest dat er in onze samenleving nog altijd mensen zijn die weggijken bij misstanden zoals deze.
34 Het is duidelijk dat er iets niet goed gaat binnen Talpa en The Voice of Holland. Er moet snel gehandeld worden om deze problemen op te lossen.
35 Ik hoop dat dit nieuwsartikel zal bijdragen aan het creëren van meer aandacht voor dit ernstige probleem.
36 Als kijker van The Voice of Holland is dit voor mij een schokkende openbaring.
37 Ik vraag me af of er in de toekomst meer aanvullende informatie naar buiten zal komen.

Figure 3: Difference between batches

Setting	F1	GPT-3.5		Human	
		Precision	Recall	Precision	Recall
ZS-1	0.912 ± 0.047	0.926 ± 0.055	0.924 ± 0.084	0.931 ± 0.065	0.912 ± 0.065
FS-1	0.936 ± 0.032	0.943 ± 0.039	0.914 ± 0.082	0.939 ± 0.065	0.940 ± 0.044
CL-1	0.925 ± 0.043	0.932 ± 0.046	0.922 ± 0.103	0.932 ± 0.080	0.928 ± 0.056
ZS-2	0.923 ± 0.024	0.937 ± 0.041	0.918 ± 0.039	0.913 ± 0.038	0.936 ± 0.045
FS-2	0.934 ± 0.027	0.962 ± 0.027	0.896 ± 0.074	0.915 ± 0.056	0.962 ± 0.030
CL-2	0.920 ± 0.034	0.888 ± 0.047	0.966 ± 0.071	0.969 ± 0.062	0.873 ± 0.090

Table 3: Metrics per setting

distinguishing human-written comments from machine-generated comments. Lower values therefore indicate better performance of that particular setting. In terms of the F1-score, which represents the overall performance of the model, there is little difference, with the ZS-1 setting having the lowest of 91.2%. However, it does not have the lowest precision and recall of both classes. In terms of precision, which measures the accuracy of positive predictions made by the classifier, CL-2 performed best for GPT-3.5 classification with 88.8% and for humans ZS-2 performed best with 91.3%. It can be interpreted that the CL-2 setting out of 100 generated comments, predicts 89 on average correctly and classifies 11 as human. The same holds for human precision. The recall measures the proportion of correctly identified instances. For GPT-3.5, FS-2 performed best at 89.6% and for humans, CL-2 did at 87.3%. Although these are relatively high, other settings are not far away. All corresponding confusion matrices are shown in Appendix A.

In Table 4, the metrics are shown for each article. Just like in Table 3, all values are close to each other. The F1 scores are again all around 90% with a small standard deviation. Overall, models which used article 1 as testset resulted in the lowest F1 score of 89.2%. Another observation is that article 1 resulted in the lowest GPT-3.5 recall and human precision, article 7 in the lowest GPT-3.5 precision and article 9 in the lowest human recall. Additionally, it also seems that in some cases there is a tradeoff between precision and recall. Specifically, articles 1, 2, 5, 6, 7 and 9 show a noticeable pattern where higher precision values are associated with lower recall values and vice versa.

Article	F1	GPT-3.5		Human	
		Precision	Recall	Precision	Recall
1	0.892 ± 0.029	0.971 ± 0.008	0.812 ± 0.085	0.843 ± 0.061	0.975 ± 0.008
2	0.928 ± 0.008	0.906 ± 0.028	0.958 ± 0.025	0.957 ± 0.024	0.898 ± 0.036
3	0.931 ± 0.023	0.947 ± 0.019	0.913 ± 0.050	0.919 ± 0.044	0.948 ± 0.019
4	0.945 ± 0.017	0.930 ± 0.031	0.965 ± 0.008	0.963 ± 0.010	0.925 ± 0.037
5	0.913 ± 0.028	0.954 ± 0.019	0.870 ± 0.083	0.887 ± 0.066	0.957 ± 0.019
6	0.937 ± 0.021	0.979 ± 0.016	0.893 ± 0.067	0.907 ± 0.055	0.980 ± 0.017
7	0.926 ± 0.014	0.886 ± 0.026	0.982 ± 0.024	0.981 ± 0.024	0.872 ± 0.034
8	0.927 ± 0.019	0.910 ± 0.037	0.952 ± 0.028	0.951 ± 0.028	0.903 ± 0.043
9	0.914 ± 0.010	0.867 ± 0.028	0.980 ± 0.017	0.978 ± 0.018	0.848 ± 0.038
10	0.948 ± 0.013	0.961 ± 0.035	0.938 ± 0.045	0.943 ± 0.0395	0.958 ± 0.039

Table 4: Metrics per article

4.3 Similarity between comments

Aside from the BERT models, we can also analyze the difference between outputs in terms of ROUGE. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score is generally used in determining the quality of text summaries, by comparing them to human ideal summaries [Lin04]. It measures the count of overlapping sequences between an ideal reference text and a generated text. Although it is typically used for analyzing summaries, in our case this metric can be utilized to evaluate the difference between outputs generated by different prompts. While there are multiple rouge scores, we consider the ROUGE-1, ROUGE-2 and ROUGE-L scores for this research. ROUGE-N refers to the overlap of n-grams between a set of generated opinions and a reference opinion. We use the ROUGE Python library⁶ in our implementation. It expresses each ROUGE score in recall, precision and F1. For ROUGE-N, the recall is calculated by dividing the number of n-matches by the number of n-grams in the reference opinion. On the other hand, precision is calculated by dividing the number of n-matches by the number of n-grams in the generated opinion. Subsequently, the F1-score can be calculated:

ROUGE-N

$$\text{Recall} = \frac{\sum \text{n-matches}}{\sum \text{n-grams in reference opinion}}$$

$$\text{Precision} = \frac{\sum \text{n-matches}}{\sum \text{n-grams in generated opinion}}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROUGE-L measures the *Longest Common Subsequence* (LCS) given two sequences X and Y of length m and n . The LCS is the longest sequence in the same relative order, without it needing to be contiguous. The advantage of ROUGE-L over ROUGE-N is that it does not depend on

⁶ROUGE - <https://pypi.org/project/rouge/>

consecutive n-gram matches, such that it can capture sentence structures more accurately. The scores for the sentence level ROUGE-L metric are computed as follows:

ROUGE-L

$$\text{Recall}_{lcs} = \frac{LCS(X, Y)}{m}$$

$$\text{Precision}_{lcs} = \frac{LCS(X, Y)}{n}$$

$$\text{F1}_{lcs} = \frac{(1 + \beta^2) \times \text{Recall}_{lcs} \times \text{Precision}_{lcs}}{\text{Recall}_{lcs} + \beta^2 \times \text{Precision}_{lcs}}$$

Where $\beta = \text{Precision}_{lcs} / \text{Recall}_{lcs}$, as described in the original paper. We utilize the F1 scores of the ROUGE-1, ROUGE-2 and ROUGE-L metrics to analyze the difference between the output of different prompts. In particular, we compare the output of the different personas (P1 and P2) on article 1. In terms of the chosen setting, the zero-shot prompt was selected, as it is most likely to have characteristic traits in the output.

By calculating the different ROUGE scores of the outputs generated by P1, we can observe the degree of similarity among the comments. To compute the average ROUGE score, we can compute the ROUGE score for every possible pair in the generated opinion set. For 100 comments, there are 4950 unique pairs, calculated as follows:

$$n \times \frac{n - 1}{2}$$

The ROUGE-1 score of P1 averaged a value of 0.28. This was calculated by dividing the highest ROUGE-1 value per opinion by the total count. Of all comparisons, the maximum ROUGE-1 score was calculated to be 0.55. The distribution of calculated ROUGE-1 scores is shown in Figure 4.

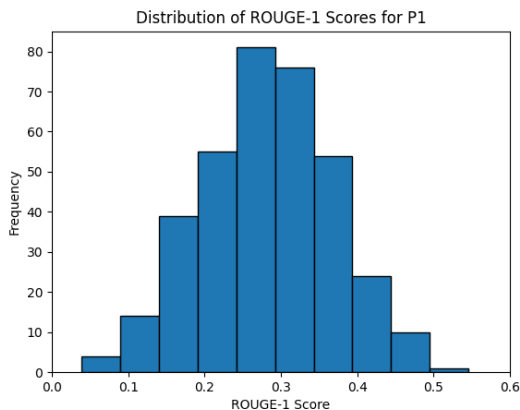


Figure 4: ROUGE-1 score P1

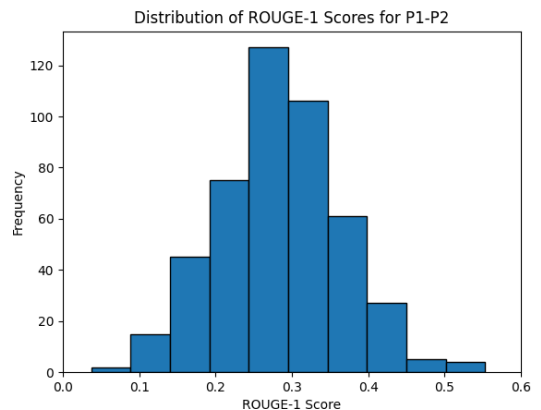


Figure 5: ROUGE-1 score P1-P2

There does not seem to be a difference between the ROUGE-1 scores of P1 and P1-P2, since the average of the highest ROUGE-1 scores between P1 and P2 also equaled 0.28, as well as the

highest ROUGE-1 score of 0.55. Its distribution is shown in Figure 5. At first sight, it looks very similar to the P1 distribution. However, the P1 distribution seems to be more normal, with larger bins in the centre. An important remark here is that the average of P1-P2 is based on twice as many comparisons as the P1 average, which may potentially be misleading and not completely representative.

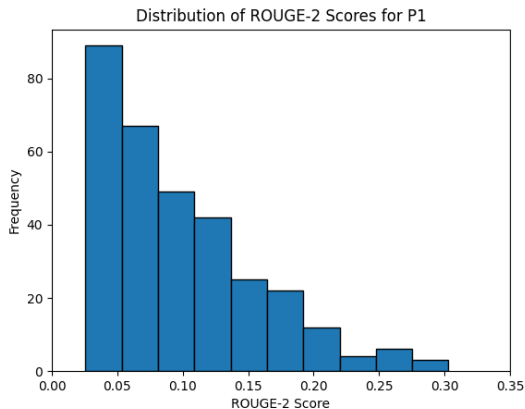


Figure 6: ROUGE-2 score P1

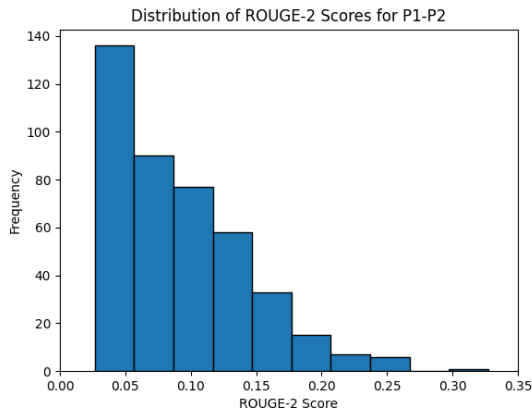


Figure 7: ROUGE-2 score P1-P2

The distribution of the ROUGE-2 scores for P1, as well as P1-P2, are shown in Figure 6 and Figure 7. Again, there does not seem to be a big difference. P1 has an average of 0.10, whereas P1-P2 averaged 0.09. Compared to the ROUGE-1 scores, the distributions are shifted much more to the left.

