ISD Powering solutions to extremism, hate and disinformation

Methodological Appendix

As part of ISD's Gen Z & The Digital Salafi Ecosystem project, this methodology paper provides an overview of ISD's research approach for our data-driven snapshot of the Salafi digital landscape. The paper outlines the inclusion criteria, platform selection, and data-gathering approaches used to conduct the research, as well as the frameworks developed to empirically analyse narratives and formats within the Salafi online ecosystem. We also outline our unique approach to understanding 'toxicity' in the Salafi context, including the scoring and coding approach employed in our analysis.

Background

Between September 2020 and September 2021, the Institute for Strategic Dialogue (ISD) undertook a one-year policy-focused research programme looking at the appeal of Salafism, with a focus on Gen-Z audiences. As part of these efforts, ISD used data analytics tools to map Salafi communities in Arabic, English and German across a range of platforms to capture the scale, nature, actors and narratives of the Salafi online ecosystem.

Through this research, ISD sought to answer a number of questions to provide an in-depth, fine-grained analysis of the Salafi online ecosystem: What are the key platforms used by Salafi communities online? How influential are different subgroups within the movement? What are their key narratives and grievances, and how, if at all, do they differ between Arabic, English and German Salafis? What are the key formats used by online influencers to convey their messages to existing and potential followers, and how do these differ between platforms? And what are the emerging trends affecting Gen-Z Muslims in particular, including the resonance of specific subcultures, narratives and youth-oriented platforms?

As the majority of Salafis do not support violence, and many are not overtly political, another question ISD sought to answer was: What's the harm? Going beyond attempts to classify whether or not individual accounts or communities fit into broad categories such as 'quietist', 'jihadist' or 'extremist', ISD decided to focus on the level of messages within these communities. In cooperation with the tech start-up Textgain, ISD co-developed an ontology through which large collections of Salafi messages could be automatically scanned and analysed. The aim was to look for 'toxic' discourse related to Salafism, such as threatening, dehumanising and othering language. This approach also acknowledges the reality that many accounts and communities mix content from a diverse range of Salafi voices.

ISD's research findings are presented in a companion paper to this Methodological Appendix. This paper does not discuss the findings, but rather the methods used to reach them. By publishing an outline of the methodological approach and describing the decisions taken during the research process, ISD is attempting to be as transparent as possible to aid policymakers, civil society activists and other researchers interested in the findings.

This paper first discusses the platforms, inclusion criteria and data-gathering approaches used to conduct the research. Subsequently, the approach used to empirically identify and analyse narratives and formats within the Salafi online ecosystem is laid out. The final section describes and explains the general approach, scores and labels employed in the toxicity analysis, and provides some examples of what constitutes 'toxic' language in a Salafi context.

Platforms

To analyse the Arabic-, English- and German-language sections of the Salafi online ecosystem, ISD focused on a number of social media platforms and Salafi websites that were judged to be particularly relevant, based on existing research. This was either due to their importance for Salafis in particular, in public discourse or to young internet users. Social media platforms analysed included Facebook, Instagram, Twitter, YouTube, Telegram and TikTok.

The researchers' ability to quantitatively access and analyse data was dependent on each platform. While it was possible to systematically gather data from Facebook, Instagram, Twitter, YouTube and Telegram to varying degrees, it was not possible to systemically gather data relating to specific accounts, keywords or hashtags on TikTok. As TikTok is hugely popular among Gen-Z users, Salafi communities were nevertheless analysed using qualitative approaches. Similar qualitative analysis was conducted on the key themes, formats and offers of stand-alone Salafi websites in Arabic, English and German.

Inclusion Criteria

Salafism is a current within Sunni Islam which advocates a return to the practices of the first three generations of Muslims (the salaf or 'ancestors') who lived immediately after the prophet Mohammed. Within Salafism, there are different strands, which differ significantly in their interpretations of the holy scriptures of Islam and their implications for political action. While quietist Salafis reject political activism, political Salafis are actively engaged in transforming society according to their ideological ideas. Jihadist Salafis use violence to impose a Salafi interpretation of Islamic law.

Actors were classified as Salafis irrespective of which subgroup of Salafism they belonged to. Therefore, the accounts identified did not necessarily need to advocate actively for the political transformation of society and the establishment of a totalitarian Islamic state, or explicitly advocate violence to that end, as long as they sought to return to the practices of the first three generations of Muslims.

