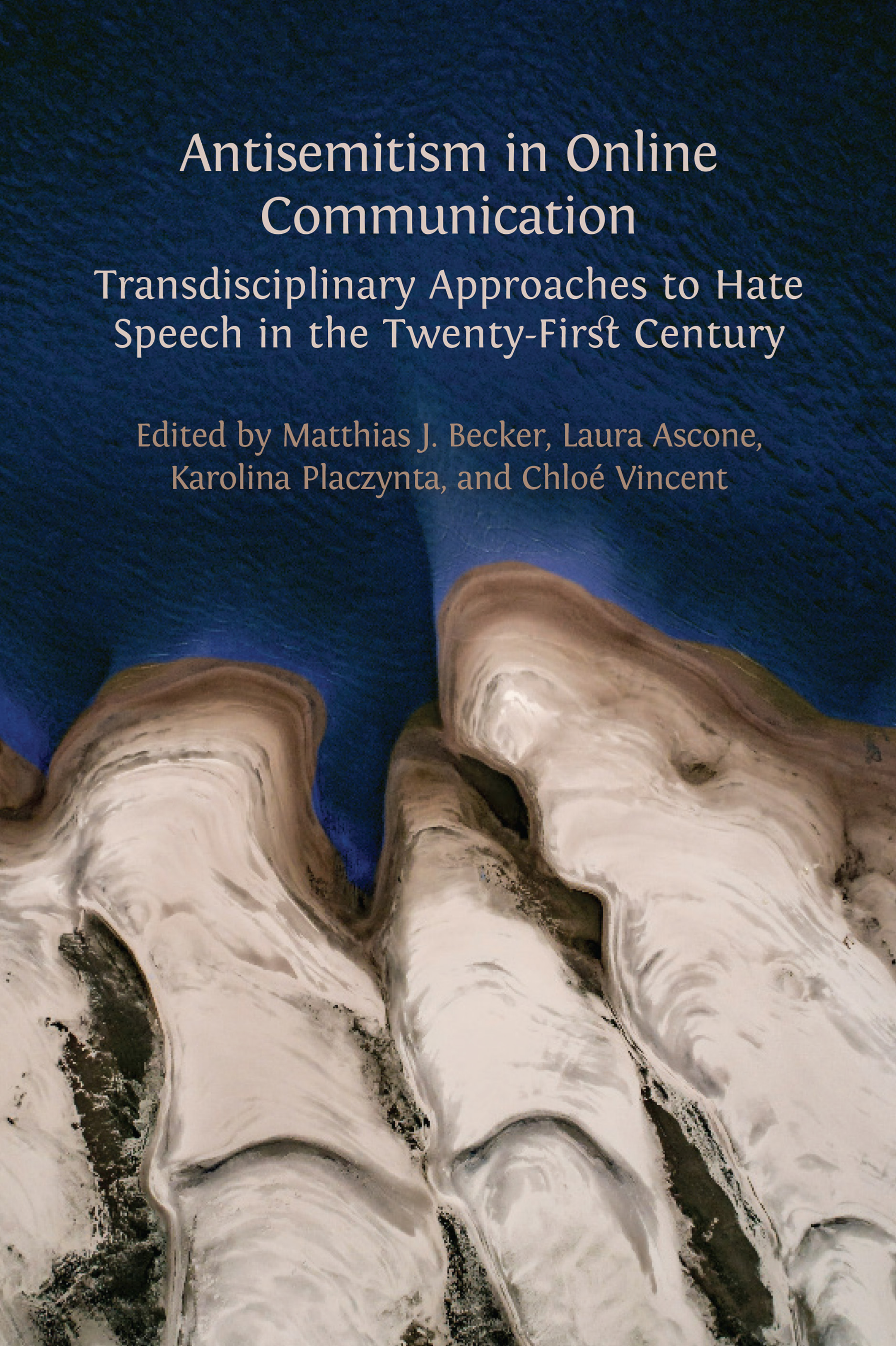
Antisemitism in Online Communication

Transdisciplinary Approaches to Hate Speech in the Twenty-First Century

Edited by Matthias J. Becker, Laura Ascone, Karolina Placzynta, and Chloé Vincent



https://www.openbookpublishers.com

Matthias J. Becker, Laura Ascone, Karolina Placzynta, and Chloé Vincent (eds), *Antisemitism in Online Communication: Transdisciplinary Approaches to Hate Speech for the Twenty-first Century*. Cambridge, UK: Open Book Publishers, 2024, https://doi.org/10.11647/OBP.0406

All external links were active at the time of publication unless otherwise stated and have been archived via the Internet Archive Wayback Machine at https://archive.org/web

Any digital material and resources associated with this volume will be available at
https://doi.org/10.11647/OBP.0406#resources

We acknowledge support by the Open Access Publication Fund of Technische Universität Berlin

ISBN Paperback: 978-1-80511-260-0
ISBN Hardback: 978-1-80511-261-7
ISBN Digital (PDF): 978-1-80511-262-4
ISBN Digital eBook (EPUB): 978-1-80511-263-1
ISBN HTML: 978-1-80511-265-5

DOI: 10.11647/OBP.0396

Cover image: Photo by Marc Bloch, 2023, CC-BY. Cover design: Jeevanjot Kaur Nagpal.

8. Algorithms Against Antisemitism?

Towards The Automated Detection of Antisemitic Content Online

Elisabeth Steffen, Milena Pustet, Helena Mihaljević

The proliferation of hateful and violent speech in online media underscores the need for technological support to combat such discourse, create safer and more inclusive online environments, support content moderation and study political-discourse dynamics online. Automated detection of antisemitic content has been little explored compared to other forms of hate-speech.

This chapter examines the automated detection of antisemitic speech in online and social media using a corpus of online comments sourced from various online and social media platforms. The corpus spans a three-year period and encompasses diverse discourse events that were deemed likely to provoke antisemitic reactions. We adopt two approaches. First, we explore the efficacy of Perspective API, a popular content-moderation tool that rates texts in terms of, e.g., toxicity or identity-related attacks, in scoring antisemitic content as toxic. We find that the tool rates a high proportion of antisemitic texts with very low toxicity scores, indicating a potential blind spot for such content. Additionally, Perspective API demonstrates a keyword bias towards words

https://doi.org/10.11647/OBP.0406.08

related to Jewish identities, which could result in texts being falsely flagged and removed from platforms.

Second, we fine-tune deep learning models to detect antisemitic texts. We show that OpenAI's GPT-3.5 can be fine-tuned to effectively detect antisemitic speech in our corpus and beyond, with F1 scores above 0.7. We discuss current achievements in this area and point out directions for future work, such as the utilisation of prompt-based models.

1. Introduction

In the third quarter of 2022, the US technology giant Meta reported that it had taken action on 10.6 million pieces of *Facebook* content considered to be hate speech. Of these posts, over 90% were found and acted on proactively, that is, prior to users reporting them (Meta 2022). Given the sheer volume of content published on social media, automatic detection of hate speech and other offensive content has become a key task for mainstream social media platforms. Similar challenges arise in the research based on empirical data and in the monitoring work of NGOs or journalists who analyse political discourses.

The technical foundation of this task is text classification, which is the process of automatically assigning categories (or classes) to a text. In the realm of political online communication, examples of such categories include various forms of hate speech, devaluation and exclusion related, for example, to misogyny, racism and antisemitism. Historically, individually formulated rules targeting particular textual aspects were used to perform text classification; however, modern approaches leverage machine learning and deep learning for superior results. This entails feeding large datasets into, for example, deep neural networks from which they learn patterns in the texts that allow them to more accurately predict classes for new, unseen data.

So far, classification of texts is usually done in a supervised manner, whereby an algorithm is trained using human-labelled data to make accurate predictions. The human annotations serve as a 'gold standard' and are used to 'teach' the algorithm. Labelled examples are also utilised to evaluate the learned model's predictions based on standard metrics. Often, so-called benchmark datasets are used to compare the performance of different machine learning models for a specific task on a common set of data, using task-specific metrics. Efforts to generate benchmark datasets for the automated detection of antisemitism have been conducted so far by only a handful of researchers (Chandra et al. 2021, Jikeli et al. 2022, Steffen et al. 2022, Jikeli et al. 2023), and have not yet resulted in datasets comparable to available scientific corpora for related phenomena, such as offensive language, toxic language and other forms of hate speech.

