Problematic Webpage Identification: A Trilogy of Hatespeech, Search Engines and GPT

Warning: The paper contains examples which the reader might find offensive.

Ojasvin Sood **Sandipan Dandapat**
Microsoft Corporation
{ojsood,sadandap}@microsoft.com

Abstract

In this paper, we introduce a fine-tuned transformer-based model focused on problematic webpage classification to identify webpages promoting hate and violence of various forms. Due to the unavailability of labelled problematic webpage data, first we propose a novel webpage data collection strategy which leverages well-studied short-text hate speech datasets. We have introduced a custom GPT-4 few-shot prompt annotation scheme taking various webpage features to label the prohibitively expensive webpage annotation task. The resulting annotated data is used to build our problematic webpage classification model. We report the accuracy (87.6% F1-score) of our webpage classification model and conduct a detailed comparison of it against other state-of-the-art hate speech classification model on problematic webpage identification task. Finally, we have showcased the importance of various webpage features in identifying a problematic webpage.

1 Introduction

Since the advent of the Internet, there has been a rapid rise of content being generated by both users and organisations, which has also expedited the rise of hateful and violent content. In this paper, we focus on the identification of such content within webpages on the internet. We define webpages promoting hate and violence against individuals and communities as *Problematic* webpages. These problematic webpages often affect various Search Engines, which index such webpages resulting in them showing up in the search results. Problematic webpages can be indexed by the search engine crawlers when crawling the web. The ranking models executing at the back-end of these search engines can end up showing these problematic webpages, when a user queries for something similar. This is not just limited to user queries, which are themselves having a problematic intent. Such problematic webpages can also show up in information seeking innocuous queries on sensitive topics as well. For example, there is a possibility that a hateful webpage towards black community ends up as a search result for a query: *data on black population in US*. This can lead to a bad experience for the end user, as well as spread of targeted hate against certain individuals and communities. Thus, problematic webpage classification has its applications in search engine indexing, and ranking to filter out such webpages in search engine results and stop spread of such problematic content on the internet. Problematic webpages also contribute to hate speech texts in social media as part of a post, comment. Aljebreen et al. (2021) have estimated that 21% of tweets in general have URLs shared within them. Hence, problematic webpage classification can also be additionally applied to improve hate speech detection as the URLs shared within the tweet can be an important feature to classify the underlying text.

A lot of research has happened on automatic hate speech detection, with several hate speech datasets (Mollas et al., 2022; Mathew et al., 2021; ElSherief et al., 2021; Ousidhoum et al., 2019; de Gibert et al., 2018), and models (Kim et al., 2022; Caselli et al., 2021a; Sarkar et al., 2021; Rajput et al., 2021). These data sets and models primarily focus on identifying hate and violence in short-text data in the form of posts, comments on various social media platforms. The existing hate speech models can be leveraged to identify user queries which might lead to problematic webpages showing up in the top results of a search engine. However, these model cannot be used to identify problematic webpages and remove them from the top results for such queries. Many hate speech detection models (Caselli et al., 2021b; Sarkar et al., 2021; Ousidhoum et al., 2019) replace URLs with a placeholder *URL* from the short-text before conducting hate speech detection. This indicates that the underlying information in the URL shared along with the corresponding piece

The 7th Workshop on Online Abuse and Harms (WOAH), pages 126–137
July 13, 2023 ©2023 Association for Computational Linguistics

of text is lost when classifying that speech as hateful. Hence the current hate speech detection techniques are focused on a part of the boarder scope of online hate and abuse. These hate speech models do not focus on classifying webpages which are prominent in spreading hate and violence in two main forms. Firstly, problematic webpages can show up as results in search engine queries, and secondly, as part of posts, comments in popular user generated content sharing platforms. Therefore, identifying a problematic webpage is very much crucial to stop online hate. To the best of our knowledge, there has been little to no research in identifying webpages which promote hate and violence of various forms.

A webpage is a very complex object as compared to short-texts and contains a variety of associated features which includes URL, Title, Body, Links, Ads. Existing state-of-the-art hate speech detection models often are limited towards detecting shorter text-based hateful content (e.g. tweets, social media posts, reviews etc.). In this paper, we show that these existing SOTA hate speech detection models are not effective in solving the problem from the perspective of detecting problematic webpages. This is due to complex structures of webpages, large amount of context present within them, and the nature of the data used to train these hate speech models. Some of these issues can be addressed with a new classification model dedicated to identifying problematic webpages. We show that such a model trained on data created from webpages containing important features, and annotated with GPT-4 does much better than state-of-the-art hate speech detection models.

There exist a lot of challenges to build such a dataset of webpages, both collecting and annotating. Hateful, violent Webpages cannot be easily discovered and mined. Webpage is also not something that can be generated synthetically, rather can only be mined from the World Wide Web. A lot of work has happened in website classification with respect to phishing, e-commerce website classification such as (Yang et al., 2019; Bruni and Bianchi, 2019). These works primarily focus on the developing classification models and often ignore the process of mining candidate webpages to build such classifiers. This calls for a strategy to discover and mine webpages on the internet to build a comprehensive data set, and eventually build a classification model.

