ISD | Powering solutions to extremism and polarisation

CASM technology

Methodological Discussion Paper

Analysing New Zealand Online Extremism

Carl Miller, Jakob Guhl and Milo Comerford

This report was delivered with support from New Zealand's Department of Internal Affairs

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Introduction

Across 2020, the Institute for Strategic Dialogue (ISD) and CASM Technology (CASM) have worked together to research the nature of online extremist activity related to New Zealand. Over that time, researchers have sought to answer a number of questions: how extremism manifests online in New Zealand, the platforms it is carried out on and any important differences between them, the scale of online extremist activity in both absolute and relative terms, and any important changes underway which are likely to alter this picture in the months or years ahead.

The findings of that research are presented in another paper which is a companion to this. The intent of this paper, however, is not to discuss what was found, but rather how it was found. In publishing the methodological, conceptual and ethical thinking and decisions that underlie the empirical findings, we hope that the lessons might be of some value and use to the research communities across the world that have similar research questions in mind.

This paper is split into two parts. In the first part, we discuss the methodologies and frameworks used to conduct the work as described above. In the second part, we discuss the research design from the perspective of research ethics and law, including approaches to ensuring anonymity, generating aggregated data and maintaining privacy online.

Part One: Research Methodology

Initial Consultation: Addressing Two Key Research Challenges

At the very beginning of the study, researchers from CASM and the ISD conducted a consultation with over twenty policy-makers from a broad range of relevant agencies and departments within New Zealand's government. Alongside these, interviews were held with select academic experts, as well as an ISD-facilitated focus group with a broad range of civil society stakeholders impacted by online extremism, selected by the New Zealand Government Department of Internal Affairs' Office of Ethnic Communities.

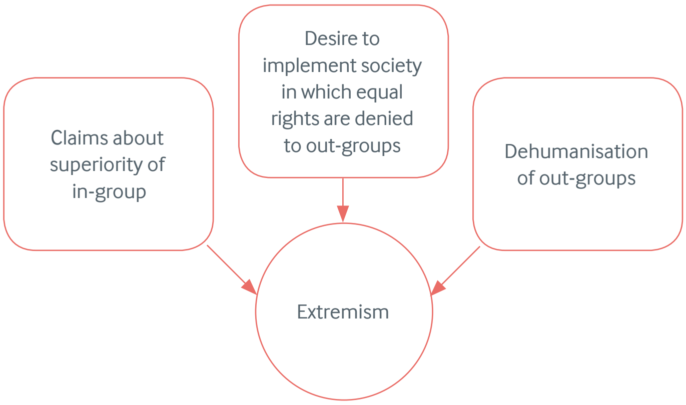
During this period of outreach and discussion, two challenges were identified as key to the sensitive and effective understanding of New Zealand and extremism:

1. The first was conceptual: the importance of robustly defining extremism. The research would only be successful if it could be isolated as both an idea and practice from the broader landscape of potentially harmful online content and behaviours, including hate speech, disinformation and conspiracy theories.
2. The second was empirical: the research needed to be able to understand 'domestic' New Zealand-specific expressions of extremism amid a highly trans-nationalised online landscape, shaped and influenced by a whole array of international extremist voices and groups. At the same time, however, the research also needed to understand that broader landscape as something that could importantly influence and shape extremism within New Zealand.