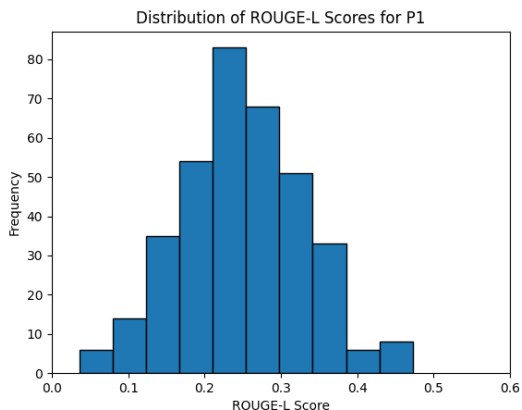


Figure 8: ROUGE-L score P1

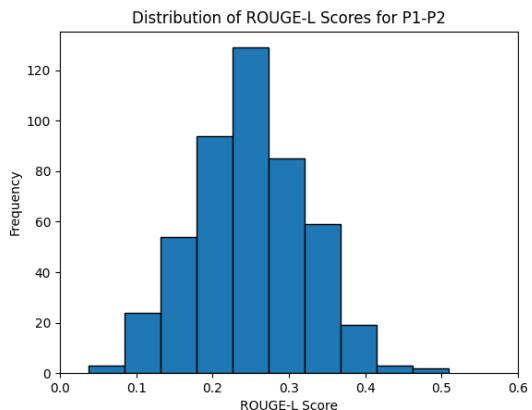


Figure 9: ROUGE-L score P1-P2

In Figure 8 and Figure 9, the distributions of the ROUGE-L scores are shown. It is noticeable that the P1 has a more left-skewed distribution as opposed to the P1-P2, meaning that P1 comments have more similarity among themselves in terms of ROUGE-L than the comments of P2. It is remarkable that the P1-P2 distribution however has a higher maximum of 0.5, indicating a pair of comments having relatively higher similarity. However, despite the lower maximum ROUGE-L score of 0.47, the P1 distribution has more instances around it. As a result, it has a higher average of 0.25 than the P1-P2 distribution, being 0.23. Again, not a substantial difference.

4.4 Lexical Diversity

Another quantitative method to analyze the different outputs is the *Type-Token Ratio* (TTR). TTR is the ratio calculated by dividing the types (t), which are the unique words occurring in a text, by its tokens (n), the total number of words. This, therefore, measures the lexical diversity, given multiple texts. After some preliminary observations of the generated output, we noticed that it seemed that GPT-3.5 used a lot of the same words. By utilizing the LexicalRichness Python package⁷, the lexical variety between human comments and GPT-3.5 generated comments can be quantitatively measured. Since not all generated comments equalled as many tokens as human comments and longer texts tend to have higher TTR values because they have more opportunities for unique words to occur, we particularly utilized the *Corrected Type-Token Ratio* (CTTR) [TC13] metric. CTTR normalizes the TTR by using the square root, providing a more accurate measure by considering the potential effect of the text length, and is calculated as follows:

$$CTTR = \frac{t}{\sqrt{2n}}$$

Where we computed the CTTR value over the total text of concatenated comments. In Table 5, the CTTR of all human and generated comments are shown. Since the number of human comments differs significantly per article, for simplicity only the first 100 are taken into account. The highest calculated GPT-3.5 CTTR per article is boldfaced. It appears the human text consistently has a higher value, compared to all the different settings. None of the settings exceeds the value of the human CTTR. From all generated completions, the few-shot completion of article 5 had the highest calculated CTTR value. If we calculate the difference between each setting to the human CTTR and average it from that, we can observe which setting is closest to the human CTTR on average. These results are shown in Table 6. In Figure 10, the averages per setting are compared to each other, as well as to the average human CTTR. While not very significant, it is evident that FS-1 and FS-2 averaged the lowest values, followed by ZS-1 and ZS-1, with CL-1 and CL-2 averaging the highest difference. In Figure 11 the CTTR values of human-written comments are compared to the average (of all settings) of the generated GPT-3.5 output per article. It is noticeable that all articles exhibit a significant difference in CTTR, except for article 4.

⁷<https://github.com/lsys/lexicalrichness>

Article	Human	ZS-1	FS-1	CL-2	ZS-2	FS-2	CL-2
1	14.422	8.892	8.707	9.104	8.664	9.794	9.218
2	13.043	9.103	8.559	7.531	7.569	8.650	7.940
3	13.192	9.062	9.893	9.833	9.442	8.973	9.196
4	10.581	7.778	8.697	8.970	8.493	7.822	8.198
5	12.457	9.018	9.503	8.302	8.302	10.191	9.641
6	13.657	8.126	9.334	7.458	8.671	10.027	7.751
7	14.436	9.043	9.056	8.620	8.370	8.247	7.837
8	12.963	7.831	7.878	7.768	7.221	8.674	6.738
9	13.552	9.296	8.643	7.341	9.334	8.133	6.862
10	13.440	6.330	7.639	7.318	7.410	8.093	6.819

Table 5: CTTR values

	ZS-1	FS-1	CL-1	ZS-2	FS-2	CL-2
Average difference	4.726	4.383	4.950	4.827	4.314	5.154

Table 6: Average difference per setting

4.5 Qualitative analysis

Aside from the quantitative analysis, we can also analyze the output qualitatively. Since our fine-tuned model classified 0 as *AI (GPT-3.5)* and 1 as *Human*, a false positive is considered a GPT-3.5-written comment classified as a human. A false negative is a human-written comment which got classified as GPT-3.5. As our main goal was to investigate human likeness, mostly false positive instances are considered. In Section 4.5.1, false positive instances from every setting are analyzed, whereas in Section 4.5.2, a false negative instance is investigated. It is worth noting that while all comments were generated in Dutch, the translations are provided for the analysis. More examples of completions are shown in Appendix B.

4.5.1 False positives

The first example is a zero-shot as persona 1 GPT-3.5 generated comment on article 1:

As an avid tennis player, I can already see the end of the season ahead of me. This is bad news not only for my own enjoyment, but also for the sports club where I play. Why are we hitting the sports and cultural sector so hard?

The model predicted that this particular instance was human-written, but in fact, was generated by GPT-3.5. Especially the first two sentences immediately stand out. The content seems rather personal and sentimental, which is most likely to cause the incorrect classification. This is the perfect example of the model utilizing the given persona, which was the initial intention of providing one. In the original Dutch output it also immediately stood out that GPT-3.5 made a spelling mistake by generating `tenniser` instead of `tennisser`.

In Figure 12, the instance is visualized with SHAP. SHAP is a game theoretic approach to

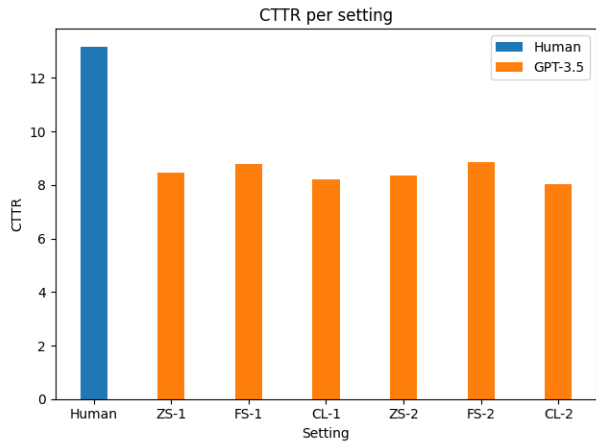


Figure 10: CTTR comparison per setting: Human vs. GPT-3.5

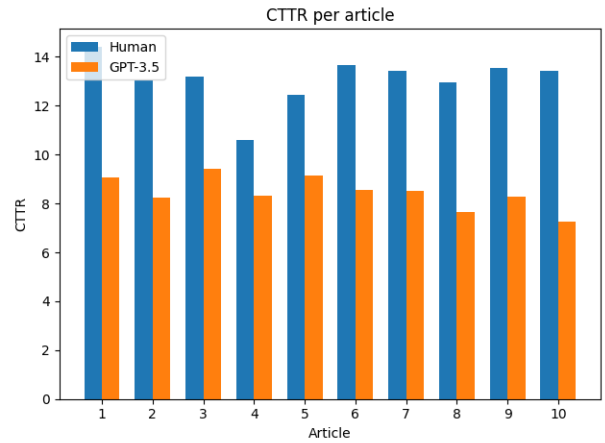


Figure 11: CTTR comparison per article: Human vs. GPT-3.5

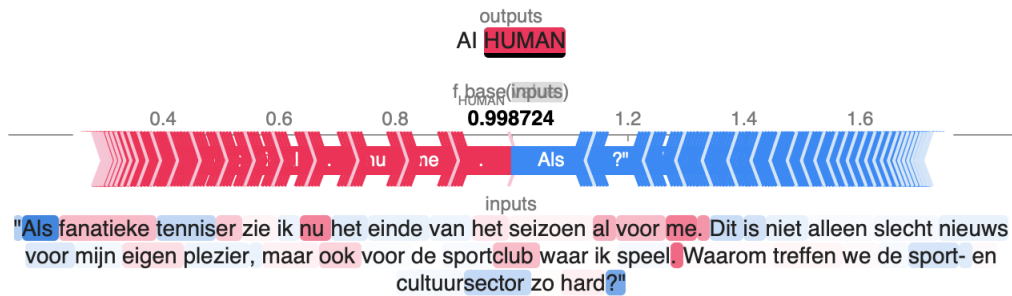


Figure 12: ZS-1 false positive

explain the output of any machine learning model.⁸ As we have selected the assigned human class at the top, the contribution of each token to the human class is shown. The individual contribution of a token is determined by calculating the difference between the total classification and the classification with a single token masked. Tokens in red suggest a positive contribution to the selected class, while tokens in blue suggest a negative contribution. It is evident that tokens in the first two sentences such as *fanatieke*, *het seizoen al voor me*. and *eigen* positively contribute to the classification. It is also clear that the third sentence, which has a rather formal tone, barely contributes positively to the classification. This is in line with our earlier observation. Our model does not associate tokens such as *als*, *tennis*, *sport*, *sector* and *?* with a human-written comment, suggesting these are more GPT-3.5 like.

The second example was generated by the few-shot prompt as persona 1 on article 2:

I understand the frustration that athletes should be punished for something that is not their fault, but on the other hand, if Russia does not play by the rules

⁸SHAP - <https://shap.readthedocs.io/en/latest/>



Figure 13: FS-1 false positive

then they do not deserve to be at the Olympic Games.

Although the personal characteristics of P1 such as valuing sports are highly relevant for this article, none of these are evident in the generated output, in contrast to the first given example. This is an example of a comment with a rather formal and factual perspective on the topic as we found in early observations. The content itself does not explicitly express an opinion based on personal beliefs, but rather a nuanced and potential perspective on the topic. Nonetheless, our model classified it incorrectly. In Figure 13, this false positive instance is also shown. The tokens that were most responsible for the positive classification were **Ik**, **snap**, **dat** and **dan verdienen ze het niet op** and **staan**. Of all tokens, these exhibit the most human likeness and outweigh the tokens that negatively impact the classification. These GPT-3.5 like tokens were respectively **de**, **atleten**, **gestraft** and **de Spelen te**. While this is a comment generated by the FS-1 prompt, the output does not seem to exhibit any characteristics the previous examples did not.

The third example is part of the output generated by the context prompt as persona 1 on article 3. It briefly translates to:

I am happy with the recommendation to use vaccination certificates to stop the spread of the virus, but I have to admit that I have additional doubts. I wonder if this recommendation is useful for people who have already had COVID-19 or have developed antibodies against the virus, and whether they also need a booster vaccination.

This is another example of a rather formal and nuanced perspective on the topic. It firstly shows a brief opinion without any personal motivation, after which it is critically nuanced. The main difference here is that the output exhibits more in-depth content. There could be two explanations for this. Firstly, it could be caused by the data GPT-3.5 was trained on. Secondly, the introduction of the article provided in the prompt could have influenced the output. The tokens **booster** and **vaccination** for example, were only mentioned in the introduction. It is likely this had an impact on the generated output. In practice, however, a combination of the two is most probable. Since this topic is relatively close to the training data cut-off date, it has relatively more context on the topic than articles which were published much later.



