Understanding and Detecting Hateful Content Using Contrastive Learning

Felipe González-Pizarro[1], Savvas Zannettou[2,3]

[1] University of British Columbia
[2] Delft University of Technology
[3] Max Planck Institute for Informatics
felipegp@cs.ubc.ca, s.zannettou@tudelft.nl

Abstract

The spread of hate speech and hateful imagery on the Web is a significant problem that needs to be mitigated to improve our Web experience. This work contributes to research efforts to detect and understand hateful content on the Web by undertaking a multimodal analysis of Antisemitism and Islamophobia on 4chan's /pol/ using OpenAI's CLIP. This large pre-trained model uses the Contrastive Learning paradigm. We devise a methodology to identify a set of Antisemitic and Islamophobic hateful textual phrases using Google's Perspective API and manual annotations. Then, we use OpenAI's CLIP to identify images that are highly similar to our Antisemitic/Islamophobic textual phrases. By running our methodology on a dataset that includes 66M posts and 5.8M images shared on 4chan's /pol/ for 18 months, we detect 173K posts containing 21K Antisemitic/Islamophobic images and 246K posts that include 420 hateful phrases. Among other things, we find that we can use OpenAI's CLIP model to detect hateful content with an accuracy score of 0.81 (F1 score = 0.54). By comparing CLIP with two baselines proposed by the literature, we find that CLIP outperforms them, in terms of accuracy, precision, and F1 score, in detecting Antisemitic/Islamophobic images. Also, we find that Antisemitic/Islamophobic imagery is shared in a similar number of posts on 4chan's /pol/ compared to Antisemitic/Islamophobic textual phrases, highlighting the need to design more tools for detecting hateful imagery. Finally, we make available (upon request) a dataset of 246K posts containing 420 Antisemitic/Islamophobic phrases and 21K likely Antisemitic/Islamophobic images (automatically detected by CLIP) that can assist researchers in further understanding Antisemitism and Islamophobia.

Introduction

The spread of hateful content on the Web is an everlasting and vital issue that adversely affects society. The problem of hateful content is longstanding on the Web for various reasons. First, there is no scientific consensus on what constitutes hateful content (i.e., no definition of what hate speech is) (Sellars 2016). Second, the problem is complex since hateful content can spread across various modalities (e.g., text, images, videos, etc.), and we still lack automated techniques to detect hateful content with acceptable and generalizable performance (Arango, Pérez, and Poblete 2019). Third, we lack moderation tools to proactively prevent the spread of hateful content on the Web (Konikoff 2021). This work focuses on assisting the community in addressing the issue of the lack of tools to detect hateful content across multiple modalities.

Most of the research efforts in the space of detecting hateful content focus on designing and training machine learning models that are specifically tailored towards detecting specific instances of hateful content (e.g., hate speech on text or particular cases of hateful imagery). Some examples of such efforts include Google's Perspective API (Perspective API 2018) and the HateSonar classifier (Davidson et al. 2017) that aim to detect toxic and offensive text. Other methods aim to detect instances of hateful imagery like Antisemitic images (Zannettou et al. 2020b) or hateful memes (Kiela et al. 2020; Zannettou et al. 2018). These efforts and tools are essential and valuable, however, they rely on human-annotated datasets that are expensive to create, and therefore they are also small. At the same time, these datasets focus on specific modalities (i.e., text or images in isolation). All these drawbacks limit their broad applicability.

The lack of large-scale annotated datasets for solving problems like hate speech motivated the research community to start developing techniques that learn from unlabeled data (a paradigm known as *self-supervised learning*). Over the past years, the research community released large-scale models that depend on huge unlabeled datasets such as OpenAI's GPT-3 (Brown et al. 2020), Google's BERT (Devlin et al. 2019), OpenAI's CLIP (Radford et al. 2021), etc. These models are trained on large-scale datasets and usually can capture general knowledge extracted from the datasets that can be valuable for performing classification tasks that the model was not explicitly trained on.

Motivated by these recent advancements on large-scale pre-trained machine learning models, in this work, we investigate how we can use such models to detect hateful content on the Web across multiple modalities (i.e., text and images). Specifically, we focus on OpenAI's CLIP model because it helps us capture content similarity across modalities (i.e., measure similarity between text and images). To achieve this, CLIP leverages a paradigm known as Contrastive Learning; the main idea is that the model maps text and images to a high-dimensional vector space and is trained in such a way that similar text/images are mapped closer in this vector space (for more details see Background section).

Focus & Research Questions. This work focuses on understanding the spread of Antisemitic/Islamophobic content on 4chan's /pol/ board. We concentrate on hateful content targeted towards these two demographics mainly because previous work indicates that 4chan's /pol/ is known for disseminating Antisemitic/Islamophobic content (Zannettou et al. 2020b; Prisk 2017). Specifically, we focus on shedding light on the following research questions:

- **RQ1:** Can large pre-trained models that leverage the Contrastive Learning paradigm, like OpenAI's CLIP, identify hateful content with acceptable performance? How does CLIP's performance compare to state-of-the-art classifiers for detecting hateful imagery?
- **RQ2:** How prevalent is Antisemitic/Islamophobic imagery and textual hate speech on 4chan's /pol/?

To answer these research questions, we obtain all the posts and images shared on 4chan's /pol/ between July 1, 2016, and December 31, 2017, ultimately collecting 66M textual posts and 5.8M images. Then, we leverage the Perspective API and manual annotations to construct a dataset of 420 Antisemitic and Islamophobic textual phrases. We retrieve 246K posts that include any of our 420 hateful phrases. Finally, we use OpenAI's CLIP to detect Antisemitic/Islamophobic images when provided as input the above-mentioned hateful phrases and all images shared on 4chan's /pol/; we find 21K images that are likely Antisemitic/Islamophobic.

Contributions & Main Findings. Our work makes the following contributions/main findings:

- We investigate whether large pre-trained models based on Contrastive Learning can assist in detecting hateful imagery. We find that large pre-trained models like OpenAI's CLIP (Radford et al. 2021) can detect Antisemitic/Islamophobic imagery with 0.81, 0.54, 0.53, 0.54, accuracy, precision, recall, and F1 score, respectively. The CLIP model outperforms two baselines that detect hateful imagery in terms of accuracy (0.11 increase), precision (0.17 increase), and F1 score (0.08 increase) (**RQ1**).
- We find that on 4chan's /pol/, Antisemitic/Islamophobic imagery appears in a similar number of posts compared to Antisemitic/Islamophobic textual hateful content. This finding highlights the need for the development and use of multimodal hate speech detectors for understanding and mitigating the problem (**RQ2**).
- We make available (upon request from researchers) a large dataset of Antisemitic/Islamophobic posts, phrases, and images shared on 4chan's /pol/.[1] The released dataset includes 892 Antisemitic/Islamophobic images and 420 Antisemitic/Islamophobic phrases that are annotated by the authors of this paper (we expect no false positives in our human-annotated set). Additionally, we release a set of 21K images that were detected by the CLIP model as being Antisemitic/Islamophobic. Note that the set of 21K images includes images that are false positives, since the CLIP model has a precision score of 0.54 for detecting Antisemitic/Islamophobic images. Therefore, researchers aiming to use the dataset and create classifiers for Antisemitic/Islamophobic images, should consider the existence of false positives in the dataset. Nevertheless, we argue that the released dataset can assist researchers in future work focusing on detecting and understanding the spread of hateful content on the Web across multiple modalities (i.e., text and images).

Ethical Considerations. We emphasize that we rely entirely on publicly available and anonymous data shared on 4chan's /pol/, hence we do not and are unable to obtain consent from users that shared posts/images anonymously on 4chan. Also, all of our study's manual annotations (i.e., given an image/phrase, annotate if the image/phrase is Antisemitic/Islamophobic) were exclusively performed by the authors of this work, hence minimizing exposure of crowd workers to potentially disturbing content. Given that we only analyze publicly available anonymous data from 4chan and that all manual annotations are undertaken by the authors of this paper, our work is not considered human's subject research by our institution's Ethical Board Committee. For our analysis, we follow standard ethical guidelines (Rivers and Lewis 2014) like reporting aggregate results and not attempting to deanonymize users.

