Can LLMs assist humans in assessing online misogyny? Experiments with GPT-3.5

Christian Morbidoni[1,*], Annalina Sarra[1]

[1]Università degli Studi G. d'Annunzio, Pescara, Italy

Abstract

Today's social media landscape is flooded with unfiltered content, which can range from hate speech to cyberbullying and cyberstalking. As a result, locating and eliminating such toxic language presents a significant challenge and is an active current research area. In this paper we focus on detecting hate speech against women, i.e. misogyny, exploiting a "prompt-based learning" paradigm with the aim of providing a first assessment of recent developed LLM (OpenAI's GPT-3.5-turbo). We experiment with a benchmark dataset of Reddit posts and evaluate different prompts types w.r.t. response stability, classification accuracy and inter-annotator agreement. Our experiments show that zero-shot detection GPT capabilities - against human annotations - outperform supervised baselines on our evaluation dataset and that ensembling different prompts possibly further improve the accuracy up to 91%. We also found that responses to specific prompts is quite stable, while slightly more variation and less agreement is observed when asking the questions in different ways.

Keywords

online misogyny detection, pre-trained language model, GPT, text classification, prompt-based learning,

1. Introduction

Concomitant with social media becoming a major means of communication, online toxic and hateful comments is increasing. Hate speech is typically defined as offensive, cruel or discriminatory statements (posts, images, comments, etc.) and actions (not just online) that show intolerance and hatred for a particular group or individual, based on their race, ethnicity, religion, sexual orientation, gender identity, disability, or other characteristic [1, 2, 3].

Online hate speech, in various forms like name-calling, slurs, threats, and harassment, has harmful effects. It spreads bigotry, causes emotional harm to targeted individuals and groups, and sometimes incites violence. It also creates a hostile and unsafe online environment, discouraging participation in digital communication. Social platform anonymity and evading legislation contribute to the growth of the phenomenon, which ultimately distorts the original purpose of social media, i.e. facilitating communication and enriching related activities regardless of geographic restrictions.

Therefore, the identification of potentially offensive language and information extraction

GENERAL'23: GENerative, Explainable and Reasonable Artificial Learning Workshop 2023, held in conjunction with CHITALY 2023

*Corresponding author.

christian.morbidoni@unich.it (C. Morbidoni); annalina.sarra@unich.it (A. Sarra)

0000-0003-0244-9322 (C. Morbidoni); 0000-0002-0974-0799 (A. Sarra)

from social media data become essential. Over the last years, there has been a growing interest in online hate speech detection. Although manual approach to the problem has the distinct advantage of rapidly capturing context and responding to new developments, the process is labour-intensive, expensive, and time-consuming, which restricts scalability and quick solutions. The automation of this task has become increasingly important as the amount of hate speech online has grown. Up-to-date applications investigating the problem of hate speech detection generally rely on pre-trained transformer-based language models, which are trained of then fine-tuned on hate speech detection tasks [4, 5, 6, 7, 8, 9]. While such supervised approaches show promising results, they need large annotated datasets to be trained on, and often require adaptation to transfer to different domains and corpora. In recent years, the growth of Large Language Models (LLMs) has led to a shift towards the "prompt-based learning" paradigm, demonstrating good capabilities on a variety of NLP tasks [10, 11]. LLMs have been shown to incorporate some general knowledge and understanding, enabling them to address unforeseen tasks, operate in few-shot of zero-shot setting, and even present emergent capabilities.

Recent AI tools like ChatGPT are quickly changing the way humans interact with machines, in a conversational fashion, similarly to communications among humans. People are in fact already using ChatGPT for a variety of everyday tasks, which open the way to great opportunities and innovations but also to dangerous threats. For example, as LLMs are pre-trained with a huge amount of (uncontrolled) human generated content, their answers can reflect inherent biases in the pre-train corpus, that might lead to unfair content being generated, e.g. penalizing minorities and perpetuating gender stereotypes [12]. However, this "biased knowledge" can possibly be leveraged for good, for example to automatically detect hate speech or to assist humans in annotating different types of hateful or unfair content [13]. Such a potentially beneficial application is further enhanced by the inherent explainability of responses. They can be easily instructed to generate a natural language motivation for the predictions themselves, along with the predictions. In this paper, we consider the largest GPT (Generative Pre-trained Transformer) language model available at the time of writing, developed by OpenAI, and presented by Brown in 2020 [10]. Specifically, we experimented with the recently released GPT-3.5-turbo model[1].

