Discourse Report | 5

Pilot Project

Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online

Centre for Research on Antisemitism,
Technische Universität Berlin

Decoding Antisemitism

April 2023

ZfA Zentrum für Antisemitismusforschung
TU berlin
htw
HATE LAB
KING'S College LONDON
ALFRED LANDECKER FOUNDATION

Decoding
Antisemitism

Principal Investigator:
Dr Matthias J. Becker
Centre for Research on Antisemitism, Technische Universität (TU) Berlin

Co-Investigator:
Prof. Helena Mihaljević
Hochschule für Technik und Wirtschaft (HTW) Berlin

Research Team TU & HTW Berlin:
Dr Laura Ascone
Dr Matthew Bolton
Alexis Chapelan
Dr Jan Krasni
Karolina Placzynta
Milena Pustet
Marcus Scheiber
Elisabeth Steffen
Hagen Troschke
Chloé Vincent

Project Manager:
Prof. Uffa Jensen
Centre for Research on Antisemitism, TU Berlin

In collaboration with Cardiff University's HateLab and the Department of Digital Humanities, King's College London

Funded by the Alfred Landecker Foundation

Project Coordination:
Dr Susanne Beer

Secretary:
Andrea Rellin

Student assistants:
Pia Haupeltshofer
Alexa Krugel
Victor Tschiskale

Technische Universität Berlin
Centre for Research on Antisemitism (ZfA)
Kaiserin-Augusta-Allee 104 – 106
10553 Berlin

Contact: info@decoding-antisemitism.eu
Web: decoding-antisemitism.eu

Citation recommendation:

Chapelan, Alexis; Ascone, Laura; Becker, Matthias J.; Bolton, Matthew; Haupeltshofer, Pia; Krasni, Jan; Krugel, Alexa; Mihaljević, Helena; Placzynta, Karolina; Pustet, Milena; Scheiber, Marcus; Steffen, Elisabeth; Troschke, Hagen; Tschiskale, Victor; Vincent, Chloé (2023). *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online.* Discourse Report 5. Berlin: Technische Universität Berlin. Centre for Research on Antisemitism.

DOI: https://doi.org/10.14279/depositonce-17105

Advisory Board

Table of Contents

Executive Summary

1. This report investigates antisemitic web comments in the context of three international events: Kanye West's antisemitic statements (focus on the UK and France), the antisemitic incidents occurred during the 2022 FIFA World Cup in Qatar (UK), and to the Israeli legislative elections (Germany).

2. For years, the rapper Kanye West's discourse has been studded with far-right and antisemitic statements. Since October 2022, his antisemitic position has become even more radical. In our study, we examined a total of 3,953 comments posted in reaction to West's case. The analysis shows that antisemitic reactions were more frequent in France (14 %) than in the UK (11 %), and that the percentages of antisemitic concepts vary according to the media outlet.

3. The analysis shows also that British and French users react in similar ways: they tend to either AFFIRM, RELATIVISE, or DENY West's antisemitic statements. Likewise, they express ideas of JEWISH POWER and antisemitic CONSPIRACY THEORIES. In supporting the singer, web users rely increasingly on detour communication, or tend to reframe the debate either by claiming VICTIMHOOD through alleged censorship or by outright DENYING THE ANTISEMITIC NATURE of the incident.

4. During the 2022 FIFA World Cup, various antisemitic incidents took place, such as the heckling of Israeli journalists during interviews with football fans. In the UK, these events were largely discussed on *Twitter*. The analysis conducted on 1,250 comments shows that 10 % of the corpus is composed of antisemitic statements, and that users often evoke the following concepts: APARTHEID ANALOGY, DENIAL OF THE JEWISH RIGHT TO SELF-DETERMINATION and EVIL.

5. Benjamin Netanyahu's victory in 2022 Israeli elections caused a major inter-national upheaval. This event was widely covered in Germany and triggered antisemitic reactions in the comments sections of the main German media websites as well as on their *Facebook*, *Twitter*, and *YouTube* profiles. The corpus, which consists of 2,111 comments, comprises 7 % of antisemitic comments. We identified a wide range of antisemitic concepts such as NAZI or APARTHEID ANALOGIES, which fuel the idea of Israel as an EVIL state.

6. Our colleagues from the field of Data Science at HTW Berlin examined existing approaches to automated detection of antisemitic posts. The existing web service *Perspective API*, which aims at detecting toxic language, was found to be biased towards specific keywords, and it could not easily detect disguised forms of antisemitism. Their approach, using transfer learning, is explained and their first results are presented. They also discuss the question of how the machine learning models, such as the one they work with, will be used.

1. Introduction

Decoding Antisemitism is a transnational and interdisciplinary research project mapping the content, structure and frequency of antisemitism in digital spaces, with the aim of understanding the full dimension of Jew-hatred online in all its complexity, significantly improving automated algorithmic recognition of antisemitic speech. Taking a dynamic approach to studying antisemitism in accordance with our framework, based on the IHRA definition of antisemitism,[1] we follow the evolution of anti-Jewish discourse within the context of real-world events. Our corpora (or datasets) include web user comments reacting to various relevant national and international news on social media. This allows our team to sample and analyse in detail the wide range of antisemitic stereotypes and topoï in a variety of discourse environments. Our analyses prove the fundamental plasticity and adaptability of the antisemitic discourse, whose symbolic grammar is constantly evolving to evade increased social awareness about hate speech, but also to mobilise new supporters. In turn, these rich and varied datasets are fed back into machine learning, providing the basis for much broader analysis of online debates with increasing accuracy.[2] The bi-annual publication of our Discourse Reports gives insights into the progress and interim results of our corpus analyses.

[1] – See Discourse Report 2: https://decoding-antisemitism.eu/publications/second-discourse-report.

[2] – For details on our research design see https://decoding-antisemitism.eu/about.

In this fifth Discourse Report, we focus on the online fallout of three major events from late 2022, as reported by British, French and German media. Alongside this, we are pleased to present the first comprehensive results from the process of AI development and testing carried out on our previously coded data by our partners at the Hochschule für Technik und Wirtschaft (University of Applied Sciences) in Berlin. These preliminary results have shown both the great potential and the substantial challenges in creating machine learning abilities to mimic the decision making of human experts.

The first part focuses on the rapper Kanye West's antisemitic statements made in the autumn of 2022. Our analysis of British and French corpora reveals that West's articulation of canonical antisemitic ideas of JEWISH POWER or CONSPIRACY THEORY was echoed and expanded upon by substantial sections of web users (11 % and 14 % respectively), often in conjunction with more recent topoï, such as TABOO OF CRITICISM or DENIAL and RELATIVISATION OF ANTISEMITISM.

The report also analyses the online response to antisemitic incidents during the 2022 FIFA World Cup in Qatar, including the heckling and shunning of Israeli journalists. Our analysis of *Twitter* responses found that more than 10 % were antisemitic, with a number of Israel-related antisemitic topoi – ranging from APARTHEID ANALOGY and DENIAL OF THE RIGHT TO SELF-DETERMINATION to INSTRUMENTALISATION OF ANTISEMITISM – being particularly prominent.

Meanwhile, Israeli legislative elections held in November 2022 led to the victory of Benjamin Netanyahu's conservative Likud party. The result of the elections caused a major international upheaval. Worries about a democratic backslide were echoed across Western media – including in Germany. Our study of a German-language corpus shows that commenters seized this fraught debate as an opportunity to promote antisemitic narratives. Israel's political history was often read through the lenses of Nazi or apartheid analogies, which demonise and radically delegitimise the state of Israel.

In the last and most comprehensive chapter, our colleagues Helena Mihaljević, Milena Pustet, and Elisabeth Steffen from the HTW Berlin discuss approaches to automated detection of antisemitic posts. They examine the capabilities and limitations of *Perspective API*, a web service designed to detect toxic language, when dealing with antisemitic speech. Their research shows that the service's effectiveness is impacted by a bias towards specific keywords, and that it has difficulty recognising disguised forms of antisemitism. They also provide first findings from their experiments using state-of-the-art approaches to text classification, explaining what transfer learning means, and how it differs from classical approaches. The authors also delve into the conflicting objectives that arise in different potential applications of these machine learning models.

2. Kanye West's antisemitic remarks in autumn 2022

The American musician, fashion designer and socialite Kanye West, now using the mononym Ye, is one of the most recognisable figures in the entertainment industry, with millions of followers worldwide. His audience is composed of younger, less politicised demographic who are directly exposed to his messaging. For years, the rapper's discourse has been studded with far-right and antisemitic statements; the admiration for Adolf Hitler and the alleged plans to name his 2018 album after him are probably the most explicit examples.[3] Since October 2022, his antisemitic position has become even more radical, expressed through several references to antisemitic conspiracy theories and tropes, and the following now-deleted tweet from 8 October last year:

> "I'm a bit sleepy tonight but when I wake up I'm going death con 3 On JEWISH PEOPLE. The funny thing is I actually can't be Anti Semitic because black people are actually Jew also You guys have toyed with me and tried to black ball anyone whoever opposes your agenda."

Although West suffers from bipolar disorder which may have caused his erratic behaviour, the statement itself is blatantly antisemitic – including a death wish expressed through a pun, DENIAL OF ANTISEMITISM, and the stereotype of CONTROLLING OPINIONS. Unsurprisingly, this and other, similar claims have led various platforms such as *Instagram* and *Twitter* to condemn his words and lock his accounts. Likewise, *Adidas* and *Balenciaga*, two of the brands collaborating with West, terminated their contracts, and Madame Tussauds Museum in London removed his wax figure.

The rapper's escalation generated substantial press coverage in the UK and France and triggered antisemitic reactions which reiterated canonical elements of the anti-Jewish mythology, such as JEWISH POWER or CONSPIRACY THEORIES, as well as more recent ones, such as TABOO OF CRITICISM or DENIAL and RELATIVISATION OF ANTISEMITISM, see Figure 1. In expressing support for the celebrity, web users relied increasingly on detour communication, or tried to reframe the debate either by claiming VICTIMHOOD through censorship or by outright denying the antisemitic nature of the incident. In the following sub-chapters, we present the results of our qualitative analysis of these conceptual features in the UK and French media.

[3] For more details about Kanye West's statements, see Wilson 2022 and Solomon 2023.