ISD researchers classified pages, groups and channels on alternative platforms as Salafi if any of the following conditions was met:

- The accounts belonged to widely known Salafi preachers, activists or organisations.
- The accounts repeatedly and affirmatively shared the content of known Salafi preachers, activists or organisations, or expressed support for them.
- The accounts posted content that clearly fell under our definition of Salafism (e.g. by arguing that the restoration of a 'caliphate' is necessary to revive the practices of the first three generations of Muslims who lived immediately after the prophet Mohammed).
- In cases where accounts shared content by previously unknown organisations or individuals, ISD researchers conducted open-source searches to find out more about their religious, political and ideological background and determine whether they would fall under our definition of Salafism.

Discovery of Salafi Accounts

ISD researchers conducted an initial phase of qualitative analysis and open-source investigation to develop a list of Salafi pages, groups, accounts, channels and websites.

The identification of these accounts proceeded in two steps. Firstly, ISD researchers conducted ethnographic monitoring on the platforms outlined above, starting from accounts and websites known to ISD from previous research and the wider literature on Salafism, subsequently proceeding through a 'snowballing' approach to build out the lists of relevant channels, accounts and groups, using recommendation suggestions and other connected accounts.

Secondly, every entity was manually reviewed by a second expert researcher to ensure that each met the inclusion criteria for Salafism outlined above and was not falsely classified as Salafi.

Data Gathering

Wherever one was available, ISD researchers used the respective platform-specific application programming interface (API) to gather posts from public pages and groups (Facebook), posts from pages (Instagram), tweets and follower information from accounts (Twitter), video titles, descriptions and comments from channels (YouTube), and posts and comments in channels and groups (Telegram). The time frame for the data gathering was between 27 October 2019 (the killing of ISIS leader Abu Bakr al-Baghdadi) and 13 July 2021.

The data from across all platforms was collated in Method52, a bespoke software package for the analysis of social media data. ISD's partners at the Centre for the Analysis of Social Media (CASM) have been developing Method52 since 2012. ISD and CASM are working together on the application and expansion of Method52 to better capture and analyse social media data. The collation of the quantitative data from Facebook, Instagram, Twitter, YouTube and Telegram in one centralised system allowed for the subsequent comparative analysis of Arabic-, English- and German-language Salafi communities across these platforms.

Narrative and Format Analysis

Two of the key research interests for this project were finding out what Arabic-, English- and German-language Salafis were talking about online, and what formats they chose to convey their message. The aim of this analysis was to begin to understand the subjective views of Arabic-, English- and German-language Salafis, and their strategic choices of specific formats.

The datasets gathered from Salafi communities across Facebook, Instagram, YouTube, Twitter and Telegram were too large to be coded manually in their entirety. Therefore, an alternative approach was needed to identify key narratives and formats within such large digital datasets.

Automated analysis approaches, such as topic modelling, can be used to identify hidden structures within large amounts of text and to identify broad topics, but do not allow for a more fine-grained assessment of the key narratives. Keywords can be used to determine the approximate frequency of narratives within a dataset. However, because keywords are based on researchers' assumptions about which topics are particularly important within the data, solely relying on a keyword-based approach risks counterintuitive but prominent narratives within the data being overlooked.

To achieve a comprehensive insight into the data collected, and to minimise the influence of pre-existing assumptions on the content analysis, ISD researchers created randomised samples of 100 Arabic-, English- and German-language Salafi messages (posts, tweets, videos and comments) from each platform (Facebook, Instagram, YouTube, Twitter and Telegram), resulting in 15 samples of 1,500 messages overall. 100 messages were coded per platform and language, with percentages equal to totals (e.g. 25% of Arabic messages on Telegram equal 25 messages).

The narrative and format coding was then carried out in two stages. During the first stage, ISD analysts inductively coded samples of Arabic, English and German messages into categories without a pre-defined list of categories for narratives and formats. The intention was to develop a list of categories without imposing pre-existing theoretical assumptions on the empirical analysis. ISD researchers then compared the narrative and format categories different analysts had come up with, identifying overlaps and common categories. Based on this exercise, a unified system of broad narrative and format categories was developed.