Our aim is to provide a first assessment of LLM's zero-shot capabilities in facilitating the detection of online misogyny. Online misogyny refers to the phenomenon of hostility, discrimination, and hatred directed towards women in online spaces. This can take various forms, including sexist comments, derogatory language, sexual harassment, cyberbullying, and online stalking. The task of categorizing online content into misogynistic or non-mysoginistic content is in fact challenging for humans as well, and prone to disagreement among distinct human annotators.

Given the impressive language understanding capabilities of OpenAI's GPT, we aim at assessing its ability to act as an annotator of misogynistic content, exploiting an expert-annotated dataset of Reddit posts, introduced by Guest et al. [14]. In our experiments, we design and evaluate slightly different prompt structures. We measure the stability of responses and classification accuracy against annotations collaboratively created by humans. Through this, we derive preliminary insights that might be useful for further investigating practical interaction

[1]https://platform.openai.com/docs/models/gpt-3-5

strategies with LLMs. We found that zero-shot GPT-3.5 prompts outperform supervised deep learning methods, including BERT, on the considered test set. Combining multiple prompts with a simple majority voting can further increase performance. Additionally, we observed that different prompt engineering choices affect both the classification metrics and the stability of results, measured by inter-annotator agreement.

The remainder of this paper is structured as follows: Section 2 gives background information. Section 3 describes the dataset and our experimental approach to zero-shot learning. Section 4 discusses results and Section 4.1 reports on a qualitative results analysis, focusing on missing predictions and disagreements between human annotators. Finally, Section 5 draws conclusions.

This study is partially supported by the ICOMIC project[2].

2. Related works

The detection of abusive language is an unsolved and difficult problem for the natural language processing community [15]. A significant portion of work already done on the abusive content automatic detection focuses on using supervised machine learning (e.g. [16, 17]).

Until recently, pre-trained language models, such as BERT, were commonly used to address a variety of NLP task by fine-tuning the model, i.e. running additional training over specific annotated datasets to adapt it to specific downstream tasks. In [14], an expert-annotated online misogyny dataset was presented, and two variants of BERT were trained and evaluated on this dataset. More recently, in Yin et al. [18], a novel variation of BERT called AnnoBERT was proposed. It incorporates annotator labels to improve training and classification performances, and it was evaluated on the same dataset. Alongside AnnoBERT, previously published methods, namely CrowdLayer [19] and Label-Embedding Attentive Model (LEAM) [20], were also evaluated on the same dataset. We based our experiments on this benchmark dataset, which is described in more detail in section 3.1.

Some studies have very recently documented how Prompt-based learning can be applied for spotting offensive or hateful content. Chiu et al. [13] use GPT-3.5 to detect hate speech using the prompts “Is this text racist?" and “Is this text sexist?". Schick et al. [21] demonstrate that toxicity in large generative language models can be avoided by employing similar prompts to self-diagnose toxicity during the decoding process. AlKhamissi et al. [22] recently framed hate speech detection as a few shot learning task. He obtained significant performance improvements by decomposing the task into its constituent parts and demonstrated that task decomposition of hate speech detection moves us closer to explainable systems.

[2]Identifying and Counteracting Online Misogyny in Cyberspace, EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021, https://csrlab.unich.it/en/icomic/

3. Experimental setting

3.1. Dataset

For this study, we selected the dataset curated by Guest et al. [14], that looks at the specific hate against women, and that is available on GitHub[3]. This expert-annotated dataset was obtained from Reddit, a social news website that is increasingly serving as a platform for numerous communities that support misogynistic ideologies.

The dataset consists of 6567 labels for posts and comments, collected from February to May 2020. We specifically chose this dataset for our experiments on misogyny detection because it benefits from a detailed hierarchical taxonomy, based on existing literature on online misogyny [23]. Misogynist Pejoratives, descriptions of Misogynistic Treatment, acts of Misogynistic Derogation, and Gendered Personal attacks against women are the four broad categories of misogyny considered in the data annotations. In order to capture more nuanced and extensive cases of misogyny and ensure a wider linguistic variety, the above mentioned authors use 12 subreddits to target the sampling. The dataset has been manually labelled by six expert annotators who have undergone thorough training in recognizing misogynistic abuse. Each annotator independently provided a personal judgement, then the the final labels were chosen through a process of group-based facilitation. For more details see [14].