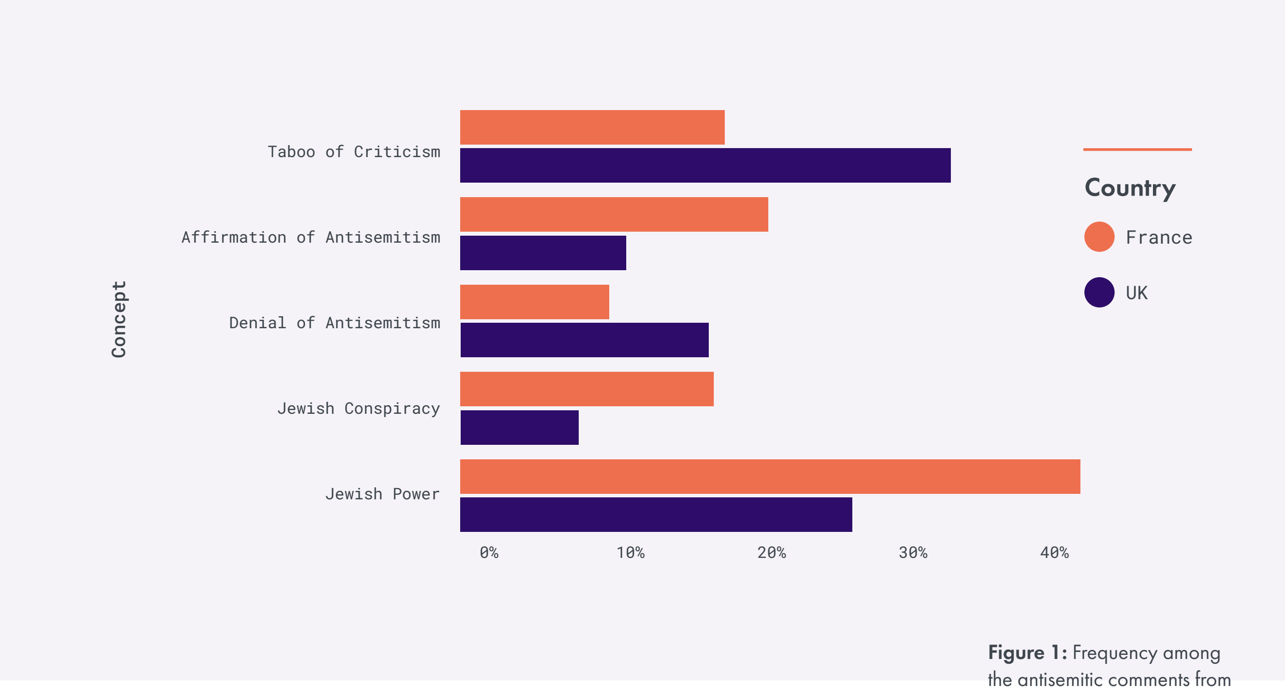

Figure 1: Frequency among the antisemitic comments from the Kanye West corpus in France and in the UK.

2.1 UK

Comment threads in the UK corpus were collected from the official social media accounts (mainly *Facebook*, with the exception of one *Daily Mail Twitter* thread) of ten major British mainstream media outlets reporting on West's statements – *BBC News*, *Daily Mail*, *The Guardian*, *The Independent*, *Metro*, *The Mirror*, *The Sun* and *The Times* – as well as two non-political entertainment publications, *OK!* and *Vice*. We selected twenty threads posted between October and November 2022, and analysed 100 comments from each. Out of the 2,000 comments, 11 % were categorised as antisemitic.

There was a marked disparity across the different outlets: antisemitic comments made up 29 % of the *Daily Mail Twitter* thread, 15 % of the three *Independent Facebook* threads, 12 % of two *Vice Facebook* threads, and around 11 % for each of *The Guardian* and *Daily Mirror Facebook* threads. By contrast, only 1 % of comments within *The Sun Facebook* thread and 4 % of the *Metro Facebook* thread were antisemitic. We can thus tentatively suggest that antisemitic reactions and support for West were a stronger feature of *Twitter* discussions, left-liberal outlets and culture-focused media than elsewhere. Across the corpus, the most frequent antisemitic concepts were AFFIRMATION of West's antisemitic comments, DENIAL OF THE ANTISEMITISM they contain, a TABOO OF CRITICISM and the accusation of JEWISH POWER, as well as the antisemitic idea of a JEWISH CONSPIRACY. As with the percentage distribution of antisemitic comments, there was a disparity across the different outlets in terms of the most common concepts as well, with the stereotype of the JEWISH POWER being particularly prominent in the *Guardian* and *Independent* threads.

The AFFIRMATION of West's antisemitism was present in 42 % of antisemitic comments and was repeatedly expressed in two main ways. The first directly referenced supposed correctness of West's words: "They can't cancel him fast enough. He's speaking too much truth for them!!" (INDEP-FB[20221009]); "I support what he says I just watched a whole interview and there was nothing I didn't agree with" (MIRROR-FB[20221021]). This type of straightforward support, without reproduction of any antisemitic stereotypes or mention of antisemitism, was particularly common. The second mode of AFFIRMATION was heroic portrayals of West. Here his stature is exaggerated to such extent that rather than West's antisemitism detracting from his reputation, the latter is used as further justification for the truth of his antisemitic statements: “Heroes don’t always wear capes. They sometimes are billionaires” (MIRROR-FB[20221021]). Some comments rejected the criticism of West as a deliberate ploy distracting from the truth of his claims, or as a cynical political weapon used to defame and silence him. They often took on the form of soundbites, such as this linguistic reversal: “truth sounds like hate, to those that hate the truth 😏” (DAILY-FB[20221019]), or “They call him Anti-Semitic but they dont call him a liar !🤔” (DAILY-TW[20221029]). This reclassification comes close to the 19th century concept of political antisemitism in its originally positive rather than pejorative connotation.

AFFIRMATION of antisemitism was often accompanied by its DENIAL or RELATIVISATION. Comments which genuinely asked why West was being accused of antisemitism were not coded as antisemitic – only those which explicitly denied the clear antisemitic content of his statements, e.g. “It’s not antisemitic to point that out” (DAILY-TW[20221029]) or questioned or negated the antisemitism while framing antisemitism as “his opinion!! Let the man live!!” (SUN-FB[20221020]). Some web users again presented West’s arguments as truth:

> “How is it anti-semitic to point out that Jews control or have power in the media? I think that it has been common knowledge for years that they do”
>
> (DAILY-TW[20221029]).

Others sought to frame the accusations of antisemitism against him as motivated by anti-Black racism – thus both implicitly denying their antisemitic content, and explicitly uttering another stereotype: “Look at all the white people in here complaining about a black man having an opinion about rich corrupt Jewish businessmen. Racist af” (GUARD-FB[20221026]).

Elsewhere, a web user ironically suggested an alternative headline to the article they commented on: “The black man that speaks out against the Jews needs to be flogged” (INDEP-FB[20221027]).

Around one fifth (21 %) of antisemitic comments revolved around the stereotypes of JEWISH POWER and INFLUENCE, inspired by West’s own focus on this idea. Many users expressed the belief that Jews dominate certain social spheres, especially media and finance: owning companies, holding monopolies, controlling the banking system and “sign[ing] the checks for everyone” (GUARD-FB[20221026]). Some users broadened their accusations, claiming the existence of a “significant overepresentation of Jews in, for example, pornography, banking, the current US cabinet, hollywood, law, etc.” (DAILY-TW[20221029]), or of a totalising JEWISH POWER, here using the three brackets, common within far-right online milieus to make an implicit reference to Jews: “(((They))) control everything” (GUARD-FB[20221026]). Jews are presented as being in charge of “even the words you’re to say.” (INDEP-FB[20221029]).

Some comments echo the TABOO OF CRITICISM trope – insisting that “The more he’s cancelled, the more his point is validated 🐑🐑🐑” (INDEP-FB[20221027]), with the sheep emoji used to criticise the public mindlessly following a media agenda, or citing a quote originating with the far right but often mistakenly attributed to Voltaire: “If you want to know who rules over you, look at who you are not allowed to criticize” (GUARD-FB[20221026]). Others pointed out the alleged existence of double standards in the treatment of Jews or Israel in comparison to other groups: “well, i guess we can say they made it clear ONLY antisemitism is intolerable, the rest can go to hell for all they care.Double triple standards as usual” (BBC-FB[20221025]), sometimes making a specific reference: “He was free to say whatever he wanted about George Floyd and whatever else.. but the moment he mentioned Jewish people he’s cancelled from everything. Why is that?!” (METRO-FB[20221025]). These and other accusations were often combined with anti-Zionist sentiment. Such comments referred to a “Zionist lobby” or “Zionist cabal” (DAILY-TW[20221029]), and asked

“Do Zionists think are are the Judges of this world ??” (TIMES-FB[20221026]).

These attributions culminate in the idea of a Jewish CONSPIRACY against non-Jews, appearing in 10 % of antisemitic comments, again using an implicit signifier: “The ‘J’ don’t like it when you call them out for controlling all the media and the banks because that means they have been caught. They want everyone who isn’t ‘J’ to be fighting with each other not knowing the reason for all their problems was started by the ‘J’” (INDEP-FB[20221009]). Other comments refer to Jews as “fanatic puppeteers” [(GUARD-FB[20221025]), reinforcing the image of a powerful group operating in secret, the “master” that should be “obey[ed]” (INDEP-FB[20221029]), or “the creeps who run the world” (INDEP-FB [20221029]). Some commenters lionised West for revealing the supposed hidden truth of JEWISH POWER: “Kanye is speaking plain verifiable truths. Simple as that. The irony is that the reaction has proved his point. They have unmasked themselves, and then some. Kanye is literally smashing the control Matrix to pieces 👊😁😎” (INDEP-FB[20221027]). Others suggested that he may be harmed as a result, positioning him as a martyr: “he might be crazy for still wanting to say the things he says even tho he has seen the result from ppl who tried in the past” (VICE-FB[20221019]), often referring to the group who would be responsible for this through coded language: “bcz his death is on the way illuminati will kill him” (MIRROR-FB[20221103]); “THE KHAZARIAN MAFIA IS TRYING TO DESTROY KANYE” (TIMES-FB[20221026). The reference to Khazars evokes an increasingly common antisemitic origin myth that presents contemporary Jewish communities as ‘fake’ or ‘imposters,’ thus chiming with West’s own Black Hebrew Israelite-influenced statements.[4] Other comments called for an action against the alleged control: “Its about time people started raising awareness about what they get up to so the masses can revolt against them” (DAILY-FB[20221020]).

2.2 France

The French corpus includes comment threads from the *Facebook* or *Twitter* profiles of ten media outlets, both political news media (*Le Point*, *Le Monde*, *Le Figaro*, *Le Parisien*, *Le Nouvel Obs*, *BFMTV*, *LCI*, *Les Echos*, *TF1*) and pop culture and tabloid media (*QG*, *Les Inrockuptibles*, *French Rap US*). 1,953 user comments were sampled and analysed, grouped into four major clusters. The first cluster of posts focuses on West’s antisemitic remarks, the second reports on the public backlash and the splits between the rapper and his partner brands, the third follows West’s subsequent escalation, in a series of statements in which he publicly supported Adolf Hitler and the Nazi regime, and the fourth covers the artist’s belated apologies. In total, 14 % or 279 comments were categorised as antisemitic.

Web users appreciative of West attempted to negotiate their support in multiple ways (while also often being openly challenged by counter speech from other users). Unlike in our previous analysis of reactions to the French comedian Dieudonné M’bala M’bala, where his defenders often used inside jokes from the comedian’s routines and shows (see Becker et al. 2021), West’s supporters do not seem to form a united “deviant community” (see Proust et al. 2020). Similar to reac-

[4] – In a now-deleted tweet, West has claimed that he “can’t actually be Anti Semitic because black people are actually Jew also” (8 October 2022). He made similar statements in an earlier *Instagram* post (“a Jew just like all so-called black people”, 6 October 2022) and a *Fox News* interview (“When I say Jew, I mean the 12 lost tribes of Judah, the blood of Christ, who the people known as the race Black really are,” 6 October 2022). See e.g. https://www.timesofisrael.com/black-people-are-actually-jews-the-origins-of-kanye-wests-inflammatory-remarks (last accessed on 27 February 2023).

tions in UK comments sections, expressions of support and AFFIRMATION OF ANTISEMITISM were often articulated in simple, straightforward ways, such as "Sending all my support to Kanye" ["Tout mon soutien a Kanye"] (LCI.F-FB[20221026]) or "More power to him 💯💯👏👏👏" ["force à lui 💯💯👏👏👏"] (BFMTV-FB[20221027]). Emoticons and icons such as hearts or clapping hands further suggested approval and praise. As in the UK corpus, some users highlighted West's artistic talent in order to present him as a misunderstood visionary and rebel who understands better than anyone the inner workings of the society "More power to Kanye West. A misunderstood genius against the thought control" ["Force a Kanye West. Un génie incompris face au dicta de la pensée"] (BFMTV-FB[20221027]);

> **"He's a true artist that does not bow to political correctness because he's a great man"**
>
> ["C est un véritable artiste s il ne s inscrit pas dans la bien-pensance c est qu il est un grand homme"]
>
> (BFMTV-FB[20221030]).