Figure 14: CL-1 false positive

In Figure 14, all SHAP values are shown for this instance. The first sentence exhibits a brief personal opinion with I am happy, the tokens `Ik`, `ben` and `blij` also show positive contributions, whereas `gebruiken`, `om`, `de`, `verspreiding` contribute negatively. This might be caused by the fact that GPT-3.5 comments showcased these more often. The second sentence, which puts the statement into perspective shows both strong positive and negative contributions. It starts off with the persona wondering whether the recommendation is useful, which the model associated with GPT-3.5. However, the more in-depth content, unlike our suspicions also contributes both positively and negatively. The tokens `antilichamen`, `hebben`, `ontwikkeld` and `virus` had a negative impact, whereas the other tokens had a positive impact on the classification. This included the `booster` token, which was part of the provided context in the prompt.

4.5.2 False negative

While our primary focus lies on false positive instances, analyzing a false negative instance might lead to additional insights. A false negative in our case is a human-written comment classified as GPT-3.5. Although it is not directly related to the setting, the false negative we analyze is an instance from the CL-2 model on article 7. It briefly translates to:

Sad that as a world, we are slipping away like this. Not even capable of dialogue. I am genuinely worried about the future of my children.

It is noticeable that just like the examples shown in Figure 13 and Figure 14. the comment has a rather formal structure. Rather than a strong opinion accompanied by personal motivation, the comment presents a rather impersonal perspective through factual statements. The last sentence exhibits the same personal characteristics of P1, which the model potentially associates with GPT-3.5-like opinions.

In Figure 15, the SHAP values are shown. The human class is again selected, meaning that in this scenario all blue tokens contribute positively and all red tokens negatively to the AI class. The first thing that stands out is that almost every token contributes positively to the AI class. It is remarkable that the first word is split into two tokens, which individually do not have a meaning. The token `riest` has the strongest negative contribution of the entire comment. Our suspicion

about the last sentence having a GPT-3.5-like characteristic seems correct. Apart from the token `echt`, it strongly contributes to the AI classification.



Figure 15: CL-2 false negative

5 Discussion

In this section, we discuss and interpret the found results, as well as some limitations and challenges we encountered during this research.

5.1 Analysis of results

The results of the fine-tuned BERT models were all very high. Although this initially seems to be good, it tells us the capabilities of GPT-3.5 are quite limited. In terms of human likeness, there does not seem to be a significant difference in the difficulty of distinguishing outputs from the settings with human-written comments. Our findings also do not suggest that a specific prompting technique like zero-shot, few-shot or context results in more human-like outputs. This suggests that capturing the complexity of human-natured comments is still a challenging task for GPT-3.5. The F1 scores are not significant enough to make statements about which setting can best reproduce human opinions on news articles. Besides, while zero-shot 1 showcases the lowest F1 score, it does not show the lowest scores in terms of precision and recall. In terms of GPT-3.5 precision, the context 2 prompt exhibited the lowest score of 88.8%. This means that in our research, the BERT models on average had the most difficulty distinguishing between actual context 2 outputs and human outputs, across all articles. However, we can not conclude that providing context in the prompt guarantees a higher level of human likeness, especially with the context 1 models not performing the same. Personally, our suspicion aligns with the preliminary observations we made. Regardless of the specific prompt, GPT-3.5 generally outputs comments which had a rather factual and boring tone of style. We tried to counter this by providing a persona, but this had little to no effect. This was probably because the generated personas were fairly narrowly described and therefore the model had a limited idea about its beliefs and motivations, especially on multiple topics.

The second metric we utilized was the ROUGE metric in order to investigate the difference in style between the generated personas. We compared the ROUGE-1, ROUGE-2 and ROUGE-L across the personas. We accomplished this by first calculating the average ROUGE of every possible

pair of comments, given all comments of a single persona. Subsequently, we did the same for every possible pair, given all comments of both personas. Our findings were that there were no clear differences between the distributions. A possible explanation for this is first that the personas did not have distinct perspectives on the subject. In particular, both specifically are social characters who valued sports in their spare time. Secondly, only the first article and zero-shot prompt were taken into account when measuring the ROUGE metrics. This did not cover all of the generated data, since it was a small part of the dataset. It would have been possible that there were other subjects where the distributions did differ. Conclusions drawn from a single setting and article might be misleading. Generating average distributions over all articles might have led to other insights.

As we observed earlier that GPT-3.5 generated comments seemed rather formal and boring and often had the same structure, we analyzed the lexical diversity by calculating the CTTR values. We found that human opinions consistently exhibited higher CTTR values, meaning that the ratio of unique words to total words was greater than that of machine-generated comments. This is in line with our observations. In particular, it was clear that few-shot 1 and few-shot 2 resulted in the highest lexical diversity, followed by zero-shot completions, with context completions averaging the lowest CTTR value. It seems very logical that the few-shot setting has the highest value since it directly learns from real-world instances and therefore copies such words more easily. Another conclusion that can be drawn is that providing the introduction of the article in the prompt, reduces the variety of words the model used. However, the difference in diversity among the output of the prompts was much smaller than the difference between human-written or machine-generated comments.

After having analyzed individual misclassifications quantitatively, we found that GPT-3.5 tends to create its comments based on more factual impersonal statements, rather than strong opinions with personal motivations or beliefs. While some GPT-3.5 instances were incorrectly classified, we still managed to observe this particular pattern. The false negative instance we observed showed a similar style.

5.2 Limitations

We encountered several limitations during our research. The first one was the OpenAI API. As mentioned in Section 3.2.5 the API, in fact, had some flaws. At the time of our research, we had to deal with access to the API. It was not always possible to utilize it, since we were dependent on the overall occupation of it. On some days, a single completion would take much longer than usual or even not even be generated due to overloaded servers. This sometimes resulted in extremely long run times, if we were lucky enough that it worked fully at all. Not only were we struggling with correctly generating completions, there was also a limitation in the number of requests we could make. Every OpenAI account was credited an \$18.00 free trial at the time, allowing the user to generate a limited amount of tokens. Another issue we had was the token per request limit, not allowing a prompt to run successfully at once. As mentioned earlier, we managed to counter this by catching possible errors in our script. At first, this does seem like a major issue. However, we found differences in the outputs of the prompt. While being prompted the same, GPT handled every prompt as an individual request. As a result, the different batches often had different styles, such as noting quotation marks or starting every comment with the same words. On one hand, it indicates that GPT-3.5 still struggles to accurately capture the complexity of human comments.

On the other hand, it is worth highlighting that human likeness does not necessarily relate to the overall performance of the model.

6 Conclusions

The goal of this research was to investigate to what extent GPT-3.5 is able to generate human-like comments on Dutch news articles and how to best generate these. We answered this question by experimenting with multiple prompting techniques, after which we could analyze the different outputs. In particular, the zero-shot, few-shot and context prompts were utilized, corresponding with a generated persona. We fine-tuned the pre-trained RobBERT-v2 model to classify whether unlabeled comments were human-written or generated by GPT-3.5. While in previous research zero-shot and few-shot learning had shown remarkable performances, it does not so in our case. We found that the BERT models we fine-tuned were able to achieve very high accuracies. In our case, the context persona 2 setting was the hardest to distinguish from human opinions. Since there was no clear distinction between the settings, we can conclude that GPT-3.5 is still limited in generating human-like comments on Dutch news articles, regardless of which setting. It suggests that capturing the complexity of human-nature comments, even with real examples, is still a challenging task. In order to analyze the difference in output, we first utilized the ROUGE metric to investigate the persona. Although the output did exhibit some personalized instances, we found that there are no significant differences in terms of ROUGE-1, ROUGE-2 and ROUGE-L scores. The lexical diversity was measured using the CTTR metric. It became evident that human-written comments generally have a much higher lexical diversity, as opposed to GPT-3.5-generated comments. Again, while not very significant, we did find that few-shot prompts averaged the highest CTTR and therefore the most human-like, with ZS and CL having a lower lexical diversity. The manual analysis of individual misclassifications led to additional insights that GPT-3.5 very often tends to generate comments in a rather formal and factual style.

During our research, OpenAI publicly announced GPT-4. According to OpenAI, GPT-4 is more creative and collaborative than ever before.⁹ This may be an advantage in future studies on opinionated text generation. Instead of the current limit of 4,096 tokens, GPT-4 is capable of handling 25,000 tokens per request. This is a major improvement and can be used to run prompts at once, instead of running them multiple times. Secondly, it can be used to provide more context such as the entire article in the prompt, which would possibly lead to more in-depth comments. At last, a lot more human-written examples can be provided in the prompt, which may positively influence the human-likeness of the output.

Other potential further research does not necessarily relate to opinionated text. It could involve the investigation of to what extent the training cut-off date of GPT-3.5 influences its performance, experimenting with subjects from different time periods before and after the cut-off date.

⁹OpenAI - <https://openai.com/gpt-4>

References

- [BMR⁺20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [CCA21] Ke-Li Chiu, Annie Collins, and Rohan Alexander. Detecting hate speech with GPT-3. *arXiv preprint arXiv:2103.12407*, 2021.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [DWB20] Pieter Delobelle, Thomas Winters, and Bettina Berendt. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online, November 2020. Association for Computational Linguistics.
- [HAS⁺23] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.
- [HBD⁺19] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- [Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [LOG⁺19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.
- [SBAG23] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. AI model GPT-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924*, 2023.
- [SQXH19] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583, 2019.
- [SSS⁺20] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*, 2020.
- [SWR⁺21] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

- [TC13] Joan Torruella and Ramón Capsada. Lexical statistics and tipological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95:447–454, 2013.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
- [WXD⁺23] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. Is ChatGPT a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.
- [ZLZ23] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Extractive summarization via ChatGPT for faithful summary generation. *arXiv preprint arXiv:2304.04193*, 2023.

A Confusion Matrices

In this section the confusion matrices with corresponding fine-tuned models are shown. All matrices have the same structure as Table 7, but have been reduced in size for convenience. Figure 16, Figure 17 and Figure 18 represent the matrices of first persona prompts. The matrices of the second persona prompts are shown in Figure 19, Figure 20 and Figure 21.

		Actual	
		Human	GPT-3.5
Predicted	Human	TP	FP
	GPT-3.5	FN	TN

Table 7: Confusion Matrix

91	9	98	2	93	7	95	5	70	30	93	7	100	0	93	7	98	2	93	7
3	97	14	86	4	96	6	94	4	96	1	99	17	83	12	88	18	82	1	99
1.1		1.2		1.3		1.4		1.5		1.6		1.7		1.8		1.9		1.10	

Figure 16: ZS-1 matrices

74	26	95	5	97	3	98	2	97	3	85	15	99	1	90	10	99	1	98	2
1	99	8	92	3	97	4	96	3	97	1	99	11	89	5	95	12	88	12	88
2.1		2.2		2.3		2.4		2.5		2.6		2.7		2.8		2.9		2.10	

Figure 17: FS-1 matrices

67	34	98	2	85	15	97	3	94	6	86	14	100	0	98	2	100	0	97	3
2	98	10	90	6	94	8	92	4	96	0	100	12	88	8	92	19	81	3	97
3.1		3.2		3.3		3.4		3.5		3.6		3.7		3.8		3.9		3.10	

Figure 18: CL-1 matrices

90	10	93	7	84	16	96	4	83	17	95	5	91	9	94	6	95	5	87	13
2	98	8	92	5	95	6	94	3	97	3	97	9	91	14	86	13	87	1	99
4.1		4.2		4.3		4.4		4.5		4.6		4.7		4.8		4.9		4.10	

Figure 19: ZS-2 matrices

88	12	92	8	90	10	97	3	83	17	77	23	99	1	96	4	96	4	88	12
4	96	4	96	3	97	3	97	2	98	1	99	9	91	3	97	9	91	0	100
5.1		5.2		5.3		5.4		5.5		5.6		5.7		5.8		5.9		5.10	

Figure 20: FS-2 matrices

77	23	99	1	99	1	96	4	95	5	100	0	100	0	100	0	100	0	88	12
3	97	17	83	10	90	18	82	10	90	6	94	19	81	16	84	20	80	0	100
6.1		6.2		6.3		6.4		6.5		6.6		6.7		6.8		6.9		6.10	

Figure 21: CL-2 matrices

B Prompts and completions

In this section a few example prompts as well as a few completions from various prompts and articles are shown. In Table 8, example prompts of P1 are shown. A few examples per setting on different articles are shown in Table 9, Table 10, Table 11, Table 12, Table 13 and Table 14.