Annotating webpages is also a very challenging problem which requires to consider various webpage features such as URL, Title, Headings, SubHeadings, Body, Links, Ads. The problematic webpages are ever-evolving, and are very subtle while promoting hate. To address these issues, sometimes it is required to look at the entire page content which can be very long. In the webpages curated as part of this work, we observed that the webpage body contains a large number of tokens (5350 ± 205). Some webpages discussing sensitive topics in the form of news and information can be misinterpreted as problematic. Similarly, some webpages which are very subtle when promoting problematic content such as political propaganda and spreading hate against a community can be misinterpreted as non-problematic. Hence, it is a complex task that requires a lot of time and attention for a human judge to annotate these webpages. This is a major challenge making it very difficult to build a large scale annotated problematic webpage data set. Therefore, we plan to use GPT-4 (OpenAI, 2023) to annotate these webpages at larger scale, as it has been observed that it exhibits strikingly close to human-level performance on complex benchmark tasks (Bubeck et al., 2023). It is difficult to use GPT-4 as a classifier on its own due the scalability issue towards annotating billions of webpages. Thus, we use GPT-4 to annotate a significant amount of data to train a reasonably accurate classifier which further can be used in scale for labelling large volume of webpages.

Therefore in this paper, we focus on creating a fine-tuned Transformer-based (Vaswani et al., 2017) webpage classification model focused on webpage text features: *URL*, *Title*, *Headings* and *Paragraph Texts* to identify problematic webpages promoting harmful content. We present a novel webpage data collection and annotation strategy, and use that to create the training, validation, and measurement set which can help future research in this area. We will release all the data publicly upon the acceptance of the paper.

The main contribution of the paper are as follows:

- We propose a novel strategy to create dataset in any webpage classification tasks using short-text dataset available (often easily) for the similar tasks and search engines.

- We also developed a precise few shot GPT-4 prompt to annotate harmful webpages using various features from the webpage.

- We have created a comprehensive and diverse data set which will be useful in future research in problematic webpage classification.
- We have fine-tuned a Transformer-based model for classifying problematic webpages with a reasonably good quality with F1-score of 87.6%.

2 Data Collection and Annotation Process

As mentioned in the previous section, there are many challenges in creating an annotated problematic webpages dataset. Some such problems include: discovering potential candidate webpages from the internet, annotating webpages which are often comprised of large volume of text and has rich meta information, and feature extraction to appropriately represent the webpage. We propose a novel solution to discover candidate webpages by leveraging popular search engines. For feature extractions, we have used popular web scraping tools, and also processed the input data to create important features (cf. Section 2.1). The webpage annotation (cf. Section 2.2) using GPT-4 takes care of the complex structure and longer token sequence. Algorithm 1 illustrates the data curation and annotation process. The details of the steps are described in the following subsections.

2.1 Webpage Data Curation

Webpage data curation starts with collecting existing hate speech short-text data sets (refereed as *HSData* in step 1 of the algorithm) dealing with different forms of hate and violence. We curated multiple data sets published in this field such as those described in (Mollas et al., 2022; de Gibert et al., 2018; ElSherief et al., 2021; Davidson et al., 2017; Kennedy et al., 2020). Each of these public data can be consider as one data point H_i. In the step 3, we therefore pre-process the data to remove unnecessary spaces, non-ASCII characters and stop words in *Preprocess*() function.

We use popular search engines to mine webpages using the short-text hate speech data. The intuition behind using search engine comes from the sophisticated indexing, ranking which often helps in retrieving relevant webpages Das and Jain (2012). This makes search engines well suited to mine relevant webpages given some related short-text data. In Step 4, the pre-processed short-text hatespeech data (H_i) is queried against popular web search engines (S_j) by leveraging their APIs from Google,[1] and Bing,[2] to mine the top 10 webpages $U_{i,j}$ for these pre-processed queries(H_i) using search engine S_j. The URLs $U_{i,j}$ mined in the above steps (step 1 to 7) are then de-duplicated in Step 8. The number of unique URLs mined from these search engines for each of the short-text datasets are shown in Table 1.

Algorithm 1 Webpage Data Collection & Annotation

```
Input:
    $HSData$ = { $H_1, H_2, \ldots$}
    $SearchEngines$ ={ $Google, Bing$ }
Output:
    $W$ = { $(U_1, T_1, B_1), (U_2, T_2, B_2), \ldots$ } //W indicates set of webpages, U indicates URL, T indicates Title, B indicates BodyText,
    $O$ = { $(W_1, L_1), (W_2, L_2), \ldots$ } //L indicates Label, 0 indicates Output

 1: for $H_i$ in $HSData$ do
 2:     for $S_j$ in $SearchEngines$ do
 3:         $H_i \leftarrow$ Preprocess($H_i$)
 4:         $U_{i,j} \leftarrow S_j(H_i)$
 5:         $U$.Add($U_{i,j}$)
 6:     end for
 7: end for
 8: $U \leftarrow$ Unique($U$)
 9: $W \leftarrow$ Scraping($U$)
10: $O_D \leftarrow$ DomainLabelling($W$)
11: $O_G \leftarrow$ GPT($W$)
12: $O \leftarrow \{O_D, O_G\}$
```

Dataset	#Texts	#Webpages
Davidson et al. (2017)	20620	80342
Mollas et al. (2022)	433	2145
de Gibert et al. (2018)	1196	6783
ElSherief et al. (2021)	8188	45321
Kennedy et al. (2020)	14170	64765

Table 1: Distribution of Curated & Annotated Webpages

We extract webpage features in Step 9 using *Scraping*(), to accurately represent a webpage. The

[1] https://developers.google.com/custom-search

[2] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

extracted features included in our work are *URL*, *Title*, *Headings*, *SubHeadings*, *Paragraph Texts*. We have leveraged open source python libraries such as Selenium,[3] Beautiful Soup,[4] to scrape these URLs. The extracted *Headings*, *Subheadings*, *Paragraph Texts* are appended together to create a feature called *BodyText*. Thus, the final data set consists of 150k unique webpage objects, *W* contains three features: *URL* (U), *Title*(T), *BodyText*(B) which have been represented in the output *W = (U, T, B)*. Note that the count of webpages mentioned in Table 1 do not sum up to 150k, many webpages appear as search engine results for more than one hate speech short-text datasets.

