Analyzing Antisemitism and Islamophobia using a Lexicon-based Approach

Moonis Ali[◇,♠], **Savvas Zannettou**[‡,♠]

[◇]University of Saarland, [♠]Max Planck Institute for Informatics, [‡]Delft University of Technology
moonis.ali@mpi-inf.mpg.de, s.zannettou@tudelft.nl

Abstract

The spread of Antisemitic and Islamophobic content in a longstanding problem, in particular within fringe Web communities. In this work, we attempt to analyze the spread of Antisemitic and Islamophobic content on 4chan's Politically Incorrect board (/pol/) using a lexicon-based approach. We use an openly-accessible knowledge graph, word embedding techniques that allow us to assess semantic similarity between terms, as well as manual annotations to create 2 lexicons. A lexicon of 48 Antisemitic terms and another lexicon of 135 Islamophobic terms. Then, by extracting all posts containing these terms from /pol/, we assess the popularity and veracity (i.e., what percentage of posts that contain these terms are actually Antisemitic/Islamophobic). We find that 93% and 81% of posts that contain terms from our lexicons are Antisemitic and Islamophobic, respectively. Also, we find that the veracity and frequency of these terms greatly varies on 4chan's /pol/. Finally, using topic modeling, we provide an overview of how popular Antisemitic and Islamophobic terms are used on 4chan's /pol/. To conclude, we make publicly available our lexicons for Antisemitic and Islamophobic terms, which are likely to be useful for researchers working on Antisemitism/Islamophobia or hate speech in general.

Disclaimer. In this paper we analyze content that we consider as derogatory, offensive or racist. We thus warn the reader(s) that in the rest of this paper, we do not censor any of the language and suggest reader discretion.

1 Introduction

The increased access to Internet and social media platforms has led to sprawling of Internet communities and cultures on "mainstream" social networks like Facebook, Twitter etc. At the same time, it has also allowed for more opportunities for congregation of niche and clandestine user communities like Gab, 8chan, 4chan which are sometimes referred to as the "fringe" web communities. Often do these communities act as meeting grounds for racist and nationalistic movements to flourish online (Breland 2020). Actors from these groups in recent times have actively carried out real world violence resulting in injuries and deaths of innocent people (Austin 2019). Reports in News media (Evans 2018) suggest, these communities in particular 4chan's /pol/ may have "radicalized" these "active users" from benign users to terrorists. Two such example is the case of the Pittsburgh synagogue shooter and the Christchurch mosque shooter (Perry 2020). Hence, we cannot understate the role of fringe web communities like 4chan's politically incorrect messaging board (/pol/) and their impact.

In this work, we attempt to provide a methodical study of Antisemitism (AS) and Islamophobia (IP) on /pol/ with our aim to draw effective analyses, from a clandestine platform that is often considered to use niche language and coded references (Zannettou et al. 2020) to attack minority religious and ethnic groups and is notorious for far-right and racist content (Papasavva et al. 2020). In order to analyze the AS and IP content on the platform we restrict our focus to the "terms" that are directly used as AS / IP tropes or indirectly connote such behavior. We thus aim to take a lexical (term/-word focused) approach to curate lexica (here plural for, lexicons) one for AS and one for IP. Additionally, we are also motivated by the need to have an effective lexicon to extract "relevant" posts and collect datasets for conducting research in this domain. Thus our study takes a lexical approach in understanding the prevalent Antisemitism and Islamophobia on /pol/ and the nuances thereof. To this end, our research work attempts to answer the following research questions:

- **RQ1:** How to create lexica for identifying Antisemitism and Islamophobia and how accurate are the terms that we use for analyzing Antisemitic/Islamophobic hate speech on /pol/?
- **RQ2:** In what context are the frequently mentioned Antisemitic and Islamophobic terms used in? Are there observable patterns/similarities/differences?

To provide answers to the above-mentioned questions, we use several methods that allow us to create lexica and analyze the contexts that terms appear. Specifically, to create our Antisemitic and Islamophobia lexica, we use a combination of publicly available knowledge graphs, word embedding models trained on 4chan's /pol/, and manual annotations. The combination of these methods allow us to generate both general-purpose lexica related to Antisemitism and Islamophobia, as well as identify terms that are specifically used in fringe Web communities like /pol/. To analyze the context and topics that our Antisemitic/Islamopho-

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

bia terms appear on 4chan's /pol/, we use a topic modeling approach that leverage Sentence Embeddings and clustering techniques (Grootendorst 2020) that allow us to extract clusters of posts that we later qualitatively analyze.