The posts are divided into train-set (5264 posts) and test-set (1303 posts) in order to train and test supervised BERT models. We ran our zero-shot experiments over the test-set, in order to obtain results comparable with supervised learning experiments reported in literature. The proportion between misogynistic and non-misogynistic posts in the test set is about 0.11 (129 misogynistic vs. 1174 non-misogynistic), meaning the the dataset is highly unbalanced towards the negative class. Within the test-set, 119 posts (about 9%) obtained disagreeing judgements from different annotators, witnessing the high subjectivity of the classification task at hand. Among the posts finally labeled as misogynistic, about 48% received disagreeing judgements from single annotators, while they disagreed only on about 6% of the posts labeled as non-misogynistic.

3.2. Prompt engineering

We started our experiments using a simple prompt composed by a direct question ("Is the following post misogynistic?"), followed by the post to be classified. However, inspecting the responses from GPT-3.5, we observed that they were often verbose and it was difficult to map them to our two classes: Misogynistic and Non-Misognynistic. Thus, we introduced an answer conditioning aiming at obtaining a simple "Yes" or "No" as responses. In particular, we experimented with two alternative answer conditioning commands: a *soft* one, but appending at after the question the text "(yes/no)" and a *hard* one, using a more precise command: "(Only "Yes" or "No" are allowed as answers)".

As discussed later, the two conditioning commands provide indeed different results, reducing the fraction of post that cannot be easily mapped to the two classes. In our experiments, we use a simple mapping function that assigns the Misogynistic class to only those responses

[3]https://github.com/ellamguest/online-misogyny-eacl2021

that begin with a “Yes", and the Non-Misogynistic class to all the other responses, including those beginning with “No". Even if we use the conditioning particle, as described above, the LLM does not always provide a clear Yes/No answer. In some cases, the answer expresses the inability to classify the post, while in other cases, it contains a positive or negative answer but is formulated in a way that requires human interpretation. In practical applications, we might want to avoid or limit such unclear or uncertain answers. As a result, we also measure the fraction of unanswered prompts (unclassified examples) provided by each alternative prompt.

Besides, with the aim of investigating how additional context influences the results, we extracted from [14] a concise description of the different types of misogynistic and non-misogynistic content. This context is similar to that provided, along with annotation guidelines, to human annotators employed to label the dataset. The intuition is that by including a more precise description of the types of content we are going to classify in the prompt, this would condition the answers to be more precise with respect to the specific classification task. The text we used is extracted from [14] and is not reported here due to space constraints. It describes misogynistic pejoratives, derogation, counter-speech, and other content types.

As LLMs are non-deterministic in their responses, it is important to understand how “stable" the answers to a given prompt are. To achieve this, we run each prompt variation five times and attempt to measure the variability of the classification metric obtained, as well as the agreement on the classification of each specific post. To address the latter, we measure inter-annotator agreement between each pair of answers to the same prompt provided by independent runs, using Cohen’s K as the metric.

Finally, we are also interested in investigating how the results change in response to different formulations of the same question. As the LLM response is essentially a calculation of the most probable words following the prompt, using different words to formulate the question might affect the results. Thus, we first asked GPT-3.5 itself to provide 6 variations over our starting question.

We then compare the results obtained from each of the proposed variations. To summarize, in Table 1, we report the 7 questions used in our experiments (Q1 to Q7) and the 2 answer conditioning (C1 and C2). Our prompts can be therefore represented as follows:

```
[Ctx] + Qx + Cx + "Post:" + {text of the post}
```

where Ctx is optional and represents the additional context text, Qx is one of the questions (Q1-Q7) and Cx is the answer conditioning particle (C1 or C2).

The following is a random example of prompt using the Q1 question and the C1 conditioning.