Manually inspecting the data W, we observed that $\sim$ 42% of non-problematic data come from the domain of sports, and e-commerce. These webpages are not related to any potentially sensitive hate or violence topics, hence its safe to annotate them as non-problematic page directly. On manually spot checking sports, and e-commerce websites, we have not seen any problematic content. However, it may be the case that some hateful, violent content may appear in these sports and e-commerce websites. In step 10, we labelled 50k webpages directly by applying a simple domain-level labelling using *DomainLabelling*(). We extract the domain name for the website and match it with a curated list of domains focused on sports [5] and e-commerce.[6] This helped to reduce the GPT-4 labelling time and cost. The remaining 90k webpages are annotated using GPT-4 in step 11.

2.2 GPT Prompt Creation, Validation, and Annotation

The first step towards the annotation process was deciding upon the various categories of problematic content that we will be targeting with our annotation process. Similar to various categories of hate speech as described in (Mollas et al., 2022), we decided to go with similar categories i.e. problematic content promoting hate based on race, religion, gender, sexual orientation and violence. The data that we have curated with our strategy will be annotated in the following classes.

- Race
- Gender Identiy
- Religion
- Sexual Orientation
- Violence
- Non-Problematic

A webpage belonging to either one or more of the first five classes is labelled as Problematic.

To develop the GPT-4 based annotation process, the foremost step is creating a gold standard annotated set of webpages, which will be leveraged to measure the accuracy of various iterations and variations of prompts. We randomly sampled a set of 1000 webpages from the set of collected data which was manually labelled by 2 in-house experts[7]. We duplicate some of the dataset among annotators to measure the inter annotator agreement. We have got a pairwise $\kappa = 0.87$ using Cohen's kappa (Cohen, 1960) indicating high quality reliable annotation.

The GPT-4 prompt needs detailed context to be able to accurately distinguish between problematic and non-problematic webpages. Hence, we make use of different webpage features (*URL*, *Title*, *BodyText*) extracted in the previous step and include them in the prompt as part of the input section. This along with detailed instructions gives the required context to GPT-4 to label a webpage.

2.2.1 Prompt Development

Our prompt is comprised of multiple sections which includes *Task Description, Instructions, Input, Examples* as shown in Figure 1. During the prompt development cycle, we tried multiple strategies of prompting with different combinations of the aforementioned sections. Following are the different prompting strategies we have explored in this work.

Basic Instructions: Figure 1 (a) shows the basic version of the prompt. Here we have a simple prompt with *Task Description*, basic *Instructions*, *Input* and ask the GPT-4 model to annotate the webpage. In the basic instructions GPT-4 is expected to give a binary label for each of the five classes of hate and violence. This means any candidate webpage is either problematic or non-problematic in each of 5 sub classes of hate and violence.

Precise Instructions: In this version of the prompt, we have a more complex prompt where