For the recognition of broader linguistic phenomena intersecting with antisemitism, such as hate speech and toxic language, openly accessible production-ready web services have been established. A prominent example is Perspective API, a free service created by Jigsaw and Google's Counter Abuse Technology team, which is widely applied for content moderation and research. For example, it has been used for analyses of moderation measures on *Reddit* (Horta Ribeiro et al. 2021), for investigations of political online communities on *Reddit* (Rajadesingan, Resnick and Budak 2020) and *Telegram* (Hoseini et al. 2021) and for identifying antisemitic and Islamophobic texts on *4chan* (González-Pizarro and Zannettou 2022). The service allows for the detection of abusive content by providing scores (between 0 and 1) for different attributes such as toxicity, insult or identity attack. The definition of what constitutes (severely) toxic or identity attacking comments in Perspective API suggests that antisemitic speech should be detectable through the service, thus offering an easily accessible approach to recognising certain forms of antisemitic speech. However, recent work on German-language communication on *Telegram* and *Twitter* (now *X*) indicates an oversensitivity to identity-related keywords such as 'jew' or 'israel', which makes the service prone to falsely classifying texts as antisemitic simply for addressing Jewishness or mentioning Israel (Mihaljević and Steffen 2022). It has been found, furthermore, that the service performs rather poorly on more subtle or encoded forms of antisemitism, often failing to recognise them as toxic (ibid.).

In this chapter, we evaluate Perspective API on a multilingual dataset comprising more than 55,000 comments from online platforms that were manually annotated by experts working on the international project Decoding Antisemitism. In our experiments, the service shows a bias towards identity-related keywords and tends to penalise expressions of counter speech. We therefore argue that the Perspective API is only of very limited use for tackling antisemitism online and is likely to produce a high number of false positives when applied in contexts with a frequent occurrence of counter speech.

With the advancement of machine learning, particularly deep learning, non-profit anti-hate organizations have expanded their focus to include large-scale analyses of online content that often entail the development of machine learning-based text classifiers. Several organisations have reported successfully establishing models for the detection of antisemitic speech. For instance, the Anti-Defamation League (ADL) has developed a model for detecting antisemitic speech across various social media platforms as part of their Online Hate Index (OHI).[1] The tool is being developed by experts in antisemitism and volunteers from the targeted community. The Institute for Strategic Dialogue (ISD) has also conducted various analyses of large social media datasets requiring automated detection of antisemitic content,[2] while Fighting Online Antisemitism (FOA) reports to have begun using an antisemitism detection model recently developed through collaboration with Code for Israel, a tech-for-good volunteer organisation, and an Israeli tech company.[3] However, these tools, while presumably offering superior effectiveness in detecting antisemitic speech compared to the generalistic Perspective API, are not readily accessible to the broader research community and are primarily utilised within the respective organisations for research and monitoring purposes. This limitation makes it challenging to employ them for custom analyses or to evaluate their performance on other datasets. For instance, the antisemitism classifier for German-language *YouTube* comments developed by the ISD and the Centre for Analysis of Social Media (CASM) involves filtering the corpus by keywords related to Judaism, Jewish people or the state of Israel, as well as other keywords derived from previously developed

[1] The Anti-Defamation League, 2022. "How Platforms Rate on Hate: Measuring Antisemitism and Adequacy of Enforcement Across Reddit and Twitter", https://www.adl.org/sites/default/files/pdfs/2022-05/How%20Platforms%20Rate%20on%20Hate%202022_OHI_V10.pdf

[2] Institute for Strategic Dialogue, 2020. "Das Online-Ökosystem Rechtsextremer Akteure", https://www.isdglobal.org/isd-publications/das-online-okosystem-rechtsextremer-akteure/ and "Mapping hate in France: A panoramic view of online discourse", https://www.isdglobal.org/isd-publications/mapping-hate-in-france-a-panoramic-view-of-online-discourse-2/

[3] The Jerusalem Post, 2023. "Israeli tech warriors code a solution to fight online antisemitism", https://www.jpost.com/diaspora/antisemitism/article-749349

classifiers.[4] These restrictions result in a higher proportion of relevant content and enable the labelling of a sufficient number of texts from all classes, particularly antisemitic ones, within a reasonable timeframe. However, classifiers trained on such datasets, which are more balanced regarding the class distribution, may not generalise well to more realistic corpora representing discourses that were not pre-filtered.

We thus trained custom classification models using the corpus of the Decoding Antisemitism project. The dataset comprises online comments in English, German and French from various sources such as news portals, *Twitter* or *Facebook*, annotated regarding a plethora of additional attributes, including rhetoric and linguistic aspects of antisemitic speech. Our experiments are focused solely on English-language data and aim to distinguish between antisemitic and non-antisemitic posts. The results demonstrate that effective models can be trained even in the more challenging scenario of a corpus that has not been pre-filtered by selected keywords related to Jewishness or Israel, where implicit expressions of antisemitism are frequent and the class of antisemitic posts is significantly underrepresented. We show that fine-tuning an openly available BERT-like model achieves satisfactory results on test data but is significantly outperformed by a fine-tuned GPT-3.5 model not only on the test data but also in discourse and domain transfer. We discuss the practical implications of these findings, potential future directions and plans for research using prompt-based approaches.

2. Dataset

The team of the project Decoding Antisemitism has annotated online comments in English, French and German from various leading media sources, including a range of news portals and social media platforms such as *Twitter* or *Facebook*, using a self-developed code schema based on the IHRA definition.[5] The resulting corpus spans a three-year period

[4] Institute for Strategic Dialogue, 2020. "Using a German-language classifier to detect antisemitism on YouTube", https://www.isdglobal.org/digital_dispatches/using-a-german-language-classifier-to-detect-antisemitism-on-youtube-background-and-methodology/

[5] The International Holocaust Remembrance Alliance (IHRA), 2024. "Working definition of antisemitism", https://holocaustremembrance.com/resources/working-definition-antisemitism

and encompasses diverse discourses that were deemed likely to provoke antisemitic reactions. The focus on mainstream political milieus while dispensing keyword filters in the corpus creation yields a broad set of covered topics as well as represented antisemitic narratives, often expressed in a rather implicit way resorting to puns, allusions or irony.

The ideation level is annotated on a comment-by-comment basis, and it comprises the classes 'not antisemitic', 'counter speech', 'antisemitic', 'contextually antisemitic', 'confirmation of antisemitism' and 'unclear ideation'. The scheme contains a plethora of additional codes; some of these are applied at the level of entire comments, while others refer to specific segments within the text in order to describe, for example, the conceptual or linguistic layer of the antisemitic statement. It should be noted that comments responding to a post or news article are organised in a tree-like manner (depending on the platform) as users can respond directly to either preceding comments or the original posts (see Chapter 7).

Nevertheless, for the experiments presented in this chapter, we consider the texts as independent units and restrict modelling to those comments that could clearly be labelled as antisemitic ('AS') or not antisemitic ('not AS') on the level of ideation. In particular, this excludes texts labelled as contextually antisemitic, wherein antisemitic content cannot be detected without further information such as the content behind a linked URL, information from the article itself, previous comments or the reader's world knowledge. For instance, the comment 'I think you have been told to do this' might be antisemitic when taking into account previous comments that make clear to what 'this' and 'you' refer. While a human annotator (or a content moderator) can usually fully resolve such ambiguities—marking this case 'AS' if the user claims that a previous commenter is expressing themselves in a certain way due to an imagined Jewish influence—this poses a significant challenge when attempting to automate the task in practice. However, as a machine learning model or a service like Perspective API would need this information in order to make a correct inference, we proceed with the described setting only.

We ran the experiments with Perspective API on a part of the multilingual data from the Decoding Antisemitism project—a subset consisting of around 3,500 comments manually labelled as antisemitic and around 53,500 texts labelled as not antisemitic. Our custom models for the detection of texts labelled as antisemitic were trained on the English sub-corpus comprising around 23,000 examples for model training and evaluation.