> Is the following post misogynistic? (yes/no) Post: """ Don’t underestimate the cluelessness of a stacy. She would say you aren’t getting that attention because you aren’t posting photos or posing well or picking the right outfit or editing well. And through her eyes that phrase is true, even if she benefits from being the beauty standard. """

To carry out our experiments, we used the GPT-3.5-turbo API. The temperature parameter, controlling how randomly the prompt responses are generated, is set to 0. In this a way, we obtain answers that were as stable as possible. However, for some prompt variations, we also measure the stability of results over 5 runs, setting the temperature to 1 (see Table 3). In all the cases the *Top-P* parameter is set to 1.

Table 1
Summary of prompts elements used in our experiments

Questions	
Q1	"Is the following post Misogynistic?"
Q2	"Does the following post expresses contempt or prejudice towards women?"
Q3	"Does the following post contain misogynistic content?"
Q4	"Does the following post contain misogyny?"
Q5	"Is this post expressing misogynistic views?"
Q6	"Is this post conveying misogynistic opinions?"
Q7	"Is this post exhibiting any discriminatory attitudes towards women?"
Answer conditioning	
C1	(yes/no)
C2	(Only "Yes" or "No" are allowed as answers)

4. Results and discussion

In Table 2 we report the classification results obtained by combining different variations of prompts using the question Q1 and Q2, as detailed later in this section, and compare them with those obtained with supervised methods as reported in [14] and in [18]. As the two studies make use of different classification metrics, in Table 2 we report all of them, namely average precision, recall, F1 for the misogynistic class and overall accuracy (as done in [14]), and Specificity (True Negative Rate) and macro average F1 score (used in [18]). We note that Sensitivity (True Positive Rate), used in [18] is equal to the recall on the positive class (misogynistic). We use the "-" symbol when it is not possible to calculate a metric from the results reported in the original papers, and we mark in bold the best value, among those obtained with automatic methods (supervised or zero-shot), for each metric.

Additionally, we report the average metrics calculated by comparing the independent judgments of human annotators with the final labels (the ground truth), appointed after discussion among annotators. One can see that our zero-shot setting outperforms all the supervised methods, including those using BERT ([24]) trained on more than 5000 manually annotated posts, with respect to recall and F1 for the misogynistic class and macro-averaged F1. In particular, we notice a sensible improvement (+ 0.11) with respect to the main metric considered in the original study ([14]), F1-score on the Misogynistic class. These results suggest that the knowledge encoded in the LLM is largely sufficient to replace the specific training set. On the other hand, all methods are still far from reaching the accuracy of human annotators' judgment in such a hard and sometimes subjective classification task.

An exception is given by AnnoBERT which equals human annotators in specificity, i.e. the ratio of correctly detected non-misogynistic posts over the total non-misogynistic entries. We have to specify, however, that the average metrics obtained by single annotators are not directly comparable to those obtained by automatic annotation systems, as each annotator considered only a portion of the test dataset (an average of 135 posts annotated by single annotators).

In Table 3 we report the mean and standard deviation of the classification metrics, averaged over 5 independent runs, obtained with different prompt variations using Q1 as question, and using a temperature of 1. The first experiment (*Q1+C1*) uses the soft answer conditioning

Table 2
Main results of our zero-shot prompt based experiments compared with supervised baseline methods using human annotation process outcomes as a ground truth.

Supervised methods						
	Precision (mis.)	Recall (mis.)	F1 (mis.)	Specificity (mis.)	Macro F1	Accuracy
BERT (unweighted)	**0.67**	0.30	0.42	-	-	**0.93**
BERT (weighted)	0.38	0.50	0.43	0.93	0.69	0.89
CrowdLayer	-	0.09	-	0.90	0.49	-
LEAM	-	0.09	-	0.94	0.51	-
AnnoBERT	-	0.37	-	**0.97**	0.69	-
Zero-shot GPT-3.5						
	Precision (mis.)	Recall (mis.)	F1 (mis.)	Specificity (mis.)	Macro F1	Accuracy
Prompts ens. Q1	0.44	**0.67**	0.53	0.93	0.74	0.90
Prompts ens. Q2	0.47	0.63	**0.54**	0.94	**0.75**	0.91
Human annotator (mean)	0.69	0.83	0.71	0.97	0.84	0.93