We also discuss some ethical implications of releasing our dataset and possible misuse of the dataset. There is a possibility that malicious actors can make use of our dataset and train models based on Generative Adversarial Networks (Goodfellow et al. 2014) with the goal of automatically generating new Antisemitic/Islamophobic imagery. Subsequently, malicious actors can share this Antisemitic / Islamophobic imagery on social media platforms, hence affecting people. To minimize this risk, we will make the dataset available only upon request and only to researchers that can provide a description of their intended use.

Disclaimer. *This manuscript contains Antisemitic and Islamophobic textual and graphic elements that are offensive and are likely to disturb the reader.*

Background

This section provides background information on Contrastive Learning and OpenAI's CLIP model, on Google's Perspective API, and 4chan's /pol/ that is our data source.

Contrastive Learning. To understand Contrastive Learning, it is essential to grasp its differences compared to traditional Machine/Deep Learning (ML/DL) classifiers. Traditional ML/DL classifiers take as an input a set of labeled data, each accompanied with a class, and aim to predict the class from the labeled data, a paradigm known as supervised learning (Caruana and Niculescu-Mizil 2006).

On the other hand, Contrastive Learning is a self-supervised technique, meaning that there is no need to have classes and models learn from unlabeled data. The main idea behind Contrastive Learning is that you train a model that relies on unlabeled data, and the model learns general features from the dataset by teaching it which input samples

[1] https://zenodo.org/record/6993868#.ZCAS4uxBwxw

are similar/different to each other (Hadsell, Chopra, and LeCun 2006). In other words, Contrastive Learning relies on a set of unlabeled data samples with additional information on which of these samples are similar to each other.

Contrastive Learning is becoming increasingly popular in the research community with several applications on visual representations (Chen et al. 2020; Kim et al. 2020), textual representations (Giorgi et al. 2021; Wu et al. 2020; Gao, Yao, and Chen 2021), graph representations (You et al. 2020; Hassani and Khasahmadi 2020), and multimodal (i.e., text/images, images/videos, etc.) representations (Radford et al. 2021; Diba et al. 2021; Yuan et al. 2021; Zhang et al. 2022).

OpenAI's CLIP. OpenAI recently released a model called Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) that leverages Contrastive Learning to generate representations across text and images. The model relies on a text encoder and an image encoder that maps text and images to a high-dimensional vector space. Subsequently, the model is trained to minimize the cosine distance between similar text/image pairs. To train CLIP, OpenAI created a huge dataset that consists of 400M pairs of text/images collected from various Web sources and covers an extensive set of visual concepts[2]. By training CLIP with this vast dataset, the model learns general visual representations and how these representations are described using natural language, which results in the model obtaining general knowledge in various topics (e.g., identifying persons, objects, etc.).

In this work, we leverage the CLIP model to extract representations for our 4chan textual/image datasets and assess the similarity between the text and image representations. The main idea is that by providing the pre-trained CLIP model with a set of hateful text phrases, we will be able to identify a set of hateful images that are highly similar to the hateful text query. To demonstrate CLIP's potential in discovering hateful imagery from hateful text-based queries, we show an example from our 4chan dataset in Fig. 1. On the left side, we show examples of images that are highly similar to a benign text query like "cute cat sleeping," while on the right, we show examples of images that are similar to an Antisemitic and toxic phrase ("gas the jews")[3]. For each image, we report the cosine similarity between the representation obtained from the text query and the representation of the image in our dataset. This example shows that the CLIP model can detect objects in images (i.e., cats) and provide relevant images to the queries (i.e., the cats are indeed sleeping according to the query). Furthermore, by looking at the images for the toxic query, we observe that CLIP can identify harmful images based on the query and can link historical persons to it (e.g., the textual input does not mention Adolf Hitler; however, the model knows that Hitler was responsible for the holocaust). Also, CLIP can detect images that share hateful ideology by adding text on memes (i.e., CLIP also performs Optical Character Recognition and can correlate that text with the text-based query). Overall, this example shows the predictive power of the CLIP model in detecting hateful imagery from hateful text phrases.

Google's Perspective API. As a first step towards identifying hateful phrases, we use Google's Perspective API (Perspective API 2018; Google 2021), which provides a set of Machine Learning models for identifying how rude/aggressive/hateful a comment is. We use the Perspective API for identifying hateful text mainly because it outperforms other publicly available hate speech classifiers like HateSonar (Davidson et al. 2017; Zannettou et al. 2020a). This work focuses on the SEVERE_TOXICITY model available from Perspective API because it is more robust to positive uses of curse words (Google 2021), and it is a production-ready model. The SEVERE_TOXICITY model returns a score between 0 and 1, which can be interpreted as the probability of the text being rude and toxic.

4chan's /pol/. 4chan is an anonymous image board usually exploited by troll users (Hine et al. 2017). On 4chan, users can create a thread by creating a post that contains an image, and other users can create replies with or without images, and they might add references to previous posts. 4chan is well-known for its anonymity and ephemerality. These are the main reasons its users are aggressive in their posts, as there is a lack of accountability (Bernstein et al. 2011). Our work focuses on 4chan, particularly the Politically Incorrect board (/pol/). /pol/ is the main board for discussing world events and politics and is known for the spread of conspiracy theories (Zannettou et al. 2017; Tuters, Jokubauskaitė, and Bach 2018) and hateful content (Hine et al. 2017; Zannettou et al. 2020b).

Dataset

We collect the data about posts on 4chan's /pol/ using the publicly available dataset released by Papasavva et al. (2020); the dataset includes textual data about 134.5M posts shared on /pol/ between June 2016 and November 2019. Our work focuses on the period between July 1, 2016, and December 31, 2017 (to match the time period of the image dataset mentioned below), including 66,383,955 posts. We complement the above dataset with the image dataset collected by Zannettou et al. (2020b). The dataset includes 5,859,439 images shared alongside /pol/ posts between July 1, 2016, and December 31, 2017. Overall, our dataset comprises all textual and image activity on /pol/ between July 1, 2016, and December 31, 2017, including 66M posts and 5.8M images.

Methodology

This section describes our methodology for detecting hateful text phrases and hateful imagery, focusing on Antisemitic and Islamophobic content.

Identifying Antisemitic and Islamophobic Phrases

Here, our goal is to identify a set of phrases that are Antisemitic/Islamophobic. To do this, we follow a multi-step semi-automated methodology. First, we use the SEVERE_TOXICITY scores from the Perspective API to iden-

[2] The exact methodology for creating this dataset was not made publicly available by OpenAI.

[3] In this work, we treat an image as similar to the text phrase if it has a cosine similarity of 0.3 or higher (see Methodology section).