West's antisemitic statements are interpreted as a refusal to sell his "dignity" for money and fame. Therefore, a distinctive feature of the French corpus was the comparison of the rapper's behaviour with the perceived SERVILITY of other mainstream French entertainers or artists from minority backgrounds: "At least Kanye didn't bend over like those wet rags, such as [the comedian Jamel] Debbouze, etc." ["Au moins Kanye a su garder son pantalon contrairement à tt ces serpillières debouze etc"] (BFMTV-FB[20221027]). West's supposed nonconformity irks the 'powers-to-be' and exposes him to retribution, as he denounces an alleged conspiracy: "Ye's free speech and indomitable spirit is an obstacle to the conspiracy" ["Cette liberté d'expression et d'esprit de YE dérange la théorie du complot"] (BFMTV-FB[20221027]). Some commenters use the slogan "Je suis Kanye," echoing "Je suis Charlie," painting West as a martyr of freedom of speech and of conscience – and his opponents indirectly as representatives or stooges of brutality and terrorism. In doing so, they implicitly AFFIRM West's ANTISEMITISM.

Casting West as an icon of free speech assailed by the establishment maps onto the antisemitic stereotype of the TABOO OF CRITICISM. A web user states that

> **"strangely, only those who criticise the J *** are done away with, treated worse than murderers or child rapists!!!"**
>
> ["Bizarrement il n'y a que ceux qui critiquent les j* * * * qui finissent au placard, présenté comme des assassins pire que les vrais violeurs de gosses !!!"]
>
> (LEPOI-FB[20221025]).

The verb "criticise" legitimises and indirectly reclassifies the antisemitic statements, while the adverb "strangely" triggers vague conspiracist associations: the writer feigns surprise at the supposed preferential treatment of Jews as compared to other minorities. Implicatures, mock-naivety and "just asking questions" are traditional elements of the conspiracist discursive grammar. The Jewish community is rarely mentioned directly, but is regularly alluded to through phrases such as the "chosen people", "the untouchables" or claims that "there are first-class citizens, and those who are not permitted to look at them or talk about them" ["Il y a les citoyens de première zone et ceux qui ne peuvent les regarder ou parler d'eux"] (LEFIG-FB[20221025]). This hyperbole is meant to highlight a radical asymmetry of power and feeds into the populist dichotomy between an alleged corrupt (possibly Jewish) elite and the pure people, kept in the dark and denied dignity and rights. Drawing on the geopolitical situation, Kanye West is even compared to the Russian president Vladimir Putin – who is said to also have been demonised and made into a pariah by a corrupt, Jewish-led media system: "When Putin spoke up, he was hated and is still hated.

Ye is being crushed by the same juggernaut. A single community wants to rule over everything" ["Quand poutine a dit cela on l'a détesté et le déteste actuellement YE ne fait que subir ce même rouleau compressor Une seule communauté veut faire le dictat"] (BFMTV-FB[20221027]).

As with the UK corpus, these concepts culminate in broad allegations of a conspiracy:

> "The dude spoke against the world order, he's getting shot down by the rulers of our rulers. Those who are in control, and who, it seems, are the ones insulted in the past, right?"
>
> ["Le mec a parlé contre l'ordre mondial, il se fait abattre par les dirigeants des dirigeants. Ceux qui contrôlent, et qui, paraît ils d'ailleurs sont ceux insultés dans l'histoire non?"] (BFMTV-FB[20221027]).

The biblical topos of the "chosen people" is again evoked in support of the alleged Jewish world domination projects: "This is what happens when one tries to control humanity, thinking they are Gods or the chosen people!...Well, one must expect people to revolt!" ["Lorsque on essaye de contrôler l'humanité, de se prendre pour dieu ou pour le peuple élu.. ! Beh il faut s'attendre à ce que les gens se révolte !" (LEPOI-FB[20221025]). Jews are accuses of pulling the strings of mainstream media outlets, but also of controlling the cultural industries, thus achieving almost complete control over public opinion. West's exclusion is thus easily explained by the fact that "it's them who control the world, the industries, Hollywood, the banks, everything [...] even social media" ["c'est eux qui contrôlent le monde les industries Hollywood les banques tous quoi [...] même des réseaux sociaux !!"] (LEPOI-FB[20221025]). Users rely heavily on innuendo to convey antisemitic meaning, for instance through the deictic term "they", often used in the language of conspiracy to imply powerful shadowy forces. Sometimes, typographical practices indicate to other users 'in the know' that the phrase refers to Jews, as in the following comment: "And then we're told it's not true that (((they))) don't control everything" ["Et après on nous dit que ce n'est pas vrai, (((ils))) ne contrôlent pas tout"] (BFMTV-FB[20221027]).

Antisemitic innuendo also relies on popular culture, such as comics or manga. For instance, Jews are repeatedly referred to as "celestial dragons," a race of evil superhumans from Eiichiro Oda's manga series *One Piece*: "He's right, he didn't say anything bad and you know that, but as they are celestial dragons..." ["Il a raison ya rien de grave et tu le sais mais comme ce sont les dragons céleste..."] (FRENC-TW[20221026]). This type of dog whistles bears the risk of bringing antisemitism to a new, younger generation, who might not be initially very politicised. Another transparent dog whistle, one peculiar to French political discourse, is the rhetorical question "But WHO?" ("Mais QUI?"), which emerged during the Covid-19 pandemic to suggest it had been orchestrated by Jews for population control and for profit (Ascone et al. 2022). It has since become a universal catchphrase that hints at alleged Jewish schemes within informal discourse. Some web users apply it to the West affair, sometimes in combination with other dehumanising antisemitic stereotypes: "Dogs and rats work together in all the Western societies, apparently. An elite of paedocriminals controls our governments and our media... They all obey the orders... But whose???" ["Les chiens et les rats se mettent d'accord dans toutes les sociétés occidentales apparemment. Une élite de pedocriminels qui décident dans nos gouvernements et nos médias..... ils sont tous 'aux ordres' .... mais de qui ????"] (BFMTV-FB[20221027]). The reference to paedophilia fits into the broader QAnon narrative, which itself subtly echoes blood libel accusations (Friendberg 2020).

3. Antisemitic incidents at the 2022 FIFA World Cup (UK)

The Palestinian flag was a conspicuous presence throughout the 2022 World Cup in Qatar. In the stands, fans – particularly, but not exclusively, from the surrounding Arab and North African region – waved the flag and sang pro-Palestinian songs and chants. On the pitch, the Moroccan team displayed the flag when celebrating both their victories. Outside the grounds, there were numerous reports of Israeli journalists being shunned or harassed by fans. The incidents were much discussed on *Twitter*, with threads following tweets by journalists and independent commentators gaining hundreds, and at times thousands of responses. The ten threads selected for analysis came from a variety of well-followed accounts, including the ESPN.co.uk chief football writer Mark Ogden (268,000 followers), to Palestinian policy analyst Dr Yara Hawari (82,000 followers) and the American-Israeli philanthropist Adam Milstein (171,000 followers). We coded the first 125 comments from each thread, giving a total of 1,250 comments. Across the corpus as a whole, 10 % of comments were categorised as antisemitic. The most frequent antisemitic concepts were APARTHEID ANALOGY, DENIAL OF THE JEWISH RIGHT TO SELF-DETERMINATION and EVIL, see Figure 2.

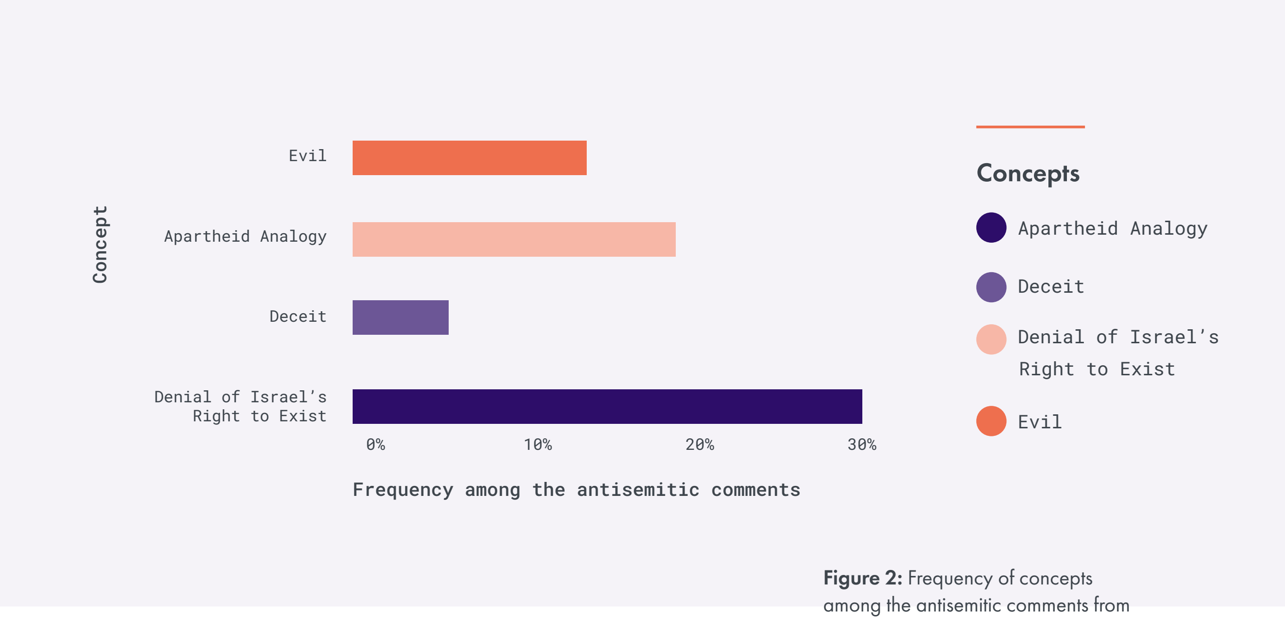

Figure 2: Frequency of concepts among the antisemitic comments from the World Cup corpus in the UK.