Persona	Je bent een 32-jarige marketingmanager die houdt van gezelligheid en reizen. Je hebt een drukke baan, maar vindt het belangrijk om tijd te spenderen met familie en vrienden. In je vrije tijd ben je graag buiten en doe je aan hardlopen en yoga. Je hebt een passie voor koken en probeert altijd nieuwe recepten uit. Je reageert op een online nieuwsplatform.
ZS	Geef een lijst van 25 gevarieerde kritische reacties als deze persona op het volgende nieuwsartikel: Minister rekent op 1.400 euro vergoeding voor studenten uit 'pechgeneratie', waarbij elke reactie ongeveer uit 72 woorden bestaat.
FS	Geef een lijst van 25 gevarieerde kritische reacties als deze persona op het volgende nieuwsartikel: Rutte biedt excuses aan voor slavernijverleden: 'Aan alle nazaten tot hier en nu', waarbij elke reactie uit ongeveer 51 woorden bestaat. Hier zijn vier voorbeelden: 1. Ik heb naar de gehele toespraak geluisterd. Ik denk dat Rutte het heel goed heeft verwoord. Hij beschreef duidelijk het leed dat destijds is aangedaan, en heeft ook goed aangegeven dat niemand die vandaag leeft verantwoordelijk is voor het leed van destijds. Hij heeft duidelijk in mijn ogen oprechte excuses gemaakt en ik denk, nee, verwacht zelfs, dat deze ook door degenen die echt excuses verwachten worden geaccepteerd. Voor de groepen die klagen en zeker voor degenen die herstelbetalingen eisen hoop ik dat er op een later moment een duidelijk bericht vanuit de overheid komt dat dergelijke betalingen er nu niet en nooit niet zullen komen. 2. Ik ga volgende week excuses aanbieden voor het uitsterven van de Dodo, daar heb ik persoonlijk ook geen fluit mee te maken maar het staat wel interessant. 3. Iemand die nu leeft, heeft geen schuld aan slavernij. Maar de Nederlandse Staat draagt verantwoordelijkheid voor het leed', zei de premier. Ik leef nu, en betaal nu belasting. Het kan dus niet zijn dat mijn belastinggeld gebruikt wordt om deze excuses af te kopen. Vreselijk wat er toen gebeurd is, maar niet iets wat je de huidige generaties kunt aanrekenen! 4. Deze excuses staan symbool voor alles wat er mis is met de politiek in Nederland. Ik vraag mij af hoeveel Nederlanders deze excuses steunen en hoeveel niet. Mijn gevoel (ik weet dat het niets zegt) zegt mij dat een grote meerderheid zich niet in deze excuses kan vinden. Het zijn in ieder geval niet mijn excuses. Ik verwacht namelijk ook geen excuses van de kinderen of kleinkinderen van de Duitse soldaten die mijn opa en overgrootvader hebben uitgehongerd en omgebracht.
CL	Geef een lijst van 25 gevarieerde kritische reacties als deze persona op het volgende nieuwsartikel: Jumbo stopt per direct met WK-reclamespot na storm van kritiek, waarbij elke reactie uit ongeveer 50 woorden bestaat. Dit is de inleiding van het artikel: Jumbo stopt per direct met zijn WK-reclamespotje waarin dansende bouwvakkers te zien zijn. De supermarktketen doet dit na de golf van kritiek die woensdag ontstond, onder meer op sociale media.

Table 8: Example prompts P1

no.	Comment
1	“Wat is het nut van een avondklok als we de rest van de dag wel alles mogen doen? Dit voelt als een halfbakken oplossing die geen echte verandering teweegbrengt.”
2	“Een avondklok is discriminerend en kan alleen maar voor meer polarisatie zorgen. Wat als je mensen hebt die nachtdiensten draaien of juist op andere tijden willen leven? Dit is geen oplossing, maar juist een probleem.”
3	“In plaats van nieuwe maatregelen te bedenken moeten we juist kijken naar bestaande maatregelen en deze beter handhaven. Dan pas kunnen we zien of er nog extra maatregelen nodig zijn.”
4	“Als accountant zie ik nu al voor me hoe de economie zal inzakken door een avondklok. Dit is geen oplossing, maar juist een probleem verplaatst naar andere gebieden.”
5	“Het is tijd om veerkracht te tonen en juist positief te blijven. De overheid moet ons hierin steunen, in plaats van te onderdrukken met nieuwe maatregelen.”
6	“Een avondklok zal alleen maar leiden tot meer verzet en onrust. We moeten in plaats daarvan leren hoe we gezamenlijk verantwoordelijkheid kunnen nemen en onze vrijheden kunnen behouden.”
7	“Waarom altijd kiezen voor de makkelijke weg? De overheid moet juist kijken naar de moeilijke, maar effectieve strategieën. Zo kunnen we de pandemie bestrijden zonder onze samenleving verder te beperken.”
8	“Als ouder maak ik me zorgen over wat voor invloed dit heeft op de kinderen. Ze hebben al genoeg aan hun hoofd, nu krijgen ze nóg meer beperkingen opgelegd.”
9	“Een avondklok biedt geen enkele garantie op het voorkomen van besmettingen. Het is tijd voor een meer intelligente en effectieve aanpak.”
10	“Dit voelt als een machtsmiddel van de overheid om controle uit te oefenen over onze levens. We moeten ons juist verzetten tegen deze inbreuk op onze vrijheid.”
11	“Wanneer stoppen we met helikopterdenken en kijken we naar de realiteit? Een avondklok is helemaal niet praktisch uitvoerbaar en zal alleen maar nutteloze stress veroorzaken.”
12	“Dit is precies wat er gebeurt als we politici onze levens laten leiden. Ze denken alleen in hokjes en regels, niet in menselijke interactie en verantwoordelijkheid.”
13	“Wat is er gebeurd met onze democratie en vrijheid? Dit is precies wat ik vreesde toen de pandemie begon. We mogen dit niet laten gebeuren op de lange termijn.”
14	“Een avondklok is een teken van zwakte van de overheid. In plaats van in te grijpen in de levens van mensen, moeten we leren samen te leven met de pandemie en verantwoordelijkheid nemen in elke situatie.”
15	“Dit voelt als paniekvoetbal. Mensen hebben al genoeg stress, angst en onzekerheid in hun leven door de pandemie. We moeten niet nog meer beperkingen opleggen, maar juist samenwerken en solidair zijn.”

Table 9: ZS-1 examples article 1

C Human output

In this section, examples of human outputs are shown. For simplicity, we show comments on the same articles as in Appendix B. The examples are respectively shown in Table 15, Table 16, Table 17, Table 18, Table 19 and Table 20.

no.	Comment
1	Het is goed dat er platforms zijn waarop politici hun meningen over internationale kwesties en evenementen kunnen delen en dit is een weloverwogen beslissing van de burgemeester.
2	Het lijkt me niet de beste manier om het probleem van oorlogsdaden en mensenrechtenschendingen aan te pakken door atleten te blokkeren, maar er moeten stappen worden ondernomen om de situatie in Rusland te verbeteren.
3	Het is verbazingwekkend dat het IOC zo lang duurt om te beslissen over de kwestie van Russische atleten op de Olympische Spelen. Het is hetzelfde probleem dat we keer op keer hebben gezien met de deelname van Russische sporters aan internationale evenementen.
4	Het vakgebied van de burgemeester is misschien niet de meest geschikte voor het uitspreken van politieke verklaringen en zij moet zich daar bewust van zijn.
5	Het lijkt een drastisch besluit om alle Russische sporters van de Olympische Spelen te weren, gezien de hoeveelheid individuele atleten die geen banden hebben met het leger of de oorlog in Oekraïne. Een meer gerichte en specifieke maatregel zou beter zijn.
6	Landen moeten verantwoordelijkheid nemen voor hun daden, zowel op politiek als op sportgebied. Het weren van Russische atleten van de Olympische Spelen draagt bij aan dit besef en brengt de aandacht op de juiste problemen en struikelblokken.
7	Hoewel ik het eens ben met het standpunt van de burgemeester, vraag ik me af of de OS wel de juiste plaats is om politieke kwesties op te lossen. Het is een sportevenement en zou dit moeten blijven om de atleten te steunen die zo hard hebben getraind.
8	Het lijkt alsof sport en politiek altijd met elkaar verweven zullen zijn, tot onze spijt. Ik ben benieuwd of de OS de Russische politiek zal beïnvloeden, zoals de burgemeester hoopt.
9	Door Rusland van de Olympische Spelen te weren stelt de wereldgemeenschap duidelijk dat zij de daden van het land niet tolereert en dat er consequenties zijn voor schendingen van het internationaal recht.
10	Het IOC moet de beslissing van de burgemeester zeker in overweging nemen en rekening houden met de argumenten die zij heeft gegeven. Het weigeren van Russische atleten vanwege oorlogsdaden is een legitiem standpunt.
11	Is dit niet gewoon een politieke afstraffing voor Rusland? Of zijn er echte maatregelen nodig om de oorlog te beëindigen en gebieden te herstellen tot stabiliteit.
12	Het is belangrijk om sport los te zien van politiek, maar in het geval van Rusland worden de twee onvermijdelijk met elkaar verweven. Ik denk dat de burgemeester goede argumenten heeft om te pleiten voor uitsluiting van Russische sporters.
13	Het is een triest feit dat de wereldgemeenschap regelmatig sporten gebruikt als politieke drukmiddel. Maar steeds meer lijkt er geen andere optie meer te zijn om landen die mensenrechten schenden verantwoordelijk te houden.
14	Hoe dan ook, ik ben blij dat de burgemeester haar stem laat horen in het debat over het weigeren van Russische sporters op de Olympische Spelen. Het is belangrijk om een duidelijk signaal af te geven dat het schenden van internationale wetten nadelige gevolgen heeft.
15	Het blokkeren van Russische atleten zou ook de atleten van andere landen beïnvloeden die samenwerken met Russische partners of sponsors. Door middel van sport moeten we deze complexe vraagstukken oplossen.