3. Antisemitism and toxicity: potentials and limitations of Perspective API

Currently, Perspective API provides scores for six attributes of textual content: *toxicity*, *severe toxicity*, *threat*, *insult*, *identity attack* and *profanity*. The most relevant of these for our study, because they are defined in a way that suggests they are capable of detecting certain forms of antisemitic speech, are *toxicity*, *severe toxicity* and *identity attack*. Content is designated *toxic* if it is considered "rude, disrespectful, or unreasonable [...], likely to make people leave a discussion", while the related attribute of *severe toxicity* is supposed to be "much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words". *Identity attack* refers to "negative or hateful comments targeting someone because of their identity" (Thain, Dixon and Wulczyn 2017; Google 2022).

Perspective API scores are computed by machine learning models (Lees et al. 2022) trained on crowd-labelled data. The underlying strategy is to create large sets of (diversely) labelled data by using simple definitions that can be understood and applied by non-experts. To counteract the subjectivity and vagueness of the definition, texts are annotated by multiple individuals and their assessments are aggregated before they are used to train the models.

We evaluated the scores for the attributes *identity attack*, *toxicity* and *severe toxicity*. Specifically, we looked at how many texts labelled as antisemitic by the human annotators were scored above 0.5 by the service, and investigated if certain keywords affected the API's performance.

3.1. Perspective API often scores antisemitic texts as little *toxic*

The distributions of all three attribute scores differ significantly between the two groups of antisemitic and non-antisemitic texts, as identified by the human annotators, with clearly higher scores for antisemitic texts (see Figure 8.1). However, 75% of antisemitic texts were scored with respect to *toxicity* or *severe toxicity* below 0.5, which is a typical threshold for assigning texts to one of two groups. This means that a high proportion of antisemitic texts would not be considered as toxic based on the assessment through Perspective API. Considering that the service currently recommends using 0.7 as a threshold, and that various existing studies even chose a threshold of 0.8, this would mean an even larger number of false negatives. The scores for the group of antisemitic comments are highest with regard to *identity attack*. However, even here, around 70% of antisemitic comments fall below 0.7 and would have been missed if one was to follow the official recommendation.

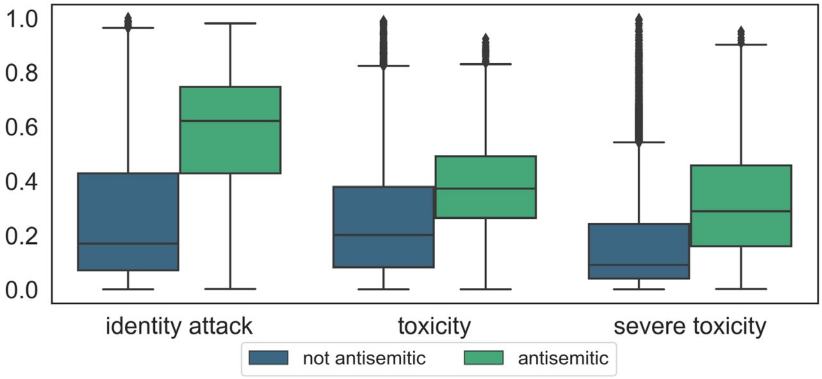


Figure 8.1: Distributions of scores for *identity attack*, *toxicity*, and *severe toxicity*, split according to the antisemitic/not antisemitic data labels. The horizontal lines of the boxes indicate the lower quartile (25%), the median (50%) and the upper quartile (75%) of the scores.

The higher scores for *identity attack* are not surprising, given the fact that antisemitism is an identity-related form of hate which involves prejudice and discrimination against Jewish people based on their perceived identity as a group. However, the high scores for this attribute might also indicate that the service is overly sensitive to certain identity-related keywords such as 'Jew(ish)' or 'Israel'. This 'false positive bias'—the system's tendency to overestimate the level of toxicity if 'minorities' are mentioned regardless of the stance expressed towards them—has been discussed by the developers of the API (Dixon et al. 2018) and confirmed by other research (Hutchinson et al. 2020, Röttger et al. 2021).

3.2. Texts containing identity-related keywords get higher scores

To explore the potential effect of identity-related keywords on *identity attack* scores, we tagged all texts that contained some variations of the keywords 'jew' and 'israel', depending on the corpus language. Figure 8.2 visualises how the scores are distributed when taking this additional information into account: comments containing identity-related keywords (green dots) tend to have higher *identity attack* scores, and this holds for the texts labelled as both antisemitic and not antisemitic. This suggests that texts with references to Jews, Jewishness or Israel, even if they do not express antisemitism, are likely to be flagged as an identity attack. Although the presence of respective keywords alone does not account for a high *identity attack* score[6] (see, for example, the first column in Table 8.1), it still shows a high positive correlation. More precisely, the median identity attack score for comments labelled as not antisemitic is 0.43 higher if the text contains one of the identity-related keywords. For antisemitic texts, the difference is less pronounced (0.15). Similar effects can be observed for the other two Perspective API attributes.

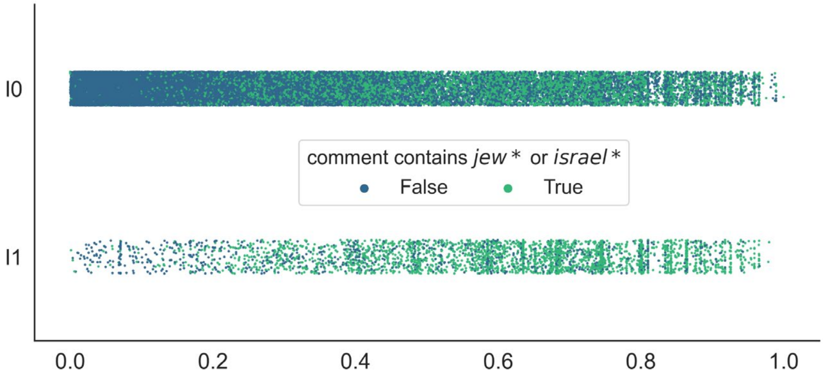


Figure 8.2: *Identity attack* scores broken down by text label and presence of identity-related keywords.

[6] This is also not to be expected as the Perspective API models utilise far more information from text than the frequencies of certain words.

	median *identity attack* score		
	texts without identity-related keywords	texts with identity-related keywords	difference
texts labelled as not antisemitic	0.15 (N=45,761)	0.59 (N=7,769)	**+0.43**
texts labelled as antisemitic	0.49 (N=969)	0.64 (N=2,522)	**+0.15**
difference	**+0.34**	**+0.07**	

Table 8.1: Median *identity attack* scores per class label and presence of identity-related keywords, rounded to 2 decimal places. Group sizes are indicated in brackets. All four differences are statistically significant (Mann-Whitney-U test, p<0.01).

This analysis does not provide a causal relation between the occurrence of keywords related to Jewishness and the state of Israel and higher scores. However, prior research for German language data has shown that adding these keywords significantly increases the scores of texts (Mihaljević and Steffen 2022), which confirms the keyword bias. Other research indicates a similar identity-related keyword bias, showing that texts using standard group labels are assigned even higher scores compared to texts using slurs for referring to different communities (Mendelsohn et al. 2023).

3.3. A (partially) shared vocabulary: varying degrees of intersection between antisemitic texts, identity attacks, and toxic statements

To further investigate the relation between the phenomenon of antisemitism and the attributes *identity attack* and *toxicity*, we determined the 100 most significant terms using a chi-squared test in the English-language sub-corpus for the following categories:

1. texts labelled as antisemitic (or not)
2. texts with an *identity attack* score above 0.7 (or below 0.3)
3. texts with a *toxicity* score above 0.7 (or below 0.3)

The word cloud in Figure 8.3 reveals that words associated with Jewishness and Israel hold considerable significance in texts that were manually labelled by experts as antisemitic. Terms related to Palestinian identity are also prominent. Negatively connotated terms such as ‘apartheid’, ‘terrorist’, ‘cleansing’, ‘occupy’ or ‘force’ likely stem from text passages containing accusations against and demonisations of Israel within the context of the Arab-Israeli conflict.