Table 3
Results obtained with variations of the prompt using Q1 as question

Prompt	Precision (mis.)	Recall(mis.)	F1 score(mis.)	Accuracy	Missing predictions	K over 5 runs
Q1+C1	0.34±0.01	**0.77**±0.00	0.48±0.01	0.86±0.00	0.17±0.00	0.96±0.010
Q1+C2	0.39±0.00	0.70±0.01	0.50±0.00	0.89±0.00	0.04±0.00	0.98±0.003
Ctx+Q1+C1	0.41±0.01	0.65±0.02	0.50±0.01	0.89±0.01	0.07±0.01	0.87±0.025
Ctx+Q1+C2	**0.49**±0.01	0.53±0.03	0.51±0.01	**0.92**±0.00	**0.02**±0.00	0.91±0.019
Prompts ens. Q1	0.44±0.01	0.67±0.01	**0.53**±0.01	0.90±0.00	0.04±0.00	0.98±0.006

and already provides a F1 score over the Misogynistic class that is higher than the supervised baselines (see Table 2). It also provides a relatively high recall (0.77) but a low Precision (0.34). A problem with this prompt is, however, the high fraction (17%) of posts with a missing prediction, where the LLM responses did not contain a clear Yes/No answer. Such a problem is mitigated by the use of a stronger answer conditioning (*Q1+C2*). In this case the percentage of missing predictions lowers to 4%, the F1 score registers a improvement of 0.2, while the recall drops by 0.7. Including more context into the prompts provides some improvement with respect to recall, at the cost of a lower recall. However,we can see that the *Ctx+Q1+C2* prompt, preposing additional context and using a stronger answer conditioning, provides a slightly higher F1 score (0.51). A further improvement in the F1-score can be obtained by combining the responses from the 4 different prompt variations (see Prompts ens. Q1 in Table 3). In our experiment, we ensembled the predictions by means of a simple majority voting. In cases where there is a equal number of predictions for the two classes we marked the post as "missing prediction". We can see that in this case the F1 reaches 0.53, while the precision, recall, overall accuracy and fraction of missing predictions are between the maximum and minimum provided by the single prompts. Table 5 provides the full classification report relative to the *Prompts ens. Q1* setting.

From Table 3 one can see that there is almost no variation of metrics over the 5 runs in the case of *Q1+C1* and *Q1+C2* prompts, with a single exception of a 0.01 standard deviation for the recall in the case of *Q1+C2*. On the other hand, the standard deviation is higher when additional context is included (*Ctx+Q1+C1* and *Ctx+Q1+C2*). This suggests that there is a higher

Table 4
Results obtained with differently phrased questions, postposing the C1 “soft" answer conditioning text.

Prompt	Precision (mis.)	Recall (mis.)	F1 (mis.) score	Accuracy	Missing predictions
Q1+C1	0.34	0.77	0.48	0.86	0.17
Q2+C1	**0.43**	0.65	**0.52**	**0.90**	0.10
Q3+C1	0.35	0.75	0.48	0.87	0.07
Q4+C1	0.36	0.78	0.49	0.87	**0.05**
Q5+C1	0.37	0.74	0.49	0.87	0.12
Q6+C1	0.36	0.77	0.49	0.87	0.11
Q7+C1	0.34	**0.81**	0.48	0.86	**0.05**

Table 5
Detailed classification metrics of the prompt ensembles with prompt variations based on Q1 and Q2 questions. Results are averaged over 5 runs.

	Q1			Q2		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Non-Misogynistic	0.97	0.92	0.95	0.97	0.94	0.95
Misogynistic	0.44	0.67	0.53	0.47	0.63	0.54
accuracy			0.90			0.91
macro avg	0.70	0.80	0.74	0.72	0.78	0.75
weighted avg	0.93	0.90	0.91	0.93	0.91	0.92