Table 10: FS-1 examples article 2

no.	Comment
1	Dit is weer een staaltje van de EU die zich weer eens bemoeit met ons privéleven. Wanneer houdt het op met al die bureaucratie?
2	Dit is goed nieuws voor de vaccinatiecampagne. Mensen worden gestimuleerd om op tijd op te komen dagen voor hun vaccinatie.
3	Dit is belachelijk, het creëert een tweedeling in reizigers, waarbij mensen zonder vaccinatie worden buitengesloten.
4	Door de QR-code slechts 9 maanden geldig te laten zijn, worden mensen gestimuleerd om snel na hun laatste vaccinatie een boosterprik te halen. Dit is goed voor de gezondheid.
5	Het is niet verstandig om mensen hun vaccinaties te laten bewaren, terwijl we weten dat de veiligheid en werkzaamheid niet voor altijd zijn gegarandeerd.
6	Belangrijker is dat mensen moeten zich beseffen dat de pandemie nog steeds bestaat. Zij moeten de moeite nemen om zich te beschermen en te laten vaccineren.
7	Het idee van het reisbewijs is om het reizen weer te normaliseren. Niemand wilt zo lang wachten omdat zijn of haar bewijs inmiddels verlopen is.
8	Door deze tijdslimiet te plaatsen op de QR-code zet men de druk op. Mogelijk worden mensen er zelfs sneller mee gechanteerd. Nee, dit is niet de juiste manier om het vaccineren te stimuleren.
9	Ik heb geen probleem met een tijdslimiet op de vaccinatie, maar dit moet worden bepaald door gezondheids-experts, niet door EU-functionarissen.
10	Het stellen van deadlines verhoogt de druk op gezondheidsorganisaties en medisch personeel. Bovendien kan deze regelgeving over de hele linie teveel afhankelijk zijn van politiek in plaats van wetenschap.
11	Het feit dat vaccinatiebewijzen maar 9 maanden geldig zijn voor reizen zet aan tot wantrouwen en mogelijk meer complotdenken.
12	Dit is belachelijk, ik ga niet om de negen maanden terug vaccineren om alleen te mogen reizen. Laat mij liever dan thuis blijven.
13	Dit lijkt een poging om mensen te dwingen zich regelmatig te laten vaccineren, ook al zorgt dit voor risico's en onzekerheden.
14	De EU-leden moeten eerst kijken naar de wetenschap. Leven we nog immuun op dat moment? En hoe lang duurt het voordat het immuunsysteem zijn werking verliest?
15	Dit is een slimme zet om druk uit te oefenen op de farmaceutische bedrijven om sneller nieuwe vaccins te ontwikkelen. Hopelijk zal dit leiden tot meer innovatie in de toekomst.

Table 11: CL-1 examples article 3

no.	Comment
1	Dit is teleurstellend om te horen. Het is belangrijk dat sporters zich uitspreken tegen discriminatie en voor inclusie, ongeacht hun religieuze overtuigingen.
2	Ik vind het jammer dat K�k�c�u deze kans mist om een statement te maken voor de LGBTQ+-gemeenschap. Sport is een geweldig medium om te tonen dat iedereen dezelfde rechten en kansen verdient.
3	Het is begrijpelijk dat K�k�c�u handelt vanuit zijn geloofsovertuiging, maar ik hoop dat hij inziet hoe belangrijk het is om op te komen voor minderheden en diversiteit te omarmen.
4	Het is teleurstellend dat we nog steeds moeten discussi�ren over LGBTQ+ rechten in de sport. Laten we ons allemaal inzetten voor gelijke totstandkoming en tegen discriminatie.
5	Ik hoop dat K�k�c�u in de toekomst van gedachten verandert. De regenboogband dragen kan een krachtig symbool zijn voor onze gemeenschappelijke waarden zoals gelijke behandeling.
6	Zowel K�k�c�u's religieuze overtuigingen als inclusiviteit zijn beiden belangrijk. Het vinden van een goede balans is de sleutel tot vooruitgang hierin.
7	Ik vraag me af of er compromissen zijn besproken om aan beide waarden te voldoen. Het zou goed zijn om dit te leren en het gesprek gaande te houden.
8	Sporters zijn rolmodellen en beïnvloeden vaak hoe jongeren denken over waarden zoals inclusie. Het is belangrijk om hierbij stil te staan.
9	Het is jammer dat regenboogbanden nodig zijn in de eerste plaats, maar ze geven wel een kans om ons te verenigen.
10	Ik hoop dat K�k�c�u in de toekomst de kans grijpt om een verenigend statement te maken. Samen staan we sterker.
11	Het zou interessant zijn om de reacties van andere spelers en het team hierop te weten, omdat inclusiviteit van belang is voor het gehele team.
12	Het is cruciaal dat we openstaan voor een discussie en naar elkaar luisteren wanneer het gaat om waarden en diversiteit.
13	Ik hoop dat dit incident leidt tot meer gesprekken over diversiteit en inclusie in de sport en verder.
14	We moeten ons realiseren dat iedereen het recht heeft om hun eigen religieuze overtuigingen te hebben, maar ook dat inclusiviteit onze gemeenschappelijke verantwoordelijkheid is.
15	Het is bemoedigend dat het nieuws hierover verspreid wordt, omdat dit het gesprek over dit soort kwesties aanmoedigt.

Table 12: ZS-2 examples article 4

no.	Comment
1	Het is goed dat Jumbo de reclamespot stopt na kritiek, maar het is zorgwekkend dat niemand in het bedrijf, noch het reclamebureau, heeft ingezien hoe smakeloos de spot was voordat deze op televisie werd uitgezonden.
2	Het is triest dat de makers van de reclamespot niet inzagen dat de associatie met overleden bouwvakkers niet gepast was en dat het boycotten van het WK in zijn geheel gerechtvaardigd is.
3	Het is belangrijk dat we ons realiseren dat Qatar sterk tekortschiet in termen van mensenrechten en dat boycotten of afkeuren van het WK daar een manier kan zijn om onze afkeuring te laten blijken.
4	Het is verontrustend dat er zoveel mensen betrokken waren bij de productie van de reclamespot en niemand heeft bedacht dat het idee dat bouwvakkers feestvierend bij een Jumbo-supermarkt te zien zijn, ongepast was. Dit toont aan dat er een gebrek aan ethische overwegingen is.
5	De beslissing van Jumbo om de reclamespot te stoppen is een positieve stap, maar het had niet zo ver hoeven komen als het bedrijf had nagedacht over de gevolgen van hun keuze om bepaalde groepen mensen en beroepen af te beelden.
6	Het is triest om te zien hoe commercieel succes boven ethiek wordt geplaatst - het is duidelijk dat Jumbo alleen reageert op de kritiek vanwege de grote publieke verontwaardiging.
7	Hoewel het begrijpelijk is dat Jumbo de reclamespot stopzet, lijkt het meer op een publiciteitsstunt dan op een oprechte verontschuldiging.
8	Het is ontgoochelend dat er nog steeds merken zijn die niet inzien dat het afbeelden van stereotypen schadelijk kan zijn en bijdraagt tot in standhouden van misvattingen.
9	De reactie van Jumbo toont aan hoe belangrijk consumenten invloed kunnen hebben op het gedrag van merken - we moeten als consumenten ons bewust blijven van onze kracht en leiderschap verwachten
10	Wat een misselijkmakende reclamespot: hoe konden ze zich niet realiseren dat het compleet ongevoelig was om bouwvakkers af te beelden die feestvierden terwijl hun collega's zijn omgekomen in een ongeluk?
11	Als vaste klant van Jumbo ben ik zeer teleurgesteld dat het bedrijf zulke afschuwelijke reclamespot heeft gemaakt. Het had nooit zo ver moeten gaan.
12	Als marketeer voor een groot bedrijf weet ik hoe belangrijk het is om ethisch verantwoord te werk te gaan. Het is verontrustend dat Jumbo dit belang blijkbaar niet inziet.
13	Gelukkig laat de erg negatieve reactie op deze reclame zien dat de samenleving ethiek hoog in het vaandel heeft staan. We moeten zulke kritiek serieus nemen en merken verantwoordelijk houden voor hun acties.
14	Jumbo's reactie op de controverse rond de reclamespot toont aan dat het bedrijf niet bekend is met belangrijke sociaal maatschappelijke problemen en dat ze echt moeten nadenken over hun verantwoordelijkheid in de samenleving.
15	Het is diep treurig dat er nog steeds zo veel schadelijke stereotypen gebruikt worden in de reclame-industrie.

Table 13: FS-2 examples article 5

no.	Comment
1	Als ik het goed begrijp, is dit bedrag slechts een voorlopige schatting. Zou het niet beter zijn om te wachten tot er meer informatie beschikbaar is voordat we conclusies trekken?
2	Het is goed om te zien dat het kabinet bereid is om studenten tegemoet te komen. Echter, dit bedrag lijkt me niets meer dan een schijntje, gezien de kosten van het levensonderhoud voor studenten.
3	Als marketingmanager ben ik mij bewust van het belang van goede communicatie en transparantie. Ik hoop dat de minister ervoor zorgt dat er duidelijke informatie beschikbaar wordt gesteld over wie in aanmerking komt voor deze vergoeding en hoe zij deze kunnen aanvragen.
4	Ik ben zelf geen student, maar ik kan me voorstellen dat deze vergoeding weinig zal betekenen voor studenten die al duizenden euro's aan studieschuld hebben.
5	Het lijkt me logischer om de manier waarop hoger onderwijs wordt gefinancierd te herzien, in plaats van alleen eenmalige vergoedingen uit te delen.
6	Als hardloper en yogabeoefenaar weet ik hoe belangrijk het is om gezond te blijven. Ik hoop dat deze vergoeding er ook voor zorgt dat studenten gezonde maaltijden kunnen kopen in plaats van noodgedwongen te kiezen voor ongezonde, goedkope opties.
7	Het is fijn dat er aandacht wordt besteed aan studenten die geen studiebeurs hebben gekregen, maar het zou nog beter zijn als er meer financiële ondersteuning beschikbaar was voor alle studenten.
8	Als fervente reiziger ben ik me bewust van de kosten die hiermee gemoeid zijn. Ik hoop dat deze vergoeding studenten ook in staat stelt om extra-curriculaire activiteiten te ondernemen, zoals reizen of internationale stage-ervaringen op te doen.
9	Ik vraag me af of het niet beter zou zijn om deze vergoeding te koppelen aan het behalen van bepaalde studieprestaties, zodat studenten worden gestimuleerd om hun best te doen.
10	Als marketeer zou ik zeggen dat dit voorstel wellicht 'window dressing' is - in plaats van daadwerkelijk structureel onderwijsbeleid te veranderen en de positie van studenten te verbeteren.
11	Het is belangrijk dat er aandacht wordt besteed aan de financiële situatie van studenten, maar het is ook van belang dat er meer wordt gedaan om de kwaliteit van het hoger onderwijs te verbeteren.
12	Als foodie vind ik het opmerkelijk dat de minister geen voorstellen doet om de kosten van studiebenodigdheden zoals boeken aan te pakken. Dit zijn vaak de grootste uitgavenposten voor studenten.
13	Het lijkt me verstandig om de vergoeding in te zetten op manieren die studenten op de lange termijn helpen, zoals het creëren van betaalbaar momentopnames voor huisvesting, in plaats van slechts een eenmalig bedrag uit te delen.
14	Het is belangrijk dat de regering de studenten tegemoet komt - maar dat zou niet ten koste moeten gaan van andere beleidsterreinen, zoals de zorg of het milieu.
15	Ik vind het goed dat de minister zich richt op 'pechgeneratie' studenten, maar laten we niet vergeten dat ook andere groepen in het hoger onderwijs worstelen met financiële problemen.