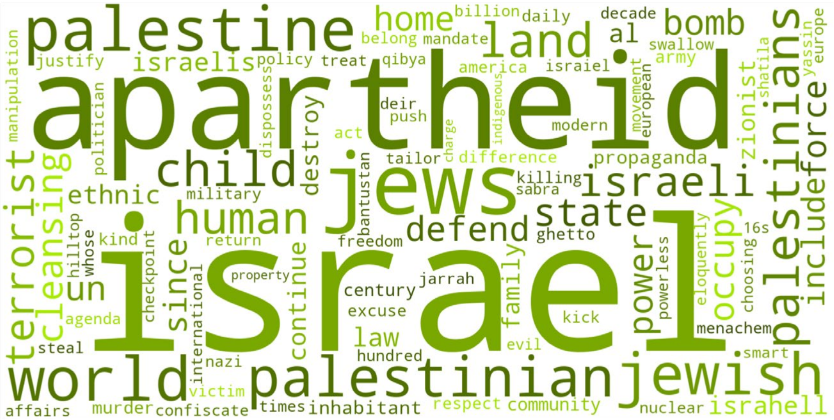


Figure 8.3: 100 most significant terms in texts manually labelled as antisemitic

There is a noticeable overlap between these terms and those found significantly often in text with high *identity attack* scores, particularly in relation to Jewishness. Interestingly, the significance of references to Palestinian identity is considerably reduced in this context. Instead, we observe a strong presence of terms relating to Muslim identities.

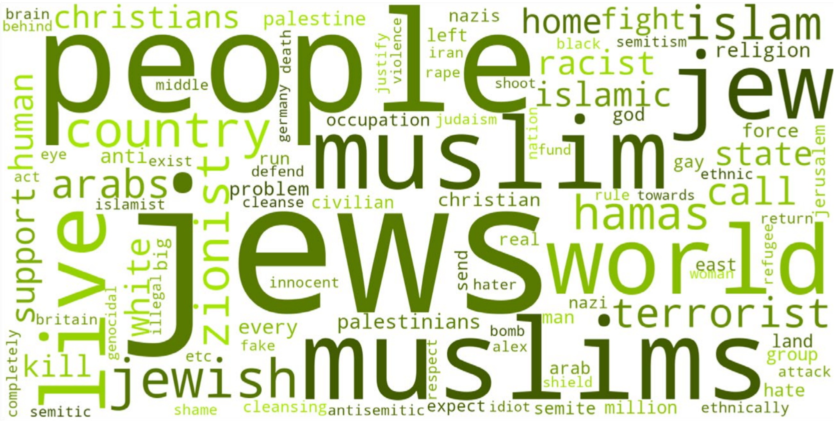


Figure 8.4: 100 most significant terms in texts with an *identity attack* score > 0.7

On the other hand, the 100 most significant terms in texts with a *toxicity* score above 0.7 overlap rather little with antisemitic texts, as shown in Figure 8.5. Note that the strong significance of terms like 'child' and 'kill' might be indicative of narratives surrounding the 'child murderer Israel' references, which are likely to appear in our corpus due to its topical focus.

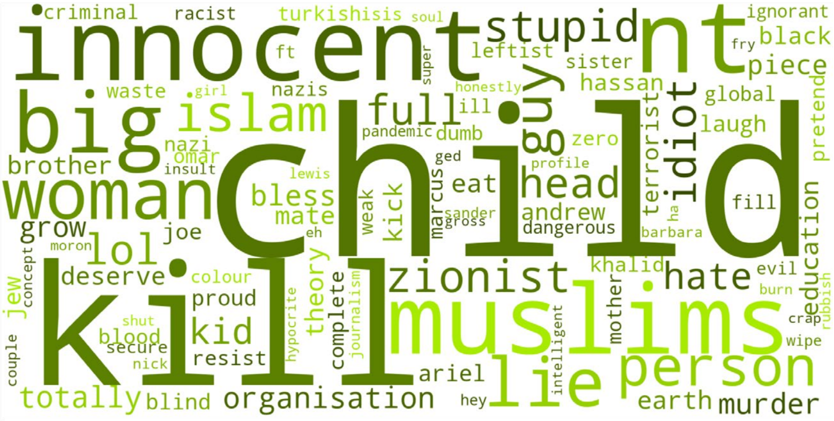


Figure 8.5: 100 most significant terms in texts with a *toxicity* score > 0.7

3.4. From comment to sentence level: exploring the API's span score feature

While it is reasonable to observe a significant presence of identity-related keywords when utilising the attribute of *identity attack*, we believe that it is crucial to conduct further examination of the API's results before employing them for content moderation or research purposes related to antisemitism. A more thorough analysis would benefit from investigation into which parts of a text are responsible for a high score. In addition to the thus-far discussed summary scores, which represent the overall score of an attribute for the entire comment, Perspective API also offers individual scores for each sentence in a comment. These so-called 'span scores' are supposed to assist moderators in identifying the exact section of a longer comment that is, for example, particularly *toxic*. It is important to note that the relation between a comment's summary score and its span scores is neither documented nor easily observable from examples. In particular, the summary score is neither the average nor the maximum or some other obvious statistic based on the span scores.

We believe that exploring this feature is valuable not only for assessing the API's performance but also because the span scores could aid in conducting text analyses at a more granular level. Understanding which specific parts of a text contribute most to its *toxicity*, *identity attack*, or similar aspects is beneficial for in-depth investigations of respective corpora.[7]

To examine the API's capabilities in this aspect, we conducted a qualitative exploration of the summary scores versus the span scores in the English-language subcorpus, focusing on the attribute *identity attack*. In the following paragraphs, we will present noteworthy examples from our findings, providing the summary score of a given comment, as well as the span scores for each sentence in square brackets. Sentences with a score > 0.7 are coloured red.

Overall, our observations support previous indications of a keyword bias, particularly towards identity-related keywords and terms indicating violence, even when used to oppose violence. The following text provides an example:

Israel has shown itself as terrorists. [0.74] All for a land grab and power. [0.04] Stop evicting and killing Palestinians. [0.88]	summary score: 0.86

The last sentence in this comment, when considered independently, can be interpreted as a call to halt acts of violence against Palestinians. It is unclear why this has been assigned such a high score. We speculate that it may be due to the presence of the term "killing", possibly in conjunction with "Palestinians", within the sentence.

The following comment can be interpreted as advocating for tolerance, acceptance and the equality of all human beings regardless of "caste, creed, and religion". Given this, the high summary score is perplexing. It is likely driven by the term "Jews", as the segment containing the term "Jews" is, in fact, assigned a significantly higher span score than all other sentences:

[7] Based on our experience as annotators, we would consider it useful if annotators would specify the parts of a text that guided their classifications, as this approach would help to avoid unintentionally calling upon contextual knowledge.

There's nothing in caste, creed, and religion. [0.56] Does the blood color change? [0.15] Does the Jews come from another planet? [0.81] After all, we stay on the same planet, and we breathe the same air. [0.14] We are all humans [0.09]	summary score: 0.78

This highlights the API's sensitivity to identity-related keywords, which can lead to unintended consequences in scoring such comments, namely incorrectly flagging counter speech.

We observe a similar pattern in the following comment, which emphasises the importance of a state for Jewish people.

USERNAME, that is why Jews need their own state. [0.80]	summary score: 0.80

The following comment presents a defence of Israel's rocket defence system, the "Iron Dome":

USERNAME, Hamas is also attacking civilians. [0.23] If it wasn't for Israel's Iron Dome, more Israeli civilians would be killed. [0.84]	summary score: 0.74

One might expect that the first sentence would yield a relatively high score due to its mention of Hamas attacking civilians, but the assigned span scores provide a different perspective. Interestingly, the second sentence receives a significantly higher span score, which might result from the occurrence of the word "killed" in the text. Such a high *identity attack* score is not plausible, though, since the actual meaning of the sentence implies that the killing of Israeli civilians should be prevented.

We encounter a further case of counter speech that is assigned unreasonably high scores in the following comment against anti-Muslim racism:

USERNAME, you clearly have no idea how many Muslims there are in the world if you believe most of them are violent would-be terrorists. [0.80] The vast majority of them want to live in peace and harmony with a roof over their head, just like the vast majority of human beings in general. [0.13]	summary score: 0.71

Once again, we believe that the API's sensitivity to identity-related keywords is at play here. The segment containing the reference to Muslims receives an unreasonably high score, despite the overall comment countering negative stereotypes. This highlights the limitations of the API's scoring system in accurately capturing the nuances and intentions behind certain comments. One could state that the Perspective API is almost incapable of correctly understanding the stance or sentiment of a text, being rather strongly guided by certain keywords.

It is important to note that our findings are exploratory in nature and should be further supported by systematic assessments of the API's span scores, which we consider an open task for future research.