variability in responses when the prompt include more details. To further investigate this issue, we measure the average inter-annotator agreement between the distinct runs of the same prompt, using the Cohen’s K coefficient. This provides an indication of the “stability" of the classification of each post. The inter-annotator agreement matrices for the 4 prompts variations are reported in Figure 1 a), b) c) d), while the mean and standard deviation of the inter-annotator agreement of each prompt over 5 distinct runs is reported in the last column of Table 3. As one can see, the introduction of additional context (Figure 1, panels (c) and d)) sensibly decreases the inter-annotator agreement if compared to the prompts without additional context (Figure 1, panels (a) and b)), thus providing less stable results. On the other hand, the use of the C2 particle (Figure 1, panels (b) and d)), increases the stability. Overall, we can observe that the responses are reasonably stable within the same prompt, as a value of the Cohen’s K greater than 0.8 is generally considered to indicate almost perfect agreement. However, if we measure inter-annotator agreements among the 4 different prompt variations using Q1 as question, we notice that the Cohen’s K value significantly drops, as shown in Figure 1, panel e).

In Table 4 we report the results obtained with different phrasings of the question Q1, which have been retrieved querying GPT-3.5 itself, as described in section 3.2 and reported in Table 1. It is evident that the classification metrics obtained vary sensibly depending on how the question is formulated. In particular, Q2 reaches the highest F1-score and precision on the Misogynistic class, while Q7 provides the higher recall. We notice that these two questions are the only ones that do not contain the word “misogyny" or “misogynistic". Looking at the inter-annotator agreement matrix (Figure 1) panel f), we can observe that the Cohen’s K varies between 0.94 (almost perfect agreement) between Q5 and Q6, which are similar, and 0.74 (moderate agreement) between Q1 and Q2, where the phrasing of the question is very different.

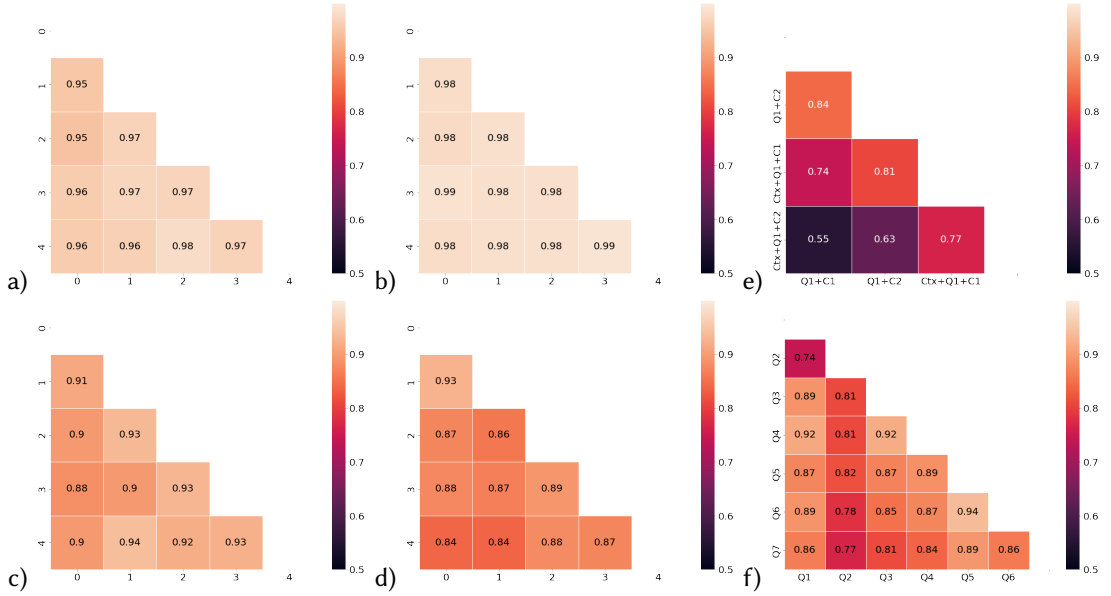

Figure 1: a) b) c) d): Inter-annotator agreement (Cohen's K) matrix over 5 distinct runs of each prompt variation based on Q1 question. a) Q1+C1, b) Q1+C2, c) Ctx+Q1+C1 and d) Ctx+Q1+C1; **e)** Cohen's K matrix for the 4 prompt variations using Q1 as question; **f)** Cohen's K matrix for the 7 different question formulations

Finally, as done for Q1, we designed 4 prompt variations using the question Q2, that provided the highest F1 score, and combined their predictions with a majority voting. Classification results over the test set are reported in Table 5. If we compare such results with those reported in Table 5, we observe an additional 0.1 increment in the F1 score and overall accuracy.