Main Findings. Our main findings are:

- **RQ1:** We find that the terms in our curated lexica (AS lexicon and IP lexicon) are effective in identifying AS/IP. By manually annotating samples of 50 posts per each term from our IP/AS datasets, we find an accuracy of 93.83 % for posts including Antisemitic terms and & 81.18% for posts including Islamophobic terms (i.e., 93% of the posts that include Antisemitic terms are indeed Antisemitic). Also, we find that users on /pol/ "prefer" to use certain specific AS / IP term(s) overwhelmingly more than others. Especially, we note that this difference is more noticeable in case of Antisemitic terms example, an AS term "kike" is referenced in around 67% of the total Antisemitic posts we study.
- **RQ2:** We find that among the most mentioned AS terms. Most (used) important terms for 'all' topical clusters yielded for each term reference names of religions, names of geo-political entities along with extensive usage of antisemitic tropes. Our meta-labeling of themes for topical clusters suggests that discussions related to 'Donald Trump as a jewish agent' is observale for posts mentioning 'kike'. Similarly, we find the theme of 'Holocaust denail' in posts mentioning terms, 'shlomo' and 'juden'. Along with this, we also observe the theme where posts refer to Jordan Peterson as Juden Peterson in an antisemitic wordplay. For most mentioned IP terms our contextual analysis shows that topical clusters related to these terms always reference names of religions, geo-political entities along with islamophobic, antisemitic or racist symbolism. We also observe that the term 'paki' is almost always used in narratives related to racism against people of color.

2 Related Work

In this section we review previous work. We differentiate between previous work as work focusing on generic hate speech (not specifically limited to an individual race or minority group like AS or IP) and work specific to AS and IP. We focus on works that analyse fringe web communities like /pol/.

Generic Hate speech Recent years have seen an increased interest in the area of hate speech research on fringe web communities in particular the recency of the related works we present here show that this interest has resulted in multiple studies that specifically focus on /pol/. Generic studies such as (Papasavva et al. 2020) which provide public access to their dataset of 134.5M posts from /pol/ (June 2016 - November 2019) observe that around 27% of these posts can be considered as "Severely Toxic" and containing extremely unpleasant and/ or hateful content. Similarly, (Hine et al. 2017) focus on /pol/ and use Hatebase as a source for 'terms' to extract and analyse 8M posts from June 2016 to September 2016. They note that 12% of these posts contain hate speech. (Zannettou et al. 2018) analyse the multi-modal aspect of /pol/ and report that among the 4.3M images they study, they report that a substantial amount are infact racist and hateful memes that are common in fringe Web communities like /pol/.

Antisemitism and Islamophobia specific studies Our study reveals that there are works that analyse antisemitism on fringe platforms such as /pol/. In a study dedicated to analysis of antisemitism (Zannettou et al. 2020) study 67.4M posts and 5.8M images, their main findings reveal that antisemitic terms are closely associated with the generic terms that reference judaism or jews in terms of cosine similarities and there exist substantial antisemitic memes with diverse themes on the platform. Similarly, (Mittos et al. 2020) analyse 1.3M posts regarding a different subject altogether, they intend to study conversations regarding "genetic testing" but their results reveal that antisemitic language is a substantial feature of such conversations. A similar study on /pol/ by (Zelenkauskaite et al. 2021) that does a comparative analysis of AS and IP around Pittsburgh and Christchurch attacks, observes that there exist a "dedicated" user base on the platform that support the said attacks. Another study by (Crawford, Keen, and Suarez-Tangil 2021) that assesses 135K images for violence and extremism on 4chan,8kun and other niche websites, shows that out of the 20 most popular images shared 12 contained an "omnibus" of AS conspiracies. Similarly (Samuel C. Woolley 2020) perform a qualitative analysis to study dissemination of islamophobia on Gab and note that around 27% of the 12k posts they compile contain islamophobic references (keywords). Works on AS like (Chandra et al. 2021) that study 3,102 and 3,509 posts on Twitter and Gab respectively, observe that the majority posts they analysed contained AS. They also lament the lack of "rigorous" works studying the phenomenon.

3 Lexica Methodology & Dataset

Constructing Antisemitism and Islamophobia Lexica Our methodology mainly relies on semantic similarity approaches to curate specialized hate speech (AS and IP) lexica. Previous work (Zannettou et al. 2020) details the coded or nuanced nature of hate speech references lend credence for reliance on semantic approaches. To this end, we formulate a detailed 3-step approach to curate our two lexica, one each for AS and IP. In detail our methodology is described in the following three steps:

Step 1: We utilize a word similarity and word analogy based paradigm to populate our lexica. For this approach, we make use of Concept Net (Speer, Chin, and Havasi 2017), a general knowledge graph that connects words and phrases of natural language (terms) with labeled, weighted edges to other terms. As noted by (Speer and Lowry-Duda 2017), instead of simply being a large gazetteer of 'named entities Concept Net aims to represent the general and known relationships between frequently-used words and phrases. We lend further credence to our choice by highlighting that Concept Net is a freely-available, open access semantic network that showcases state-of-the-art results for word similarity and analogy tasks (Speer and Lowry-Duda 2017). In

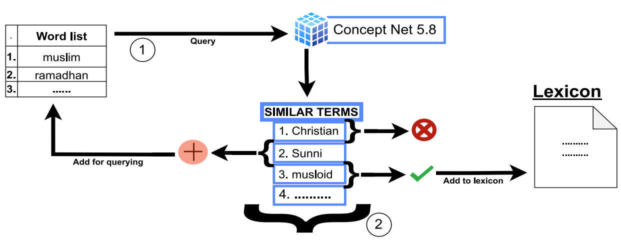


Figure 1: Concept Net methodology **(Step 1)** *Figure best viewed in color and when zoomed.*

Word	Similarity	Word	Similarity
muzzie	0.8239	rapefugee	0.7087
mudslime	0.8097	shitskin	0.6992
arab	0.7778	somalian	0.6987
somali	0.7550	sandnigger	0.6907
islamic	0.7106	sunni	0.6801

Table 1: Top ten words that are similar in their cosine similarities with the term "muslim" and their respective cosine similarities in /pol/ dataset (July 2016 - October 2019)

addition to its performance and open accessibility, we note the sheer volume of "diverse" matrices of embeddings it utilises, precisely, the word2vec embeddings trained on 100 billion words of Google News using skip-grams with negative sampling (Mikolov et al. 2013) and the GloVe 1.2 embeddings trained on 840 billion words of the Common Crawl (Pennington, Socher, and Manning 2014). This reflects the point about it being able to capture how terms relate to one-another over the Internet. We use Concept Net 5.8 via its online web interface[1] between April 2021 and June 2021. Initially, we source a set of general *seed terms* to query into Concept Net that reference terms related to the phenomena we study i.e., general terms about Judaism/Jews and Muslims/Islam available on the Internet. For this we collect, 33 general terms (like Aliyah, Bar mitzvah, etc.) which are used about or related to, Jews/Judaism from (RMC 2019) and 18 terms (like Allah, Salah, etc.) related to Muslims/Islam from (Luckie 2011). Thereon, we use Figure 1 to showcase how we use Concept Net in our work with an example. For each phenomenon under question (IP or AS), a word list consisting of the above mentioned seed terms is queried into Concept Net. For each queried term a list of *"Similar Terms"* are returned by Concept Net. These similar terms are in fact terms that are, (a) semantically related or (b) the derived form of the queried term that Concept Net returns to us. In Stage 2, the similar terms are checked and subsequently either rejected, or if found relevant to the phenomenon under question (i.e., related to Islam/Muslims or Jews/Judaism) they are iteratively added to the "Word list" and used for further querying. In case a term is found to be Antisemitic or Islamophobic it is inserted into the respective Lexicon (shown as Lexicon in Figure). This process is exhibited in the Stage 2 of the Figure 1. Wherein from the 'Similar Terms' returned, the word *"sunni"* is marked using manual assessment as being relevant to the phenomenon being studied and hence added to the list of words to be queried further, also the word *"musloid"* that we consider to be an Islamophobic pejorative is added to the IP lexicon. Ultimately, we discover 40 terms for IP lexicon and 16 for AS lexicon using Concept Net in the first step of our curation methodology.

Step 2. In this step we aim to robustify our lexica against coded language and niche referencing of minority groups in hateful contexts which seems to be a feature of /pol/ (Zannettou et al. 2020). We do this by incorporating terms and techniques from previous research work. Specifically for each lexicon, we proceed with the following approach: **AS:** For robustifying our AS lexicon we make use of the cosine similarity results presented by (Zannettou et al. 2020) in the section related to "Text Analysis" wherein they train a word2vec model on /pol/ data collected by them between July 2016 to January 2018. As model parameters they choose the context window size of 7 and consider only words that appear at least 500 times in the dataset. From their results about the top 10 most similar words to the word 'jew', we choose 6 terms and add them to our list. These 6 terms are: (((jew))) , kike , goyim , jewri , juden, heeb. Furthermore, we cross-reference these terms with Hatebase and discovered that all these terms exist on the platform in some form or another thus lending further credence to this curation approach. **IP:** Similar to above, we aim to discover coded IP terms on /pol/ from previous works, but as maybe noted from our Related Work (section), to the best of our knowledge we note that no such previous work exists. Therefore, we ourselves undertake the same approach as that of (Zannettou et al. 2020) as described above, to find coded islamophobic terms. Subsequently, we train a word2vec model (Mikolov et al. 2013) to generate word vectors for words from a publicly available /pol/ dataset (Papasavva et al. 2020) which contains all posts from /pol/ from July 2016 to October 2019. Our results are shown in Table 1. Among these 10 words we choose 5 words: muzzie, mudslime, rapefugee, shitskin, sandnigger our selection is again based on the knowledge of the domain. Similar to the step related to AS (above), we cross reference these 5 terms with Hatebase and note that all the terms appear on the platform.