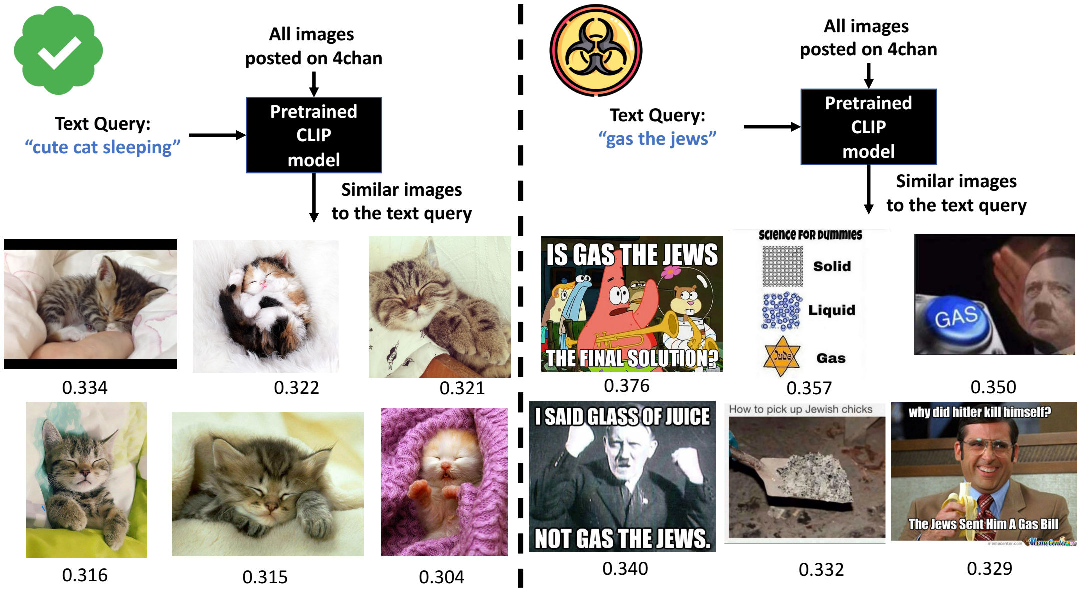
All images posted on 4chan
Pretrained CLIP model
Text Query: "cute cat sleeping"
Similar images to the text query
0.334
0.322
0.321
0.316
0.315
0.304
All images posted on 4chan
Pretrained CLIP model
Text Query: "gas the jews"
Similar images to the text query
IS GAS THE JEWS
THE FINAL SOLUTION?
0.376
Science For Dummies
Solid
Liquid
Gas
0.357
GAS
0.350
I SAID GLASS OF JUICE
NOT GAS THE JEWS.
0.340
How to pick up Jewish chicks
0.332
why did hitler kill himself?
The Jews Sent Him A Gas Bill
0.329

Figure 1: Example of images similar to text queries on 4chan (i.e., cosine similarity between the text CLIP-representation and the image CLIP-representation equals 0.3 or more). On the left side, we show a benign text query ("cute cat sleeping"), while on the right, we show the results for a toxic and Antisemitic query ("gas the jews").

Dataset	Textual		Visual	
	# Phrases	# Posts	# Images	# Posts
Antisemitism	326	209,224	15,711	143,506
Islamophobia	94	37,354	5,548	29,978
Total	420	246,578	21,259	173,484

Table 1: Overview of our Antisemitism/Islamophobia Textual and Visual datasets. The number of phrases is based on lemmatized versions, and the number of images is based on the unique pHash values. We consider only images associated with at least 10 phrases to reduce the number of false positives.

tify posts that are toxic/offensive without considering the target (e.g., if it is Antisemitic). Specifically, we consider all posts that have a score of 0.8 or more as toxic, following the methodology by Ribeiro et al. (2021). Out of the 66M posts in our dataset, we find 4.5M (6.7%) toxic posts.

Having extracted a set of toxic posts from 4chan's /pol/, we then aim to identify the main targets of hate speech on /pol/ by extracting the top keywords. To do this, we preprocess the data to remove HTML tags, stop words, and URLs, and then we create a term frequency-inverse document frequency array (TF-IDF). Next, we manually inspect the top 200 words based on their TF-IDF values and identify the words related to Jews or Muslims. As a result, we find seven keywords: "jews," "kike," "jew," "kikes," "jewish," "muslims," and "muslim." Then, based on these keywords, we filter the toxic posts obtained from the previous step, hence getting a set of 336K posts with a SEVERE_TOXICITY score of 0.8 and include at least one of the seven keywords. Note that we decide to focus on the top 200 terms because we want to focus on popular targeted groups in the set of toxic posts. Moreover, while terms such as "Islam" or "Islamic" might appear crucial to include in our list of keywords, we not include them because their TF-IDF scores are far from the top ones; the term "islam" occupies the position 275th, "islamic" the 1014th position, "mohammed" the position 1553th, and "prophet" the position 2505th.

Since our goal is to create a set of Antisemitic/Islamophobic phrases, we need to break down the toxic 4chan posts into sentences and then identify the ones that are Antisemitic/Islamophobic. To do this, we apply a sentence tokenizer (NLTK 2021b) on the 336K posts, obtaining 976K sentences. To identify common phrases used on 4chan's /pol/, we apply WordNet lemmatization (NLTK 2021a), excluding all sentences that appear less than five times. We obtain 4,582 unique common phrases; not all of these sentences are Antisemitic/Islamophobic. We note that some phrases are contained in longer phrases. However, we do not treat them as duplicates, given that the text encoder of the CLIP model encodes them differently.

Identifying whether a phrase is Antisemitic/Islamophobic is not a straightforward task and can not be easily automated. Therefore, we use manual annotation on the 4,582 common phrases to annotate the common phrases as Antisemitic/Islamophobic or irrelevant. Two authors of this paper independently annotated the 4.5K common phrases. On average, these phrases include 11.10 words ($\sigma = 20.82$). We discard long phrases (over seven words) during the annotation since our preliminary experiments showed that OpenAI's CLIP returns a considerable amount of false positives when provided with long text queries. We also consider as irrelevant phrases that target multiple demographic groups (e.g., hateful towards Muslims and Jews like "fuck jews and muslims" or hateful towards African Americans and Jews like "fuck niggers and jews"). To ease the annotation process, we create a spreadsheet that includes a clear description of our labels, as well as all the information that an annotator needs in order to inspect and correctly annotate a phrase (e.g., number of terms per phrase). Phrases labelled as Antisemitic express hostility to, prejudice towards, or discrimination against Jews (Dictionaries 2021). Phrases labelled as Islamophobic express fear of, hatred of, or prejudice against the Islam or Muslims in general (Merriam-Webster 2021). The two annotators agreed on 91% of the annotations with a Cohen's Kappa score of 0.69, which indicates a substantial agreement (Kvalseth 1989). After the independent annotations, the two annotators discussed the disagreements to come up with a final annotation on whether a phrase is Antisemitic/Islamophobic or irrelevant. After our annotation, we find 326 Antisemitic and 94 Islamophobic phrases. The list of the Antisemitic/Islamophobic phrases is available (González-Pizarro and Zannettou 2022).

Finally, we search for these Antisemitic/Islamophobic phrases on the entire dataset. We extract all posts that include any of the Antisemitic/Islamophobic phrases (*Textual* dataset), finding 247K posts. Note that we remove 864 (0.35%) posts that contain both Antisemitic and Islamophobic phrases. Overall, we find 209K (84.85%) Antisemitic posts and 37K (15.15%) Islamophobic posts (see Table 1).

Identifying Antisemitic and Islamophobic Images

Our goal is to identify Antisemitic and Islamophobic imagery using the pre-trained CLIP model (Radford et al. 2021). To do this, we encode all images in our dataset using the image encoder on the CLIP model, hence obtaining a high-dimensional vector for each image. Also, we encode all the Antisemitic/Islamophobic phrases (extracted from the previous step), using the text encoder on the CLIP model, obtaining a vector for each phrase. Then, we calculate all the cosine similarities between the image and text vectors, which allows us to assess the similarity between the phrases and the images. The main idea is that by comparing a hateful phrase to all the images, images with a high cosine similarity score will also be hateful. To identify a suitable cosine similarity threshold where we treat a text and an image similarly, we perform a manual annotation process.

Identifying a suitable threshold. First, we extract a random sample of ten Antisemitic/Islamophobic phrases (eight Antisemitic and two Islamophobic to match the percentage of Antisemitic/Islamophobic phrases in our dataset). Then, we extract a random sample of 200 images for each phrase while ensuring that the images cover the whole spectrum of cosine similarity scores. Specifically, we extract 50 random images with cosine similarity scores for each of the following ranges: [0.0, 0.20), [0.2, 0.25), [0.25, 0.3), [0.3, 0.4]. To select these ranges, we plot the Cumulative Distribution Function (CDF) of all cosine similarity scores obtained by comparing the ten randomly selected phrases and all the images in our dataset (we omit the figure due to space constraints). We find that 40% of the scores are below 0.2, and we expect these images to be entirely irrelevant to the phrase. To verify this, we select the [0.0, 0.20) range. Additionally, we select the [0.2, 0.25] because it has a considerable percentage of the scores (50%), and we expect that the images will not be very similar again. Finally, we select the [0.25, 0.3) and [0.3-0.4] ranges because we expect that the ideal threshold is somewhere in these two ranges, and devoting half of the selected images in these ranges will help us identify a suitable threshold.