Table 14: CL-2 examples article 6

no.	Comment
1	Dus in 2 weken tijd daalt het aantal besmettingen van 12000 naar 5500 per dag. Een daling van 60%. Maar de regering vindt dat blijkbaar aanleiding om nog zwaardere maatregelen door te voeren. Waar zit daar de logica in?
2	'Ik snap de weerstand tegen een avondklok niet zo goed. Als we die avondklok zetten tegenover de huidige situatie en de mogelijke gevolgen, dan is die avondklok echt niet zo'n probleem. Moet je nergens heen? Dan blijf je dus gewoon binnen. Pak een boek, zet de TV aan en blijf thuis. Zo'n lastige opgave kan dat toch niet zijn?'
3	Weer van die kansloze symboolpolitiek. Een avondklok heeft zin als er 's avonds nog steeds iets te doen is. Alles is echter al dicht. Je kunt al nergens heen. Geen restaurants, geen sportclubs, geen bioscopen, geen kroegen. Dit is wederom een maatregel die helemaal niets gaat oplossen. Het is enkel voor de Bühne: 'Kijk lieve Nederlandse bevolking. We doen er echt alles aan. En dan moeten jullie in maart wel weer even op ons stemmen.' Hoe lang gaan we dit soort onzin nog accepteren? - Dit denk ik.
4	Ik vind het invoeren van een avondklok een uitermate slecht idee. Het beperkt de vrijheid van het individu mijn inziens te veel. Ik werk 9 uur per dag en ik loop 's avonds na het eten altijd nog een stuk. Op die manier kom ik er nog even uit en krijg ik wat beweging, aangezien de sportscholen ook gesloten zijn. Een avondklok zou voor mij dus erg vervelend zijn, en horeca die nu vooral inkomsten genereren uit afhaalen (mogen bezorgers nog wel de straat op?) Raken die inkomstenbron ook nog kwijt. Er wordt nu alles ingezet op het bestrijden van het coronavirus (wat volkomen begrijpelijk is) maar er wordt volledig voorbij gegaan aan de mentale gevolgen voor veel mensen. Ik vind dat hier meer aandacht aan besteed moet worden. Bij velen is de rek er nu wel een beetje uit, en dan zijn dergelijke maatregelen naar mijn mening te invasief.
5	Ik wil me best aan een aantal regels houden ten gunste van de gezondheid van een beperkte risicogroep. Maar ik laat me niet onderdrukken, dit is gewoon absurd.
6	'Ik wil best mijn best doen om verspreiding te voorkomen, maar dan verwacht ik op z'n minst een beetje inzet om gebedsruimtes te sluiten. Het is een feit dat de gelovige gemeentes de meeste besmetting hebben, en er wordt daar niks aan gedaan.
7	Als er een avondklok komt, dan ga ik de straat op! Dat gaat mij persoonlijk te ver! In onze rechtsstaat leveren wij allemaal bewust een stukje van onze eigen autonomie in en daartegenover staat dat wij o.a. een stuk veiligheid van de overheid krijgen. Ik ben een veteraan en heb gevochten voor onze vrijheid op missie en besef heel goed wat vrijheid betekend en inhoud, maar niet tegen elke prijs!
8	Voor bijna elke genomen maatregel van dit kabinet de afgelopen jaren geven ze geen logische redenen waarom iets genomen wordt. Logisch ook als je twee leraren aan de macht hebt. Die zijn goed in lesgeven niet in een land besturen. Dat zag je de afgelopen 10 jaar.
9	Zojuist bij WNL een OMT lid die om 05:30 al in de auto zat om op tijd in de studio te zijn. Doet z'n beklag dat er zoveel drukte is op de snelweg en mensen niet thuiswerken. Om 05:30 zijn het hoofdzakelijk busjes met mensen die niet thuis kunnen werken. En wat doe je zelf in de studio? Zo'n gesprekje van 10 minuten kan toch ook prima via teams? Dit OMT lid ziet op 1 dag meer mensen dan ik in een week tijd.
10	Zolang er vanuit Londen elke dag 20 volle vliegtuigen landen op Schiphol allemaal gevuld met personen met een test van 72 uur oud dan heeft elke regel in Nederland totaal geen zin. Sluit zoveel mogelijk de grenzen.
11	Hou toch eens op met dat gezeur over besmettingen, dat is totaal niet relevant, het gaat om het daadwerkelijke aantal ziekenhuis opnames en of patiënten op ic komen. Besmet wordt je toch, een virus hou je niet tegen. We hebben nu de middelen om het effect van het virus te remmen, we gaan vaccineren, meer kan je niet doen. Zelfs de landen met de strengste lockdown zijn code oranje of rood. Een avondklok is niet te handhaven, is geen prioriteit van de politie, en heeft totaal geen nut. Wat dan met vrachtwagen chauffeurs die na een lange rit thuis komen? Ik ga mij daar zeker niet aan houden.
12	Onze regering is bereid om de persoonlijke vrijheden nog meer te beperken om enkel en alleen een signaal af te geven. Zo is het nu gesteld. En een bataljon witte jassen stelt het voor en is ook in de positie gemanouvreerd om het voor te stellen. En iedere avond is een andere witte jas in een opinie programma om de angstmening verder te verspreiden. En kritiek wordt niet geaccepteerd. Dan ben je sociaal. Wat is het sociale aan het feit dat bedrijven en ondernemers financieel aan de afgrond staan, dat kinderen niet naar school mogen, dat burgers niet meer het zelfbeschikkingsrecht hebben om zelf te bepalen hoe het eigen leven wordt ingericht. De regering en alle lockdown roeptoeters claimen steeds het morele en absolute gelijk. Dat er gedacht wordt over een avondklok en er mensen zijn die dan JA roepen beangstigd. Dat laat maar weer zien dat overheidsbeleid niet per definitie zuiver is en dat teveel burgers blindelings en kritiekloos volgen. Geschiedenis herhaalt altijd.
13	Weer een maatregel die vrijheid beperkt maar praktisch geen nut heeft. Wie is er buiten tussen 22:00 en 04:00? Klinkt hetzelfde als alcohol tot 20:00 verkopen i.p.v. 21:00. Mooi voor de Bühne maar praktisch nut of resultaat.....
14	Nog meer beperkingen? De jeugdzorg krijgt op dit moment 60% meer aanvragen van jongeren met (psychische) problemen. De eenzaamheid onder ouderen loopt op, in Enschede gaan nou een aantal ouderen naar de rechter om het voor elkaar te krijgen dat zij weer bezoek mogen ontvangen. Ondernemers die steeds wanhopiger worden, want wanneer krijgen ze weer grien licht? Een kabinet dat twee weken geleden al riep dat de lockdown waarschijnlijk verlengd ging worden, maar over de compensatie moet men nu nog gaan nadenken? Waarom vraagt niemand wat het ware probleem is? Er liggen 2700 mensen in het ziekenhuis. Nederland heeft 120 echte ziekenhuizen, dit komt neer op 22,5 personen per ziekenhuis. Wij zijn toch heel ver afgedegen met onze zorg dat 22,5 personen per ziekenhuis een probleem is! Ondertussen doen alle landen dezelfde maatregelen, maar nergens gaan de cijfers echt naar beneden. Is het dan geen tijd voor een andere aanpak?
15	Misschien eens heel serieus naar Japan kijken? Geen lockdown, wel heel serieus ventileren. Minste doden in de geïndustrialiseerde wereld. Zie artikel in the Economist. Natuurlijk, Japan heeft ook de meest slanke en gezonde oudere bevolking ter wereld. Maar waarom blijven we de oplossing zoeken in een richting die tot nu toe weinig tot niets oplevert?

Table 15: Human examples article 1

no.	Comment
1	Sport en politiek zijn zeker in het geval van Rusland niet los te zien van elkaar. Voorbeeld: de Russen wonnen in Tokio 71 medailles. 45 daarvan werden gewonnen door atleten die lid zijn van de Central Sports Club van het Russische leger. Van het leger dat dagelijks onschuldige burgers bombardeert, verkracht en vermoordt. Hulde voor het standpunt van Hidalgo en ongelofelijk dat het IOC hier nog over na moet denken. Zolang kogelkoppen en speervangen niet zijn verheven tot Olympische sporten hebben de Russische atleten niks te zoeken op de OS.
2	Hulde aan deze burgemeester, een land dat oorlog voert op de meest weezinwekkende is niet welkom op de Olympische Spelen. Daar hoeft je geen ellenlange betogen te spenderen!
3	Zowel Rusland als Belarus hebben niks te zoeken buiten hun eigen grenzen. Het is ook moedig van de burgemeester om zich zo expliciet uit te spreken tegen deze landen en hun sporters. Ik ben het er ook volledig mee eens. Zolang Rusland oorlog voert zijn ze nergens welkom.
4	Eerder werden atleten uit deze landen al geweerd vanwege grootschalige dopingprogramma's die door de staat werden gefaciliteerd (of zelfs opgedrongen als atleten niet wilden). Dat probleem is nooit echt verholpen en het werkt ook niet om ze dan onder neutrale vlag te laten meedoen. Los daarvan is de houding die Rusland en Belarus nu hebben in de invasie van Oekraïne en de grootschalige oorlog die Rusland daar voert, meer dan genoeg reden om ze van elk internationaal platform (of het nu politiek overleg of een sportwedstrijd is) te weren. De steun van de regimes in deze landen zal pas gaan wankelen wanneer de burgers zelf op meerdere vlakken de nadelige gevolgen ervan ervaren.
5	Hidalgo ziet het scherp. Zolang er bommen vallen op Oekraïne: geen Russen op de OS in Parijs. Klasse vrouw
6	Politiek is onlosmakelijk verbonden met internationale professionele sportmanifestaties en degenen die dat ontkenen leven in het stenen tijdperk. Meedoen onder neutrale vlag? Welk volkslied wordt dan gespeeld bij winst? Ik ben het dan ook volstrekt eens met de Parijse burgemeester.
7	Ik vind dit zo'n onzin. Sport en politiek dient gescheiden te zijn. We hebben de Amerikanen en Europese landen voor hun oorlogen ook nooit hun sporters belemmerd.
8	De atleten moeten hun regering, zeg Poetin, maar onder druk zetten om oorlog te beëindigen. Wanneer de oorlog nog voort duurt niet toelaten op de spelen.
9	Goed standpunt, zolang de oorlog woedt geen Russische afvaardiging toelaten. Probeer op deze manier maar een beetje druk op Poetin te houden.
10	Prima beslissing om Russische atleten niet toe te laten. Het kan toch niet zo zijn dat er dadelijk weer eentje rondloopt met een grote Z op zijn shirt. Rusland moet in alles uitgesloten worden. En kom niet met sport en politiek moeten gescheiden blijven want Rusland doet dat zeker niet.
11	Heel goed dat de burgemeester zich zo duidelijk uitspreekt. Het gaat erom dat alle Russen ervaren, dat wat nu gebeurt in Oekraïne onacceptabel is en dat je er als Rus nu gedwongen wordt erover na te denken en zelfs dat je er last van ervaart. Het is niet zo dat we elke Rus verantwoordelijk houden! Daarnaast kan op dezelfde manier Poetin en zijn klik geen gebruik maken van eventuele successen van de Russische sporters ter ere van het moederland!
12	Gewoon geen enkele Russische sporter mee laten doen., Zo ingewikkeld lijkt me dit niet. Niet mee eens? Moet je aan je leider vragen of hij wil stoppen met bommen gooien., Kan je er niks aan doen? Naja, dat kan je wel, maar je kiest ervoor om er niks aan te doen. En natuurlijk snap ik het menselijke aspect van zo'n sporter: Super vervelend als hij niet mee kan doen na alle training/voorbereiding, maar er zijn nu eenmaal belangrijkere zaken.
13	Symbool politiek en wegstrijkend van andere conflicten in de wereld. Deze burgemeester laat in alles niet zien waar de olympische gedachte over zou moeten gaan. Parijs is dus niet de geschikte locatie voor dit evenement.
14	Wat een waardeloos besluit. Sport dient als verbroedering en heeft niks met politiek te maken.
15	Laten we dan ook gewoon *alle* Russen weren van de Spelen, sporters, trainers etc. En ze dus ook niet onder een andere vlag laten uitkomen.