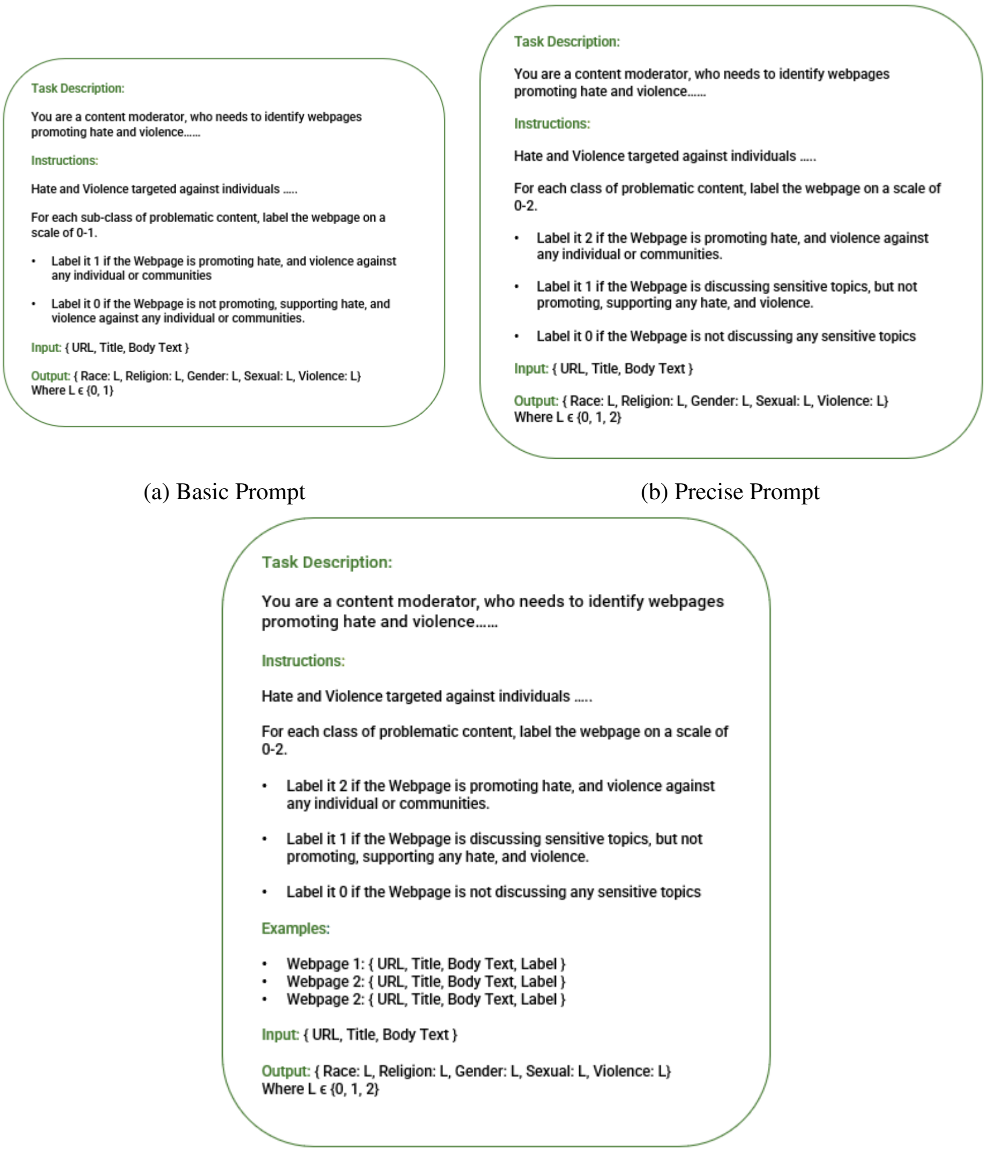
[3] https://pypi.org/project/selenium/

[4] https://pypi.org/project/beautifulsoup4/

[5] https://www.similarweb.com/top-websites/sports/sports/

[6] https://www.kaggle.com/datasets/wiredwith/websites-list

[7] The annotators are very proficient in english and have done all their formal education in english, and have been doing these kind of annotations for last 4 years



Task Description:

You are a content moderator, who needs to identify webpages promoting hate and violence……

Instructions:

Hate and Violence targeted against individuals …..

For each sub-class of problematic content, label the webpage on a scale of 0-1.

- Label it 1 if the Webpage is promoting hate, and violence against any individual or communities
- Label it 0 if the Webpage is not promoting, supporting hate, and violence against any individual or communities.

Input: { URL, Title, Body Text }

Output: { Race: L, Religion: L, Gender: L, Sexual: L, Violence: L}
Where L ϵ {0, 1}

(a) Basic Prompt

Task Description:

You are a content moderator, who needs to identify webpages promoting hate and violence……

Instructions:

Hate and Violence targeted against individuals …..

For each class of problematic content, label the webpage on a scale of 0-2.

- Label it 2 if the Webpage is promoting hate, and violence against any individual or communities.
- Label it 1 if the Webpage is discussing sensitive topics, but not promoting, supporting any hate, and violence.
- Label it 0 if the Webpage is not discussing any sensitive topics

Input: { URL, Title, Body Text }

Output: { Race: L, Religion: L, Gender: L, Sexual: L, Violence: L}
Where L ϵ {0, 1, 2}

(b) Precise Prompt

Task Description:

You are a content moderator, who needs to identify webpages promoting hate and violence……

Instructions:

Hate and Violence targeted against individuals …..

For each class of problematic content, label the webpage on a scale of 0-2.

- Label it 2 if the Webpage is promoting hate, and violence against any individual or communities.
- Label it 1 if the Webpage is discussing sensitive topics, but not promoting, supporting any hate, and violence.
- Label it 0 if the Webpage is not discussing any sensitive topics

Examples:

- Webpage 1: { URL, Title, Body Text, Label }
- Webpage 2: { URL, Title, Body Text, Label }
- Webpage 2: { URL, Title, Body Text, Label }

Input: { URL, Title, Body Text }

Output: { Race: L, Religion: L, Gender: L, Sexual: L, Violence: L}
Where L ϵ {0, 1, 2}

(c) Few-shot Prompt

Figure 1: Different GPT-4 prompts used for webpage annotation. For all the actual prompts, please refer to Figure 2 in Appendix

the basic *Instructions* are modified to the version that can be seen in Figure 1 (b). Instead of a binary label, we have asked GPT-4 to label each webpage on a three point label for each subclass of hate and violence. This helps GPT-4 model make more precise annotations, and better understand the decision boundary between problematic and non-problematic webpages. In the domain of hateful and violent content, it can be misleading for models to understand content which is discussing sensitive topics but not promoting any problematic intent. Introducing a three point labelling mechanism removes this confusion and clarifies the decision boundary. This enables GPT-4 make more precise judgements.