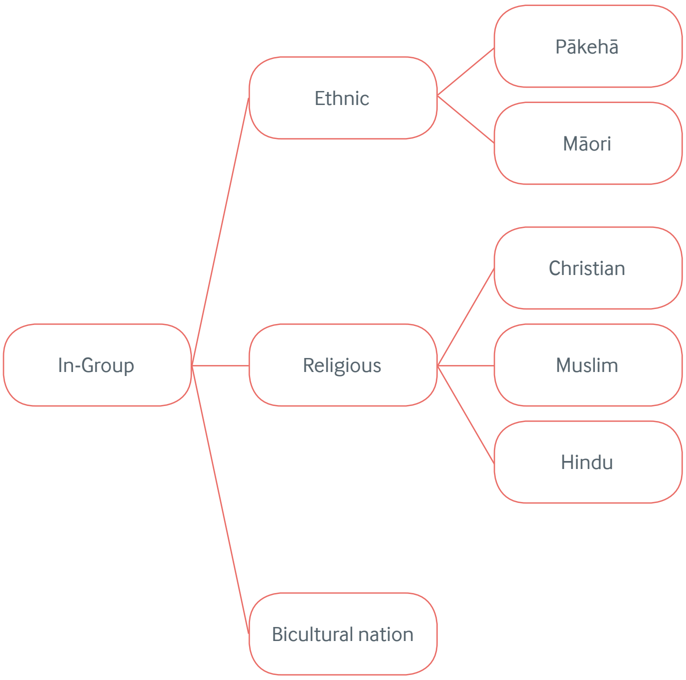
Defining Extremism

Extremism as a Concept

The boundary around exactly what is extremism, and what is not, is often nebulous, blurry and yet extremely important to draw. Across a large number of projects, ISD has developed a definition to help guide its classification of extremism which is based on 'social identity': the active pursuit and advocacy of systemic political and societal change, to reflect an ideology that claims the supremacy of one 'in-group' over all 'out-groups' and propagates the dehumanisation of that out-group. This understanding of extremism forms the basis of this research.


Claims about superiority of in-group
Desire to implement society in which equal rights are denied to out-groups
Dehumanisation of out-groups
Extremism

Within the New Zealand context, any support for political and social change that would advocate for the dominance of one specific, ethnically, politically or religiously defined in-group that would deny out-groups the right to equal rights, participation and belonging in New Zealand's society would fall under this project's definition of extremism. While this is not an exhaustive list, potential in-groups of extremist groups in New Zealand could be defined in terms of ethnicity, religion or an exclusionary interpretation of bicultural nationhood that dehumanises out-groups perceived as a "threat" to that status quo. Out-groupings might include Muslims, Jews, migrants, Māori, 'elites' and the rainbow community. It is notable that Māori were found to be framed as both an in-group and out-group by different far-right extremist constituencies in the course of our research.



Extremism and Adjacent Phenomena

The boundary between extremism and adjacent phenomena such as hate speech, conspiracy theories and disinformation online can also be nebulous, with content often falling within an ambiguous 'grey zone', which is not necessarily illegal, violent or in contravention of platform terms of service, but which may nonetheless be considered harmful or represent a precursor, permissive background or 'gateway' into extremism. While hate speech, conspiracy theories and disinformation can certainly be highly relevant to extremism according to this definition, we only analysed these phenomena when these have a demonstrable relationship to explicitly extremist narratives or communities, as according to the definition above.

This definition simultaneously narrows and broadens the scope of potentially relevant actors. It avoids a narrow focus on solely political violence and terrorism, by considering the non-violent promotion of extremist ideologies through politics, media and culture. In its emphasis on the advocacy for the superiority of one identity-based 'in-group', it does however draw a line between radical - but not necessarily supremacist or authoritarian - critiques of the status quo, and extremists who seek systemic political and societal change to subjugate of all 'out-groups.'

Extremism in Practice

The consultation guided the selection of the extremist ideologies that the report would focus on. Three were clearly identified by a range of experts having an established history and community within New Zealand: far-right, far-left, and Islamist.

The initial scoping drawn from the consultation was then subject to an initial phase of ethnographic research and discovery, to empirically confirm the presence of these extremisms across the platforms of interest, and to further validate the possibility of associating members of these communities with a high degree of confidence to New Zealand. This phase also exposed some notable absences of some forms of extremism with any kind of substantial and identifiably New Zealand presence: whilst evidence of misogynistic narratives was present in the data, researchers did not find evidence of cohesive New Zealand online communities of extremism-related phenomena such as Incel movements during the course of our study, although these could be the subject of further research.

Within far-right, far-left and Islamist extremist movements, there are considerable tactical and ideological differences. All these different sub-groups prioritise different issues, target different communities, use different methods and pose different threats. The question of the legitimacy of political violence is particularly controversial, with some violent extremists advocating for its use, while other extremists instead rely on non-violent political, educational and cultural means.

Far Right Extremism

'Far-Right' is an umbrella term for a broad range of radical right and extreme right ideologues, communities, and organisations. Various far-right actors have diverging views on whether a culturally, ethnically or racially defined group forms the basis of their in-group. This divide often correlates with diverging attitudes towards Judaism and Islam. While cultural nationalists often present themselves as defenders of a 'Judeo-Christian Western civilisation' which must be defended from Islam and Muslim immigration, antisemitism is central to most racial extreme right movements, including Neo-Nazis and white supremacist groups.