The threads contained frequent references to Israel as an "apartheid regime" (TW-JENNINE[2022 1126]), "Zionist apartheid" (TW-OGDEN [20221206]), or "bloodthirsty apartheid state" (TW-JASKOLL[20221126]). Several comments attempted to give the APARTHEID ANALOGY more substance by saying the view is widely shared by "the general public" who "wholeheartedly dislikes apartheid Israel" (TW-JASKOLL[20221126]) or appealing to the authority of the international community and institutions: "Israel is a systemic apartheid regime no different to White South Africa rule. 98 countries agree at the recent UN resolution" (TW-JASKOLL[20221126]). When antisemitism of such statements was pointed out, the reaction was often contempt or hostility: "It's weird that your surprised that no one outside their apartheid bubble likes bullies and mass murderers" (TW-JENNINE[20221126]), here compounded by a NAZI ANALOGY:

> "You're losing your mind trying to defend a genocidal apartheid regime.. step out of your Zionist echo chamber for a second and you'll realize that you've long lost the public's opinion. 😂You're the oppressor and nobody likes or respects you. Modern day Nazis, but weaker"
>
> (TW-JASKOLL[20221126]).

Israel's legitimacy was called into question through its description as a "zionazi entity" (TW-HARAWI [20221127]), "Israeli LLC company" (TW-MILSTEIN[20221127]), "not even a country" (TW-CARTER[20221210]) or, ominously, "a 73 years old colony that won't be a thing before it turns 100" (TW-AMRO[20221206])".

Other comments positioned Israel as prototypically EVIL and an enemy of humanity, arguing that "Hatred for Israelis will come automatically if you are a human being" (TW-MILSTEIN[20221127]). Some comments made the age-old allegation of propensity to CHILD MURDER – with a blunt "baby killer bitches" (TW-OGDEN[20221206]), sarcastic "u Got bullied in kid so u endroce killing of babies. How cute mr kleine schwanze"[5] (TW-CARTER[20221204]), or defiant "Murder of children and illegal occupation is wrong and you can label me ANYTHING YOU WANT for saying that" (TW-JENNINE[20221126]). On occasion, this claim was combined with further analogies between Israel and Nazism: "Look at what Israel is doing to little children and families every day that can't live in peace then think who the real nazis are" (TW-MILSTEIN[20221127]), or even Jews and Nazism: "How do you defend killing thousands of children. Displacing them from homes and butchering them in the streets. (...) Jews are worst than Nazis. 21 st century holocaust" (TW-AMRO[20221206]).

When reacting to fans refusing to be interviewed by an Israeli reporter, commenters often presented the latter as DECEITFUL, claiming that "these israeli journalists are there for the clickbait. These are the exact interactions and answers they're looking for so that they can play victim and cry wolf" (TW-HARAWI[20221127]), or that "[t]hey're on a mission to report antisemitism to support and further push their own agenda (TW-JENNINE[20221126]), looking to "create an entire 'the whole world hates us' anti-Semitic narrative" and "try to get themselves physically assaulted" (TW-HARAWI[20221127]), in other words – using fraudulent behaviour in order to INSTRUMENTALISE ANTISEMITISM for their own gain.

[5] – The German-language insult at the end of the comment plays on the name of the user it replies to. It is characteristic of the discourse to mix serious antisemitic accusations with petty personal jibes, while direct antisemitic slurs are largely absent (possibly as a result of moderation).

4. The Israeli elections in November 2022 (Germany)

On November 1, 2022, the legislative elections in Israel were held and the conservative Likud party, led by Benjamin Netanyahu, won the majority of votes. While German media started reporting about the elections in Israel well before the election day, articles that triggered most of the comments highlight the presence of a right-wing or even radical right-wing government in Israel. The media headlines often focused on the ideology of the winning parties with terms such as 'right-wing conservative', 'rightward shift', 'a national-religious alliance' – descriptions which also appeared in most of the comments critical of Israel, including the antisemitic ones.

We collected comment threads from websites of the media outlets *Zeit*, *Spiegel*, *ZDF*, *DW*, *Arte*, *Welt* and *Tagesschau* and their pages on social media platforms *Facebook*, *YouTube* and *Twitter*. The corpus of 19 threads and 2,111 annotated comments had almost 7 % antisemitic comments. There was also a relatively high percentage of 4 % of comments expressing counter speech that were obviously referring to comments deleted by moderation. The high number of deleted comments and counter speech statements indirectly proves the intensity with which this discourse event triggers antisemitic (or generally hateful) comments. The antisemitic concepts found in the annotated threads are EVIL, the idea that JEWS HAVE NOT LEARNED FROM THE PAST, the alleged TABOO OF CRITICISM, NAZI ANALOGY, APARTHEID ANALOGY and BLAMING JEWS FOR ANTISEMITISM, see Figure 3.

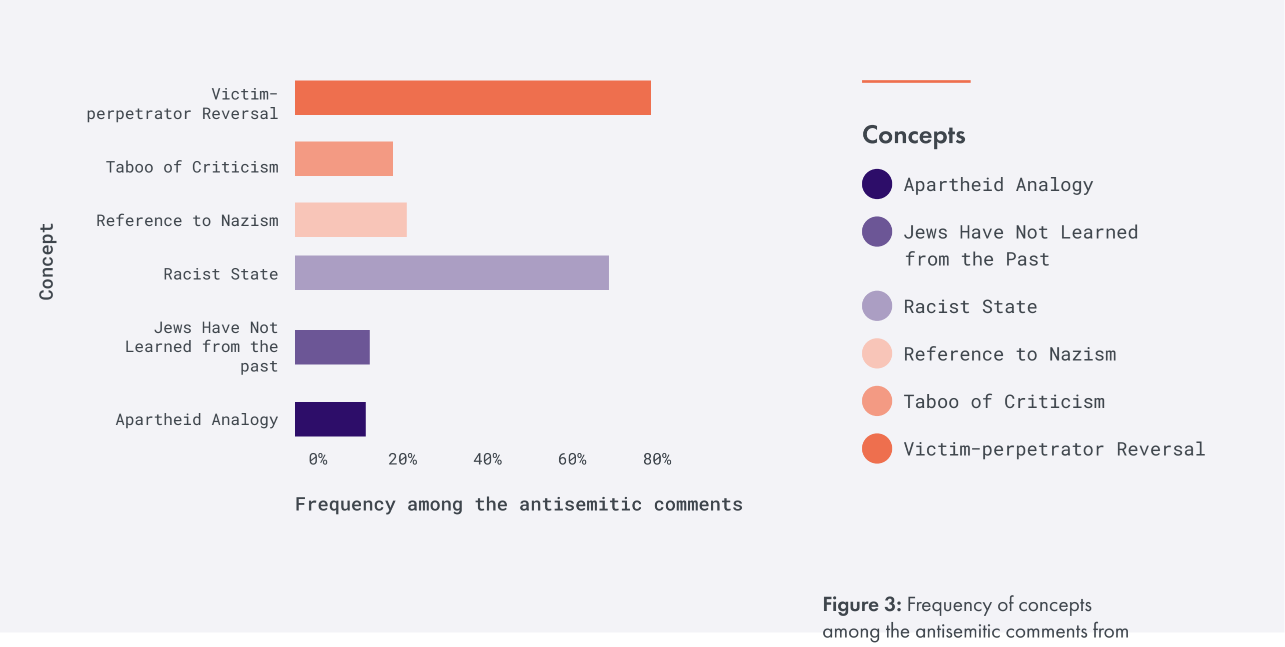

Figure 3: Frequency of concepts among the antisemitic comments from the Israeli Elections corpus in Germany.

The expression of the antisemitic stereotype of EVIL is connected to the image of Jews being bloodthirsty and acting without any rational reason. In line with this concept, a comment states: “Well, Israel is so busy at the moment throwing bombs en masse again and coaching, monitoring and accompanying their drops” [“Naja, Israel ist ja auch derzeit derart beschäftigt wieder mal Bömbchen en masse zu werfen und deren Abwürfe zu coachen, monitoren und begleiten”] (TAGES-FB[20221215]). Furthermore, EVIL appears combined with other concepts – as in the following example with an implied VICTIM-PERPETRATOR REVERSAL.

> “Ghettos are built and it is decided according to daily mood how many scraps and development flows into the ghetto. If there is a rebellion, the supply is cut off and a few houses are bombed as an example”
>
> [“Ghettos werden errichtet und es wird nach tagesstimmung entschieden wieviel Almosen und Entwicklung ins Ghetto fließt. Wenn aufgemuckt wird, wird die Versorgung abgeschnitten und exemplarisch ein paar Häuser zerbombt”] (ZEIT[20221101]).

This commenter implicitly claims that Israel would repeat the crimes of Nazi Germany. This idea is activated by recourse to world knowledge and triggered on the one hand by the allusion "ghetto," on the other hand by phrases such as 'giving scraps' and 'bombing houses as retaliation.' An alleged disproportional reaction to rebellion additionally evokes the stereotype of JEWISH VENGEFULNESS.

The concept of EVIL can equally be linked to the notion of TABOO OF CRITICISM that would supposedly emanate from Jews and Jewish institutions. The following comment adds an unfounded allegation to the critique laid out in the media articles on ultra-Orthodox parties and, by doing so, connects the latter with the antisemitic ideas of evil and taboo:

> “This obviously applies only to Israel. There, evil may not be called ‘evil’ directly. Just as ‘ultra-religious’ people in other countries are called backward, religious fundamentalists”
>
> [“Gilt offensichtlich nur für Israel. Da darf das Böse nun mal nicht direkt ‘böse’ genannt werden. So wie auch ‘Ultra-Religiöse’ in anderen Ländern als rückständische, religiöse Fundametalisten bezeichnet werden”] (ZEIT[20221104]).

The comments drawing on the concept JEWS HAVE NOT LEARNED FROM THE PAST are sometimes surprisingly explicit: “Nothing learned from their own ancestral history” [“Nichts aus der eigenen Geschichte ihrer Vorfahren gelernt”] (ZEIT-IG[20221102]). Sometimes, however, the concept lies behind an expressed concern about the ‘true’ danger: “If the radical forces come to power, we can only hope that the West and especially America has the integrity to prevent a genocide by the people who should know from history what it means to be persecuted because of absurd ideologies” [“Falls die Radikalen Kräfte an die Macht kommen bleibt nur zu hoffen das der Westen und vor allem Amerika die Integrität hat einen Genozid zu verhindern von dem Volk was eigentlich historisch wissen sollte was es bedeutet verfolgt zu werden aufgrund von absurder Ideologien”] (ZEIT[20221101]).

Likewise, users hide this stereotype behind expressions of compassion and concern:

> "I sometimes wonder about the Israelis, even though they have experienced so much suffering themselves, how much suffering they bring themselves, here too you see, as with all cultures, that people are not able to learn from history"
>
> ["Ich wundere mich manchmal über die Israelis, obwohl sie selbst so viel Leid erfahren haben, wieviel Leid sie selbst bringen, auch hier sieht man, wie bei allen Kulturen, dass die Menschen nicht in der Lage sind aus der Geschichte zu lernen"] (ARTED-YT[20221229]).

The following comment fits the stereotype of BLAMING THE JEWS FOR ANTISEMITISM as it construes the connection between the election results and hate towards Jews: "The headquarters of antisemitism makers is Netanyahu's office in Jerusalem" ["Die Zentrale der Antisemitismusmacher ist Netanyahu's Büro in Jerusalem"] (SPIEG[20221113]). This argumentation is, however, inspired by the (misunderstood) message from the media. On 17 November, *Spiegel* promoted an article on *Twitter* with a post declaring: "If a radical right-wing government comes to power in Israel, a new wave of antisemitism threatens – against Jews in Europe and Germany" ["Wenn in Israel eine rechtsradikale Regierung an die Macht kommt, droht eine neue Welle des Antisemitismus – gegen Juden in Europa und Deutschland"] (SPIEG-TW[20221113]). The article itself argued that the new government's ideology would be used as an alibi for hatred towards Jews. By making a connection between Israeli politics and European antisemitism (thus, the former misused as a justification for the latter), the tweet left itself open to web users' interpretations which sought to blame antisemitism as such on Israeli and – more broadly – on Jewish behaviour.