Precise Instructions with Few-shot Examples: In the final version of the prompt as seen in 1 (c), we add the *Examples* section to help GPT-4 identify problematic webpage content (Liu et al., 2021; Brown et al., 2020). Giving examples in case of webpage annotation with GPT-4 can be challenging. The Body Text feature as observed in our dataset is very long (5350 ± 205 tokens). In such a scenario, including the entire *BodyText* feature will make the final prompt too long. This can lead to recency bias, and loosing the entire context for the GPT-4 annotation model. We propose to leverage text summarisation technique to create a summarised body text feature to be included in few shot examples. We have leveraged GPT-4 model (Please refer to Figure 2 in Appendix for Webpage Content Summarisation prompt) itself to summarise few chosen examples from the gold set. These are leveraged as few shot examples in this prompt version.

2.3 Prompt Performance

The detailed results of GPT-based annotation on the 1000 webpages using different prompts is reported in Table 2. The results observed on identifying problematic webpages in each hate sub-class are inline with general observations reported in (Chiu et al., 2022; Mollas et al., 2022) that addition of detailed instruction and few-shot examples generally yield better classification results for hate speech detection. We find that adding precise instructions increase the F1-Score by 5.7 absolute points compared to the Basic prompt. points overall. Furthermore, adding few-shot examples to the Precise prompt increase the F1-score by 10.6 absolute points compared the the Basic prompt. We found higher F1-score for sexual harm and violence compared to other 3 sub classes. This can be explained by the fact that often explicit words (profane, adult words, harmful words) are available in the surface form for sexual and violence categories. Our observations suggests that these sub classes of problematic content tend to be promoting more explicit form of hate with language which is not very subtle. In contrast, webpages promoting Gender, Race and Religion hate are more implicit in nature with subtle tonality.

In Step 11 of Algorithm 1, we use the best performing prompt (precise instructions with few-shot examples) to annotate the 90k webpages with *GPT*(). GPT-4 labelling identifies 21k problematic webpages. The class wise distribution of the final dataset with the following prompt strategy is given in Table 3. Note that a webpage can belong to multiple hate categories. For example, a webpage which is problematic in gender hate class might be problematic in sexual hate class too. The aggregated label after GPT annotation for a webpage is: *Problematic, Topically Sensitive, Clean* The aggregated labelling strategy is as follows:

- Webpage is **Problematic**, if it is labelled as problematic in terms of any one of the the hate classes.
- Webpage is **Topically Sensitive** if its aggregated label is not problematic, and is labelled as topically sensitive in at least one hate class.
- Webpage is **Clean** if its aggregated label is neither problematic nor of sensitive topic.

2.4 Data set Description

The final annotated data consists of three broader labels: *Problematic, Topically Sensitive, Clean*. Annotating the 90k domain filtered webpages, we have 21k problematic, 44k non-problematic and topically sensitive, 25k clean webpages. To prepare the final data set, we decided to maintain a rough ratio of 1:2:2 for problematic, topically sensitive and clean classes, respectively. This was to ensure that the final model should be robust and does not have any bias to a particular class due to the data distribution. Hence, we randomly sampled some more clean webpages (17k) from previous domain-level filtered data. Our final data set comprised of 21k Problematic, 44k Topically Sensitive, and 42k Clean webpages. Note, that the *Topically Sensitive* data can be effective use as counterfactual data to make the model more robust (Wu et al., 2021). Each webpage is represented by three features – *URL, Title, BodyText.*

3 Experiments & Results

As mentioned in Section 1, latency and cost are the major challenges to leverage GPT-4 to annotate webpages at scale. This is important point to consider during our experimentation as there are billions of webpages in a search engine index, and millions of webpages embedded in social media posts and comments. Thus, to solve the problem of identifying problematic webpage classification, we need a lighter model.

3.1 Model Training

We build the problematic webpage classifier using Transformer-based (Vaswani et al., 2017) models and fine tune the same using our labelled dataset. We have experimented with various pre-trained transformer base models such as BERT (Devlin et al., 2019), HateBERT(Caselli et al., 2021a), Longformer(Beltagy et al., 2020), and compared the results. Longformer models have been included in our experiment specially because we observed that the input sequence can be very long in a webpage (5350 ± 205 tokens). BERT and HateBERT limit the maximum input sequence length to 512

Class	Basic			Precise			Few-shot		
	P	R	F1	P	R	F1	P	R	F1
Race	71.3	75.6	73.4	83.9	80.2	82.0	89.4	87.6	**88.5**
Gender	72.5	74.1	73.3	80.6	79.9	80.2	87.0	87.4	**87.2**
Religion	66.6	72.3	69.3	78.5	75.1	76.8	86.9	81.4	**84.1**
Sexual	79.4	82.5	80.9	87.5	88.3	87.9	93.5	87.6	**90.5**
Violence	78.7	83.5	81.0	89.2	84.4	86.7	92.7	89.3	**91.0**
Overall	76.3	79.9	78.1	85.3	82.3	83.8	90.1	87.3	**88.7**

Table 2: Prompt Accuracy per Hate sub-classes (P: Precision, R: Recall, F1: F1-Score)

Class	**#Problematic**	**#Sensitive Topic**
Race	5312	11512
Gender	3414	7631
Religion	3451	5904
Sexual	4513	10467
Violence	6718	14871
Overall	**21712**	**44512**

Table 3: Distribution of labelled webpages. Note that one instance may occur in multiple classes thus the overall number may not be equal to the sum of the individual classes.