Using this list as a reference point, ISD analysts returned to the data and coded the 15 samples of Arabic-, English- and German-language Salafi messages from the five social media platforms into the shared category framework. This process ensured that findings would be comparable across platforms and linguistic contexts. To enable an even more nuanced description of narratives across the Salafi online ecosystem, analysts not only coded broad narratives based on the initial category framework, but also described more specifically what each post was about. For example, the broad narrative of a post could be 'discussion of religious concepts', 'discussion of out-groups' or 'political grievances', while the corresponding specific narratives might be 'sharia' ('discussion of religious concepts') 'Shia' ('discussion of out-groups') or 'Palestine' ('political grievances'). As Salafi messages often touch on multiple themes within a single piece of content, particularly in longer videos, researchers were able to code each message into three broad and three specific narratives.

Narrative categories

Based on the first stage of coding, ISD researchers identified the broad narrative categories outlined below. If a message did not fall under any of these categories, its broad narrative was coded as 'other'.

- **Discussions of religious concepts and activities**: Specific narratives that would fall under this category include, but are not limited to, discussions of the unity of God (tawhid), divine predestination, eschatology, Islamic history, morality, the afterlife, the relationship between religious law (Sharia) and democracy, calls to prayer, festivities such as Ramadan, pillars of Islam such as Zakar and Hajj, or the conversion of individuals to Islam.
- **Discussions of out-groups:** Specific narratives that would fall under this category include, but are not limited to, discussions of women engaging in supposedly sinful behaviour, hostility towards LGBTQ+ people and rights, attacks on national, racial and ethnic groups (e.g. Kurds, Israelis, blacks), attacks on Westerners, attacks on apostates, atheists, secularists, liberals and followers of other religions, hostility towards other Muslim sects (e.g. Shi'ites, Ahmadiya) or other Muslims for supposedly not behaving in accordance with their religion by either being too extreme (including terrorism, Kharijites) or too liberal and progressive.
- **Discussions of political grievances**: Specific narratives that would fall under this category include, but are not limited to, perceived domestic anti-Muslim racism (e.g. Islamophobia, the far-right, 'assimilationist' integration policies), accusations that 'the media' is biased against Islam, the conflicts and humanitarian crises in Syria, Myanmar and Palestine, blasphemy (especially calls to 'boycott France'), support for prisoners held on extremism/terrorism-related charges, refugees and historical grievances (e.g. colonialism, the Reconquista of Spain, the abolition of the Ottoman Caliphate).
- **Gender**: Specific narratives that would fall under this category include, but are not limited to, discussions of gender, gender roles, sex, relationships, family life and marriage.
- **COVID-19**: Any piece of content in which the COVID-19 pandemic is a key theme.
- **Commercial and fundraising**: Specific narratives that would fall under this category include, but are not limited to, promotion of specific shops, sales of alternative medicine, adverts for classes, courses and lectures, and calls for donations.

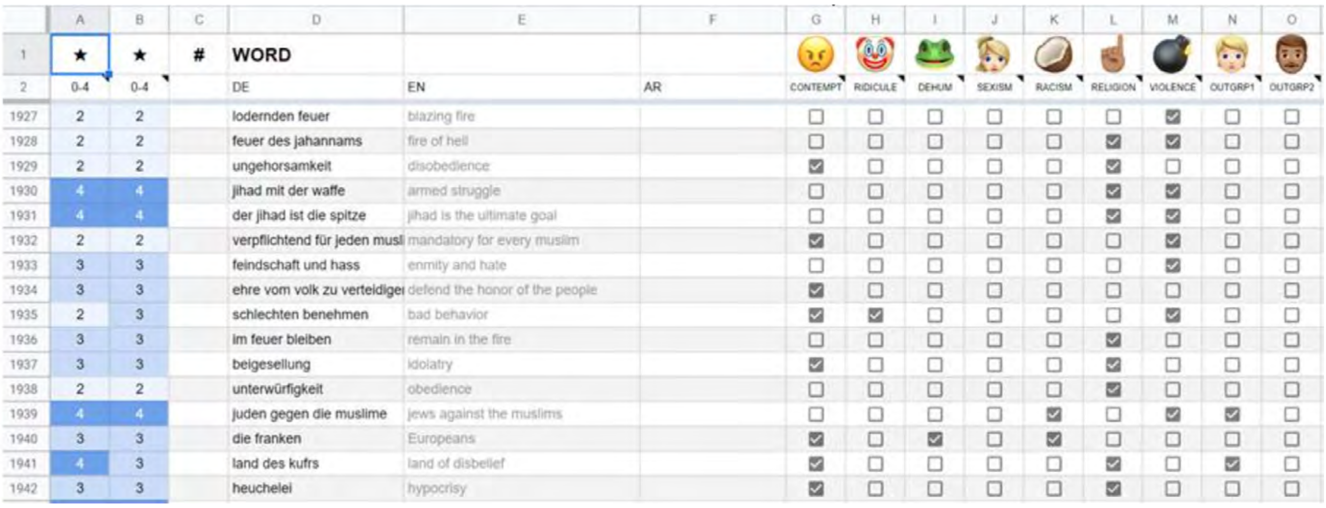
Some of these narratives may overlap to some extent, which was one of the reasons why ISD researchers were able to assign up to three broad narratives for each post. For example, calls to boycott France would likely constitute both the expression of a political grievance and a discussion of the religious concept of blasphemy. Similarly, historical grievances in discussions of Islamic history might fall under both political grievances and discussions of religious concepts, depending on the context.