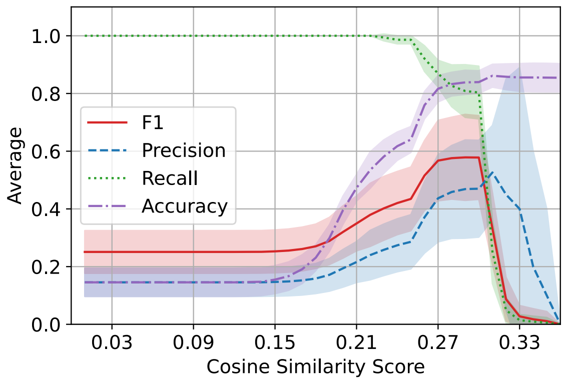

Figure 2: Performance of the CLIP model in identifying Antisemitic/Islamophobic imagery for varying cosine similarity thresholds. The lines refer to the average metric for ten random phrases (2K images), while the area refers to the standard deviation across the ten phrases.

Then, two authors of this paper independently annotated the 2,000 images to identify which are Antisemitic/Islamophobic or irrelevant. In a similar fashion to our annotations for toxic phrases, we labeled images as Antisemitic, those that clearly express hostility, prejudice towards, or discrimination against Jews (Dictionaries 2021). Images labeled as Islamophobic clearly express hatred of or prejudice against Muslims (Merriam-Webster 2021). The annotators agreed on 94% of the annotations with a Cohen's Kappa score of 0.75, which indicates a substantial agreement. Again, the two annotators solved the disagreements by discussing the images and deciding a final annotation on whether the image is Antisemitic/Islamophobic or irrelevant.

As a result, our initial ground truth dataset of Antisemitic and Islamophobic imagery includes 291 (14.55%) hateful images: 239 (82.13%) of them are Antisemitic and 52 (17.87%) are Islamophobic. Having constructed an initial ground truth dataset of Antisemitic and Islamophobic imagery, we then find the best performing cosine similarity threshold. We vary the cosine similarity threshold, and we treat each image as Antisemitic/Islamophobic (depending on the phrase used for the comparison) if the cosine similarity between the phrase and the image is above the threshold. Then, we calculate the accuracy, precision, recall, and F1 score, for each of the ten phrases. We report the average performance across all phrases and the standard deviation (as the area) in Fig. 2. We observe that the model performs best with a cosine similarity threshold of 0.3 as we achieve 0.84,

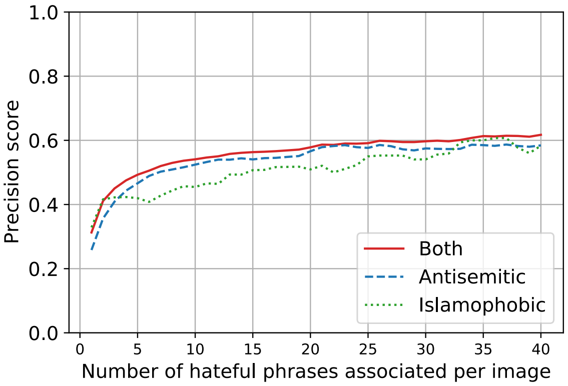

Figure 3: Precision of the CLIP model in identifying Antisemitic/Islamophobic imagery by varying the number of hateful phrases that have a cosine similarity of 0.3 or more with each image.

0.47, 0.80, and 0.58 for accuracy, precision, recall, and F1 score, respectively. Indeed, the 0.3 threshold is also used by previous work by Schuhmann et al. (2021) that inspected CLIP's cosine similarities between text and images and determined that 0.3 is a suitable threshold.

To construct our initial Antisemitic/Islamophobic image dataset, we extract all images that have a cosine similarity of 0.3 or higher with any of the Antisemitic/Islamophobic text phrases. We label each image as likely Antisemitic or likely Islamophobic depending on whether the textual phrase is Antisemitic or Islamophobic. To identify unique images, we use the Perceptual Hashing (pHash) algorithm (Monga and Evans 2006) that calculates a fingerprint for each image in such a way that any two images that look similar to the human eye map have minor differences in their hashes. Similar to the Textual dataset, we remove all images labeled as both Antisemitic and Islamophobic (3,325 images), mainly because our manual inspections indicate that most of them are noise. Overall, we find 69,610 likely Antisemitic and 22,519 likely Islamophobic images that are shared in 472,048 and 101,465 posts, respectively.

Evaluating performance to the entire dataset. To evaluate the quality of our Antisemitic/Islamophobic detection approach in the entire dataset (and not limited to a few phrases as before), we perform an additional manual annotation on 2,000 randomly selected images (from our 92K likely Antisemitic/Islamophobic images mentioned above). We obtain 1,507 (75.4%) potential Antisemitic images and 493 (24.7%) potential Islamophobic images. Two authors of this paper independently annotated these images to identify which are actually Antisemitic/Islamophobic. The two annotators agreed on 84.9% of the annotations with a Cohen's Kappa score of 0.64, which indicates a substantial agreement (Kvalseth 1989). The annotators identify 551 (27.6% out of total annotated images) Antisemitic/Islamophobic images: 389 (70.6%) are identified as Antisemitic and 162 (29.4%) are identified as Islamophobic.

Improving performance. Given the relatively small percentage (27.6%) of the images that are actually Antisemitic/Islamophobic, we set out to investigate how we can improve this performance. We hypothesize that we can improve the detection performance by considering the number of hateful phrases that have high cosine similarity with the detected images. Indeed, based on our annotated dataset, we find that images associated with a higher number of hateful phrases are more likely to be Antisemitic/Islamophobic (see Fig. 3). For instance, by considering only images associated (i.e., cosine similarity between the phrase and the image at least 0.3) with ten or more hateful phrases, 54,1% of them can be identified as Antisemitic/Islamophobic (see Fig. 3). Considering this threshold, the CLIP model has a precision score of 0.57 when identifying Antisemitic imagery and a precision score of 0.43 when identifying Islamophobic imagery. For the rest of the analysis, we use this threshold as it greatly reduces the number of false positives that are generated. Table 1 shows the final number of posts in our Visual dataset. Our final visual dataset contains 21K likely Antisemitic/Islamophobic images.

	Accuracy	Precision	Recall	F1
MMBT-Grid	0.70	0.37	0.61	0.46
MOMENTA-C w/o OCR	0.56	0.27	0.63	0.38
MOMENTA-C	0.60	0.27	0.51	0.35
MOMENTA-P w/o OCR	0.46	0.24	**0.73**	0.36
MOMENTA-P	0.57	0.29	0.69	0.40
CLIP Model	**0.81**	**0.54**	0.53	**0.54**

Table 2: Performance comparison between the CLIP model and the two baselines. MOMENTA-C and MOMENTA-P correspond to the MOMENTA model pre-trained on a COVID-19 and US Politics dataset, respectively.

Distance Metric & Dimensionality. We also investigate other ways to improve performance by using different distance metrics or applying dimensionality reduction techniques. In particular, we experiment with Euclidean distance and Mahalanobis distance (Mahalanobis 1936), with both performing substantially worse than cosine distance. Also, we try reducing the dimensionality of the CLIP embeddings to 64, 128, and 256 dimensions using the Uniform Manifold Approximation and Projection approach (McInnes et al. 2018), without any performance gains. Note that we do not include the actual performance with different distance metrics and after dimensionality reduction due to space constraints. Based on these results, for our detection and analysis, we use the cosine distance metric on the original embeddings obtained from the CLIP model.