Class	**Model**	**F1**
Race	BERT	80.8 ± 0.7
	HateBERT	81.9 ± 0.1
	Longformer	**87.6 ± 0.4**
Gender	BERT	76.9 ± 0.2
	HateBERT	78.7 ± 0.3
	Longformer	**83.1 ± 0.9**
Religious	BERT	75.3 ± 0.9
	HateBERT	75.6 ± 0.5
	Longformer	**78.0 ± 0.2**
Sexual	BERT	83.5 ± 0.2
	HateBERT	85.7 ± 0.9
	Longformer	**89.1 ± 0.4**
Violence	BERT	84.7 ± 0.2
	HateBERT	84.3 ± 0.4
	Longformer	**88.7 ± 0.9**
Overall	BERT	82.9 ± 0.8
	HateBERT	83.6 ± 0.9
	Longformer	**87.6 ± 0.4**

Table 4: Webpage Classification Performance

tokens. Longformer on the other hand has a maximum limit of 4096 tokens which can fit our webpage data without much truncation.

To create the input text for the tokenizers, we have leveraged the same three features (*URL*, *Title*, *Body Text*) as was previously used in GPT-4 prompt. These text features were appended together, separated by corresponding separator tokens. The maximum input sequence length limitations require us to choose and send the most relevant context to the model. Pre-processing of the input text was done to ensure that only relevant tokens are used to fine-tune and infer the model. Basic pre-processing steps involves the removal of (i) unnecessary spaces, non-ASCII characters, numbers (except the number *18* due to its frequent occurrence in the adult pages) (ii) common Webpage related tokens like "www", "https", "php" and (iii) tokens which either contain greater than 15 characters, or only a single character. We train a binary classification model with two classes: Problematic, Non-Problematic. Here, Non-Problematic includes both *Sensitive Topic* and *Clean*.

We have considered 80% of the data set as training data in fine-tuning the pre-trained models, 10% as validation data to measure the out-of-sample performance of the model during training, and hyper-parameter tuning, and 10% as test data to measure the out-of-sample performance after training. To prevent over-fitting, we have used stratified sampling to select 0.8, 0.1, and 0.1 portions of the data from each class (Race/Gender/Religion/Sexual/Violence) while creating train, validation, and test set.

To understand the importance of different features, we have experimented with three models trained on increasing level of contexts – *URL*, (*URL* + *Title*), (*URL* + *Title* + *BodyText*). We have also trained a HateBERT-based classification model fine-tuned on baseline short-text hate speech

Feature	F1
URL	51.7
Title	58.1
BodyText	76.9
URL + Title	71.8
URL + Title + BodyText	**87.6**

Table 5: Webpage Feature Importance

datasets in Table 1. This is to show the importance of webpage specific data collection instead of solely using short-text hate speech data.

3.2 Model Performance & Results

Table 4 presents the details of our experimental results across all categories using 3 SOTA models tuned and tested on our GPT-4 annotated dataset. Longformer based webpage classification model outperforms and reaches an overall F1-score of 87.6%. We find that Longformer models have much better accuracy in all 5 categories and have 4.7 and 4 absolute point improvement in F1-score compared to BERT and HateBERT models, respectively. HateBERT model performs slightly better than BERT with an overall F1-score of 83.6% and 82.9%, respectively.

Furthermore, we evaluated the best performing Longformer-based webpage classification model with different combinations of features to understand the importance of the features. Table 5 details the results, which show that all the three features are very important to provide detailed context to the model to classify a webpage. Each feature on its own has much lower performance compared to the combined feature. *BodyText* on its own has the highest accuracy compared to the other two features (*URL* and *Title*). This essentially indicates that a lot of useful information is there in the Body Text for webpage classification but the URL and Title also provides additional information to improve the overall performance.

Finally, in Table 6, we present the comparison of best performing Longformer based problematic webpage classification model (L-PWC) against the Hate speech classification model trained using only short-text data (S-HSC). L-PWC model outperforms the S-HSC model in all classes and has an overall gain of 13.7% compared to the S-HSC model.

Class	Model	F1
Race	S-HSC	72.9 ± 0.2
	L-PWC	**87.6 ± 0.4**
Gender	S-HSC	69.2 ± 0.8
	L-PWC	**83.1 ± 0.9**
Religious	S-HSC	66.7 ± 0.4
	L-PWC	**78.0 ± 0.2**
Sexual	S-HSC	75.4 ± 0.5
	L-PWC	**89.1 ± 0.4**
Violence	S-HSC	74.1 ± 0.6
	L-PWC	**88.7 ± 0.9**
Overall	S-HSC	72.9 ± 0.9
	L-PWC	**87.6 ± 0.4**

Table 6: Performance comparison between L-PWC and S-HSC models. L-PWC: Longformer based Problematic Webpage Classification, S-HSC: Hate Speech Classification trained on Short-Text data

4 Conclusion

This paper presents a novel way of collecting and annotating problematic webpages which is important for building a problematic webpage classification model. We have shown that easily available short-text data along with the knowledge of SOTA generative models (GPT-4) can help in building annotated datasets for a complex task such as problematic webpage classification. We also report and re-establish the fact that writing precise prompt along with a few examples is effective and achieve very high quality annotation. We compare different pre-trained models and fine-tune them with our dataset and report comparative results. We also report an ablation study and show that the different features used in our experiment are together effective for webpage classification. Finally, we show empirically that our data set is effective for building a problematic webpage classifier.

The work can be further extended by leveraging additional features for webpage classification such as Ads, Link Connections to other webpages, Authority of the domain. The work can also be extended towards creating multilingual dataset for problematic webpage classification and subsequently build a model for the same.