Step 3. We further strengthen our lexica for AS and IP by scraping the available list of AS and IP terms found on Hatebase between April 2021 to June 2021. We limit our selection of hateful terms to english terms. This results in selection of 90 terms for IP lexicon and 26 for AS lexicon. Our final lexica include 48 terms for AS and 135 terms in our IP lexicon.[2] We show an overview of all the lexica curation steps in Figure 2.

Dataset. To study the extent and behavior of the usage of Antisemitic and Islamophobic terms on /pol/ we use a dataset of all /pol/ posts between July 2016 to February

[1] https://conceptnet.io

[2] The final lexica can be found anonymously via https://github.com/schetudiante/asip/

Lexicon	Step 1	Step 2	Step 3	Total
AS	16	6	26	**48**
IP	40	5	90	**135**

Table 2: Overview of the composition of our 2 lexica after each step of our curation.

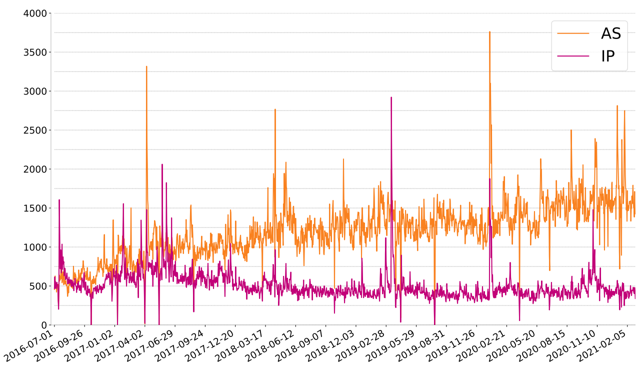


Figure 2: Frequency of posts for each day in our datasets.

	#posts	%posts containing AS/IP
Antisemitism	1,995	93.83%
Islamophobia	4,238	81.18%

Table 3: Results from manual annotation showing the percentage of posts that contain AS/IP.

2021 which include $204,097,331$ posts. We acquire this data from the authors of (Papasavva et al. 2020). We then parse each post from this dataset into 2 large-scale datasets each for AS and IP. We do this by searching for terms from our final lexica. Specifically, before searching the posts we first pre-process the content of each post by removing numbers, HTML tags, links, user handles, 4chan specific quote markings and Unicode. Hereon, we refer to these 2 parsed datasets as AS dataset and IP dataset. Throughout our study we utilize these datasets for the analyses we make. In terms of the volume of posts, the AS dataset contains $2,015,753$ posts and the IP dataset contains $849,252$ posts. In Figure 2 we show the frequency of daily posts for both our datasets. We observe that AS posts for each day generally outnumber the IP posts on /pol/. This result is not surprising and validates previous works like (Zannettou et al. 2020) that note the overwhelming Antisemitic character of /pol/.