Baseline models. Here, we aim to compare the performance of the CLIP model (that considers an image as Antisemitic/Islamophobic if it has a cosine similarity of 0.3 or more for at least ten hateful phrases), using our final ground truth dataset, which combines the two above-mentioned annotation procedures (4K images). Our ground truth dataset includes 678 Antisemitic images, 214 Islamophobic images, and 3,158 non-hateful images. We compare our method of identifying Antisemitic/Islamophobic imagery with two hateful detection models (see Table 2).

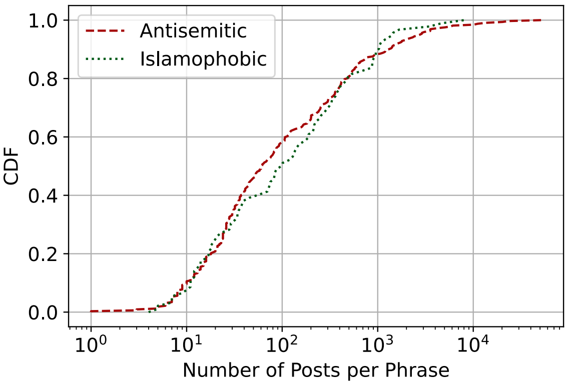

Figure 4: CDF of the number of Antisemitic/Islamophobic posts containing each phrase.

First, **MMBT-Grid** (Kiela et al. 2019) is a multimodal architecture that consists of supervised multimodal transformers using Image-Grid features. We use the pre-trained weights released by Kiela et al. (2020) for hateful image detection. Second, we use **MOMENTA** (Pramanick et al. 2021), which analyzes the input's local and global perspective for detecting harmful memes and their targets. Several experiments show that it outperforms several robust approaches. **MOMENTA** can identify images that have the potential to cause harm to individuals, organizations, and communities, which is also the focus of our work. This model is pre-trained with two datasets related to COVID-19 and US politics. While our 4chan dataset is different from those, we use MOMENTA as it is a generalizable model (Pramanick et al. 2021).

Table 2 shows the results for Antisemitic/Islamophobic imagery detection. We observe that the CLIP model outperforms all baselines in accuracy, precision, and F1 score, with an improvement of 0.11, 0.17, and 0.08, respectively (compared to the second-best performing model). Also, we find that the CLIP model has a lower recall score than the baselines. Nevertheless, in this work, we favor precision over recall, as we aim to reduce the number of false positives generated by our method.

Results

This section presents our results from analyzing the Antisemitic/Islamophobic Textual and Visual datasets.

Popular Phrases in the Textual Dataset

We start our analysis by looking into the most popular phrases in our Antisemitic/Islamophobic textual datasets. Fig. 4 shows the Cumulative Distribution Function (CDF) of the number of posts per each Antisemitic/Islamophobic phrase. We observe that these hateful phrases tend to appear in a considerable amount of posts. For instance, 90.4% and 92.47% of the Antisemitic and Islamophobic phrases appear in at least ten posts. Furthermore, we identify that the percentage of Antisemitic phrases (41.8%) that appear in at least 100 posts is slightly lower than the percentage of Islamophobic phrases (49.46%). At the same time, we observe that a small percentage of phrases (11.54%) is shared in more than 1000 posts on the Antisemitic/Islamophobic textual datasets combined.

Antisemitic phrases		**Islamophobic phrases**	
Phrase	**# Posts**	**Phrase**	**# Posts**
a kike	51,216	fuck muslim	7,993
fuck kike	23,604	kill muslim	4,639
fuck jew	20,241	fuck islam	3,464
gas the kike	13,353	kill all muslim	1,672
fuck off kike	11,108	muslim be terrorist	1,445
kike shill	10,134	i hate muslim	1,379
gas the kike race war now	6,105	muslim shithole	1,208
kill jew	5,815	muslim shit	1,085
you fuck kike	5,007	all muslim be terrorist	1,039
filthy kike	4,111	muslim be bad	1032
jew fuck	3,557	ban all muslim	958
kike faggot	3,540	fuck mudslimes	951
kike on a stick	3,537	muslim cunt	907
gas the jew	3,081	i hate islam	881
faggot kike	2,905	fuck sandniggers	876

Table 3: Top 15 phrases (lemmatized versions), in terms of the number of posts, in our Antisemitic and Islamophobic Textual dataset. For each phrase, we report the number of posts that contain it.

We also report the top 15 phrases, in terms of the number of posts, in our Antisemitic and Islamophobic Textual dataset (see Table 3). In the first dataset, we observe that 12 out of the 15 most frequent phrases contain the term "kike," a derogatory term to denote Jews. We also identify three phrases related to the extermination procedure in the gas chambers during the holocaust. Indeed, 16,433 (7.85%) of the Antisemitic posts contain at least one of these phrases: "gas the kike," "gas the jew," or "gas the kike race war now." Phrases accusing jews of being accomplices ("kike shill") or alluding to a supposed good social-economic status ("filthy kike") are also trendy, appearing in 10,134 and 4,111 posts, respectively.

We also show the top 15 most popular Islamophobic phrases in Table 3. Here, we observe many posts with phrases calling Muslims as terrorists. For instance, "Muslims be terrorist" and "All Muslim be terrorist" appear in 1,445 and 1,039 posts, respectively. The second and fourth most popular phrases are calls for attacks targeting Muslims; "Kill Muslim" and "Kill all Muslim" appear in approximately 4.6K and 1.6K posts. We also find phrases against Islam; "Fuck Islam" appears in 3.4K posts and "I hate Islam" in 881 posts. Finally, we also identify phrases containing the terms "mudslimes" (Urban Dictionary 2006a) and "sandniggers" (Urban Dictionary 2006b), which are derogatory names to refer to Muslims and Arabs.

Popular Images in the Visual Dataset

We also look into the popularity of images in our Antisemitic/Islamophobic datasets (in terms of the number of posts they shared). Fig. 5 shows the CDF of the number of posts for each Antisemitic/Islamophobic image. We observe

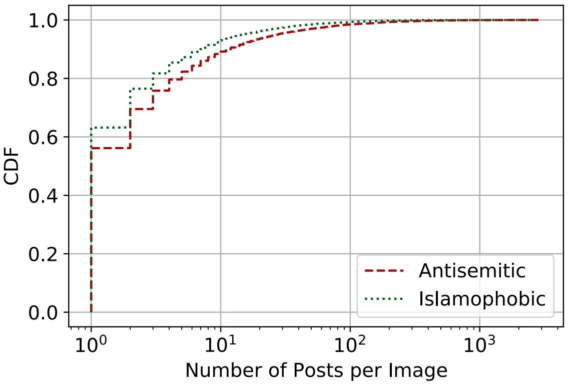

Figure 5: CDF of the number of Antisemitic/Islamophobic posts containing each image.

that Antisemitism and Islamophobia imagery is a diverse problem, with 56.13% and 63.16% of the images appearing only in one post for Antisemitism and Islamophobia, respectively. At the same time, we have a small percentage of images that are shared many times on 4chan's /pol/; 10.62% of all the Antisemitic/Islamophobic imagery are shared at least ten times. Overall, we observe a similar pattern in the distribution of the number of posts per image for Antisemitism and Islamophobia.

Next, we look into the most popular images in our Antisemitic and Islamophobic visual datasets. We avoid showing the images since they are highly offensive and are likely to disturb the readers, and only discuss the main insights from inspecting the most popular images. We identify the top ten Antisemitic images. We find that the Happy Merchant meme appears in five out of the top ten Antisemitic images. Other interesting examples of popular images are hinting that members of the Jewish are allegedly the masterminds of lousy stuff happening or conspiracy theories (i.e., shut down the Jewish plan or Rabbi painting Nazi symbols). We also find two false positives among the top ten Antisemitic images; the one shows Pepe the Frog wearing a t-shirt with a swastika symbol, while the other shows again Pepe the frog dressed as a crusader with the text "KEK WILLS IT."