Format categories

Based on the first stage of coding, ISD researchers identified the content formats outlined below. If a message did not fall under any of these categories, its format was coded as 'other'.

- **Q&A sessions**: Preachers and influencers providing answers to audience questions. These often centre around specific political and religious concerns, especially in relation to the activities and beliefs that are permissible within a Salafi view of Islam.
- **Presentations**: Speeches, sermons and lectures, or excerpts from them, usually focusing on religious or political topics.
- **Citizen journalism**: Reports from specific events, such as demonstrations, which Salafi activists frequently post. The use of the term 'journalism' is not meant to indicate that they are necessarily adhering to professional journalistic standards.
- **Conversations**: Interviews or debates between scholars, activists or influencers discussing religious and political topics.
- **Street dawah**: Missionary activities by Salafi scholars and activists in public spaces, aimed at converting people to Islam or persuading Muslims to adopt a Salafi worldview.

- **Animations**: Animated images or videos presenting a specific Salafi message.
- **Quranic verses and hadith**: Quotes and references to the Quran, hadith or scholars seen as authoritative sources (e.g. Ibn Taymiyyah).
- **Anonymous tales:** Stories presented with an invisible narrator focusing on themes such as Islamic doctrine, historical figures from Islam, or morality.
- **Nasheeds**: Islamic hymns sung without instruments which convey a specific religious message. While nasheeds are not tied to a particular interpretation of Islam, research has documented their popularity within Islamist and Salafi-jihadist movements.[1]



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	★	★	#	WORD			😠	🤡	🐸	👱‍♀️	🥥	☝️	💣	👱	🧔
2	0-4	0-4		DE	EN	AR	CONTEMPT	RIDICULE	DEHUM	SEXISM	RACISM	RELIGION	VIOLENCE	OUTGRP1	OUTGRP2
1927	2	2		lodernden feuer	blazing fire		☐	☐	☐	☐	☐	☐	☑	☐	☐
1928	2	2		feuer des jahannams	fire of hell		☐	☐	☐	☐	☐	☑	☑	☐	☐
1929	2	2		ungehorsamkeit	disobedience		☑	☐	☐	☐	☐	☑	☐	☐	☐
1930	4	4		jihad mit der waffe	armed struggle		☐	☐	☐	☐	☐	☑	☑	☐	☐
1931	4	4		der jihad ist die spitze	jihad is the ultimate goal		☐	☐	☐	☐	☐	☑	☑	☐	☐
1932	2	2		verpflichtend für jeden musl	mandatory for every muslim		☑	☐	☐	☐	☐	☐	☑	☐	☐
1933	3	3		feindschaft und hass	enmity and hate		☐	☐	☐	☐	☐	☐	☑	☐	☐
1934	3	3		ehre vom volk zu verteidigei	defend the honor of the people		☑	☐	☐	☐	☐	☐	☐	☐	☐
1935	2	3		schlechten benehmen	bad behavior		☑	☑	☐	☐	☐	☐	☑	☐	☐
1936	3	3		im feuer bleiben	remain in the fire		☐	☐	☐	☐	☐	☑	☐	☐	☐
1937	3	3		beigesellung	idolatry		☑	☐	☐	☐	☐	☑	☐	☐	☐
1938	2	2		unterwürfigkeit	obedience		☐	☐	☐	☐	☐	☑	☐	☐	☐
1939	4	4		juden gegen die muslime	jews against the muslims		☐	☐	☐	☐	☑	☐	☑	☑	☐
1940	3	3		die franken	Europeans		☑	☐	☑	☐	☑	☐	☐	☐	☐
1941	4	3		land des kufrs	land of disbelief		☑	☐	☐	☐	☐	☑	☐	☑	☐
1942	3	3		heuchelei	hypocrisy		☑	☐	☐	☐	☐	☑	☐	☐	☐