3.5. Concluding remarks on usage of Perspective API for antisemitic speech detection

In summary, our findings suggest that the Perspective API could be useful for conducting corpus analyses on a broad level, particularly when using the *identity attack* attribute to detect texts related to Jewishness. However, it cannot be automatically assumed that these texts express explicit antisemitism. It is crucial to recognise that the service may not be as helpful for content-moderation efforts that aim to address more complex forms of antisemitism encoded within texts. It provides the ground for actors who strategically utilise linguistic codes, emojis or irony and sarcasm in order to bypass keyword-based automated detection methods. Presumably, the overall labelling approach of Perspective API is not suitable for the incorporation of antisemitic types of *toxic* content, given the difficulty even for experts in labelling short texts typical of online and social media communication. Furthermore, through various experiments, we have observed that the API tends to be overly sensitive to certain identity-related keywords and counter speech, which may impact its accuracy and effectiveness in certain contexts.

Thus, automatic detection of antisemitic speech is still needed and requires careful modelling based on high-quality labelled data.

4. Training of custom models to detect antisemitic comments

In recent years, the so-called 'pre-training and fine-tuning' approach has substantially improved the training of classification models. Fine-tuning leverages language models that were pre-trained using massive amounts of diverse data from corpora such as *Wikipedia* or *Google Books* on generalistic language tasks, such as predicting the next word or a masked word in a sentence. The pre-trained large language models (LLMs) made available in the last years—such as BERT, RoBERTa, GPT-2 or XLM—have learned rich representations of language that capture a variety of linguistic phenomena such as word- and sentence-level semantics, syntactic structures, discourse-level phenomena, as well as subtleties of human language like sarcasm or slang. A pre-trained LLM is adapted in the fine-tuning step to a specific task such as tagging each token in a sentence with respect to a grammar scheme, or, as in our case, to classify texts regarding antisemitism.

A plethora of pre-trained language models are available for fine-tuning different downstream tasks, including text classification. They differ, for example, in terms of the data source used for training (e.g., *Wikipedia* vs. *Twitter*) and its language(s), architecture (e.g., the type and number of layers), training task (e.g., predicting a masked token or the next token), or pre-processing of the text (e.g., lowercasing all words). One of the most popular models (and architectures) employed is BERT, first published in 2018, which has achieved the state of the art for a range of NLP applications, especially classification-oriented tasks. BERT-like pre-trained language models are typically used in recent research to build text classifiers for various text classification tasks, including hate speech (Basile et al. 2019, Aluru et al. 2020, Mathew et al. 2022), offensive language (Wiegand, Siegel and Ruppenhofer 2018, Zampieri et al. 2019 and 2020, Mandl et al. 2021) or (pre-specified) conspiracy theories (Pogorelov et al. 2020, Moffitt, King and Carley 2021, Elroy and Yosipof 2022, Phillips, Ng and Carley 2022). The majority of these benchmark datasets are in the English language and were drawn primarily from *Twitter* (cf. Poletto et al. 2021), in part because of the platform's popularity but also because it offered easy technical access to the data for researchers. As already mentioned, antisemitism has, so far, only been addressed in a handful of efforts for text classification. In addition to BERT-like models, other, significantly larger, models that have been developed for auto-regressive text generation, such as GPT-3, are increasingly being used for classification tasks.

In essence, the fine-tuning step makes use of the rather domain-independent general knowledge encoded by the source model, while 'only' needing to learn the particulars of the target categories/classes. Technically, this can be thought of as extending the source model with a comparatively small set of application-specific parameters that must be learned from the target task data (and modifying the existing model parameters slightly). Fine-tuning allows for the production of efficient classification models with a relatively small number of labelled data samples, which is often all that is available for texts in the political sphere. The approach also better handles out-of-distribution data (that is, data examples that differ from those in the training set) and, in general, provides higher level of generalisation.

However, the amount of text examples required to successfully train a classification model depends on several factors, including the complexity of the classification task, the variability of the text data and the algorithm used to train the model. Although data quality and relevance play a crucial role and can make up for a smaller size of a dataset, it generally makes sense to include as much training data as possible. As a rule of thumb, it is often recommended to provide at least 1,000 labelled examples per class during training.

4.1. Experimental results for English-language comments

In our experiments, we fine-tune BERT-like models as well as GPT-3.5. There are various differences between these two model families. Firstly, BERT, RoBERTa, etc., are open models, while GPT-3.5 is a closed model owned by OpenAI. Because the latter incurs monetary costs that can become substantial when applied on a large scale, many stakeholders might not be able to afford to use GPT-3.5 (or its successor GPT-4) for monitoring, content moderation and analyses. However, GPT-3.5 has been trained on a substantially larger dataset, yielding a model that is orders of magnitude larger than BERT. As such, it is expected to provide superior performance in many tasks and serves in this study as an ‘upper bound’ for what can be achieved in such a scenario if monetary constraints are not considered.

When fine-tuning BERT-like models, we explore the influence of aspects such as the choice of the pretrained language model, standard architecture-related hyperparameters (e.g., learning rate and attention dropout) and data-related settings (e.g., handling of particularly short texts). These hyperparameters determine the overall capabilities of a machine learning model, so combinations of different values are evaluated to find the optimal one.[8] However, since the hyperparameter space can be quite large, there is a need to balance exploration and exploitation for efficient hyperparameter tuning. To address this, we employ Bayesian optimisation, which maintains a probabilistic model that predicts the performance of different hyperparameter configurations. This allows us to exploit the best parameters while still exploring new options to make sure the best parameters are found. As fine-tuning GPT-3.5 is costly, we limited our fine-tuning experiments to using only the standard hyperparameters and fine-tuned the model for up to 2 epochs.[9]

We make use of around 23,000 English-language comments classified as either ‘AS’ or ‘not AS’. It is typical for many text classification tasks, in particular when attempting to classify with respect to different political ideologies or stances, to be confronted with imbalanced data, for which one class is significantly more prevalent than the other(s). In our case, almost 90% of the comments were labelled as not antisemitic (class ‘not AS’), leaving us with only about 10% of texts annotated as antisemitic (class ‘AS’). After cleaning the data, including deduplicating texts and removing empty messages, we ended up with 2,410 samples in class ‘AS’ and 20,684 in class ‘not AS’. We used 80% of data for training (16,539 records in class ‘not AS’ and 1,936 in class ‘AS’), 10% for validation—which serves the identification of the best-performing hyperparameters—and 10% for testing the model yielding the lowest errors on the validation set.

[8] The following hyperparameters were considered: model (roberta-base, bert-base-uncased), number of epochs, downsampling of the negative class, learning rate, batch size, weight decay, attention_probs_dropout_prob and hidden_dropout_prob.

[9] The number of epochs refers to the number of times the model is presented with all of its training data in order to update its parameters based on the value of the loss function, which is being minimised during training.

Class	Records	Precision	Recall	F1-score	Accuracy
AS	225	0.75 / **0.76**	0.65 / **0.79**	0.7 / **0.77**	0.94 / **0.95**
not AS	2,084	0.96 / **0.98**	**0.98** / 0.97	0.97 / 0.97	

Table 8.2: Evaluation of the best performing fine-tuned BERT-like model and the fine-tuned GPT-3.5 model, separated by /, on the test data. The best score is highlighted in bold.

The performance metrics of both the best-performing fine-tuned BERT-like model and GPT-3.5 on the test set are displayed in Table 8.2. The scores for the best BERT-like model (represented by the first number per table cell) can be interpreted as follows: 96% of all texts predicted by the model as not being antisemitic were indeed labelled by the human annotators as such (precision class 'not AS'), and the model finds 98% of texts in this class (recall class 'not AS'). On the other hand, among the texts predicted as antisemitic, 75% were labelled as such (precision for class 'not AS'), while the model managed to find 65% of texts labelled as antisemitic by the annotators. To make this easier to grasp: if a content moderator was to apply this model to 1,000 comments, where 100 are assumed to be antisemitic, the model would find 65 of the 100 antisemitic texts and miss 35 of them. This could be seen as a low rate from the perspective of keeping the comments section free of antisemitic speech. However, the number of false alarms would be low, at 22, limiting the manual efforts required. This example highlights the trade-off between two types of errors: while one would want to increase the recall of class 'AS', it would also be desirable to keep the number of false alarms low. Thus, from an application perspective, one needs to decide which kind of error (false positives vs. false negatives) should be prioritised, and, for example, what minimum recall needs to be achieved for class 'AS' and what precision could be accepted in return.[10]

[10] To illustrate this, let us assume that we want to achieve a recall of at least 0.8 while keeping the precision as high as possible. One simple option would be to adjust the probability threshold for assigning a prediction to a class label. The classifiers we train are probabilistic, thus for each text they produce probabilities of belonging to either of these classes. By default, the threshold for binary classification is set to 0.5, meaning the class with higher probability wins. However, the threshold can be changed in order to increase the value of a desired metric. By using the validation set to find out which threshold satisfies a recall of at least 0.8 while maximising the precision, we can identify a threshold that achieves a recall of 0.81 and a precision of 0.52 on the validation set. Thus, we would capture nearly 80% of all antisemitic texts, albeit with almost every second flag being a false alarm.