Accordingly, the stereotype that a TABOO OF CRITICISM would protect Israel from any kind of justified critique takes both explicit and implicit forms. Complaints about the wrongly understood critique are typical for this stereotype: "Because every criticism of Israel is immediately understood as antisemitism" ["Weil jede Israel-Kritik gleich als Antisemitismus verstanden wird"] (ZEIT-IG[20221102]). Furthermore, comments representing the fear of a threat, potential SELF-VICTIMISATION, and desire to deliver 'subversive critique' also rely on the stereotype of TABOO OF CRITICISM and, as in the following example, assume that criticising Israel leads to a danger of being abducted by a secret service:

> "No idea, but we don't want to say anything wrong and then be picked up by the terror commando. You always hit sore points with this topic"
>
> ["Keine Ahnung aber wir wollen ja jetzt auch nichts falsches sagen und dann gleich vom Terrorkomando abgeholt werden. Man trifft bei diesem Thema ja dauernd wunde Punkte"] (SPIEG-TW[20221113]).

The comment's vaguely conspiratorial tone (created by not even naming where the danger comes from) is used to underline the possibility of such danger. At the same time, as there is no direct mention of Jews or Israel, the meaning of the comment is contextual. Following similar patterns, such statements can appear supporting other antisemitic concepts such as REFERENCE TO FASCISM, and constructing the guilt for hatred in the object of hate – in this case Netanyahu: "antisemitism... let me guess, it's... when you call a fascist who happens to be of Jewish faith a fascist!" ["antsemitismus,,,lassen sie mich raten ist,,,, wenn man einen Faschisten der zufällig jüdischen Glaubens ist,, einen Faschisten nennt!"] (ZEIT[20221101]).

Openly comparing or even equating the Jewish state to Nazi Germany – a dictatorship where Jews were systematically targeted – represents a form of victim-perpetrator reversal as well as a form of blatantly trivialising Nazi crimes. In order to avoid sanctions for expressing such antisemitic ideas, commenters obfuscate their statements and draw on similarities between historical fascist and Nazi scenarios and Israel today (paralogism). The Nazi analogy is often hidden in longer descriptions of the German historical context that are indirectly used in order to justify the comparison between both scenarios:

> "After all, Germans elected Hitler too, only with about 42 % and that also varied regionally. The Israeli people are very heterogeneous, but have been voting with increasingly nationalistic, racist tendency for years"
>
> ["Das deutsche Volk hatte Hitler auch gewählt ,immerhin nur mit ca 42% und das auch regional sehr unterschiedlich .Das israelische Staatsvolk ist sehr heterogen ,wählt aber seit Jahren mit immer nationalistischer ,rassistischer Tendenz"] (WELT[20221102]).

Some commenters approached the analogy more openly, by using rhetorical questions and open allusions: "Israel should revisit Germany's history and ask itself if this is the right way to go?" ["Israel sollte sich die Geschichte Deutschlands nochmals vor Augen führen und sich fragen, ob dies der richtige Weg ist?"] (ZEIT[20221216]) (on allusions and indirect speech acts in the context of the Nazi analogy, see Becker 2021).

Although sometimes combined with stereotypes such as taboo of criticism, the apartheid analogy is always expressed clearly: "just criticise it properly when people oppress other people, Israel is an apartheid" ["einfach mal richtige Kritik üben wenn Menschen andere Menschen unterdrücken Israel ist eine Apartheid"] (SPIEG-TW[20221113]). The object of antisemitic hatred in the apartheid analogy is often Israel, but it can also be some other Jewish institution or person. Likewise, it is used in the context of calls to action in which users presuppose a policy of segregation in Israel: "end apartheid & zionism!" (ARTED-YT[20221229]). Consistent with the idea behind this analogy is also the statement that Israel has been driven by racism: "The main thing is that racism and apartheid work. Then it's all good" ["Hauptsache Rassismus und Apartheid funktionieren. Dann ist ja alles gut"] (ZEIT[20221101]). Furthermore, these utterances are not always only against Israel, but also demonstrate support for Palestine: "Good luck and strength to the Palestinians in resisting this terrorist apartheid state. 🙏" ["Viel Glück und Kraft den Palästinensern beim Widerstand gegen diesen terroristischen Apartheidstaat. 🙏"] (ZDF-YT[20221229]).

Our analysis shows that the often critical media coverage of the Israeli politics – and especially elections – in the German media is regularly used as a trigger for demonising Israel and spreading antisemitic ideas in the comment sections.
Commentators invoke values ascribed to democratic societies, while at the same time projecting their arsenal of antisemitic concepts – such as institutionalised racism, apartheid, fascism or Nazism – onto the Israeli state. These attributions are used to delegitimise Israel as an undemocratic state and society and marginalise it globally. In cases such as the *Spiegel* tweet, the antisemitic approaches were possibly strengthened by the ambiguity in media coverage.

5. Towards the automatic detection of antisemitic discourse online

In the third quarter of 2022, *Meta* reported that it had taken action on 10.6 million pieces of content considered to be hate speech on *Facebook*. Of these posts, over 90 % were found and acted on proactively, 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 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, e.g. related to misogyny, racism and antisemitism. Text classification is a core task of Natural Language Processing (NLP), the computer-based processing of large amounts of natural language data. Historically, individually formulated rules targeting particular textual aspects were used to perform tasks such as text classification; however, modern approaches leverage machine learning for superior results. This entails feeding large datasets into algorithms which learn patterns in the texts to accurately predict classes for new, unseen data. Common applications of classification, some of which we use on a daily basis, include sentiment analysis, language identification, or spam detection.

One of the most significant challenges in text classification is the adequate operationalisation of the task. This includes determining what classes to use, deciding what constitutes a text, how to preprocess it, and at what granularity the classification should be performed (e.g. at document, paragraph, or sentence level).

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 not only to 'teach' the algorithm but also 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 are so far conducted by only a handful of researchers (Jikeli/Cavar/Miehling 2019, Chandra et al. 2021, Jikeli et al. 2022, Steffen et al. 2022), and have not yet resulted in datasets comparable to available corpora for related phenomena such as offensive language, toxic language, and other forms of hate speech.

Challenges of text operationalisation

In order to enable computers to process human language, numerical representations of the data must be generated. This encoding process, however, is challenging, as it strives to maintain as much information as possible. A simple approach known as a bag-of-words model involves representing a text by counting how often each word occurs in it. Training classification models based on this text representation can produce satisfactory results, so long as the task does not require a more complex understanding of the narrative and context. In more complex cases, it is rather employed to build baseline models that serve as reference points for future improvements.

To increase the performance of classification models, we require text representations able to capture the semantic dimension of human language such as similarity of words and concepts, and thus contextual information. Capturing the meaning of texts written by humans can be a challenging task, in particular for short messages, which are commonly found in online and social media communication. Authors may use subtle, coded, implicit expressions of their opinions, for instance to attain a certain level of ambivalence in order to 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 denial (Falkenberg/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 recent statement of 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.[6]

An issue closely related to capturing information from text is the amount of data. Labelling data is typically time- and cost-consuming, and often requires experts to execute the work, as in the Decoding Antisemitism project. This poses significant challenges as algorithms are expected to learn manifold levels of interaction between words from relatively small amounts of data. In practice, this does not yet yield satisfactory outcomes, resulting in models that tend to perform poorly when applied in scenarios (slightly) different from the training situation. In our context, this would mean that a model trained on the existing corpus might show a (significantly) decreased performance when confronted with examples of antisemitic speech in a novel discourse.

[6] – 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?," and then uses the cookie analogy for several statements of Holocaust denial. For the respective livestream scene see https://mobile.twitter.com/CalebJHull/status/118959437103
0695937 (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). In case you wonder whether ChatGPT would spot the antisemitic character of this statement: no, it would not.

Transfer learning with transformer architectures

In recent years, two major developments have been instrumental in addressing the challenges posed by the supervised learning paradigm and context-unaware language representations: (1) transfer learning, a paradigm in which knowledge acquired from solving one task is transferred to another, potentially more difficult, task and (2) transformer architectures, which are capable of capturing intricate contextual information in text.

Transfer learning involves training a model to solve a 'source' task, then adapting it for a sufficiently similar 'target' task. This is especially useful when there is little labelled data for the target task, as is often the case with classification of texts in the political sphere, but a vast set of training data for the source task. In the NLP domain, this is exemplified by the use of large digital corpora such as *Wikipedia* or *Google News* for the rather generalistic primary task of predicting a next or missing word from the preceding or surrounding context, respectively. Such language models,[7] trained without any human labelling,[8] are even able to capture a variety of linguistic phenomena such as word- and sentence-level semantics, syntactic structures and discourse-level phenomena from their training data, as well as subtleties of human language like sarcasm or slang.

Once a language model has been trained, it can be fine-tuned for various use cases, such as classification of texts into 'hate speech' and 'no hate speech'. In essence, the classification model 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.

Most recently, so-called **transformer architectures** have been leveraged to build language models that solve the source task. Transformers represent a clear shift from prior model architectures relying (entirely) on a mechanism called (self-)attention, which allows them to represent each word with respect to its current context (cf. Vaswani et al. 2017). This enables them to learn how words relate to each other, even across long distances within a text.

A plethora of pre-trained language models are available for fine-tuning for different downstream tasks including text classification. These language models differ in aspects such as data source used for training (e.g. *Wikipedia* vs. *Twitter*), language, architecture (e.g. type and number of layers[9]), or pre-processing of the text (e.g. lowercasing all words).

[7] – A language model computes how likely a given sequence of words will appear in a given language such as German or English. It can be used to predict the next word in a sequence and thus to generate text.

[8] – If we remove a word from a sentence, the rest of the sentence serves as the context to predict the missing word, thus transforming originally unlabeled data (e.g. sentences from Wikipedia) into labelled data. This type of learning is known as self-supervised learning.

[9] – Transformers, and neural networks in general, are organised in layers that consist of computational units typically called neurons, which connect inputs and outputs of the model. There are various types of layers addressing specific computational needs, and a neural network architecture can be composed of a varying number of layers. For instance, the widely used BERT language models consist of so-called transformer blocks which can be decomposed into other layers such as self-attention and normalisation layers. The base variant of BERT consists of 12 transformer blocks, while BERTlarge has twice as many. An architecture with more layers is, in general, more complex (and 'more deep') and consists of more parameters that need to be learned during training, thus requiring more data for training.

One of the most popular architectures employed is BERT, which has achieved the state of the art for a range of NLP applications. BERT-like pre-trained language models are typically used in current 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/Ruppenhofer 2018, Zampieri et al. 2019 and 2020, Mandl et al. 2021), or (pre-specified) conspiracy theories (Pogorelov et al. 2020, Moffitt/King/Carley 2021, Elroy/Yosipof 2022, Phillips/Ng/Carley 2022). The majority of these benchmark datasets is in English language, and heavily focused on *Twitter* as data source (cf. Poletto et al. 2021) which is due to the platform's popularity but also the easy technical access for researchers to the data. Antisemitism has not yet been addressed in many efforts for text classification.