Figure 1: Screenshot of the ontology used to analyse Salafi social media messages

Analysis of Toxicity in Salafi Social Media Messages

In cooperation with Textgain, a start-up specialising in language technology and artificial intelligence (AI), ISD researchers supported the development of a ranking tool that recognises expressions frequently used by Salafis in English, German, Arabic and Latin Arabic. It currently contains about 1,000 expressions in Latin Arabic and their equivalents in English, German and Arabic, along with about 500 expressions from each of the different language cultures. The accompanying ranking algorithm can accurately deal with spelling variation (e.g. kaffir = kaffeer, kaffier, kafirs), resulting in over 10,000 linguistic forensic fingerprints, each with up to ten fine-grained labels.

The tool can be used to automatically scan large collections of texts for 'toxic' discourse related to Salafism. In this context, 'toxic' discourse means harmful and exclusionary language use on a scale from unnecessary to unpleasant to unacceptable, with a spectrum of manifestations such as profanity, ridicule, verbal aggression and targeted hate speech. In contrast, for example, to using the term 'hate speech', this range does not imply that there is necessarily a specific target being hated or abused. Using this ranking algorithm, Textgain conducted a toxicity analysis on nearly 3.5 million Salafi social media messages collected by ISD between October 2019 and July 2021 across Facebook, Instagram, Twitter, YouTube and Telegram (see the 'Data Gathering' section earlier in this Appendix).

Toxicity detection in the context of Salafism and Salafi-jihadism

Textgain and ISD created an ontology of words and word combinations linked to Salafism and Salafi-jihadism, with experts assigning a score and labels to every word or word combination. The ranking algorithm could then scan large collections of texts and assign a toxicity score between 0 (not at all toxic) and 100 (most toxic) to each text, based on the words and word combinations that it recognised.[2] The approach is fast, scalable and interpretative, and can help human analysts obtain insights by automatically sorting large datasets by level of toxicity, highlighting keywords. It is a form of explainable AI (XAI) that is arguably preferable to so-called 'deep learning black boxes' in high-stakes decision-making, as the decisions taken at each step and how they impact on the overall results can be transparently explained.[3]

The ontology had over 5,000 expressions in English, German, Arabic and Latin transliterations of Arabic. For English and Arabic, expressions were collected from the ISIS magazines 'Dabiq' and 'Rumiyah',[4] and from Wikipedia articles explaining concepts and terminology in Salafism.[5] For German, expressions were collected from Salafi manifestos that were shared in German-language Salafi Telegram channels. For the most part, these documents were translations of Salafi literature, though a minority of them were original writings by German Salafis. Ideologically, they covered a wide spectrum of Salafi scholars, from the quietist al-Albani to the jihadist al-Awlaki. In brief, for each of the three datasets (English, Arabic, German), Textgain then used an AI technique called 'word embedding' to train a model of word similarity (words that occur in similar sentences in the training data).[6] For a list of, say, 100 keywords one can then automatically discover ten times as many related words. For example, using kuffar as a search keyword will expose kufr and kafir but also murtad etc. as being used in similar sentences in extremist propaganda.

Word scores

These words were added to the ontology and manually annotated by domain and language experts. The English and German lists were annotated by two experts, and the Arabic by three, reviewing each other's work. For every word in the list, a score from 0 to 4 was assigned by the experts, representing the word's level of toxicity from neutral to very toxic. Words with a score of 0 typically included names of people and places, and references to religion (e.g. *fiqh*). These words are not inherently toxic, but add context about who or what is being discussed.