4 Results

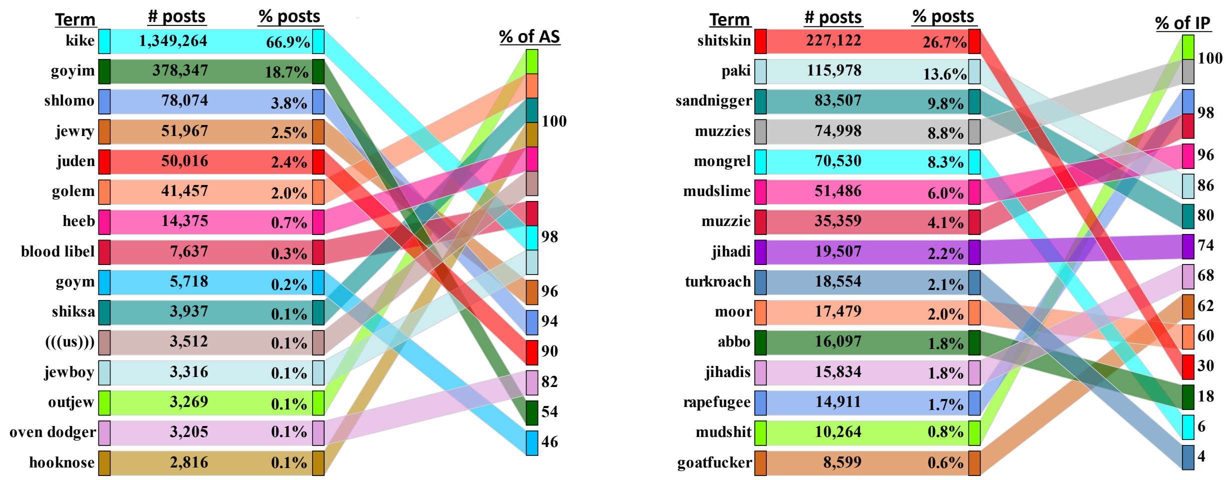
Term Effectiveness

In this section, we attempt to answer **RQ1** which seeks to explore: *"How accurate are the terms from our curated lexica to analyse AS and IP hate speech on /pol/?"*. We resolve the notion of accuracy of a term (and by extension all terms in our lexica) by the following two factors: 1) *Frequency* that refers to how many posts reference a specific AS/IP term and 2) *Veracity* that refers to how many posts that include AS/IP terms are indeed sharing Antisemitic/Islamophobic hate speech.

For the former, we simply count the number of posts that reference each term in our lexica. For the latter, we use human evaluation, specifically, for each of the 48 AS terms and 135 IP terms in our respective lexica, we choose a random sample of 50 posts for each term. One of the authors of this work then manually annotates each post to assess if the post is sharing AS or IP. Table 3 reports the number of annotated posts for AS and IP, as well as the percentage of posts that are considered as sharing Antisemitic/Islamophobic hate speech. We can observe that 93.83% of posts that contained terms from our AS lexicon are annotated as Antisemitic, while at the same time around 81% posts featuring IP terms are annotated as Islamophobic.

When establishing veracity, we focus on the results of manual annotation for each specific term. We show the results from our manual annotation specific to veracity of each term in Figure 3, as we are constrained by space limitations, here, we only display the 15 most mentioned terms that are sorted by the number of posts that mention them. We observe in Figure 3 that there is an overwhelming discrepancy in the usage of terms in the datasets. In AS lexicon, the term "kike" which occurs as antisemitic in 98% of the annotated posts is used in almost 67% of total posts in our AS dataset, this fact alone establishes that this term is an overwhelmingly "preferred" AS term (among the terms from our AS lexicon) on /pol/. Similarly, it maybe observed from the results in Figure 3b that the most commonly occurring term from the IP lexicon, "shitskin" that accounts for 26.7% of total IP posts occurs as Islamophobic in only 30% of posts in our manual annotation however, we impress upon the "anecdotal evidence" from our annotations that suggests that this term is often deployed in a "generic" racist hate speech against diverse groupings of ethnically "non-white" communities if not specifically for IP. Here, we provide two examples (among a much larger set of available posts), as: (A) *"no doubt a nigger, they would have stated it was a white guy if possible. Any time race is left out you know it is a shitskin"*. (B) *"that is rich coming from a shitskin pikey"*.

We note a similar behavior in posts referencing the terms: "mongrel" , "turkroach", "moor","abbo" & "mudshit". We also observe that in AS dataset the percentage of posts that contain at least one term from the 15 highest occurring AS terms is 99% while the same distribution of posts in IP dataset that refers to top 15 IP terms is a bit more diverse and stands around 91.8%. Finally, we note that the results from this section with regard to **RQ1** do reflect that our curated lexica are "effective" when it comes to capturing AS and IP posts. We also gain a semblance that among the terms from our lexica, users on the platform overwhelmingly prefer specific AS and IP terms more than others and that there are certain terms that are also deployed in general racist hate speech.



(a) Top 15 AS terms

(b) Top 15 IP terms

Figure 3: Usage of individual AS or IP terms in our datasets. We report the number and percentage of posts that include each term in /pol/. We also report the percentage of the posts that are actually considered Antisemitic or Islamophobic (based on our manual annotations on 50 randomly selected posts for each term). *Figure best viewed in color*.

word	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
kike	jews, kike, jew, god, jesus	hitler, kike, germany, german, did	israel, kike, iran, israeli, jews	trump, puppet, kike, israel, biden	kike, nose, like, looks, fucking
goyim	white, whites, goyim, jews, race	kikes, kike, goyim, just, fucking	christianity, jews, jesus, god, christians	shut, know, goyim, oy, vey	jews, goyim, information, people, like
shlomo	kike, fuck, shlomo, kikes, oven	white, whites, shlomo, race, people	shlomo, pls, time, stop, coming	try, nice, shlomo, harder, digits	hitler, gas, holocaust, shlomo, jews
jewry	christianity, jews, jesus, god, jewry	trump, jewry, just, like, jews	hitler, germany, jewry, poland, international	jewry, level, anon, like, just	white, whites, jewry, people, jews
juden	peterstein, juden, peterson, like, just	die, und, das, der, den	juden, kys, rat, turbo, bantz	extermination, jews, million, ich, juden	white, whites, niggers, juden, race