4.1. Qualitative results analysis

4.1.1. Missing predictions

In our experiments we reported a number of missing predictions. These can happen for different reasons: a failure in mapping the response to "yes" or "no", or, in the case of ensemble prompts, an equal number of positive and negative predictions. To further investigate this issue, we manually analysed the comments causing missing predictions. In general, as expected, the LLM seems to perform better on questions with more obvious labels, and worse on questions with less obvious labels. In particular, missing prediction occurs frequently when there is insufficient information in a comment to formalise an answer. Automatically categorizing text, like human language, is dependent on the complex and minute contextual associations between words and phrases. In these cases, GPT-3.5 often produces a hedged response of the form: "As an AI language model, I cannot accurately determine the context of the post..." This happens in 1.5% of the posts, always in a very short post categorised by annotators as non-misogynistic. Here are few examples: *"That wasnt a feminist - it was a misyognistic transwoman "*, *"On a thread about Billie Eillish body shaming"*, *"So true. Saw this in my own family."* Such posts mentioning

a lack of context, once detected, can be mapped to the non-misogynistic class (following the annotation practice implemented in our test dataset), or even serve as suggestions for humans to gain more context (e.g. looking at previous posts or authors activity) to come to a final judgement. In other cases, e.g. *"based and pinkpilled"*, *"Over for catcels"*, GPT's answer stressed how important it is to consider the intention and tone of the post, as well as the broader societal and cultural context in which it was made, as in the following:
"[...] this post [...] lacks context and clarity. However, the term "catcels" may allude to the incel community, which is known for promoting misogyny and hostility toward women."

Other posts that are difficult to understand are those that promote gender stereotypes. In these cases, the model frequently lacked the sensitivity required to label the comment as misogynistic. The following comment: *"Oh hell nah. Women of all races can be crazy, there are just some cultural differences on the style of crazy"*, is one example. According to the GPT-3.5 response, the statement does not appear to be directly misogynistic, but it could be interpreted as stereotyping women and perpetuating harmful cultural generalisations.

4.1.2. Controversial posts

We complement the analysis by looking at comments where there was a disagreement among human annotators. The reasons behind disagreement can be various. The literature has frequently noted how annotators' sociodemographic characteristics, moral principles, and lived experiences shape their interpretations. This is especially true when they must recognize online abuse and hate speech [25]. For instance, Waseem and Hovy [26] found that feminist and anti-racist activists frequently disagree with the annotations of hate speech made by crowd workers.

In the dataset used in our experiments, subjective beliefs of the annotators are frequently the source of disagreement. Few anecdotal examples are listed below: —*"I dont hate myself. I hate everyone else (men and women)"*; —*"Imagine the police believed her given the current climate over there. He would have been fucked over majorly. People like her disgust me. Trash."*; —*"I swear men can be heartless *obligatory not all. But I would overhear some guys being like damn shes ugly but we banged. Cant be seen with her. Ehhh "*. Around 10% of the post received contrasting annotations from distinct annotators at the first round. Such proportion is smaller (6%) when the post was finally marked as non misogynistic in the final round, but it increases up to 48% if we consider the misogynistic posts. As expected, GPT struggles with these kinds of posts and often fails to align with final human judgement: prediction accuracy of our models drops to 0.53 when dealing with such controversial cases. However, this is mostly due to negative samples (accuracy of 0.34), while the 80% of such controversial misogynistic samples are correctly detected.

5. Conclusions

In this paper, we presented the results of our experiments with GPT-3.5 in the hard task of automatically detecting online misogyny. Our aim was to provide a first assessment of the alignment with human annotators and of the stability of the predictions obtained via different types of prompts. Our preliminary results suggest that the use of pre-trained large language models, known to possibly generate content which reflects widespread prejudices, can, in turn,

support detection of misogynistic social media content. We also observed, however, that stability of the annotations is sensibly dependent on the prompt phrasing and context. We hope this study can be of help for further research towards understanding how emerging AI models can be integrated with human efforts in countering online toxicity. Possible next research steps include investigating the detection of specific types of misogyny (e.g. pejorative and derogation) as well as challenging non-misogynistic content (e.g. counter-speech, use of irony). While in this study we experimented with OpenAI's GPT in pure zero-shot capabilities, different and possibly open language models[4] need to be evaluated, and few-shot learning and instruction-tuning techniques could be investigated to improve human alignment.