Table 16: Human comments article 2

no.	Comment
1	En hiermee wordt stapsgewijs toch mensen bijgestuurd om continu een periodieke prik te nemen. De huidige situatie kan niet beter voor 'Big Pharma'. Willen we echt eindeloos blijven booster? Op een gegeven moment moet het toch stoppen? Een prik was toch een vrije keuze? Nu worden dus toch dwang- en drangmaatregelen overduidelijk ingezet om mensen te prikken, want anders kan je simpelweg niet naar het buitenland. Dat hoeft niet eens vakantie te zijn, want ook voor zaken zul je dus geprikt moeten zijn. Laat staan dat de vrij verkeer van personen wordt ingeperkt door EU zelf. Je vrijheid om ergens heen te gaan, wordt hiermee begrensd. Dit wordt goedgepraat met termen als volksgezondheid en IC's die overbelast raken, maar we raken continu vrijheden kwijt voor het groter goed. Klinkt wel communistisch bijna.
2	Dus toch waar we allemaal bang voor waren; de boosterprik wordt de norm. Dit gaat dus een nieuwe groep mensen opleveren, waarvan ik er misschien een uit ga maken: De half-vaxers. Wel gevaccineerd maar niet geboost.
3	Dit zag iedereen toch wel van mijlver aankomen? Voor 2030 zit er ook de CO2 tax aan vast gekoppeld let maar op! Vlieg-vakantie naar Mexico zegt U? Rest van het jaar aan de tofu want niet meer genoeg punten voor het milieu vervuilende vlees.
4	Ik vind het bijzonder, ik heb Covid én twee vaccins gehad en zou dan ook nog eens binnen 9 maanden een derde prik moeten halen om mee te kunnen doen aan de maatschappij? Nou ja op basis van het vaccin dan, mijn herstellbewijs is dan wel weer een jaar geldig. Zijn ze helemaal gek geworden! Investeer in de zorg in landen waar de vergrijzing een feit is, gebruik desnoods het leger of wat dan ook wat nodig is om dat te bewerkstelligen, alle investeringen moeten gestuurd zijn om op de korte maar ook lange termijn de zorg-capaciteit drastisch te verhogen. Een QR maatschappij is in mijn ogen geen oplossing met name door de discriminerende en totalitaire effecten hiervan.
5	Heeft er iemand een lijst van "complottheorieën" die de afgelopen 2 jaar zijn uitgekomen? Of wacht, misschien is het handiger het andersom te doen, een lijst maken van "complottheorieën" die *niet* zijn uitgekomen. Die lijst is een stuk korter, denk ik.
6	Journalistiek wordt ook wel de waakhond van de democratie genoemd. Echter, is diezelfde sector momenteel uitgegroeid tot de grote blaffende bijthond van de zittende macht. Nergens lees je kritiek op het beleid, nergens wordt een controlerende rol aangenomen, nieuwsberichten van waarde (demonstraties, quarantainekampen e.d.) worden onder de mat geschoven het is te zot voor woorden.
7	Dus om de 9 maanden een prik halen als je wilt reizen? Gaat dit nog wel om de volksgezondheid of om zijn we gewoon de bankrekening van de farmaceuten aan het spekken? Kunnen we de contracten inzien die zijn afgesloten met deze bedrijven.
8	Een bepaald Kamerlid van ons parlement waarschuwt hier al maanden voor. Een never ending story. Je kunt hem niet mogen maar heeft achteraf heel vaak gelijk gekregen.
9	Gaat het nu nog over de gezondheid of is er wat anders aan de hand. Eerst was 2 prikken voldoende en nu mag je waarschijnlijk pas met een booster de landsgrenzen over. Waar eindigt dit? We zijn hier voor de start van de campagne al voor gewaarschuwd en nu lijkt het dan toch werkelijkheid te worden.
10	Maandenlang uitgekotst toen we beweerden dat het coronacertificaat en de QR-code geen tijdelijke maatregelen waren. Trouw laat vandaag weten dat die 'tijdelijke' coronapas en QR-code toch niet "tijdelijk" zijn. Het zijn blijvertjes. Daar zorgt de Europese Unie wel even voor. Wij riepen dat natuurlijk al tijden, maar dat was allemaal zottigheid van ons. Gestoorde complotten. Wappies. Tot nu. Want opeens wordt het 'gewoon' toegegeven. Want jawel, "de Europese Commissie werkt aan een permanent coronacertificaat." We zullen er voor de rest van ons leven dus aan vast zitten. Trouw over het certificaat. "Twee vaccinaties zijn waarschijnlijk niet genoeg om een QR-code te behouden." Je kunt er dus vanuit gaan dat je je straks elk half jaar moeten laten volspuiten van de EU. Doe je dat niet, dan krijg je gewoon geen werkende QR-code en kun je dus niet reizen, naar het restaurant, naar het café, of straks zelfs naar het werk of school. Wie zijn er nú gek he?
11	Er zijn Kamerleden die voor complotdenkers uit werden gemaakt toen ze dit soort beleid maanden geleden voorspelde. Dat voorspeld niet veel goeds voor de toekomst. Ik ben niet perse tegen vaccinaties, maar met elke prik verhoog je het risico van negatieve effecten van de prik ook. Voor de mensen die een laag risico lopen van het virus zelf is elke 9 maanden een booster dus uiteindelijk een groter risico.
12	De prachtige Europese Unie, vrij reizen tussen lidstaten zonder controle, dat was eenmaal. En zal helaas niet het enige zijn waar het bij blijft voor mensen die geen heil zien in vrijheden op basis van medische status (en wat er nog meer in de toekomst aan de QR-code gehangen gaat worden, want dat gaat gebeuren).
13	Dit kon je van mijlver aan zien komen. Maar enige tijd geleden beweerde Hugo toch echt dat bij steeds meer prikken de kans op bijwerkingen groter zou zijn. Of is dat nu opeens niet meer zo?
14	Als dit geen oorlog oplevert, dan weet ik het ook niet meer
15	Dat verklaard direct waarom de EU 4x zoveel vaccins had ingekocht als inwoners voordat de eerste prik gezet werd. Denk dat ze het allang wisten dat de prik toch niet zo goed was, of misschien toch wat anders aan de hand.

Table 17: Human comments article 3

no.	Comment
1	Wat mensen niet snappen is dat deze armband dragen niet enkel een signaal van een aanvoerder is naar de lgbt+ gemeenschap, waarin hij zegt dat deze groep er mag zijn (ook binnen het voetbal). De aanvoerder vertegenwoordigd iedereen op het veld (en achter de schermen). Als club laat je hier zijn dat het dragen van deze armband blootgesteld wordt aan de persoonlijke keuzes van de aanvoerder. Hij draagt dus niet de normen en waarden van de club. Dan mag je je als Feyenoord gaan afvragen of dit je aanvoerder moet zijn.
2	De verplichting om een symbool te moeten dragen van een bepaalde overtuiging is een ontoelaatbare beperking van de individuele vrijheid. Ik vind dat we hier een grens over gaan die we niet moeten passeren. Als de KNVB of een club zijn logo wil aanpassen naar een regenboog kleur dan is dat hun keuze als organisatie, maar ga niet individuen verplichten om die kleuren ook te dragen, wat het ook is.
3	'Ik zie het probleem niet zo. Het staat iedereen vrij om wel of geen statements te willen maken.
4	'Wat ik hier aan commentaar lees is ronduit belachelijk. Mensen suggereren voor hevige consequenties voor Kokcu's keuze. Maar waar ze compleet de plank misslaan is dat Kokcu de vrijheid heeft om te kiezen. Dat betekent niet perse dat zijn keuze een 'aanval' is op de LGBT-groep. Hij is er vrij in. Dat je er niet mee-eens bent is prima. Maar waarom dan ineens consequenties willen zien? Dan ben je eigenlijk helemaal niet vrij in je keuze. Gaat dat niet tegen ons gedachtegoed in? '
5	Kokcu heeft de vrijheid om dit vandaag niet te doen, laat ik daar mee beginnen, maar als Feyenoorder ben ik behoorlijk teleurgesteld in de keuze van onze aanvoerder. Bij Feyenoord zou iedereen welkom moeten zijn, dat kan een aanvoerder uitdragen. Kokcu draagt dat naar mijn mening hiermee niet uit.
6	Prima actie. Na mijn weten is het de bedoeling om iedereen te respecteren met de keuzes die hij/zij/het maakt. Maar je kunt niet van iedereen verwachten dat hij/zij/het een levenswijze van een ander gaat promoten.
7	Dit is zijn goed recht. Geen armband dragen staat niet gelijk aan geen respect hebben voor de gemeenschap. Hetzelfde gold voor Max. Max weigerde te knielen, toen de rest van de rijders een statement voor gelijkheid maakten. Velen juichten dat hier toe. Ze gaven aan dat Max hiermee niet voor ongelijkheid is, maar niet mee wil doen met een act. Velen vonden dat toen mooi. Dat geldt dan nu ook voor Kokcu.
8	Bij Feyenoord is op dit moment een groepje onder de naam Roze Kameraden hard aan het strijden voor erkenning binnen de club. En dan maakt de aanvoerder van de ploeg een statement tegen ze. Het is diep triest. Hij kan zich verschuilen achter van alles en nog wat, dat betekent gewoon dat hij zijn prioriteiten niet op orde heeft. Wat mij betreft is hij nu gewoon geen aanvoerder meer en zoekt hij in de winterstop maar een nieuwe club
9	Feyenoord moet hem de aanvoedersband definitief ontnemen. Je staat daar niet als privé-persoon, maar als aanvoerder die zijn club vertegenwoordigd, met alle normen en waarden van die club
10	Deze discussies horen niet thuis in de sportwereld vind ik. Als ik naar een sportwedstrijd ga is dat voor mij een uitje en ontspanning en wil ik niet worden meegezogen in allerlei politieke statements. Hoe goed deze misschien ook bedoeld zijn.
11	Keurig verwoord! Kan niks anders dan dit respecteren!
12	Hij geeft aan voor iedereen respect te hebben. Daarmee geeft deze voetballer een uitstekend signaal af. Dat hij geen regenboog armband wenst te dragen, is zijn eigen beslissing die niemand hem kan en mag ontzeggen.
13	Hij zou er zelfs geen reden voor hoeven hebben. Niemand hoeft daarin mee te gaan als hij niet wil. Ook al ben ik niet zo van geloofsovertuigingen enz. Hij heeft gelijk.
14	ziek om te zien hoe intolerant alle zogenaamde tolerante mensen zijn. als je een andere mening uitdraagt om welke reden dan ook is dat kennelijk niet geaccepteerd? deze man heeft zijn eigen gelijk, laat hem dan in zijn waarde en laten we stoppen met de dwang naar tolerantie.
15	De verplichting om je uit te spreken met het risico dat als je dat niet doet men ervan uitgaat dat je tegen bent (met alle gevolgen van dien) is een buitengewoon verontrustende ontwikkeling. Net als dat het mij geen reet aangaat hoe iemand zich identificeert kan een ander mij niet opleggen dat ik me uitspreek over bepaalde kwesties.