(a) AS dataset

word	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
shitskin	white, whites, shitskin, things, people	nigger, niggers, shitskin, fuck, fucking	jews, israel, jew, shitskin, jewish	shitskin, ahahahaha, know,like, lol	muslim, islam, muslims, shitskin, religion
paki	paki, based, hes, like, accent	paki, looks, like, fucking, girl	nigger, niggers, paki, kike, chink	paki, stfu, spotted, pak, intellectuals	shop, food, paki, shops, kebab
sandnigger	islam, sandnigger, muslim,muslims, kys	nigger, niggers, sandnigger, kike, spic	sandnigger, detected, butthurt, seething, crying	jews, jew, sandnigger, jewish, just	fuck, sandnigger, country, shut, cares
muzzies	christians, muzzies, christianity, christian, god	muzzies, uk, london, british, brits	jews, muzzies, jew, hate, jewish	india, sikhs, muzzies, hindus, pakis	israel, jews, muzzies, israeli, palestine
mongrel	jews, jew, mongrel, jewish, israel	poland,germany, german,germans,mongrel	mongrel, fucking, fuck, illiterate, read	arab, mongrel, arabs, iran, iranian	greek, greeks, turks, mongrel,turk

(b) IP dataset

Table 4: Top five topical clusters for posts regarding five most mentioned AS and IP terms on /pol/. Colors indicate meta-labels for clusters.

Contextual Analysis

In this section, we attempt to answer **RQ2** which seeks to explore: *What context(s) are the frequently mentioned AS and IP terms used in? Are there observable patterns/similarities/differences?*. First we explain our methods and processes by which we perform this analysis.

Topic modelling methodology. In order to analyze the context associated with the most frequently used AS and IP terms, we use a Bidirectional Encoder Representations from Transformers (BERT) based topic modeling method by (Grootendorst 2020) (herewith, referred to as Topic Modeler). We make this choice based on its competitive performance on benchmarks, 'Topic Coherence' and 'Topic Diversity' (Grootendorst 2022) as we believe that these are important to us given that our datasets contain 'niche' language and diversely varied length of posts. To gain better results we follow the advice of the authors of (Grootendorst 2022) who suggest that leveraging state-of-the-art language (embedding) can increase the performance. We utilize all available language embedding models made available by (Reimers and Gurevych 2019) in tandem with our Topic Modeler. Specifically, we test all the pre-trained embedding models with competitive performance on semantic similarity tasks from (Reimers and Gurevych 2019) that are openly available[3] to use. Among all the embedding models we tested on our dataset, we select 'all-mpnet-v1' based on the above mentioned benchmarks as evaluated using measures from (Řehůřek and Sojka 2010) in comparison to other models and the 'quality' of topics yielded that were evaluated by human evaluation. After furnishing the embedding(s) for our posts using our Topic Modeler, to reduce the dimensionality of the high dimensional embedding space, the Uniform Manifold Approximation and

[3]https://www.sbert.net/

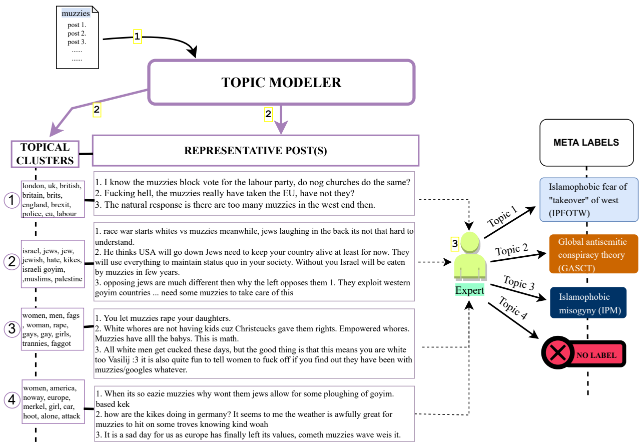


Figure 4: Methodology for assigning meta labels

Projection (UMAP) approach (McInnes and Healy 2018) is utilised. Then, our Topic modeler uses the HDBSCAN algorithm (McInnes, Healy, and Astels 2017) to cluster the embeddings. Our Topic modeler yields: (a) Topical clusters consisting of terms (or topics) that occur consistently in posts that the model considers to be similar using a class based TF-IDF approach (Grootendorst 2022). Each topical cluster shows terms per cluster that can be considered as the description of the topics within a cluster according to (Grootendorst 2020). (b) Representative posts for each topical cluster which are the "exemplar" data points that the clustering algorithm (HDBSCAN) considers to be most representative of a cluster (McInnes, Healy, and Astels 2017). HDBSCAN "considers" a post(s) an "exemplar(s)" by focusing on places in a cluster where it is most dense. Our Topic modeler utilises the approach by (Grootendorst 2020) for extracting representative posts for each topical cluster. We then manually observe each topical cluster and its respective representative post(s) and assign them meta-labels if observable themes are found in the representative post(s) of each topical cluster.