We also look at the top ten most popular Islamophobic images. We find two images that are pretty graphic and insult Prophet Muhammad and the Holy Quran (again we do not include the images since they are highly offensive). These two Islamophobic images are included in 1.1K posts within 4chan's /pol/, indicating that graphic images that insult Islam as a religion are used a lot on 4chan. Moreover, we find images that include sarcasm and link Muslims to terrorism; for instance, CLIP links the phrase "muslims shihole" with an image of a Muslim dressed as a terrorist, likely indicating that the CLIP model thinks that Muslims are terrorists. We also find an image linking Muslims to the Happy Merchant meme; i.e., the Happy Merchant dressed as a Muslim. Among the top 10 Islamophobic images, we find one image that is a false positive. This image is showing a meme that compares Americans to Europeans and is likely considered as related because it includes the word "Muhammad," however upon manual examination, we do not find this image Islamophobic.

Antisemitic/Islamophobic Content Over Time

This section presents our temporal analysis that shows the distribution of Antisemitic/Islamophobic content over time. Fig. 6 shows the number of hateful posts per day in the Antisemitic Textual/Visual datasets. We run Kendall's tau-b correlation to determine the relationship between the number of posts in the Antisemitic Textual and Visual dataset. We find a strong, positive, and statistically significant correlation ($\tau = .635, p < .001$), indicating that Antisemitic content is spread both using text and images in a similar fashion. We also observe the highest volume of textual and image content between April 6, 2017, and April 9, 2017, with 4,132 (1.97% of the dataset) posts in the Textual dataset and 2,223 (1.55%) posts in the Visual dataset. This finding confirms previous findings from Zannettou et al. (2020b) that identified a spike in the spread of the Happy Merchant memes on April 7, 2017.

By inspecting the top 15 most frequent images during that period (we omit the figure due to space constraints), we identified that those images are related to the decision of Donald Trump to remove Steve Bannon from the National Security Council Post on April 5, 2017 (Costa and Phillip 2017) and a missile attack in Syria on April 7, 2017 (Rosenfeld 2017). According to newspapers (Baker, Haberman, and Thrush 2017; Haberman, Peters, and Baker 2017), Jared Kushner, the Jewish Trump's son-in-law, seemed to be acting as a shadow secretary of state visiting and taking Middle East portfolios after that event. This political decision spread a volume of image content with the face of Jared Kushner. Also, there are some references to Donald Trump that indicate that he is controlled by Israel (e.g., most popular images are associated with the phrases "fuck trump and fuck jews", and "fuck trumpstein and fuck jewish people").

We also evaluate the distribution of Islamophobic posts over time. Fig. 7 shows the number of Islamophobic posts per day in our Textual/Visual datasets. We also find a statistically significant, strong, and positive correlation ($\tau = .393, p < .001$). In both datasets, we find a peak of activity on May 23, 2017, with 482 and 361 posts in the Textual and Visual datasets, respectively. By manually inspecting the top 15 images shared that day, we identify that the high volume of posts is related to the Manchester Bombing; on May 22, 2017, a British man detonated a suicide bomb in the foyer of the Manchester Arena as people were leaving a concert by pop singer Ariana Grande. On May 23, ISIS claimed responsibility for the attack. (Cobain et al. 2017). This event raised hateful online narratives defining Muslims as terrorists (Downing, Gerwens, and Dron 2022). We find images that contain explicit references to this attack and images questioning whether Islam is a religion of peace. Overall, our findings highlight that both textual and visual hateful content is likely influenced by real-world events, with peaks of hateful activity observed during important real-world events that are related to the demographic groups we study.

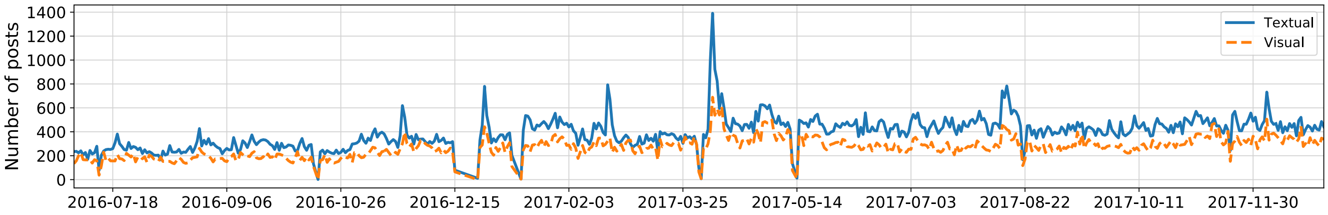

Figure 6: Number of Antisemitic posts per day in our Textual/Visual datasets.

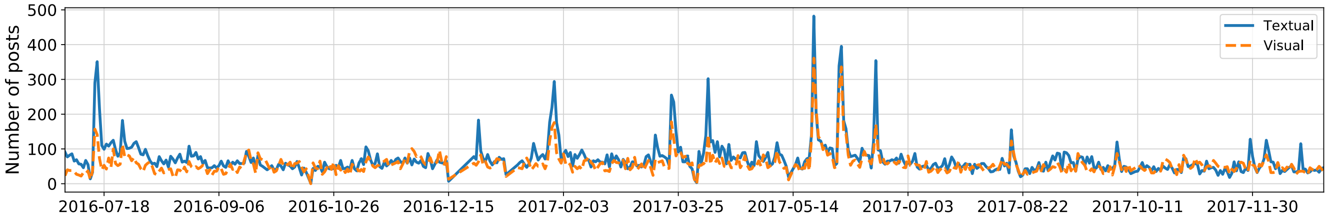

Figure 7: Number of Islamophobic posts per day in our Textual/Visual datasets.

Related Work

Hate Speech detection. Hate speech has recently received much research attention, with several works focusing on detecting hate speech in online social media. Initial research on hate speech analysis is typically oriented toward monolingual and single-classification tasks due to the complexity of the task. They used simple methods such as dictionary lookup (Guermazi, Hammami, and Hamadou 2007), bag of words (Guermazi, Hammami, and Hamadou 2007), or SVM classifiers (Malmasi and Zampieri 2017; Senarath and Purohit 2020). Recent efforts are proposing multilingual and multitask learning by using deep learning models (Wang et al. 2020; Ousidhoum et al. 2019; Glavaš, Karan, and Vulić 2020; Vitiugin, Senarath, and Purohit 2021). While previous approaches to characterize and identify hate speech focus purely on the *content* posted in social media, some research efforts shift the focus towards detecting hateful users by exploiting other contextual data (Ribeiro et al. 2018; Ahmed, Vidgen, and Hale 2021; Chaudhry and Lease 2022; Waseem and Hovy 2016). Furthermore, other research efforts investigate to what extend the models trained to detect general abusive language generalize between different datasets labeled with different abusive language types (Karan and Šnajder 2018; Meyer and Gambäck 2019; Rizoiu et al. 2019; Salminen et al. 2020; Nejadgholi and Kiritchenko 2020). While less explored, some work focuses on multimodal settings formed by text and images (Das, Wahi, and Li 2020; Kiela et al. 2020). Gomez et al. (2020) build a large dataset for multimodal hate speech detection retrieved from Twitter using specific hateful seed keywords, finding that multimodal models do not outperform the unimodal text ones.

Antisemitism. Antisemitism has grown and proliferated rapidly online and has done so mostly unchecked; Zannettou et al. (2020b) call for new techniques to understand it better and combat it. Ozalp et al. (2020) train a scalable supervised machine learning classifier to identify Antisemitic content on Twitter. Chandra et al. (2021a) propose a multimodal system that uses text, images, and OCR to detect the presence of Antisemitic textual and visual content. They apply their model on Twitter and Gab, finding that multiple screenshots, multi-column text, and texts expressing irony, and sarcasm posed problems for the classifiers. To characterize Antisemitism, Enstad (2021) propose an analytical framework composed of three indicators: Antisemitic attitudes, incidents targeting Jews, and Jew's exposure to Antisemitism. Their results show that attitudes vary by geographic and cultural region and among population sub-groups.