The anti-Muslim movement is a loose network of groups and individuals who believe that their cultures are threatened by an Islamic 'takeover'. They view Islam as a backward, homogeneous, static and unreformable threat to peaceful coexistence in Western societies, and legitimise rejection of and discrimination against Muslims on this basis. Anti-Muslim actors often seek to present themselves as 'non-racist', and opposed to antisemitism.

Ethnonationalism is a form of nationalism wherein the nation is defined in terms of ethnicity. Central to ethnonationalism is the belief that nations are tied together by a shared heritage and culture. Ethnonationalist youth movements such as the Identitarians are inspired by the French Nouvelle Droite (New Right), which developed concepts such as 'ethnopluralism' (the belief that people of different ethnicities should live strictly separated from each other to preserve 'pluralism') and 'the great replacement', which inspired the 2019 Christchurch attack.

White supremacists - who believe in the superiority of whites over non-whites, and advocate that white peoples should be dominant over non-white - were relatively rare in the context of New Zealand online extremism, though such ideas were expressed on Gab, Stormfront and Iron March. The ethnonationalist group Action Zealandia also shared individual pieces of content that suggest a belief in scientific racism (a pseudoscientific concept which posits the superiority of white people at a genetic level, especially in relation to intelligence). White supremacist groups tend to be overtly antisemitic, and view Jews as their main enemy, controlling perceived negative societal developments.

Far Left Extremism

Like 'far-right', 'far-left' is an umbrella term for a broad range of radical left and extreme left ideologues, communities, and organisations. Within the far-left, there are major disagreements about the legitimacy of political violence, the need for a state, whether to prioritise nationalism or internationalism, the relative importance of economic versus 'post-materialist' values (such as race, gender, LGBT-issues, identity politics, culture, individual autonomy and environmentalism) and their attitudes towards anti-Western authoritarian governments.

While both the radical and the extreme left aim at a systematic change of the capitalist system, the radical left does not explicitly oppose democracy, but often seeks to strengthen direct and local forms of democratic decision-making. In contrast, left-wing extremism is characterised by opposition to liberal democracy, sympathies for authoritarian regimes and conspiracy theories spread by them.

Our analysis of the digital far-left in New Zealand mainly focused on traditional communist elements supportive of a totalitarian state (unlike anarchist groups), showing ambivalent attitudes towards ('revolutionary') violence and leaning towards sympathies with anti-Western authoritarian governments that are perceived as 'anti-imperialist'. In addition to traditional communist groups, we identified a number of far-left 'Antifa' groups (short for 'anti-fascist'), a loose network often expressing a readiness to use violence against the far-right and/or the police.

Islamist extremism

According to our definition, Islamist extremists are united in their desire to create an exclusionary and totalitarian Islamic state. However, there is disagreement amongst Islamist extremists about the means through which to achieve this aim. While non-violent Islamist extremists attempt to use political activism to systematically change society and, in the long term, establish a totalitarian Islamic state, they do not use violence to achieve these goals. In the context of online extremism in New Zealand, the identified accounts were situated in the 'gray zone' between non-violent Islamist extremism and Salafi-jihadism. While some accounts analysed shared general Islamist extremist content, ISD also identified content supportive of Salafi-jihadist groups such as ISIS, as well as al-Qaeda-linked preachers such as Anwar al-Awlaki.

Conspiracy theories

Our consultation also revealed significant interest in the intersection of conspiracy theories with these forms of extremism, which led to the inclusion of this as a standalone analytical category for the research. Here, we used an overarching 'in-group and out-group' definition of extremism-relevant conspiracies, to help clarify the distinction between recognisably extremist versions of conspiracy theories and those which are not. This model was presented to DIA stakeholders, discussed and agreed before the empirical phase of the research commenced.

Extremism About New Zealand, and Extremism By New Zealanders

The second challenge identified by the consultation focused on an important complexity: many New Zealander extremists were likely well integrated into extremely international groups, subcultures and spaces online. However, it was also necessary to understand online extremist behaviour that was distinctly relevant to New Zealand.