Services for content moderation

The lack of large annotated corpora results in a lack of services for the automated detection of antisemitic content. However, progress has been made regarding production-ready web services for the recognition of other linguistic phenomena intersecting with antisemitism, such as hate speech and toxic language. 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, e.g. for analyses of moderation measures on *Reddit* (Horta Ribeiro et al. 2021), investigations of political online communities on *Reddit* (Rajadesingan/Resnick/Budak 2020) and *Telegram* (Hoseini et al. 2021), and for identifying antisemitic and Islamophobic texts on *4chan* (González-Pizarro/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 threat. The scores are computed by machine learning models[10] 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. For instance, content is supposed to be labelled as toxic if it is considered "rude, disrespectful, or unreasonable [...], likely to make people leave a discussion" (Thain/Dixon/Wulczyn 2017, Google 2022). To counteract the subjectivity and vagueness of the definition, texts are labelled by multiple individuals and their assessments are aggregated before training models.

In theory, *Perspective API* could provide an easily accessible approach to detecting certain forms of antisemitic speech. However, recent work on German-language communication on *Telegram* and *Twitter* indicates certain limitations when using the service for this task, namely an oversensitivity to certain 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/Steffen 2022). It has furthermore been found that the service performs rather poorly on more subtle or encoded forms of antisemitism, often failing to recognise them as toxic (ibid.).
To obtain a more comprehensive picture, we ran the *Perspective API* on a part of this project's multi-lingual data, consisting of around 3,500 comments manually labelled as antisemitic[11] and around 53,500 texts labelled as not antisemitic, yielding 57,021 records in total. We evaluated the scores for the attributes 'identity attack,'[12] 'toxicity,' and 'severe toxicity.' We spe-

[10] – For more details on the transformer-based architecture, see e.g. https://arxiv.org/pdf/2202.11176.pdf.

[11] – Texts labelled as contextual antisemitism have been excluded from the dataset because the service predicts scores only for the text itself and is not able to consider additional contextual information.

[12] – Referring to "negative or hateful comments targeting someone because of their identity" (Google 2022).

cifically 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.

The distributions of all three scores differ significantly between the two groups of antisemitic and non-antisemitic texts, as visualised in Figure 4, with clearly higher scores for antisemitic texts. 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 various existing studies 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, 75 % of antisemitic comments fall below 0.8 and would have been missed by the above-mentioned research designs.

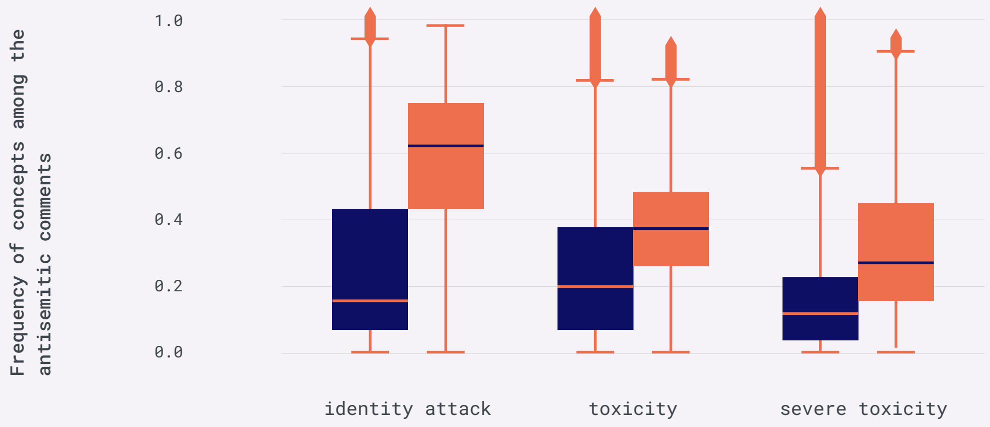

Figure 4: Distributions of scores for identity attack, toxicity, and severe toxicity, split according to antisemitism label of the data. The horizontal lines of the boxes indicate the lower quartile (25 %), the median (50 %), and the upper quartile (75 %) of the scores.

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' , i.e. 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).

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'. Figure 5 visualises how the scores are distributed if we take this additional information into account: Comments containing identity-related keywords (orange 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 scored as an identity attack. Although the presence of respective keywords alone does not account for a high identity attack score[13] (see e.g. the first column in Table 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 (0.15) is less pronounced. Similar effects can be observed for the other two *Perspective API* attributes.

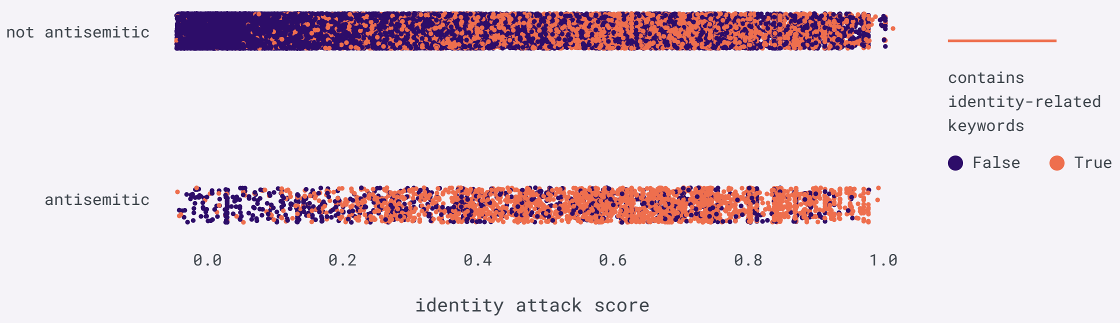

Figure 5: Identity attack scores broken down by text label and presence of identity-related keywords.

[13] – This is also not to be expected as the *Perspective API* models utilise far more 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	0.152659 (N=45761)	0.585239 (N=7769)	**+0.43258**
texts labelled as antisemitic	0.492150 (N=969)	0.642324 (N=2522)	**+0.150174**
difference	**+0.339491**	**+0.057085**	

Table 1: Median identity attack scores per class label and depending on the presence of identity-related keywords. Group sizes are displayed in brackets.

This analysis does not provide a causal relation between the occurrence of keywords related to Jewishness and the state of Israel and higher scores. It is plausible, for instance, that texts discussing Israel or Jewishness can be toxic or otherwise abusive without being antisemitic. This means that it may not be the keyword-related false positive bias, but different aspects of the texts that produce the high score. However, prior research has shown that adding these keywords significantly increases the scores of texts (Mihaljević/Steffen 2022), which confirms the keyword bias. Further exploration of results is needed, and for this, our future research will include the examination of API scores on a span level to examine which segments of a text trigger the API.

The experiments so far present the *Perspective API* service as rather limited for moderation of content with regard to antisemitic statements and for research on online antisemitism. The significant positive correlation of identity-related keywords with higher scores suggests a higher risk of falsely considering a text as antisemitic, while the presence of antisemitic codes can substantially hinder the identification of antisemitic toxic content. The latter 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 incorporation of antisemitic types of toxic content, given the difficulty even for experts in labelling short texts as are typical for online and social media communication. Thus, automatic detection of antisemitic speech is still needed and requires careful modelling based on high quality labelled data.

Initial experiments using transfer learning and transformer architectures

Previous work in this project explored the training of a logistic regression classification model (a classical approach from statistics) based on bag-of-words text representations to distinguish antisemitic from non-antisemitic texts (Ascone et al. 2022). One great advantage of using these simple models is their higher level of model transparency, allowing one to easily find out which words contribute most to the decision for each of the two classes. The word with the strongest contribution to the class of antisemitic texts is 'apartheid' (weight=13.72), followed by 'genocide' (weight=9.24). Among the twenty most

[14] – In fact, the sum of the weights of the two variants 'israhell' and 'israhel' is higher than that of 'apartheid.'

influential words for predicting a text to be antisemitic are also the codes 'israhell'[14] and 'satanyahu,' words related to violence, such as 'murderers.' the word 'lobby,' indicating conspiracy theories, or the word 'devil,' a demonising rhetorical element. While these words are reasonable for the given corpora, they also indicate potential issues with the model: for instance, a sentence defining 'apartheid' would be predicted as antisemitic, with very high probability. Additionally, a bias for the word 'Israel' is noticeable, with sentences containing it having a higher chance of being labelled as antisemitic.[15] Nevertheless, the described model, trained on a part of the current English-language corpus, achieved first promising results[16] that can be used as a starting point for further developments.

We approached the task by fine-tuning transformer-based language models for a classification task, as described previously. We decided to make some adaptations regarding the assignment of texts to classes: the Decoding Antisemitism project distinguishes between texts whose antisemitic character can be detected without further information and those that are 'contextually antisemitic,' i.e. additional context such as the content behind a linked URL, information from previous comments or the reader's world knowledge is required to recognise antisemitic content. For instance, the comment "I think you have been told to do this" cannot be fully interpreted without resolving the ambiguity of what 'this' and 'you' refer to. A machine learning model would need this information, too, in order to make correct inference, but providing it is not trivial in a practical application scenario. While a human annotator (or a content moderator) can usually fully resolve such ambiguities – namely that the user claims that another user would express themselves in a certain way due to an imagined Jewish influence – this poses a non-trivial challenge when attempting to automate the task. Thus, we consider only texts labelled as antisemitic without requiring additional contextual information to avoid the necessity to resolve potential ambiguities. However, this has the disadvantage that our already imbalanced dataset, with about 85 % of comments being labelled as not antisemitic (negative class or class 0) becomes even more skewed, with only about 10 % of texts annotated as antisemitic (positive class or class 1).

There are various approaches to dealing with strongly imbalanced data during training, such as downsampling the majority class, augmenting the minority class (e.g. through small variations of existing texts), or placing stronger penalties on errors for the minority class. To explore the influence of additional aspects, we also consider the choice of the pretrained language model,[17] standard hyperparameters for fine-tuning transformer models (e.g. learning rate and attention dropout), and data-related settings (e.g. handling of particularly short texts and removal of emojis). These hyperparameters determine the overall capabilities of a machine learning model, so combinations of different values are evaluated to find the optimal one. 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.

[15] Even totally harmless ones like "Israel is a country rich in culture and history, and its vibrant cities are full of life and energy."

[16] In the previous report, an F1 score of 0.75 was reported. However, we replicated the model and cross-validated it on different splits of the current state of the corpus, establishing an average F1 score of 0.63. This indicates that the model is quite unstable which is not surprising given the small amount of data in the positive class.

[17] We use BERT-base and RoBERTa-base.

We used 80 % of data for training (16,539 records in class 0 and 1,936 in class 1), 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. The described experiments yield a model with an F1 score of 0.7 for the positive class, and 0.97 for the negative class. Further metrics are displayed in Table 2:

	precision	recall	F1-score	number of records	accuracy
class 1 (AS)	0.75 (0.73)	0.65	0.7 (0.69)	225 (249)	0.94
class 0 (non-AS)	0.96	0.97	0.97 (0.96)	2084 (2061)	

Table 2: Evaluation of the best performing model on test and validation data, with validation data results displayed in brackets if different.