As presented in Table 8.2, the fine-tuned GPT-3.5 model outperforms the BERT model in terms of the F1 score, defined as the harmonic mean of precision and recall, on the class 'AS', primarily due to its higher recall of antisemitic texts. In the hypothetical scenario described above, the model would only miss 21 out of the 100 antisemitic texts while maintaining a very low number of false alarms. This confirms the initial hypothesis that the larger model can be trained more effectively, albeit at substantially higher monetary cost.

To enhance the performance of the fine-tuned BERT model, we conducted several experiments targeting the strong class imbalance in the dataset. We reduced the number of examples from the 'not AS' class that are easily correctly classified. During model training, all data points contribute in the same way to the computation of the loss that guides the training process. Thus, reducing such examples, or penalising them in a different way, can potentially enhance the detection of the positive class. While these strategies led to a higher recall for the class 'AS', and thus an increased identification of antisemitic texts, it came at the cost of lower precision and a comparable F1 score. We additionally employed various strategies to augment the 'AS' class, including generating new texts by substituting some words with others having a similar meaning or by adding words that are assumed not to significantly alter the overall meaning of the sentence. Additionally, we translated texts labelled as antisemitic from the German and French corpora to English and performed forward-and-backward translation with English-language records. A random sample of translations was manually inspected. However, these strategies did not result in a noteworthy improvement of the F1 score for the class 'AS'. A significant challenge stems from the fact that standard pre-trained models as those we used to identify similar words for replacement may not effectively capture the nuanced context in corresponding messages. For instance, words like 'Israel' and 'Palestine' might be deemed similar from the perspective of a generic language model, but they are not interchangeable in the context of the Middle East conflict. Models that have undergone additional fine-tuning on a corpus reflecting such nuances would be more suitable, as well as other more sophisticated text augmentation strategies that we plan to explore in future research.

4.2. Domain generalisation: discourse and domain shift

Classifiers trained on a given corpus should ideally be able to generalise and, to a certain extent, transfer their 'knowledge' to other domains. In other words, they should be able to carry out the same task when made to encounter the same phenomenon but in a potentially different distribution of data. A difference in distribution is to be expected when a model is applied to data from a time range, platforms or discourses distinct from those represented in the training data. Phenomena such as antisemitism constitute a 'moving target' in the sense that codes, narratives and forms of expression evolve with time and differ from community to community. Against this background and considering the fact that our training corpus is rather small, especially with regards to class 'AS' it is rather to be expected that trained models will struggle with domain transfer.

We have evaluated the performance of the fine-tuned models (BERT-based and GPT-3.5-based) in two settings: (1) two new discourse events that were not represented in the training data and (2) a corpus from *Twitter* that was created and annotated using a different approach.

The two discourse events not represented in our training dataset, both from 2022, were the antisemitic incidents that occurred during the FIFA World Cup in Qatar and the discussion about Kanye West's radical antisemitic statements. These two resulted in a total of 2,612 text examples in English, only 107 of which were labelled as antisemitic by human annotators.

Jikeli et al. (2023) recently published a corpus containing tweets from 2019 to 2021. The corpus was obtained through a multi-step procedure that involved filtering a 10% *Twitter* sample from the Indiana University's Observatory on Social Media database using the keywords 'Jews' and 'Israel'. The texts were annotated by two individuals, using an annotation scheme based on the IHRA definition of antisemitism. The annotators were asked to apply one of five categories to each tweet according to whether it was antisemitic and their level of confidence in each case. They marked 6,941 texts overall. Two categories, 'probably antisemitic' and 'confident antisemitic', were merged into the overarching category 'antisemitic', while the other three ('confident', 'probably not antisemitic' and 'uncertain/neutral') were merged into 'not antisemitic'.

This process resulted in 1,250 (~18%) texts being included in the positive class.

Table 8.3 presents the evaluation of both fine-tuned models on each of the two datasets. As expected, the performance of both models declines when confronted with a data distribution shift, with the difference being more pronounced on the keyword-based *Twitter* dataset. The performance of GPT-3.5 is more robust on both of the datasets, especially with regard to class ‘AS’. More precisely, both models manage to recognise texts in the class ‘not AS’ from the two new discourse events with an F1 score in the same range as before, while the F1 score for class ‘AS’ drops to 0.6 for the BERT model but remains high, at 0.76, for GPT-3.5. This suggests that, in contrast to GPT-3.5, the BERT model is strongly affected by the topics of the discourses it has seen during training and that it struggles more with recognising antisemitic speech related to a different topic.

Dataset	Class	Records	Precision	Recall	F1-score	Accuracy
(1) new discourse events	‘AS’	107	0.63 / **0.73**	0.57 / **0.79**	0.6 / **0.76**	0.97 / **0.98**
	‘not AS’	2,504	0.98 / **0.99**	0.99 / 0.99	0.98 / **0.99**	
(2) Twitter dataset by Jikeli et al.	‘AS’	1,250	0.54 / **0.62**	0.52 / **0.8**	0.53 / **0.7**	0.83 / **0.88**
	‘not AS’	5,691	0.9 / **0.95**	**0.9** / 0.89	0.9 / **0.92**	

Table 8.3: Evaluation of the fine-tuned BERT-like model and the fine-tuned GPT-3.5 model, separated by /, on (1) the dataset comprising two discourse events absent from training data, and (2) the *Twitter* dataset compiled by Jikeli et al. (2023). The best score is highlighted in bold.

The performance of the BERT model drops further on the second dataset, with an F1 score of 0.9 for class ‘not AS’ and an F1 score of 0.53 for class ‘AS’ (and an overall accuracy of 0.83). A similar tendency is visible for the GPT-3.5 model as well, however it still yields a solid F1 score of 0.7 for the class ‘AS’. The performance drop between test data (Table 8.2) and this dataset, however, is not surprising, and it showcases well the effect of the corpus and annotation scheme used for training. The annotators of the *Twitter* dataset were allowed to use the surrounding context and references to external resources when labelling a tweet. One would therefore expect that, conceptually and empirically, the comments labelled as antisemitic have substantial intersection with the 'contextually antisemitic' comments in our corpus that were excluded from the training and test set.[11] Furthermore, the distribution of the two corpora is quite different: despite the fact that our training dataset also contains tweets,[12] the topic distributions differ significantly. Our corpus reflects certain discourses, while the *Twitter* corpus is a combination of random messages related to Jewishness and Israel and messages containing antisemitic slurs. In particular, the slur 'ZioNazi' was used as one of the filter keywords. This expression, however, occurs in 529, and thus almost 90%, of texts labelled as antisemitic in the *Twitter* corpus, but only about 20 times in our entire (and significantly larger) English-language corpus.

4.3 Concluding remarks on training custom models for the detection of antisemitic speech

We have fine-tuned different state-of-the-art large language models to distinguish antisemitic speech in an English-language corpus sourced from various online platforms spanning a time period of multiple years. In particular, the corpus was not created using keyword filters but instead reflects diverse topics and discourses likely to trigger antisemitism. Stemming from mainly mainstream platforms, the corpus contains a rather high amount of implicitly formulated antisemitic speech and displays a substantial class imbalance. These aspects contribute to an increased challenge when it comes to build effective classification models.

We have shown that openly available model architectures like BERT can be effectively leveraged to detect antisemitic speech in the described corpus. An F1 score for the class 'AS' of 0.7 can be considered satisfactory considering the complexity of the dataset. In practical-application scenarios, such as content moderation, it would be sufficient for a model to identify discussions with an alarming amount of antisemitic

[11] This is supported by the fact that "lack of understanding of the context" is identified as one of the main reasons for annotator disagreement (Jikeli et al. 2023).