Score	Description	Examples
0	neutral, but sometimes leading	*Al-Tabari, America, Syria, soldiers*
1	tendentious or politicised	*crusade, kill, khilafah, Ibn Qayyim*
2	demeaning	*hypocrite, foolishness, Khariji*
3	demeaning and discriminatory	*filthy capitalist, kaafir, sodomy*
4	demeaning and discriminatory (extremely so)	*Jewish pigs, Safawī dogs, slave girls*

Figure 2: Outline of toxicity scoring with example words

Word labels

Each word could also have one or more labels: CONTEMPT, RIDICULE, DEHUMANIZATION, RACISM, SEXISM, OTHERING, THREATENING, and (not necessarily toxic) RELIGION. For example, a word like *fiqh* would have a score of 0 and be labelled as RELIGION. A word such as *kuffar* would have a score of 3 and be labelled as OTHERING and RELIGION. A word combination such as *filthy dogs* would have a score of 4 and be labelled as CONTEMPT and DEHUMANIZATION.

Label	Description	Examples
CONTEMPT	negative value judgements	*hypocrisy, lies, rafida, tawāghīt*
RIDICULE	insults of intelligence	*deviant, ignorant, apostate puppet*
DEHUMANIZATION	comparisons to animals etc.	*sheep, quburiyyun, kuffars & coconuts*
RACISM	relating to race and ethnicity	*American crusaders, filthy French*
SEXISM	relating to gender or sexuality	*mutahayyirah, zinah*
NON-MUSLIM OTHERING	relating to out-groups (e.g. Jews, Christians)	*heathens, infidels, qurān 9:29*
MUSLIM OTHERING	relating to out-groups (e.g. Shia Muslims)	*nusayrī, rejectionists, apostates*
THREATENING	verbal aggression, violence	*burn, punish, mujāhid*
RELIGION	any kind of religious vocabulary	*atheism, idols, rahimahullāh*

Figure 3: Outline of toxicity labelling with example words

- The CONTEMPT label related to negative value judgements, either expressing the author's intrinsic dislike (*anger, disgusted*), extrinsic dislike (*disgusting*) or targeted dislike (*hypocrites, liars*).
- The RIDICULE label related to insults of intelligence (*foolish, ignorant*), behaviour (*deviant, devilish*) and status (*puppets*). These were typically words with a score of 2 that were used to highlight or mock how out-groups act, more than attacking who they are.
- The DEHUMANIZATION label related to comparing humans to animals (*dogs, pigs, sheep*) or objects (*coconuts*). These were typically words with a score of 1 (e.g. *dogs* on its own could be ambiguous) or word combinations with a score of 3 or 4 (*filthy dogs*). Dehumanisation has been used throughout history to strip out-groups of their humanity and pave the way for violence against them.7
- The RACISM label related to discrimination on the basis of ethnicity (*Western hypocrites*), nationality (*American crusaders*), race or looks. These were typically words with a score of 2 or 3.
- The SEXISM label related to discrimination on the basis of gender (*your wives*), gender stereotypes (*chastity*) or sexuality (*sodomites, zina*). These were typically words with a score of 2 or 3.
- The OTHERING label distinguished between two subgroups: non-Muslims and Muslims. The NON-MUSLIM OTHERING label related to contempt for perceived Western out-groups, typically on the basis of religion: Christians, Jews and atheists (*crusaders, zionists, unbelievers*). The MUSLIM OTHERING label related to contempt for perceived Islamic out-groups: Shia, Sufi etc. Such words could have a score anywhere between 1 (*atheists*) and 4 (*filthy rafida*).
- The THREATENING label related to violence in general (*burn, destroy, kill*), verbal aggression (*die in your rage*), and targeted incitement and threats (d*estroy america, kill the mushrikun*). These words could have a score anywhere between 1 (*war*) and 4 (*war against the murtaddin*).
- The RELIGION label related to any kind of religious vocabulary, whether usually non-toxic, as in *Allah, Islam, muslim, christian*, which all had a score of 0, or otherwise, as in *kuffar*, which had a score of 3.

Threshold for 'toxic' messages

The **normal threshold** for messages was based on the average (AVG) and standard deviation (SD). The AVG was a single representative toxicity score for all messages, and the SD indicated by how much most messages deviated from this score. To err on the conservative side, anything within one SD of the mean was defined as normal. Statistically, the remainder of messages were outliers, with a toxicity score significantly higher than normal.