Meta-label assigning process: We further explain our approach of assigning meta-labels using an example in Figure 4 where each stage of the process is illustrated by a stage number that is shown in a yellow box in the figure. In this example, our Topic modeler takes all the posts mentioning the word 'muzzie' in stage 1. After ingesting the posts, the Topic modeler uses the above mentioned methodology to model posts into topical clusters and their respective representative posts in stage 2. Subsequently, in stage 3, a human evaluator examines the content of the topical clusters and their respective representative posts and decides if a label can be assigned to a particular topical cluster(s) based on the consistency of the "theme" (subject) being exhibited in the representative posts.

Based on the methodology described in Figure 4 we perform topical clustering on posts that contain five most mentioned AS and IP terms individually. We show our results in Tables 4a & 4b where each row depicts results from topic modeling for posts containing individual terms and columns such as Topic 1, Topic 2 up to Topic 5 refer to top five 'topical clusters' respectively. The topical clusters are assigned topic numbers based on how many posts are part of that particular cluster. Example, Topic 1 is a topical cluster composed of highest number of posts followed by Topic 2 and so on. Each cell in the column of a topical cluster shows the "most important terms" belonging to that respective topical cluster. The colors of the topical cluster (cells) indicates that using manual assessment a meta-label for the cluster was assigned.

In Table 4a we observe the results from topic analysis of five most mentioned AS terms. We see that the topics related to these terms almost always reference names of countries, religious traditions, names of antisemitic archetypes and tropes like hitler, holocaust. Specifically, based on the meta labels we observe that a theme (colored blue) related to term 'kike' signifies the theme of posts which reference 'Donald Trump as a secret jewish person or a jewish agent'. We also observe the theme (colored in green) related to 'kike','goyim' and 'jewry' that often references religious Antisemitic tropes by referring to 'Christianity as a religion created and controlled by jewish people', 'Jesus Christ as being jewish agent'. Similarly, we observe that the terms 'shlomo' and 'juden' feature topics (in yellow) that refer to the subject of holocaust denial. The theme (colored pink) connotes ethno-racial AS tropes such as 'White people / white countries being superior because of their non-Jewish history'. We also note that the terms 'goyim' and 'jewry' are highly spoken in theme (colored in brown) that refers to the 'controlling of non-jewish people (especially political governments/media) across the world by a (supposed) jewish elite'. Additionally, 'juden' also features a topic (in grey color) that contains posts that refer to 'Jordan Peterson' in an antisemitic word play replacing the word Jordan with the word 'Juden' (word for jewish people in German). Overall, it is noticeable that many topical clusters do not feature meta labels. We believe this reflects the diversity of contexts and themes in which a particular term is used in the /pol/ community.

From Table 4b we observe the results from topic analysis of five most mentioned IP terms. We see that the topics related to these terms almost always reference names of religious traditions, countries, islamophobic, antisemitic or racist symbolism (with usage of terms like "nigger", "sandnigger" consistently). Specifically, based on meta labels, we can see that for the term 'shitskin', features multiple topical clusters regarding the theme (orange color) about 'White supremacy and racism against people of color' (RAPOC). There also occurs a topic with theme (cyan color) that references religious IP trope focused on "muslims as a non western and non-white people" (MNWP). Similarly, for the term 'paki' we note that four topics out of five (in orange color) shows themes of RAPOC. This signifies that islamophobic tropes that reference the term 'paki' are often composed of RAPOC. We also note that the topical clusters for the term 'sandnigger' dominantly features the themes about RAPOC. This signifies a preference about the racist usage of this term among the users on /pol/. This term also features themes related to MNWP and a theme (red color) features

Antisemitic tropes together with racial hatred against middle eastern people (arabs). Finally, we note that the term ‘muzzies’ features a topic (in purple color) that signifies “worry about the increasing muslims population in the UK” which is concurs with the increase in anti-muslim hatred in the UK (Townsend 2019).

Finally, we note with regard to RQ2, our results do reflect occurrence of contexts and themes associated with the most mentioned AS and IP terms. However, with exception of the term “paki” we find that this does not occur very prominently in our dataset. As results from the Table(s) 4 indicate that a particular term can also occur in many general contexts in antisemitic or islamophobic conversations on /pol/.