[12] Note that our corpus also contains data from Twitter. We did not check for contamination of our dataset since, statistically, the chances are very low.

speech that need a closer look by human experts. At the same time, the performance of the model declines substantially when confronted with unseen discourses or a different dataset. This implies the necessity of a continuous effort in labelling a sufficient amount of data and further fine-tuning of the model.

At the same time, a fine-tuned GPT-3.5 model shows superior performance not only on the test data but also in discourse and domain transfer. As expected, the larger model is capable of providing better results, with F1 scores for class 'AS' above 0.7 and almost 1 for class 'not AS', using standard hyperparameters only and within 2 epochs of training. This model, however, incurs higher monetary costs for fine-tuning and application as it cannot be run without using OpenAI's API. Thus, the decision as to which approach might be more suitable depends on the specific application scenario and available resources.

To facilitate real-world application, we have established an inference service featuring our best BERT-based model within a web app. This service enables users to input text, receive predictions and view corresponding scores. A feedback loop has been implemented, allowing users to express agreement or disagreement, thereby enhancing our understanding of the model's performance and aiding in the collection of additional training data. The trained models can be provided upon request. Similarly, the code for the web service is available for sharing, facilitating the implementation of similar setups in other projects.

5. Future directions

5.1. Rethink the object of classification

Capturing the meaning of texts written by humans can be a challenging task. This is particularly the case for short messages, such as those commonly found in online and social media discussions. Authors may use subtle, coded, implicit expressions of their opinions, for instance, to attain a certain level of ambivalence and thereby avoid content-moderation measures. Examples of this can be found in fragmented expressions of beliefs in conspiracy theories (Steffen et al. 2022), implicit climate-change denials (Falkenberg and Baronchelli 2023) or the usage of codes in antisemitic narratives. Furthermore, references to world knowledge add to the difficulty of a model to 'comprehend' the content of a text. An extreme example of this is a statement by Nicholas J. Fuentes, a white supremacist political commentator and live streamer, who denied the Holocaust by 'jokingly' doubting the possibility of baking six million batches of cookies within five years.[13]

Moreover, comments are typically part of a longer thread, and this context is often needed to fully resolve the meaning of the individual post and its author's intention. Similarly, posts often make references to linked or embedded content that is increasingly multi-modal, as well as to current (political) events. The attempt to make such additional context available to the models is quite challenging, as, for example, relevant references can be made to any previous comment in a thread. This raises the question of whether it might be more appropriate to consider sub-threads or threads as entities instead of single comments. Because the dataset collected by Jikeli et al. (2023) took all this information into account when it was annotated, a model should have this context available as well in order to assess its comparative abilities fairly. The Decoding Antisemitism annotation scheme distinguishes between contextually antisemitic and antisemitic texts, but one might argue that annotators might not be able to fully exclude context when looking at an entire thread in sequential manner.

It is noteworthy that the Perspective API has announced plans to include conversation context—which may encompass additional text, URLs or even images—for comment evaluation.[14] When this feature becomes available, it would be intriguing to investigate whether the service's overall performance improves in scoring antisemitic speech as *toxic*.

In future research, we aim to explore various methods of providing context to individual comments within a classification model, as well

[13] In one of his live streams, Fuentes reads the following text: "If I take one hour to cook a batch of cookies and the cookie monster has 15 ovens working 24 hours a day, every day for five years, how long does it take cookie monster to bake 6 million batches of cookies?" He then uses the cookie analogy in several subsequent statements of Holocaust denial. For the livestream, see https://mobile.twitter.com/CalebJHull/status/1189594371030695937 (last accessed on 23 February 2023). For more information, see e.g, https://www.adl.org/resources/blog/nicholas-j-fuentes-five-things-know (last accessed on 14 February 2023).

[14] https://developers.perspectiveapi.com/s/about-the-api-key-concepts?language=en_US

as to develop models capable of classifying sub-threads instead of individual messages. The latter necessitates defining what should constitute an appropriate sub-thread.

5.2. Text classification with prompt-based generative models

Recently, OpenAI's further development of their Generative Pre-Trained Transformers, namely GPT-3.5 and the multi-modal advancement GPT-4, has received wide public attention because of their abilities to generate human-like responses to a given input. These models have been made publicly available through services such as ChatGPT, which allows users to easily interact with chatbots based on respective models via their web browser or API. While the introduction of these models has led to intense debates concerning the risks and potentials of so-called 'artificial general intelligence' (AGI), it also opens up new opportunities to approach the task of text classification.

In this chapter, we presented the results of fine-tuning a GPT-3.5 model for the detection of antisemitic texts. Because fine-tuning and applying OpenAI's models through their API incurs monetary cost, and the models remain with OpenAI, it would be of interest to explore the capabilities of comparable open models such as Meta's Llama-2 or Mistral AI's models Mistral and Mixtral.

Importantly, models such as GPT-3.5 and its competitors were built to facilitate few-shot learning or even zero-shot learning—scenarios in which the model is asked to classify texts into categories for which it has seen only few, or even no, in-context examples. This implies that the model is not fine-tuned, as in our experiments. Instead, it learns additional information from the task description and, perhaps, a few examples of antisemitic and not antisemitic texts provided as part of the textual instructions, the so-called prompt. In this context, design of the prompt has become a crucial task for engineers and researchers. Prompts influence the model's behaviour; they can restrict the form of its response, ask it to focus on certain aspects, or provide it with supplemental information to carry out the task, such as definitions or training examples (Liu et al. 2023; White et al. 2023).

Initial empirical evaluations indicate the huge potential of these models for increasing the efficiency of text classification. In a recent experiment, the zero-shot accuracy of ChatGPT exceeded that of crowdworkers in four out of five tasks related to content moderation, while being about twenty times cheaper (Gilardi, Alizadeh and Kubli 2023). Researchers are examining the potential of GPT-3 models for the classification of hateful content (Chiu, Collins and Alexander 2022; Wang and Chang 2022; Huang, Kwak and An 2023). Li et al. (2023) conduct extensive prompting experiments and compare the performance of ChatGPT to that of crowdworkers for the task of classifying texts as hateful, offensive or toxic (HOT). They find that ChatGPT achieves an accuracy of roughly 80% when compared to crowdworkers' annotations. While the abovementioned works address the more general phenomena of hateful speech or toxic language, the work of Mendelsohn et al. (2023) examines the performance of GPT-3 models for identifying and understanding the specific linguistic phenomenon of dog whistles, that is, "coded expressions that simultaneously convey one meaning to a broad audience and a second one, often hateful or provocative, to a narrow in-group" (Mendelsohn et al. 2023). Their experiments include antisemitic dog whistles and find that the performance of the model "varies widely across types of dog whistles and targeted groups" (Mendelsohn et al. 2023).

Our initial experiments with OpenAI's GPT-3.5 and GPT-4 and the open alternatives Llama-2 and Mistral suggest that prompting is not as effective as fine-tuning in detecting antisemitic speech. We are currently conducting experiments to explore the potential of prompting models for the detection of antisemitic comments in our corpus. These include investigating the impact of the different definitions of antisemitism, incorporating discourse event-related information and exploring various output constraints, such as allowing the model to differentiate the texts predicted to be antisemitic based on their antisemitic narratives. Additionally, we aim to further analyse the explanations generated by the model to justify its classification decisions. Moreover, we plan to investigate the potential benefits of including relevant context, such as preceding comments, in the prompt to enhance the detection accuracy.

It is important to acknowledge that perfection in automated classification is unattainable; even aiming for F1 scores substantially above those already achieved in our fine-tuning efforts might be unreasonable for corpora obtained without filtering by specific keywords. Antisemitism presents a particularly complex challenge compared to other hate ideologies. It is often conveyed using coded language that carries specific meanings for certain audiences while appearing innocuous to others. Antisemitic expressions may reference historical events, rendering them difficult to identify without contextual comprehension, and they are found in multiple political spheres or subcultures (Lauer and Potter 2023), each with distinct rhetorical nuances and argumentative strategies. Even for human experts, straightforward binary categorisation of texts as antisemitic or not antisemitic can prove challenging. Therefore, it is crucial to define realistic and appropriate application scenarios for such models and determine how they can best assist in this task.