In the dataset of Salafi social media messages, the normal toxicity score was therefore 0–30/100 (AVG ± SD). This was the normal threshold for the purposes of the analysis. Accordingly, messages with a score lower than 30/100 were normal. Messages with a higher score were defined as toxic. Based on the analysis, about 8% of the discourse was more toxic than normal, i.e., having a score assigned by the ranking algorithm > 30/100. Very toxic messages were defined by the researchers as messages that scored > 90/100 to identify the most extreme toxic expressions among Salafis.

Methodological reflections and limitations

Messages on Facebook and Telegram tend to be longer than on Instagram, YouTube or Twitter, with the latter imposing a limit of 280 characters. This requires some further methodological reflection.

Long messages made up the majority of toxic messages. These were more ambiguous: a Quran quote about punishing unbelievers triggered the algorithm, but the quote could be part of a narrative to discredit such statements. Long messages could contain many different viewpoints, and reducing these to a binary toxic vs. non-toxic decision was a challenge for classification in general, as well as for AI systems specifically. Many of these messages contained long, winding narratives that felt incongruous, perhaps because they pushed one viewpoint favoured by some online scholars while disregarding other opinions.

Short messages were more direct. If the format allowed for ten words, and two of those words were *murtadd* and 😧, then the message was clearly more likely to be harmful.

To illustrate the challenge of ambiguity, the top ten of the most toxic messages included:

- A list of words that every Muslim should know, including multiple words marked as very toxic in the ontology. That in itself did not mean that the message was very toxic, but it did present words that many Muslims would previously have been unfamiliar with, and hence seemed to push a biased worldview.
- A long account of the Battle of Badr.[8] Its references to swords and slaughter triggered the AI. The message was a non-toxic portrayal of history, but its presence on social media could be problematic, fostering polarisation rather than reconciliation in the Muslim community.
- A long rant directed against Atatürk (Turkish progressive reformer), slandering him as a homosexual, the direct cause of all problems in the Arab world, and an example of all Turkish people being perverted. The message triggered the algorithm, but the challenge was that the AI lacked any deep insight into the historical context. A deep learning system might be able to deal with this, but a recent policy debate by the European Commission has considered curtailing such AI in high-stakes decision-making, while new explainable alternatives (XAI) are still in development.[9]
- A long analysis that refuted all warlike, polarising citations from the Quran, while at the same time explicitly mentioning every polarising excerpt to its audience. This was not unlike the academic debate about whether or not hate speech studies should be explicit or rather described (cf. 'the N-word'). Hence, it was challenging to classify this message. Was it toxic? No. Did it contain toxic language? Yes.
- A minority of messages also contained counterspeech, i.e., non-Salafis posting toxic messages in the channel or group (e.g. 'Ban the filthy Islam before its too late') mostly on YouTube. Unfortunately, this was also a challenge that the AI could not resolve automatically.

Endnotes

[1] Said, B. (2012) 'Hymns (nasheeds): A contribution to the study of the jihadist culture', Studies in Conflict & Terrorism, 35(12): 863–79; Gråtrud, H. (2016) 'Islamic State nasheeds as messaging tools', Studies in Conflict & Terrorism, 39(12): 1050–70; Hegghammer, T. (ed.) (2017) Jihadi Culture. (Cambridge: Cambridge University Press).

[2] De Smedt, T. et al. (2020) Profanity & Offensive Words (POW). Textgain, June.

[3] Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nature Machine Intelligence, 1(5): 206–15.

[4] Azman, N.A. (2016) '"Islamic State"(IS) propaganda: Dabiq and future directions of "Islamic State"', Counter Terrorist Trends and Analyses, 8(10): 3–8.

[5] Wikipedia (n.d.) Salafi Movement. Available online at: https://en.wikipedia.org/wiki/Salafi_movement

[6] Mikolov, T. et al. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv.

[7] Resnick, B. (2017) 'The dark psychology of dehumanization, explained', Vox, 7 March. Available online at: https://www.vox.com/science-and-health/2017/3/7/14456154/dehumanization-psychology-explained

[8] Akhter, N. & Munir, A. (2016) 'The Battle of Badr: From challenge to chance', FWU Journal of Social Sciences, 10(1): 157.

[9] European Commission (2021) The Digital Services Act Package. Available online at: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

ISD | Powering solutions to extremism, hate and disinformation

PO Box 75769 | London | SW1P 9ER | UK
www.isdglobal.org