5 Limitations

The dataset created as part of our contribution leverages hate speech datasets focusing on the English language. Therefore, the model has neither seen, nor been evaluated in other languages.

References

Abdullah Aljebreen, Weiyi Meng, and Eduard Dragut. 2021. Segmentation of tweets with urls and its applications to sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12480–12488.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Renato Bruni and Gianpiero Bianchi. 2019. Website categorization: a formal approach and robustness analysis in the case of e-commerce detection. *Expert Systems with Applications*, 142:113001.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021a. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021b. Hatebert: Retraining bert for abusive language detection in english.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2022. Detecting hate speech with gpt-3.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Abhishek Das and Ankit Jain. 2012. Indexing the world wide web: The journey so far. In *Next Generation Search Engines: Advanced Models for Information Retrieval*, pages 1–28. IGI Global.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2018. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex &amp Intelligent Systems*, 8(6):4663–4678.

OpenAI. 2023. Gpt-4 technical report.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Gaurav Rajput, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. Hate speech detection using static BERT embeddings. In *Big Data Analytics*, pages 67–77. Springer International Publishing.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Peng Yang, Guangzhen Zhao, and Peng Zeng. 2019. Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access*, 7:15196–15209.

A Appendix

Problematic webpage classification is even more skewed as compared to classification of hate speech. Prominent search engines have already developed certain filtering mechanisms to remove problematic webpages from their search engine results. Therefore, we also sampled data from the given hate speech data sets to pick short text which are more problematic in nature in terms of hate, violence and hence, more likely to yield a problematic webpage when queried in these search engines. Hence, with each data set we have chosen a threshold to get the most problematic phrases.

The datasets that have been used for mining short text data corresponding to hate speech are:

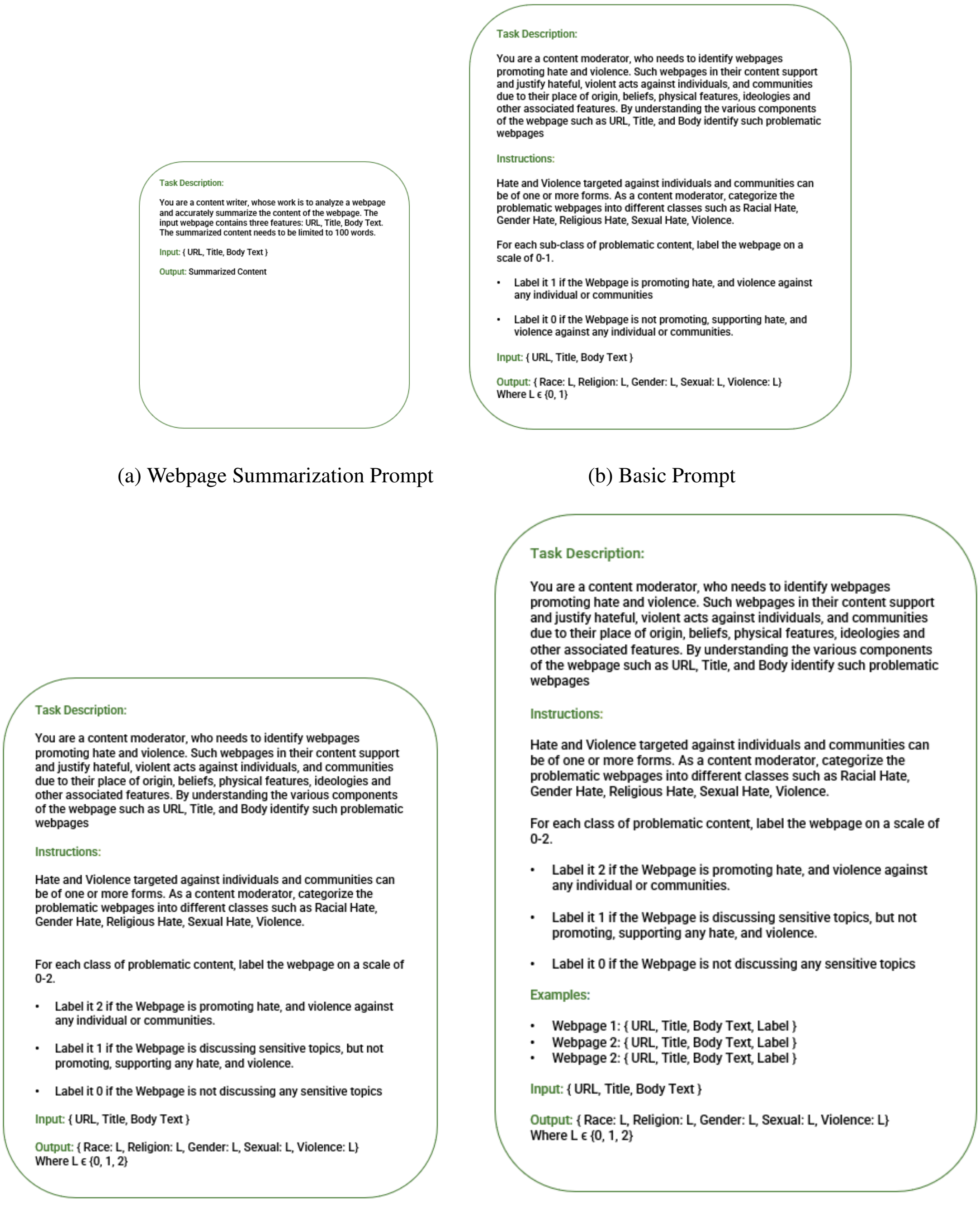
(Davidson et al., 2017): A crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords, which contains data labelled into three classes hate, offensive, neither. For our dataset preparation, we have filtered out short text data belonging to hate and offensive classes, which is roughly $\sim$ 20k in quantity.