These scores 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 0), and the model finds 97 % of texts in this class (recall class 0). On the other hand, among the texts predicted as antisemitic, 75 % were labelled as such, 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 commentary section free of antisemitic speech. However, the number of false alarms would be low at 22, keeping manual efforts relatively limited. This example highlights the trade-off between two types of errors; while one would want to increase the recall of class 1, 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 1 and what precision should be accepted in return. 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. Per 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. We use the validation set to find out which threshold satisfies a recall of at least 0.8 while maximising the precision. With a really low threshold of 0.06 we achieve a recall of 0.81 and a precision of

0.52 on the validation set. The values for the test set are shown in Table 3, implying that we would capture almost 80 % of all antisemitic texts, albeit with almost every second alarm being a false alarm.

	precision	recall	F1-score	accuracy
class 1 (AS)	0.51	0.79	0.62	0.90
class 0 (non-AS)	0.98	0.92	0.95	

Table 3: Evaluation of the best performing classification model after moving the lower threshold for the positive class from 0.5 to 0.06.

What next?

As evident from the presented evaluations, there is certainly room for improvement in building classification models, trying to establish robust models with both higher precision and recall scores. Our next step in the remainder of the project will be a detailed evaluation of the current model's performance. This involves a qualitative inspection of texts where the model makes the biggest errors, as well as statistical evaluations such as correlation of errors with rhetorical aspects in order to find out whether the model recognises some types of antisemitic content better than others (e.g. certain stereotypes or hateful language). These insights will guide the entire project team in building better models and perhaps adapting the annotation scheme accordingly.

Furthermore, as the amount of training samples is always one of the most important factors for training performant models, we will explore different strategies to augment the training data, e.g. by including annotated texts in other languages and utilise multilingual models or applying automated translation. Moreover, we will test for domain adaptation by reserving part of the discourses for training and use others for testing, repeating this procedure multiple times. Since a significant part of messages containing antisemitic content require additional context that lies outside the text itself, we plan to work on concepts for handling such data in practice. In a slightly longer run, we believe that it is important to think about the utilisation of classification models in practice, beyond academic usage, as e.g. in moderation of news outlets and social media platforms.

Currently, it is difficult to imagine well working models for the detection of antisemitic speech without manual annotation of data. Since this work is time-consuming and requires a certain level of expertise, we believe that concepts are required for scalable long-term strategies. This includes questions regarding possibilities to join forces between related research and activism projects, as well as labelling by individuals with less expertise.

References

Aluru, Sai Saketh/Mathew, Binny/Saha, Punyajoy/Mukherjee, Animesh, 2020. Deep Learning Models for Multilingual Hate Speech Detection. https://doi.org/10.48550/arXiv.2004.06465.

Appandurai, Arjun, 1990. Disjuncture and Difference in the Global Cultural Economy. In: Theory, Culture & Society, Vol. 7, No. 2/3.

Ascone, Laura/Becker, Matthias J./Bolton, Matthew/Chapelan, Alexis/Krasni, Jan/Placzynta, Karolina/Scheiber, Marcus/Troschke, Hagen/Vincent, Chloé, 2022. Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online. Discourse Report 3. Berlin: Technische Universität Berlin. Centre for Research on Antisemitism. https://doi.org/10.14279/depositonce-14976.

Ascone, Laura/Becker, Matthias J./Bolton, Matthew/Chapelan, Alexis/Krasni, Jan/Placzynta, Karolina/Scheiber, Marcus/Troschke, Hagen/Vincent, Chloé, 2022. Decoding Antisemitism: An AI-Driven Study on Hate Speech and Imagery Online. Discourse Report 4. Technische Universität Berlin. Centre for Research on Antisemitism. https://doi.org/10.14279/DEPOSITONCE-16292.

Basile, Valerio/Bosco, Cristina/Fersini, Elisabetta/Nozza, Debora/Patti, Viviana/Rangel Pardo, Francisco Manuel/Rosso, Paolo/Sanguinetti, Manuela, 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, Minnesota, USA: Association for Computational Linguistics, 54–63. https://doi.org/10.18653/v1/S19-2007.

Becker, Matthias J., 2021. Antisemitism in Reader Comments: Analogies for Reckoning with the Past. London: Palgrave Macmillan.

Becker, Matthias J./Ascone, Laura/Bolton, Matthew/Chapelan, Alexis/Krasni, Jan/Placzynta, Karolina/Scheiber, Marcus/Troschke, Hagen/Vincent, Chloé, 2021. Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online. Discourse Report 2. Berlin: Technische Universität Berlin. Centre for Research on Antisemitism.

Chandra, Mohit/Pailla, Dheeraj/Bhatia, Himanshu/Sanchawala, Aadilmehdi/Gupta, Manish/Shrivastava, Manish/Kumaraguru, Ponnurangam, 2021. "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. http://arxiv.org/abs/2104.05947.

Dixon, Lucas/Li, John/Sorensen, Jeffrey/Thain, Nithum/Vasserman, Lucy, 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, OrAbraham 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, Heidelberg: Springer-Verlag, 186–196. https://doi.org/10.1007/978-3-031-15931-2_16.

Falkenberg, Mark/Baronchelli, Andrea, 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
(last accessed on 5 February 2023).

Friendberg, Brian, Wired, July 31st, 2020. The Dark Virality of a Hollywood Blood-Harvesting Conspiracy. https://www.wired.com/story/opinion-the-dark-virality-of-a-hollywood-blood-harvesting-conspiracy
(last accessed on 5 February 2023).

González-Pizarro, Felipe/Zannettou, Savvas, 2022. Understanding and Detecting Hateful Content Using Contrastive Learning. http://arxiv.org/abs/2201.08387.

Horta Ribeiro, Manoel/Jhaver, Shagun/Zannettou, Savvas/Blackburn, Jeremy/Stringhini, Gianluca/De Cristofaro, Emiliano/West, Robert, 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/Melo, Philipe/Benevenuto, Fabricio/Feldmann, Anja/Zannettou, Savvas, 2021. On the Globalization of the QAnon Conspiracy Theory Through Telegram. http://arxiv.org/abs/2105.13020.

Hutchinson, Ben/Prabhakaran, Vinodkumar/Denton, Emily/Webster, Kellie/Zhong, Yu/Denuyl, Stephen, 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/Awasthi, Deepika/Axelrod, David/Miehling, Daniel/Wagh, Pauravi/Joeng, Weejoeng, 2021. Detecting Anti-Jewish Messages on Social Media. Building an Annotated Corpus That Can Serve as A Preliminary Gold Standard. In: Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media. US: ICWSM. https://doi.org/10.36190/2021.14.

Jikeli, Günther/Cavar, Damir/Jeong, Weejeong/Miehling, Daniel/Wagh, Pauravi/Pak, Denizhan, 2022. Toward an AI Definition of Antisemitism? In: Hübscher, Monika/von Mering, Sabine (eds.). Antisemitism on Social Media. London: Routledge, 193–212.

Jikeli, Günther/Cavar, Damir/Miehling, Daniel, 2019. Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism. https://doi.org/10.5967/3r3m-na89.

Mandl, Thomas/Modha, Sandip/Shahi, Gautam Kishore/Madhu, Hiren/Satapara, Shrey/Majumder, Prasenjit/Schaefer, Johannes, et al., 2021. Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. https://doi.org/10.48550/arXiv.2112.09301.

Mathew, Binny/Saha, Punyajoy/Yimam, Seid Muhie/Biemann, Chris/Goyal, Pawan/Mukherjee, Animesh, 2022. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 17, 14867–14875. https://doi.org/10.48550/arXiv.2012.10289.

Meta, 2022. Community Standards Enforcement | Transparency Center. https://transparency.fb.com/data/community-standards-enforcement (last accessed on 14 February 2023).

Mihaljević, Helena/Steffen, Elisabeth, 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), 1–12. Potsdam, Germany.

Moffitt, J. D./King, Catherine/Carley, Kathleen M., 2021. "Hunting Conspiracy Theories During the COVID-19 Pandemic." In: Social Media + Society, Vol. 7, No. 3. https://doi.org/10.1177/20563051211043212.

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

Pogorelov, Konstantin/Schroder, Daniel Thilo/Burchard, Luk/Moe, Johannes/Brenner, Stefan/Filkukova, Petra/Langguth, Johannes, 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/Basile, Valerio/Sanguinetti, Manuela/Bosco, Cristina/Patti, Viviana, 2021. Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review. In: Language Resources and Evaluation, Vol. 55, No. 2, 477–523. https://doi.org/10.1007/s10579-020-09502-8.

Proust, Serge/Michalon, Jérôme/Maurin, Marine/Noûs, Camille, 2020. Dieudonné: Anti-Semitism, moral panics and deviant community, Deviance and Society, Vol. 44, No 3.

Rajadesingan, Ashwin/Resnick, Paul/Budak, Ceren, 2020. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. In: Proceedings of the International AAAI Conference on Web and Social Media, 14, 557–568. https://ojs.aaai.org/index.php/ICWSM/article/view/7323.

Röttger, Paul/Vidgen, Bertram/Nguyen, Dong/Waseem, Zeerak/Margetts, Helen/Pierrehumbert, Janet B., 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.

Solomon, Daniel, K. La Revue, January 5th, 2023. Kanye and the new Far West of American Antisemitism. https://k-larevue.com/en/kanye-and-the-new-far-west-of-american-antisemitism/
(last accessed on 14 February 2023).

Steffen, Elisabeth/Mihaljević, Helena/Pustet, Milena/Bischoff, Nyco/Varela, María do Mar Castro/Bayramoğlu, Yener/Oghalai, Bahar, 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. http://arxiv.org/abs/2210.07934.

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/ Gomez,Aidan N./Kaiser, Łukasz/Polosukhin/Illia, 2017. Attention is all you need. In: Advances in Neural Information Processing Systems 30.

Wiegand, Michael/Siegel, Melanie/Ruppenhofer, Josef, 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
(last accessed on 23 February 2023).

Wilson, Jason, Southern Poverty Law Center, December 7th, 2022. Kanye's Antisemitic Hate Speech Platformed by Enablers in Tech, Media, Politics. https://www.splcenter.org/hatewatch/2022/12/07/kanyes-antisemitic-hate-speech-platformed-enablers-tech-media-politics
(last accessed on 14 February 2023).

Wodak, Ruth, 2020. The Politics of Fear: The Shameless Normalization of Far-Right Discourse. London: Sage Publications.

Zampieri, Marcos/Malmasi, Shervin/Nakov, Preslav/Rosenthal, Sara/Farra, Noura/Kumar, Ritesh, 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, Minnesota, USA: Association for Computational Linguistics, 75–86.
https://doi.org/10.18653/v1/S19-2010.

Zampieri, Marcos/Nakov, Preslav/Rosenthal, Sara/Atanasova, Pepa/Karadzhov, Georgi/Mubarak, Hamdy/Derczynski, Leon/Pitenis, Zeses/Çöltekin, Çağrı, 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). http://arxiv.org/abs/2006.07235.

Sources

Kanye West's antisemitic remarks in autumn 2022

UK

BBC-FB[20221025] BBC News, October 25th, 2022, “Adidas cuts ties with Kanye West,” https://www.facebook.com/bbcnews/posts/pfbid0FwjyxqnrwXz4j5kbGTh2n9TKkE3fwpezrmSxhRDS41adSWpBT5KyrmezfsTXGwyfl.

DAILY-FB[20221019] Daily Mail, October 10th, 2022, “Howard Stern compares Kanye West to HITLER following troubled rapper's stream of anti-Semitic remarks,” https://www.facebook.com/DailyMail/posts/pfbid02wx4bNiso7yGDEkVww2PpPGUAHkSrQ34uchJRs9Qf3eavjJZzQ347YUBx46QhPp7cl.