Table 18: Human comments article 4

no.	Comment
1	Toch heel apart dat niemand, zowel binnen de directie van Jumbo, als van het reclamebureau heeft bedacht dat deze spot wel heel erg fout is! Zulke professionals, dan verwacht je toch meer dan zoiets smakeloos.
2	Wel treurig dat klaarblijkelijk niemand bij Jumbo, productiehuis en acteurs de link met de dode bouwvakkers heeft weten te leggen. Sowieso zou iedereen dit WK moeten boycotten!
3	Wat een hoop schijnheiligheid weer. Het kabinet gaat wel, het kabinet gaat niet. Iran mag wel, Iran mag niet. We weten allemaal dat Qatar wars is van mensenrechten en dat slavernij daar nog heel gewoon is. Officials van de FIFA hebben zich vorstelijk laten betalen om de organisatie naar Qatar te laten gaan. Maar ondertussen gaat alles door in Qatar. En gaat zowel Iran, het kabinet als het Nederlands elftal heen. En dan zijn we verontwaardigd over het feit dat 50 fans zich laten feteren door Qatar en dat Jumbo bouwvakkers opvoert in haar commercial. Ik laat het WK voor wat het is. Ik ben voetbalfan, maar ik zal geen wedstrijd kijken. Helpt dat? Waarschijnlijk niet. Maar het WK is mijn aandacht niet waard, juist omdat het in Qatar is.
4	Ik vraag me toch echt wel af hoe dat gegaan is. Iemand bedenkt ineens dat het bouwvakkers moeten zijn die de polonaise lopen. Dan wordt dat idee voorgelegd, niemand die door heeft hoe stom dat is. Dan worden de opnames geplanned en wordt Donny gebeld. Die krijgt ook te horen dat ie met bouwvakkers moet dansen en zegt ook niks. Dan vinden de opnames plaats op een set die op een bouwplaats lijkt en krijgen de extra's bouwvakkerskledij aan. Nog steeds niemand die wat zegt. Dan wordt de reclame gemonteerd. Niemand die het doorheeft. Krijgen alle bobo's de reclame te... niemand die wat zegt. Dat zijn een hele boel gaten in de kaas die op dezelfde lijn zitten zeg.
5	Mensen die de reclame verdedigen, met argumenten waarom we wel naar het WK moeten gaan slaan echt volledig de plank mis. Kritiek is hier niet gericht op het wel of niet gaan naar het WK. Zelfs niet op de vraag of je überhaupt reclame moet maken voor het WK. De kritiek is gericht op de inhoud van de reclame, waar bouwvakkers vrolijk de polonaise lopen terwijl er 6500 bouwvakkers overleden zijn om dit evenement mogelijk te maken. Dan kan je hier komen betogen dat er overal in de wereld erge dingen gebeuren, of dat je anders helemaal nergens meer reclame voor mag maken, maar dat zijn allemaal non-argumenten in deze. Kan je reclame maken met vrolijke bouwvakkers voor een evenement met 1.5 keer het Excelsior stadion aan dode bouwvakkers? Dat is de vraag. Niet: moeten we naar het WK, zijn mensen hypocriet omdat ze een ook telefoon hebben, gebeuren er nog meer erge dingen, of andere ridicule drogredenen vermomd als 'slimme' vragen.
6	Negatieve reclame is ook reclame. Daarnaast was er met de reclame niets mis. Het was een stuk leuker dan de meeste reclames. Dat men de link gelijk legt met de bouwomstandigheden in Qatar is begrijpelijk, hoewel dat niet zo bedoeld zal zijn. Hopelijk komt er een nog leukere reclame voor terug
7	Dit begint langzaam de gedachtenpolitie te worden. Er gebeurt iets wat mogelijk in verband kan worden gebracht met iets anders. Vervolgens explodeert de situatie en komt het vingertje steeds maar weer van mensen die schijnbaar nog nooit iets verkeerd hebben gedaan in hun leven. Misschien was de actie van de Jumbo onhandig. Maar ik maak me meer zorgen om de backlash en moraalridders. Zeker als ze dit als een overwinning gaan zien.
8	Ik heb gratis en voor niks een nieuwe slagzin voor ze: Hallo, Dumbo! Niet te geloven dit. Hoe wereldvreemd kan een bedrijf zijn? Ik neem toch aan dat maken en uitvoeren van reclames heel wat mensen passeert.
9	Het zijn bouwvakkers, er zijn geen dubbele bedoelingen richting de stadion bouwers. Deze maatschappij ziet echt overal beren op de weg. Naast bouwvakkers zie ik ook andere mensen in het spotje, moeten we dat ook maar niet doen want die stadion bouwers zijn ook mensen. Dat er in Qatar misstanden zijn staat buiten kuif maar men moet is stoppen met het ziekelijk zoeken naar dingen die er niet zijn (niet alleen in deze situatie).
10	3 seconden van de 50 sec lopen ze over een steiger. Verder gewoon super leuk clipje en deuntje. Je kan overal wel spijkers op laag water zoeken. Iedereen heeft weer lange tenen en het wordt mijn inziens overdreven.
11	Dus het nederlands elftal gaat naar qatar. Er wordt eindeloos in sportprogramma gesproken over welke keeper mee mag. Een regeringsdelegatie komt kijken en een legioen fans ook. De televisierechten zijn aangekocht zodat de helft van het land kan kijken. En dan is men verontwaardigd om een spotje? Ik twijfel nog even over schijnheilig of hypocriet, het wordt hypocriet.
12	"Dansende bouwvakkers" - op zich er niets mis mee. Maar hoe verzin je het in deze context..?!? Ongevoelig, en ik druk me zachtjes uit.
13	'Hoeveel mensen zijn er wel niet betrokken bij het bedenken, produceren en live zetten van een reclame? Ik snap oprecht niet dat niemand binnen het marketing team en het management bedacht heeft, misschien is dit niet zo'n slim idee. Je zou bijna denken dat het bewust is.'
14	Ik verbaas me er vooral over dat niemand van zo'n reclamebureau de link legt die een gemiddeld intelligente Nederlander wel gelijk ziet. Je zou toch zeggen dat je langer dan een seconde nadenkt over een idee. Heel bijzonder.
15	Laten we eerlijk zijn: Het is op zich prima als een bedrijf de keuze maakt gewoon marketing te maken rondom het WK. Maar om dan zo enorm toondoof te zijn dat je in die reclame een grote rol weglegt voor vrolijke bouwvakkers, terwijl juist in die beroepsgroep zoveel doden zijn gevallen in Qatar... Dan ben je echt een apart soort blind voor wat er in de maatschappij leeft.

Table 19: Human comments article 5

no.	Comment
1	'Ik vind dit echt een schoffering. Kan er geen ander woord voor vinden. Je schaft de basisbeurs af en zegt onder het mom 'investeren in jezelf' dat deze groep duizenden euro's moet gaan lenen terwijl de generatie ervoor gewoon 10.000-15.000 euro over hun gehele studietijd ontving. Daartegenover zou dan een verbetering van het onderwijs staan. Deze heeft gewoon niet meetbaar plaatsgevonden. Daarnaast zeg je dat het geen impact heeft op het verkrijgen van een marktconforme hypotheek, voor deze jongeren dan de tweede grote schuld die zij aannemen. Wat blijkt in de praktijk. Dit wordt gewoon wel meegewogen. Vervolgens kom je er als overheid achter, misschien is het er allemaal toch niet eerlijker op geworden. Iets wat van te voren ook wel te redeneren viel. Deze studenten en afgestudeerden inmiddels voor 10.000 euro's het bos in gestuurd, krijgen nu dan 1400 euro compensatie. Pechstudent zou dus beter flesstudent genoemd kunnen worden. Want je bent gewoon gefleest door de staat.'
2	Beste politiek Den Haag, Met deze veel te lage compensatie schaden jullie het vertrouwen in de politiek over een generatie van 8 jaar. Mij ontgaat de hele redenatie waarom gekozen wordt voor zo'n extreem laag bedrag. Studenten uit de generatie 2015-2023 lopen gemiddeld zo'n 12.000 euro studiebeurs mis, niet 1400 euro. Mij is het inmiddels wel duidelijk. De politiek vertegenwoordigd vooral de belangen van de leeftijd 35 ~65. Val je daar buiten, dan ben je altijd de klos.
3	Elke euro is er 1 teveel. Je studeerde in die periode tegen de toen geldende regels. Andrrs heb ik ook nog een aantal wensen: 1) ik heb 22.000 euro overdrachtsbelasting betaald. Nu afgeschaft. 2) mijn vrouw is 6 jaar jonger. Toen ik trouwde kreeg men volledig AOW als 1 van de 2 65 werd. Nu niet kost mij $800 \times 12 \times 6$ dus ongeveer 60k 3) Eerdere generatiis mochten met 57,5 met de VUT. Nu afgeschaft.. Bij een van VUT 2.000 per maand was geweest kost mij dat $2.000 \times 12 \times 10$ Jr = 240k 4) Vroeger had je eindloonregeling als pensioen nu middelloonregeling. Kost toch gauw 300k. 5) ik droeg voo2 2 kinderen bij voor kinderopvang. Worft het nagenoeg gratis. Schade 25k Vroegere regels over volledige hypotheek renteaftrek, ouwe lullendagen, betaald ouderschapsverlof etc reken ik maar niets eens mee. Waar kan ik als echte pech generatie dit claimen? 1/8 zou al heeeel prettig zijn. Een kleine ton Moraal van het verhaal. Soms zit het mee, soms zit het tegen. Deal met de regels van het moment.
4	1 miljard euro, wat een grap, a 1000 euro pp. Waarom krijgen alle studenten uit het leenstelsel niet het bedrag op hun schuld in mindering dat zij onder het nieuwe stelsel hadden gekregen? Dus 255 per maand studeren (voor een afgeronde studie). Voor 5 jaar studeren komt dat (a 255eu) rond de 15.300 euro uit. Dat is wat de studenten voor en na het leenstelsel kregen en gaan krijgen. Als je eerlijk wilt compenseren, doe dat dan ook in plaats van deze pechgeneratie belachelijk te maken met zulke voorstellen.
5	Het misgelopen bedrag ligt rond de €13000 voor 4 jaar uitwonende basisbeurs. Ik begrijp echt niet hoe ze bij een tegemoet koming van €1400 zijn uitgekomen.
6	De zoveelste "blunder" binnen dit Kabinet. Wie bedenkt het bedrag van 1400 euro aan tegemoetkoming? Welke niet wiskundig onderleigd persoon heeft dit berekend? Ook onze dochter heeft geld moeten lenen aangezien zij op een veel te dure kamer woont en nu voor zeker 10000 euro benadeeld wordt ten opzichte van studenten waarvoor dit allemaal gaat veranderen. Zo niet eerlijk en pure diefstal als je het mij vraagt!
7	Ik zie weer een toeslagen affaire aankomen, maar dan voor de studenten . Die worden benadeeld door door de zoveelste misser van kabinet Rutte . Volgend jaar weer naar stembus.
8	Ik begrijp niet waarom ze er niet voor kiezen om dan een gedeelte van de schuld kwijt te schelden. Die €1.400 nu nog geven slaat nergens op.
9	wat is dit voor een schijntje. Basis beurs was ik geloof 290 euro pm. Dat is 3480 euro per jaar !. En zeg 4 jaar studeren dat is 13920. En dan gaan ze daar 10% can compenseren. Zo dat is een dikke vetpot voor mensen die ook al slechter onderwijs hebben gekregen door de lockdown.
10	Als ik in Leiden rond loop ziet het er niet naar uit dat de studenten aan de bedelstaf zijn. Terrassen, cafés en restaurants puilen uit van de studenten. Als ik dan ook nog lees dat ze op wintersportvakanties kunnen. Dan valt het allemaal wel mee. Met de studentjes
11	1400€ als je 4 jaar hebt gestuurd, dat is net geen 30€/maand. Zat je net een generatie hiervoor of erna, dan kreeg je het 3-voudige als je thuis woonde en al snel het 9-voudige als je uitwonend was. Schandalig.
12	Geen woorden voor, iemand met een uitwonende basisbeurs krijgt dit bedrag in 4 maanden. Deze pechgeneratie heeft gemiddeld 5 jaar moeten lenen met als resultaat 40000 euro schuld. Dat koop je als overheid af met 1400 euro. Degene die dit verzonnen heeft is in mijn ogen niet helemaal wijs.
13	'Geef gewoon helemaal niks terug. Iedereen heeft weleens pech met zoiets als dit. Toen mijn kinderen naar school gingen was er opeens een veel hogere kinderopvangtoeslag bijvoorbeeld. Mijn vader is met z'n 60e met pensioen gegaan. Ik moet door tot 67-70? Pech hebben hoort erbij.'
14	Wat een fooi. Jongeren hebben zich in de schulden gestoken of hebben helemaal geen studie kunnen doen die ze wilden omdat ze het onverantwoord vonden.
15	Ik snap dat het vervelend is dat er een gezellig tussenperiode is geweest en dat elke verandering niet perse een verbetering is, maar zolang er van belasting geld terrassen gevuld kunnen worden met studenten die zich drankjes van drie euro per stuk kunnen veroorloven is het misschien niet zo gek dat dit wordt aangepast. Ik denk dat studeren op zich voldoende wordt gesponsord en dat er nade studie er voldoende inkomen komt om een en ander te compenseren. Bovendien begint het bijdragen aan de belasting ruimschoots later dan een timmerman of stratenmaker en is het inkomen ook ruimschoots hoger en dus ook de pensioen opbouw ed. Bovendien zijn de werkzame jaren ook minder tot aan dat pensioen. Wat mij betreft bestaat de arme student niet, maar is er alleen maar een nieuwe politieke beslissing te maken.

Table 20: Human comments article 6