References

Aluru, Sai Saketh, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee, 2020. "Deep Learning Models for Multilingual Hate Speech Detection". Preprint, https://arxiv.org/abs/2004.06465

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso and Manuela Sanguinetti, 2019. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, MN, USA: Association for Computational Linguistics, 54–63, https://doi.org/10.18653/v1/S19-2007

Chandra, Mohit, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava and Ponnurangam Kumaraguru, 2021. "'Subverting the Jewtocracy': Online Antisemitism Detection Using Multimodal Deep Learning". In: *Proceedings of the 13th ACM Web Science Conference 2021* (WebSci '21), Virtual Event, United Kingdom, 148–157, https://doi.org/10.1145/3447535.3462502

Chiu, Ke-Li, Annie Collins and Rohan Alexander, 2022. "Detecting Hate Speech with GPT-3". Preprint, http://arxiv.org/abs/2103.12407

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain and Lucy Vasserman, 2018. "Measuring and Mitigating Unintended Bias in Text Classification". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* New Orleans LA USA: ACM, 67–73, https://doi.org/10.1145/3278721.3278729

Elroy, Or and Abraham Yosipof, 2022. "Analysis of COVID-19 5G Conspiracy Theory Tweets Using SentenceBERT Embedding". In: *Artificial Neural Networks and Machine Learning – ICANN 2022*: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part II. Berlin: Springer-Verlag, 186–196, https://link.springer.com/chapter/10.1007/978-3-031-15931-2_16

Falkenberg, Mark and Andrea Baronchelli, 2023. "How Can We Better Understand the Role of Social Media in Spreading Climate Misinformation?" Grantham Research Institute on Climate Change and the Environment. January 2023, https://www.lse.ac.uk/granthaminstitute/news/how-can-we-better-understand-the-role-of-social-media-in-spreading-climate-misinformation

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli, 2023. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". In: *Proceedings of the National Academy of Sciences 120*, No. 30, e2305016120, https://doi.org/10.1073/pnas.2305016120

González-Pizarro, Felipe and Savvas Zannettou, 2022. "Understanding and Detecting Hateful Content Using Contrastive Learning". In: *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media* (ICWSM 2023). June 5–8, 2023, Limassol, Cyprus. Palo Alto, CA: AAAI Press, 257-268, https://doi.org/10.1609/icwsm.v17i1.22143

Horta Ribeiro, Manoel, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro and Robert West, 2021. "Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels". In: *Proceedings of the ACM on Human-Computer Interaction 5* (CSCW2), 1–24, https://doi.org/10.1145/3476057

Hoseini, Mohamad, Philipe Melo, Fabricio Benevenuto, Anja Feldmann and Savvas Zannettou, 2023. "On the Globalization of the QAnon Conspiracy Theory Through Telegram". In: *Proceedings of the 15th ACM Web Science Conference 2023* (WebSci '23). Association for Computing Machinery, New York, USA, 75–85, https://doi.org/10.1145/3578503.3583603

Huang, Fan, Haewoon Kwak and Jisun An, 2023. "Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech". In: *Companion Proceedings of the ACM Web Conference 2023*, 294–97. Austin, TX: ACM, https://doi.org/10.1145/3543873.3587368

Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong and Stephen Denuyl, 2020. "Social Biases in NLP Models as Barriers for Persons with Disabilities". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491– 5501, https://doi.org/10.18653/v1/2020.acl-main.487

Jikeli, Günther, Sameer Karali, Daniel Miehling and Katharina Soemer, 2023. "Antisemitic Messages? A Guide to High-Quality Annotation and a Labeled Dataset of Tweets". Preprint, http://arxiv.org/abs/2304.14599

Jikeli, Günther, Damir Cavar, Weejeong Jeong, Daniel Miehling, Pauravi Wagh and Denizhan Pak, 2022. "Toward an AI Definition of Antisemitism?" In:

Monika Hübscher and Sabine von Mering (eds). *Antisemitism on Social Media*. Abingdon: Routledge, 193–212

Lauer, Stefan and Nicholas Potter (eds.), 2023. *Judenhass Underground. Antisemitismus in emanzipatorischen Subkulturen und Bewegungen*. Berlin / Leipzig: Hentrich & Hentrich Verlag

Lees, Alyssa, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler and Lucy Vasserman,, 2022. "A New Generation of Perspective API: Efficient Multilingual Character-level Transformers". In: *KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2022*. 3197–3207, https://doi.org/10.1145/3534678.3539147

Li, Lingyao, Lizhou Fan, Shubham Atreja and Libby Hemphill, 2023. "'HOT' ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media". *ACM Transactions on the Web 18* (2), Article No. 30, 1–36, https://doi.org/10.1145/3643829

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi and Neubig, Graham, 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". *ACM Computing Surveys* 55, No. 9, Article No. 195, 1–35, https://doi.org/10.1145/3560815

Mandl, Thomas, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Schaefer, Johannes, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini and Amit Kumar Jaiswal, 2021. "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages", http://arxiv.org/abs/2112.09301

Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal and Animesh Mukherjee, 2022. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (17), 14867–14875

Mendelsohn, Julia, Ronan Le Bras, Yejin Choi and Maarten Sap, 2023. "From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Toronto, Canada, 15162–15180, https://doi.org/10.18653/v1/2023.acl-long.845

Meta, 2022. Community Standards Enforcement | Transparency Center, https://transparency.fb.com/data/community-standards-enforcement

Mihaljević, Helena and Elisabeth Steffen, 2022. "How Toxic Is Antisemitism? Potentials and Limitations of Automated Toxicity Scoring for Antisemitic Online Content". In: *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis* (CPSS-2022), KONVENS 2022, 1–12. 01 January 2022. Potsdam, Germany

Moffitt, J. D., Catherine King and Kathleen M. Carley, 2021. "Hunting Conspiracy Theories During the COVID-19 Pandemic". *Social Media + Society*, 7 (3), https://doi.org/10.1177/20563051211043212

Phillips, Samantha C., Lynnette Hui Xian Ng, Kathleen M. Carley, 2022. "Hoaxes and Hidden Agendas: A Twitter Conspiracy Theory Dataset: Data Paper". In: *Companion Proceedings of the Web Conference 2022*. WWW '22. New York: Association for Computing Machinery, 876–880, https://doi.org/10.1145/3487553.3524665

Pogorelov, Konstantin, Daniel Thilo Schroder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkukova and Johannes Langguth, 2020. "FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020". In: *Working Notes Proceedings of the MediaEval 2020 Workshop*, http://ceur-ws.org/Vol-2882/paper64.pdf

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco and Viviana Patti, 2021. "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review". *Language Resources and Evaluation*, 55 (2), 477–523, https://doi.org/10.1007/s10579-020-09502-8

Röttger, Paul, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts and Janet B. Pierrehumbert, 2021. "HateCheck: Functional Tests for Hate Speech Detection Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), 41–58, https://doi.org/10.18653/v1/2021.acl-long.4

Steffen, Elisabeth, Helena Mihaljević, Milena Pustet, Nyco Bischoff, María do Mar Castro Varela, Yener Bayramoğlu and Bahar Oghalai, 2022. "Codes, Patterns and Shapes of Contemporary Online Antisemitism and Conspiracy Narratives — an Annotation Guide and Labeled German-Language Dataset in the Context of COVID-19". In: *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media* (ICWSM 2023). June 5–8, 2023, Limassol, Cyprus. Palo Alto, CA: AAAI Press, https://doi.org/10.1609/icwsm.v17i1.22216

Wang, Yau-Shian and Yingshan Chang, 2022. "Toxicity Detection with Generative Prompt-Based Inference". Preprint, http://arxiv.org/abs/2205.12390

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith and Douglas C. Schmidt, 2023. "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT". Preprint, http://arxiv.org/abs/2302.11382

Wiegand, Michael, Melanie Siegel and Josef Ruppenhofer, 2018. "Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language". In: *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing* (KONVENS 2018), https://epub.oeaw.ac.at/0xc1aa5576_0x003a10d2.pdf

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra and Ritesh Kumar, 2019. "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, MN, USA: Association for Computational Linguistics, 75–86, https://doi.org/10.18653/v1/S19-2010

Zampieri, Marcos, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis and Çağrı Çöltekin, 2020. "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, https://doi.org/10.18653/v1/2020.semeval-1.188