(de Gibert et al., 2018): These files contain text extracted from Stormfront, a white supremacist forum. A random set of forums posts have been sampled from several sub-forums and split into sentences. Those sentences have been manually labelled as containing hate speech or not, according to certain annotation guidelines. We have filtered the dataset and used $\sim$ 1k posts labelled as hateful.

(Mollas et al., 2022): ETHOS is a textual dataset based on YouTube, Reddit comments validated using a crowd-sourcing platform with two variants: binary and multi-label. We have sampled data from this dataset where the binary label for that instance is hate speech. This helped us get 433 short-text hate speech.

(ElSherief et al., 2021): This dataset focuses on presenting a benchmark for implicit hate speech data. While other datasets focus on explicit hate, abuse towards a individual or community, this specially focuses on implicit hate or indirect hate, which helps in the mining of diverse webpages promoting subtle hate or opinionated content. We have filtered the dataset and only taken the implicit & explicit hate posts which is $\sim$ 8k posts.

(Kennedy et al., 2020): The dataset released in their work is mined using comments from various social media platforms. Originally containing 39k comments, it also presents a continuous hate speech score for the data. To mine relevant webpages for our use case, we have filtered it and taken only those comments where the hate speech score value is greater than 0.5. This leaves us with $\sim$ 14k social media comments.



Task Description:

You are a content writer, whose work is to analyze a webpage and accurately summarize the content of the webpage. The input webpage contains three features: URL, Title, Body Text. The summarized content needs to be limited to 100 words.

Input: { URL, Title, Body Text }

Output: Summarized Content

(a) Webpage Summarization Prompt

Task Description:

You are a content moderator, who needs to identify webpages promoting hate and violence. Such webpages in their content support and justify hateful, violent acts against individuals, and communities due to their place of origin, beliefs, physical features, ideologies and other associated features. By understanding the various components of the webpage such as URL, Title, and Body identify such problematic webpages

Instructions:

Hate and Violence targeted against individuals and communities can be of one or more forms. As a content moderator, categorize the problematic webpages into different classes such as Racial Hate, Gender Hate, Religious Hate, Sexual Hate, Violence.

For each sub-class of problematic content, label the webpage on a scale of 0-1.

- Label it 1 if the Webpage is promoting hate, and violence against any individual or communities
- Label it 0 if the Webpage is not promoting, supporting hate, and violence against any individual or communities.

Input: { URL, Title, Body Text }

Output: { Race: L, Religion: L, Gender: L, Sexual: L, Violence: L}
Where L ϵ {0, 1}

(b) Basic Prompt

Task Description:

You are a content moderator, who needs to identify webpages promoting hate and violence. Such webpages in their content support and justify hateful, violent acts against individuals, and communities due to their place of origin, beliefs, physical features, ideologies and other associated features. By understanding the various components of the webpage such as URL, Title, and Body identify such problematic webpages

Instructions:

Hate and Violence targeted against individuals and communities can be of one or more forms. As a content moderator, categorize the problematic webpages into different classes such as Racial Hate, Gender Hate, Religious Hate, Sexual Hate, Violence.

For each class of problematic content, label the webpage on a scale of 0-2.

- Label it 2 if the Webpage is promoting hate, and violence against any individual or communities.
- Label it 1 if the Webpage is discussing sensitive topics, but not promoting, supporting any hate, and violence.
- Label it 0 if the Webpage is not discussing any sensitive topics

Input: { URL, Title, Body Text }

Output: { Race: L, Religion: L, Gender: L, Sexual: L, Violence: L}
Where L ϵ {0, 1, 2}

(c) Precise Prompt

Task Description:

You are a content moderator, who needs to identify webpages promoting hate and violence. Such webpages in their content support and justify hateful, violent acts against individuals, and communities due to their place of origin, beliefs, physical features, ideologies and other associated features. By understanding the various components of the webpage such as URL, Title, and Body identify such problematic webpages

Instructions:

Hate and Violence targeted against individuals and communities can be of one or more forms. As a content moderator, categorize the problematic webpages into different classes such as Racial Hate, Gender Hate, Religious Hate, Sexual Hate, Violence.

For each class of problematic content, label the webpage on a scale of 0-2.

- Label it 2 if the Webpage is promoting hate, and violence against any individual or communities.
- Label it 1 if the Webpage is discussing sensitive topics, but not promoting, supporting any hate, and violence.
- Label it 0 if the Webpage is not discussing any sensitive topics

Examples:

- Webpage 1: { URL, Title, Body Text, Label }
- Webpage 2: { URL, Title, Body Text, Label }
- Webpage 2: { URL, Title, Body Text, Label }

Input: { URL, Title, Body Text }

Output: { Race: L, Religion: L, Gender: L, Sexual: L, Violence: L}
Where L ϵ {0, 1, 2}

(d) Few-shot Prompt

Figure 2: Actual Webpage Annotation Prompt