DAILY-TW[20221029] Daily Mail, October 29th, 2022, “Kanye doubles down on anti-Semitic claim Jews control the media by sharing SPREADSHEET ‘filled with names of Jewish execs,’” https://twitter.com/MailOnline/status/1586452232844754945.

GUARD-FB[20221025] The Guardian, October 25th, 2022, “Kanye West reportedly no longer a billionaire as companies cut ties,” https://www.facebook.com/theguardian/posts/pfbid0bxs2WNVbXKzGuaqMDQRTmBCEq9VouFdR3MCR7TkNTsmgejj7BpC95F161qCh4Xxol.

GUARD-FB[20221026] The Guardian, October 26th, 2022, “Kanye West's wax figure removed from public view,” https://www.facebook.com/theguardian/posts/pfbid0GzMRz62gLzuZmjqVPPLHQ3B5MtvUFdFfQvvTnUKCLMfPqYnZhEqdBqWqFxu6mGsal.

INDEP-FB[20221009] The Independent, October 9th, 2022, “Calls for Kanye West to be banned from Twitter,” https://www.facebook.com/TheIndependentOnline/posts/pfbid02wUyh1E4b3cMdKD9Q941c89hef5fxtjBwa9Hsip7oWckAA5pWn3TQdvhsxhSn9trrl.

INDEP-FB[20221027] The Independent, October 27th, 2022, “Opinion: Kanye West's comeuppance is too little, too late,” https://www.facebook.com/TheIndependentOnline/posts/pfbid026XadeqpgthwfwhZzyEecfpd7SSTpY8fTm7nHZHdGh33KjYi1hUz4kCcxFhmp9HgKl.

INDEP-FB[20221029] The Independent, October 29th, 2022, “Former Kanye fan burns Yeezy shoe collection after rapper's antisemitic remarks,” https://www.facebook.com/TheIndependentOnline/posts/pfbid0gas9LxFdanSgDw1F2M4jvZX2zysM5yZxX4K8SbJwrkXH2iwdeBj7zZhTdQw6U1iWl.

METRO-FB[20221025] Metro, October 25th, 2022, “Kanye West ‘no longer a billionaire’ after being dropped by Adidas,” https://www.facebook.com/MetroUK/posts/pfbid029SS15zCBVt2z2JUPmU1UT8rFFiGpGqxuqiPaj7z3fFa5Gh1bS3ddZoBJGaenUiTRl.

MIRROR-FB[20221021] Daily Mirror, October 21st, 2022, “Kanye West dumped by major fashion chain after string of major outbursts,” https://www.facebook.com/MirrorCeleb/posts/pfbid0UfZkC4nF19Bae7nfvYBPJgS95j9fAnp32FAzX2FRtB9nJhZcNdeEL2cqWSr9BYbul.

MIRROR-FB[20221103] Daily Mirror, November 3rd, 2022, “Kanye West announces ‘verbal fast’ which will see him avoid speaking for 30 days,” https://www.facebook.com/MirrorCeleb/posts/pfbid0gdvW4dPKaZSCzfSHo5HCoPXcmDExArmkjvNYJ8BRKyqQyYhuWrSJeW9nugJP5WrMl.

SUN-FB[20221020] The Sun, October 20th, 2022, “Kanye West storms out of Piers Morgan TalkTV show after furious row,” https://www.facebook.com/thesun/posts/pfbid02w8mv4CiUymG8uev6CeuHXFp6eDPwq2nSfromKVtZpVhxYbJ4AJQ5WWgCez5xiL8Vl.

TIMES-FB[20221026] The Times, October 26th, 2022, “Kanye West's school Donda Academy has closed,” https://www.facebook.com/timesandsundaytimes/posts/pfbid0yWt9XLq5V1JuWyp7TeHAq9BKRqwCHnfhtdJLyVD3Y2WuoJs2bpATGKDtq7whUqgdl.

VICE-FB[20221019] Vice UK, October 19th, 2022, “Why I've Finally Given Up on Kanye West,” https://www.facebook.com/viceuk/posts/pfbid035riKXHngJmQzrn6uXMNH3fA5ejyv5Xbajy13x72gwtEgmwqh1eS3mMnGh6QRqvUsl.

France

BFMTV-FB[20221027] BFMTV, October 27th, 2022, "La statue de cire de Kanye West retirée du musée Madame Tussauds à Londres," https://www.facebook.com/BFMTV/posts/pfbid0pgh35wsyzC7ekaQxe1iKEr7ntp7nU3SAX23ep9QsWFC4soqbFy4e4R5igi94fgJGl.

BFMTV-FB[20221030] BFMTV, October 30th, 2022, "Kanye West s'excuse après ses propos sur George Floyd et ses remarques antisémites," https://www.facebook.com/BFMTV/posts/pfbid0Rsq4cGR2V2HYmNVUnE2hvDBNh7fAYnwXtAaS82RUCKkD2g6rL2yTC4k9MtiVo45bl.

FRENC-TW[20221026] FrenchRapUS, October 26th, 2022, "🚨 La statue de cire de Kanye West vient d'être retiré au musée Madame Tussauds ! [...]," https://twitter.com/FrenchRapUS/status/1585341880438784001.

LCI.F-FB[20221026] LCI, October 26th, 2022, "Propos antisémites : et si c'était (aussi) la fin de la carrière musicale de Kanye West ?," https://www.facebook.com/LCI/posts/pfbid0tL1iFLvzDchxRts2re5RnNimFfe6t5ub8ia5N8FoJYfYiKgEUMW253iEDk9N84MFl.

LEFIG-FB[20221025] Le Figaro, October 25th, 2022, "Adidas rompt son partenariat avec Kanye West après des remarques antisémites," https://www.facebook.com/lefigaro/posts/pfbid0TF7f6AiJeA48jvRaoxNLZx2kTxwQUBE3rh43r7UBZKZgpH5HjxGtPmmX6CA7ca95l.

LEPOI-FB[20221025] Le Point, Octobre 25th, 2022, "Propos antisémites – Adidas met fin à sa collaboration avec Kanye West," https://www.facebook.com/lepoint.fr/photos/a.389816860702/10158755608000703.

LESIN-FB[20221027] Les Inrockuptibles, October 27th, 2022, "Kanye West – Un drame américain," https://www.facebook.com/lesinrockuptibles/posts/pfbid028FjMUGx8SffdLYys261dFrgoMHdyxizoCBAWMwLMWw89d3chWa5dkgwCCDQ5QTFkl.

Antisemitic incidents at the 2022 FIFA World Cup in UK media

TW-AMRO[20221206] Twitter, December 6th, 2022, "Morocco celebrates their victory by raising the Palestinian flag 🇵🇸 [...]," https://twitter.com/_amroali/status/1600193400598253568.

TW-CARTER[20221204] Twitter, December 4th, 2022, "England football fan chants 'FREE PALESTINE' in Israel TV interview following win over Senegal," https://twitter.com/Bob_cart124/status/1599534417755897856.

TW-CARTER[20221210] Twitter, December 10th, 2022, "Just imagine if Morocco wins the World Cup in Qatar! [...]," https://twitter.com/Bob_cart124/status/1601643552630415360.

TW-HARAWI[20221127] Twitter, November 27th, 2022, "🧵1/4 Israeli journalists at the Qatar World Cup are being shunned left, right & center by fans, particularly from the Middle East but also beyond," https://twitter.com/yarahawari/status/1596879671169527808.

TW-JASKOLL[20221126] Twitter, November 26th, 2022, "People enjoying FIFA in Qatar, a wild abuser of human rights, oppressor of homosexuals & women, master of slave labor, funder & harborer of terrorists [...]," https://twitter.com/skjask/status/1596596107357790208.

TW-JENNINE[20221126] Twitter, November 26th, 2022, "more people rejecting the isr@eli news channels. small acts of boycott by the people ❤️," https://twitter.com/jennineak/status/1596625234752598016.

TW-KOMUGISHA[20221201] Twitter, December 1st, 2022, "Morocco 🇲🇦 hoist the Palestine 🇵🇸 flag after their 2-1 victory over Canada in solidarity with Palestine," https://twitter.com/UsherKomugisha/status/1598364139801419776.

TW-MILSTEIN[20221127] Twitter, November 27th, 2022, "#Qatar's disastrous #WorldCup is so xenophobic that some fans are now attacking Arab journalists accusing them of being Israelis [...]," https://twitter.com/AdamMilstein/status/1596740211060273152.

TW-OGDEN[20221206] Twitter, November 26th, 2022, "Morocco celebrate their win against Spain with a Palestinian flag [...]," https://twitter.com/MarkOgden_/status/1600189047812390917.

TW-SAKIB[20221202] Twitter, December 2nd, 2022, "This is killing me," https://twitter.com/mertesakib/status/1598698752574996480.

The Israeli Elections in December 2022 (German media)

ARTED-YT[20221229] ARTE, December 29th, 2022, "Israel: Demokratie in der Sackgasse?," https://www.youtube.com/watch?v=x9ok02QXKgk.

SPIEG-TW[20221113] Spiegel, November 13th, 2022, "Wenn in Israel eine rechtsradikale Regierung an die Macht kommt, droht eine neue Welle des Antisemitismus – gegen Juden in Europa und Deutschland," https://twitter.com/SPIEGiegel/status/1591850652262862849.

SPIEG[20221113] Spiegel, November 13th, 2022, "Sieg der Rechtsparteien in Israel: Eine neue Welle des Judenhasses," https://www.spiegel.de/ausland/israel-eine-neue-rechte-regierung-wird-eine-welle-des-antisemitismus-ausloesen-a-e2388318-3c2b-46c4-a0f6-c7a253164b17.

WELT[20221102] Welt, November 2nd, 2022, "Netanjahus autoritäre Züge sind für Israels Demokratie gefährlich Rechtsruck bei Israel-Wahl – Netanjahu laut Prognosen vor Comeback," https://www.welt.de/politik/deutschland/article241918157/Wahl-in-Israel-Netanjahus-autoritaere-Zuege-sind-fuer-die-israelische-Demokratie-gefaehrlich.html.

ZDF-YT[20221229] ZDF, December 29th, 2022, "Rechts und religiös - Israels neue Regierung | auslandsjournal," https://www.youtube.com/watch?v=mgPY8PapgUc.

ZEIT[20221101] Zeit, November 1st, 2022, "Eine gefährliche Wahl," https://www.zeit.de/politik/ausland/2022-11/israel-parlamentswahl-benjamin-netanjahu-rechtsextremismus-faq.

ZEIT[20221216] Zeit, December 16th, 2022, "Warum die Ultrarechten in Israel so stark sind," https://www.zeit.de/politik/ausland/2022-11/parlamentswahl-israel-benjamin-netanjahu-extremismus.

ZEIT[20221217] Zeit, December 17th, 2022, "Sehr rechts und sehr religiös," https://www.zeit.de/politik/ausland/2022-11/israel-regierungsbildung-koalitionsverhandlung-benjamin-netanjahu.

ZEIT-IG[20221102] Zeit, November 2nd, 2022, "Parlamentswahl. Warum sind die Ultrarechten in Israel so stark?," https://www.instagram.com/p/CkdMUM4q2Rp/.