Islamophobia. Surveys show that Islamophobia is rising on Web communities (Hafez et al. 2019). Vidgen and Yasseri (2020) build an SVM classifier to distinguish between tweets non-Islamophobic, weak Islamophobic, and strong Islamophobic with a balanced accuracy of 83%. Cervi (2020) use clause-based semantic text analysis to identify the presence of Islamophobia in electoral discourses of political parties from Spain and Italy. Chandra et al. (2021b) apply topic modeling and temporal analysis over tweets from the #coronajihad to identify the existence of Islamophobic rhetoric around COVID-19 in India. Civila, Romero-Rodríguez, and Civila (2020) apply content analysis over 474 images and texts from Instagram posts under the hashtag #StopIslam. Alietti and Padovan (2013) conduct telephone surveys on 1.5K Italians on Antisemitic and Islamophobic attitudes, finding an overlap of ideology for both types of hate speech.

Discussion & Conclusion

In this work, we explored the problem of Antisemitism/Islamophobia on 4chan's /pol/ using OpenAI's CLIP model. We devised a methodology to identify Antisemitic/Islamophobic textual phrases using Google's Perspective API and manual annotations and then used the CLIP model to identify hateful imagery based on the phrases. We found that the CLIP can play a role in detecting hateful content; using our methods, the CLIP can detect hateful content with an accuracy of 81%. Also, we found that Antisemitic/Islamophobic imagery exists in a similar number of posts when compared to Antisemitic/Islamophobic speech on 4chan's /pol/. Additionally, our work contributes to research efforts focusing on understanding and detecting hateful content by making a dataset of 420 Antisemitic/Islamophobic phrases, 246K textual posts, and 21K images available (upon request). Below, we discuss the implications of our findings for researchers focusing on detecting hate speech and for researchers working on large pre-trained models like OpenAI's CLIP.

Prevalence of Antisemitic/Islamophobic Imagery. Our findings show that images play a significant role in the spread of hateful content. This is likely because 4chan is an imageboard and a fringe Web community; hence, a large volume of hateful content is disseminated via images. Nevertheless, the problem of hateful imagery exists on other mainstream platforms (e.g., Twitter), hence it is of paramount importance to develop better and more accurate systems for the detection of hateful content across multiple modalities. For instance, we argue that the spread of hateful content via videos is an unexplored problem, and there is a need to develop models across text, images, and videos.

Performance and Sensitivity of CLIP model. Our experiments indicate that large-pre-trained models like CLIP are pretty powerful and have general knowledge that can be used for various tasks. When considering the hateful content detection task, the CLIP model should be used with caution. This is because the CLIP model highly depends on how the input text query is written, influencing the number of false positives returned. When CLIP is used for moderation purposes, we argue that it is essential to have humans in the loop to ensure that the automated model works as expected. Additionally, we observed that the CLIP model performs worse when considering input text queries that comprise many words. This poor performance also occurs when images contain text that are long (from our annotations we observed that many false positives are screenshots of images with a lot of text). This indicates that we need more powerful text encoders that can capture the primary meaning of textual phrases, irrespectively of how long they are. Also, we emphasize that CLIP's performance on detecting hateful content yields a substantial number of false positives, which is expected given the nature of the problem (i.e., hate speech is sometimes hard to identify and subjective). The same applies to all the baselines that we experimented with, in particular, to a larger extent since their precision score is poorer compared to CLIP. CLIP's poor precision score (0.54) is also reflected in the dataset that we are releasing. Researchers that aim to use the dataset for other downstream tasks (e.g., implementing classifiers for hateful content) should have this limitation in mind and potentially make additional manual annotations to decrease the number of false positives in the dataset.

Biases on CLIP model. Large pre-trained models like OpenAI's CLIP are trained on large-scale datasets from the Web, and these datasets might include biases, hence some of the bias is transferred to the trained model. From our experiments and manual annotations, we observed some instances of such biases; e.g., the CLIP model identifying an image showing a terrorist as similar to a text phrase talking about Muslims (i.e., the model is biased towards Muslims, thinking they are terrorists). When considering that these models can be used for moderation purposes (e.g., detecting and removing hateful content), such biases can result in false positives biased towards specific demographics. This can cause users to lose trust in the platform and its moderation systems and may cause them to stop using the platform. Overall, given the increasing use of such models in real-world applications, there is a pressing need to develop techniques and tools to diminish such biases from large pre-trained models.

Limitations. Our work has several limitations. First, we rely on Google's Perspective API to initially identify hateful text, which has its limitations (e.g., might not understand specific slurs posted on 4chan) and biases when detecting hateful text. Second, our analysis focuses on a small number of short textual phrases (at most seven words), mainly because our preliminary results showed that CLIP does not perform well in detecting hateful imagery when considering long phrases. Therefore, our approach is likely to miss some Antisemitic/Islamophobic text and imagery because of the small number of phrases that we consider. Third, we rely entirely on a pre-trained CLIP model; this is not ideal since the CLIP model is trained on a public dataset obtained from multiple Web resources and is not specific to our platform of interest (i.e., 4chan). This might result in the model not recognizing some 4chan slurs or slang language. Fourth, our work and analysis focuses on a single data source (4chan's /pol/), a limitation that does not allow us to investigate the performance of CLIP on other more mainstream communities like Twitter, Facebook, and Reddit (e.g., how CLIP performs on content shared on mainstream platforms). As part of our future work, we plan to investigate CLIP's performance on other platform and we intend to fine-tune the CLIP model with datasets obtained from mainstream social networks such as Twitter and Reddit as well as fringe Web communities that are often associated with the alt-right (e.g., Gab). Finally, as discussed above, the released dataset includes a substantial number of false positives, which indicates that researchers should consider the existence of false positives in the dataset when using it.

References

Ahmed, Z.; Vidgen, B.; and Hale, S. A. 2021. Tackling Racial Bias in Automated Online Hate Detection: Towards Fair and Accurate Classification of Hateful Online Users Using Geometric Deep Learning. *CoRR*, abs/2103.11806.

Alietti, A.; and Padovan, D. 2013. Religious racism. Islamophobia and antisemitism in Italian society. *Religions*, 4(4): 584–602.

Arango, A.; Pérez, J.; and Poblete, B. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *SIGIR*.

Baker, P.; Haberman, M.; and Thrush, G. 2017. Trump Removes Stephen Bannon From National Security Council Post. https://www.nytimes.com/2017/04/05/us/politics/national-security-council-stephen-bannon.html. Accessed: 2023-03-31.

Bernstein, M.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the international AAAI conference on web and social media*, volume 5, 50–57.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NIPS*, volume 33, 1877–1901.

Caruana, R.; and Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. In *ICML*.

Cervi, L. 2020. Exclusionary Populism and Islamophobia: A Comparative Analysis of Italy and Spain. *Religions*, 11(10): 516.

Chandra, M.; Pailla, D.; Bhatia, H.; Sanchawala, A.; Gupta, M.; Shrivastava, M.; and Kumaraguru, P. 2021a. “Subverting the Jewtocracy”: Online Antisemitism Detection Using Multimodal Deep Learning. In *WebSci*.

Chandra, M.; Reddy, M.; Sehgal, S.; Gupta, S.; Buduru, A. B.; and Kumaraguru, P. 2021b. ”A Virus Has No Religion”: Analyzing Islamophobia on Twitter During the COVID-19 Outbreak. In *HT*.

Chaudhry, P.; and Lease, M. 2022. You Are What You Tweet: Profiling Users by Past Tweets to Improve Hate Speech Detection. In *IConference*, 195–203.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Civila, S.; Romero-Rodríguez, L. M.; and Civila, A. 2020. The Demonization of Islam through Social Media: A Case Study of #Stopislam in Instagram. *Publications*, 8(4): 52.

Cobain, I.; Perraudin, F.; Morris, S.; and Parveen, N. 2017. Salman Ramadan Abedi Named by Police as Manchester Arena Attacker.”. *The Guardian*.