5 Conclusion

In this work we curated lexical terms related to Antisemitism and Islamophobia on /pol/ using a methodology that heavily relies on semantic similarities using Concept Net and word embeddings. Using these terms, we collected a large dataset from 4chan’s /pol/ and we assessed the effectiveness of our lexica and the context of which specific Antisemitic/Islamophobic terms appear on /pol/. Our results to measure how effective are the terms to capture AS and IP signify that in terms of overall accuracy of posts labelled our datasets contain overwhelmingly high AS and IP. We also show that specific terms are ‘preferred’ by users on /pol/ vis-a-vis the amount of mentions. Along with this we also note that certain terms are mentioned in Antisemitic and Islamophobic posts ‘more’ than others. Results from our contextual analysis and meta labelling of topical clusters results in a low level understanding of ‘in what context (theme) is a term used’. We found usage of certain terms tied to specific themes of conversations.

Limitations. In this work we have majorly relied on manual evaluations. Given the limitations of manual annotations based methodologies we cannot claim to have captured the “full” discourse regarding AS and IP on /pol/. Specifically, we acknowledge that future works may undertake our approaches and utilize multiple annotators and qualitatively assess larger samples. This can help in improving the work that aims to understand pejoratives related to AS and IP or other hateful phenomena at a lower level. Also, our analysis focuses on a single fringe Web community, namely, /pol/. Future work should evaluate the use of our lexica on other communities and potentially expand our lexica by using semantic similarity techniques like the ones presented in this work.

References

Austin. 2019. What Is 8chan, and How Is It Related to This Weekend’s Shootings? Here’s What to Know. https://time.com/5644314/8chan-shootings/.

Breland. 2020. 4Chan Trolls Target Scandinavian Airlines With Racist Harassment Campaign. https://www.motherjones.com/politics/2020/02/4chan-trolls-target-scandinavian-airlines-with-racist-harassment-campaign/. Accessed: 2021-03-18.

Chandra, M.; Pailla, D.; Bhatia, H.; Sanchawala, A.; Gupta, M.; Shrivastava, M.; and Kumaraguru, P. 2021. “Subverting the Jewtocracy”: Online Antisemitism Detection Using Multimodal Deep Learning. In *13th ACM Web Science Conference 2021*, WebSci ’21, 148–157. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383301.

Crawford, B.; Keen, F.; and Suarez-Tangil, G. 2021. Memes, Radicalisation, and the Promotion of Violence on Chan Sites. In *ICWSM*.

Evans. 2018. How the MAGAbomber and the Synagogue Shooter Were Likely Radicalized. https://www.bellingcat.com/news/americas/2018/10/31/magabomber-synagogue-shooter-likely-radicalized/.

Grootendorst, M. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.

Grootendorst, M. R. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*, abs/2203.05794.

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. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 92–101.

Luckie. 2011. Glossary of Islamic terms. https://www.washingtonpost.com/local/politics/glossary-of-islamic-terms/2011/06/07/AGwm2oQH_story.html.

McInnes, L.; and Healy, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, abs/1802.03426.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2: 205.

Mikolov, T.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Mittos, A.; Zannettou, S.; Blackburn, J.; and Cristofaro, E. D. 2020. ”And We Will Fight For Our Race!” A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. In *ICWSM*.

Papasavva, A.; Zannettou, S.; Cristofaro, E. D.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. *14th International AAAI Conference On Web And Social Media*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Perry. 2020. Report finds lapses ahead of New Zealand mosque attack. ReportfindslapsesaheadofNewZealandmosqueattack.

Řehůřek, R.; and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

RMC, U. 2019. Factsheet: Glossary of Jewish terms. https://religionmediacentre.org.uk/factsheets/glossary-of-jewish-terms/.

Samuel C. Woolley, N. M., Roya Pakzad. 2020. Incubating Hate:Islamophobia and Gab.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. AAAI'17, 4444–4451. AAAI Press.

Speer, R.; and Lowry-Duda, J. 2017. ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 85–89. Vancouver, Canada: Association for Computational Linguistics.

Townsend. 2019. Rise in UK use of far-right online forums as anti-Muslim hate increases. https://www.theguardian.com/world/2019/mar/16/rise-far-right-online-forums-anti-muslim-hate-wave.

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. IMC '18, 188–202. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356190.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *ICWSM*.

Zelenkauskaite, A.; Toivanen, P.; Huhtamäki, J.; and Valaskivi, K. 2021. Shades of hatred online: 4chan duplicate circulation surge during hybrid media events. *First Monday*, 26.