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys 51 (2018) 1–85. doi:10.1145/3232676.

[2] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523. doi:10.1007/s10579-020-09502-8.

[3] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Computer Science 7 (2021). doi:10.7717/peerj-cs.598.

[4] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, ime of your hate: The challenge of time in hate speech detection on social media, Applied Sciences 10 (2020) 4180. doi:10.3390/app10124180.

[5] M. Uzan, Y. HaCohen-Kerner, Detecting hate speech spreaders on twitter using LSTM and BERT in english and spanish, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to -24th, 2021, G. Faggioli and N. Ferro and A. Joly and M. Maistro and F. Piroi, CEUR-WS.org, 2018, pp. 2178–2185. doi:https://ceur-ws.org/Vol-2936/paper-194.pdf.

[6] S. Banerjee, M. Sarkar, N. Agrawal, P. Saha, M. Das, Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages, 2021. arXiv:2111.13974.

[7] L. E., S. R., K. G., M. K., TheNorth @ HaSpeeDe 2:BERT-based Language Model Fine-tuning for Italian Hate Speech Detection., in: EVALITA Evaluation of NLP and Speech Tools for Italian., Accademia University Press., 2020.

[8] S. Das, P. Mandal, S. Chatterji, Probabilistic impact score generation using ktrain-bert to identify hate words from twitter discussions, 2022. arXiv:2111.12939.

[9] H. Nghiem, F. Morstatter, "stop asian hate!" : Refining detection of anti-asian hate speech during the covid-19 pandemic, 2022. arXiv:2112.02265.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language

[4] https://huggingface.co/models

models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems., Curran Associates, Inc., 2021.
[11] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre, Training compute-optimal large language models, 2022. arXiv:2203.15556.
[12] L. Lucy, D. Bamman, Gender and representation bias in GPT-3 generated stories, in: Proceedings of the Third Workshop on Narrative Understanding, Association for Computational Linguistics, Virtual, 2021, pp. 48–55.
[13] K.-L. Chiu, A. Collins, R. Alexander, Detecting hate speech with gpt-3, 2022. arXiv:2103.12407.
[14] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An expert annotated dataset for the detection of online misogyny, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online. Association for Computational Linguistics., 2021, p. 1336–1350.
[15] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6193–6202.
[16] A. Gaydhani, V. Doma, S. Kendre, L. Bhagwat, Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach, 2018. arXiv:1809.08651.
[17] T. Kanan, A. Aldaaja, B. Hawashin, Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in arabic social media contents, Journal of Internet Technology 21 (2020) 1409–1421. doi:https://jit.ndhu.edu.tw/article/view/2376.
[18] W. Yin, V. Agarwal, A. Jiang, A. Zubiaga, N. Sastry, Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection, 2023. arXiv:2212.10405.
[19] F. Rodrigues, F. C. Pereira, Deep learning from crowds, 2018, p. 1611 – 1618.
[20] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, L. Carin, Joint embedding of words and labels for text classification, volume 1, 2018, p. 2321 – 2331.
[21] T. Schick, S. Udupa, H. Schütze, Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP, Transactions of the Association for Computational Linguistics 9 (2021) 1408–1424. doi:10.48550/arXiv.1910.10683.
[22] B. AlKhamissi, F. Ladhak, S. Iyer, V. Stoyanov, Z. Kozareva, X. Li, P. Fung, L. Mathias, A. Celikyilmaz, M. Diab, Token: Task decomposition and knowledge infusion for few-shot hate speech detection, 2022. arXiv:2205.12495.
[23] M. Anzovino, P. Fersini, E.and Rosso, Automatic identification and classification of misogynistic language on twitter, in: Natural Language Processing and Information Systems. NLDB 2018. Lecture Notes in Computer Science(), vol 10859., Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F., Springer, Cham., 2018.
[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, volume 1, 2019, p. 4171 – 4186.

[25] D. Patton, W. Blandfort, M. Gaskell, S. Karaman, Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators, in: Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[26] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93.