Costa, R.; and Phillip, A. 2017. Stephen Bannon removed from National Security Council. https://www.washingtonpost.com/news/post-politics/wp/2017/04/05/steven-bannon-no-longer-a-member-of-national-security-council/. Accessed: 2023-03-31.

Das, A.; Wahi, J. S.; and Li, S. 2020. Detecting Hate Speech in Multi-modal Memes. *arXiv preprint arXiv:2012.14891*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*, 4171–4186.

Diba, A.; Sharma, V.; Safdari, R.; Lotfi, D.; Sarfraz, S.; Stiefelhagen, R.; and Van Gool, L. 2021. Vi2CLR: Video and Image for Visual Contrastive Learning of Representation. In *ICCV*, 1502–1512.

Dictionaries, O. L. 2021. anti-Semitism. https://www.oxfordlearnersdictionaries.com/definition/american_english/anti-semitism. Accessed: 2023-03-31.

Downing, J.; Gerwens, S.; and Dron, R. 2022. Tweeting terrorism: Vernacular conceptions of Muslims and terror in the wake of the Manchester Bombing on Twitter. *CTS*, 1–28.

Enstad, J. D. 2021. Contemporary Antisemitism in Three Dimensions: A New Framework for Analysis. *SocArXiv*, 28.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, 6894–6910.

Giorgi, J. M.; Nitski, O.; Bader, G. D.; and Wang, B. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *ACL*.

Glavaš, G.; Karan, M.; and Vulić, I. 2020. XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages. In *COLING*.

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2020. Exploring hate speech detection in multimodal publications. In *WACV*, 1470–1478.

González-Pizarro, F.; and Zannettou, S. 2022. Dataset: Understanding and Detecting Hateful Content using Contrastive Learning. https://zenodo.org/record/6993868#.ZCAS4uxBwxw. Accessed: 2023-03-31.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, volume 27.

Google. 2021. Perspective. https://www.perspectiveapi.com/. Accessed: 2023-03-31.

Guermazi, R.; Hammami, M.; and Hamadou, A. B. 2007. Using a Semi-automatic Keyword Dictionary for Improving Violent Web Site Filtering. In *SITIS 2007*, 337–344.

Haberman, M.; Peters, J. W.; and Baker, P. 2017. In battle for Trump's heart and mind, it's Bannon vs. Kushner. https://www.nytimes.com/2017/04/06/us/politics/stephen-bannon-white-house.html. Accessed: 2023-03-31.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

Hafez, F.; Bayrakli, E.; Faytre, L.; Easat-Daas, A.; Younes, A.-E.; Kutuzova, N.; Karčić, H.; Emin, H. A.; Meškić, N. K.; Dizdarevic, S. M.; et al. 2019. European Islamophobia Report 2019. https://setav.org/en/assets/uploads/2020/06/EIR_2019.pdf. Accessed: 2023-03-31.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *ICML*.

Hine, G.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*.

Karan, M.; and Šnajder, J. 2018. Cross-Domain Detection of Abusive Language Online. In *ALW2*, 132–137.

Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; and Testuggine, D. 2019. Supervised multimodal bitransformers for classifying images and text. In *ViGIL-NAACL*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *NIPS*.

Kim, S.; Lee, G.; Bae, S.; and Yun, S.-Y. 2020. MixCo: Mix-up Contrastive Learning for Visual Representation. In *NeurIPS Workshop on Self-Supervised Learning*.

Konikoff, D. 2021. Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *P&I*.

Kvalseth, T. O. 1989. Note on Cohen's Kappa. *Psychol. Rep.*

Mahalanobis, P. C. 1936. On the generalized distance in statistics. National Institute of Science of India.

Malmasi, S.; and Zampieri, M. 2017. Detecting Hate Speech in Social Media. In *RANLP*.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Merriam-Webster. 2021. Islamophobia. https://www.merriam-webster.com/dictionary/Islamophobia. Accessed: 2023-03-31.

Meyer, J. S.; and Gambäck, B. 2019. A Platform Agnostic Dual-Strand Hate Speech Detector. In *ALW*, 146–156.

Monga, V.; and Evans, B. L. 2006. Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs. *TIP*, 15: 3452–3465.

Nejadgholi, I.; and Kiritchenko, S. 2020. On Cross-Dataset Generalization in Automatic Detection of Online Abuse. In *ALW-EMNLP*, 173–183.

NLTK. 2021a. NLTK Lemmatization. https://www.nltk.org/_modules/nltk/stem/wordnet.html. Accessed: 2023-03-31.

NLTK. 2021b. NLTK Tokenization. https://www.nltk.org/api/nltk.tokenize.html. Accessed: 2023-03-31.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *EMNLP-IJCNLP*, 4675–4684.

Ozalp, S.; Williams, M. L.; Burnap, P.; Liu, H.; and Mostafa, M. 2020. Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Soc. Media Soc.*

Papasavva, A.; Zannettou, S.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *ICWSM*.

Perspective API. 2018. https://www.perspectiveapi.com/. Accessed: 2023-03-31.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *EMNLP*, 4439–4455.

Prisk, D. 2017. The hyperreality of the Alt Right: how meme magic works to create a space for far right politics. *SocArXiv preprint doi:10.31235/osf.io/by96x*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ribeiro, M. H.; Calais, P. H.; Santos, Y. A.; Almeida, V. A.; and Meira Jr, W. 2018. Characterizing and detecting hateful users on twitter. In *ICWSM*.

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and West, R. 2021. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. In *CSCW*.

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3.

Rizoiu, M.-A.; Wang, T.; Ferraro, G.; and Suominen, H. 2019. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*.

Rosenfeld, E. 2017. Trump launches attack on Syria with 59 Tomahawk missiles. https://www.cnbc.com/2017/04/06/us-military-has-launched-more-50-than-missiles-aimed-at-syria-nbc-news.html. Accessed: 2023-03-31.

Salminen, J.; Hopf, M.; Chowdhury, S. A.; Jung, S.-g.; Almerekhi, H.; and Jansen, B. J. 2020. Developing an online hate classifier for multiple social media platforms. *HCIS*, 10(1): 1–34.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Data-Centric AI Workshop*.

Sellars, A. 2016. Defining hate speech. *Berkman Klein Center Research Publication*, (2016-20): 16–48.

Senarath, Y.; and Purohit, H. 2020. Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media. In *ICSC*, 199–202.

Tuters, M.; Jokubauskaitė, E.; and Bach, D. 2018. Post-truth protest: how 4chan cooked up the Pizzagate Bullshit. *M/c Journal*, 21(3).

Urban Dictionary. 2006a. Mudslime definition. https://www.urbandictionary.com/define.php?term=mudslime. Accessed: 2023-03-31.

Urban Dictionary. 2006b. Sandniggers definition. https://www.urbandictionary.com/define.php?term=sand\%20niggers. Accessed: 2023-03-31.

Vidgen, B.; and Yasseri, T. 2020. Detecting weak and strong Islamophobic hate speech on social media. *J. Inf. Technol. Politics*, 17(1): 66–78.

Vitiugin, F.; Senarath, Y.; and Purohit, H. 2021. Efficient Detection of Multilingual Hate Speech by Using Interactive Attention Network with Minimal Human Feedback. In *WebSci*.

Wang, K.; Lu, D.; Han, C.; Long, S.; and Poon, J. 2020. Detect All Abuse! Toward Universal Abusive Language Detection Models. In *COLING*, 6366–6376.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL*, 88–93.

Wu, Z.; Wang, S.; Gu, J.; Khabsa, M.; Sun, F.; and Ma, H. 2020. Clear: Contrastive learning for sentence representation. In *CoRR*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Adv. Neural Inf. Process. Syst.*, 33: 5812–5823.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal Contrastive Training for Visual Representation Learning. In *CVPR*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the origins of memes by means of fringe web communities. In *IMC*.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *IMC*.

Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020a. Measuring and characterizing hate speech on news websites. In *WebSci*.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020b. A quantitative approach to understanding online antisemitism. In *ICWSM*.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022. Contrastive learning of medical visual representations from paired images and text. In *MLHC*.