To respond to this challenge, our research design is predicated on two different data pipelines, each intended to draw a different dimension of the overall situational awareness that we wanted to create. The first, the 'New Zealand pipeline' attempts to measure the nature and scale of activity from New Zealand extremists. The second, the 'international collection', attempts to measure relevant activity about New Zealand places, people and organisations from extremists located anywhere else in the world.

New Zealand Pipeline

To build the New Zealand Pipeline, the following steps were undertaken:

1. Qualitative Survey.

ISD researchers led an initial phase of qualitative analysis and open source investigation to develop a list of extremist accounts, channels and influencers which we had a high certainty were both based in New Zealand and, according to the definition offered above, extremist. The identification of these accounts proceeded in two steps. First, ISD researchers conducted ethnographic monitoring on platforms identified during the consultation as being broadly relevant to extremist groups in New Zealand. Following this sweep of platforms, they used a 'snowballing' approach to build out the lists of relevant channels, accounts and groups, using recommendation suggestions and other connected accounts. Secondly, each entity was manually reviewed by multiple expert researchers to ensure that their behaviour met the definitions of extremism that the project used. Entities were categorised as either aligning with a specific extremist ideology or a broader conspiracy theory-based worldview.

It should be noted that the thresholds for any account to be included in this list was high. Multiple researchers reviewed each entity to confirm that the account both pursued a social and political agenda of in-group supremacy, and also explicitly identified as a New Zealander.

2. Account-based collection

Data was collected from each platform of interest using the 'application programming interfaces', or APIs, that they make available for that purpose. Each of these APIs acts as a technical gatekeeper for data which, along with their associated Terms of Service, govern how data from each platform can be collected and used. These APIs were used to collect the following behaviour from each platform of interest:

Platform	Users	Records
Twitter	172	398,828
YouTube comments*	16,294 (commenting users)	100,678 (comments)
Facebook	76	79,092
Gab	16	16,488
Parler	27	7,091
Websites, forums	16 (websites)	4,906 (posts)
YouTube videos	6	942
Reddit commentators	109	234
Telegram channel	1	76

* N.B. 'YouTube comments' are comments made on videos sent by extremist New Zealanders, rather than comments made by extremist New Zealanders. In this sense, they are distinct from the other figures in this table. 'Reddit commentators' includes commenters to a single New Zealand extremist Reddit account, and so are also not necessarily extremist New Zealanders.

3. Thematic Classification

The third step was to categorise the collected records into one of a number of relevant themes. A common challenge that social media research faces is that the quantities of data returned are too large to manually read. Automated forms of analysis are therefore needed. Hence, once data was collected, an analytical architecture was built in order to identify themes in an automated way.

First, a form of 'topic modelling' was conducted. An unsupervised clustering algorithm was used to attempt to identify clusters of language that might represent a distinct topic or theme. This process was conducted iteratively, with a researcher manually checking the results and then re-applying the clustering algorithm.[1] This process identified six key themes. While these are themes widely discussed among extremists, this does not automatically imply that the nature of discussion of these themes is inherently extremist:

- 'The economy'; including discussion of conspiracies surrounding economic policies and suggestions of preferential treatment to some groups.
- 'The environment'; including references to environmental conspiracies such as those surrounding the use of 1080 poison, UN Agenda 21 and climate change denial.
- 'Health'; including discussion of health-related conspiracy theories, such as those surrounding Covid and the rollout of 5G and Covid denialism.
- 'Politics'; including polarising discussion around elections, referenda, the glorification of political violence, and mention of groups directly related to political conspiracies or extremism.
- 'Race'; including the excusing or justification of racial violence, references to white supremacy, 'positive discrimination', reverse racism, references to no-go zones.
- 'Religion'; including anti-Muslim speech, references to religious groups controlling media, and references to the persecution of Christianity and Christians.

Next, a keyword-based filter was constructed. These were wide lists of language that correlated to each of the themes identified above.

Construction of these keyword lists was a multi-stage approach. Analysts came up with an initial list for each theme, based on the results of the clustering analysis (see above). High precision keywords were selected from these results, to form the initial keyword lists for each theme. To augment these lists, we then took these keywords and looked for words that commonly appeared together with them in documents belonging to the same cluster. This used a statistical measure of how important each word is to each document within a larger corpus, called TF-IDF.[2]

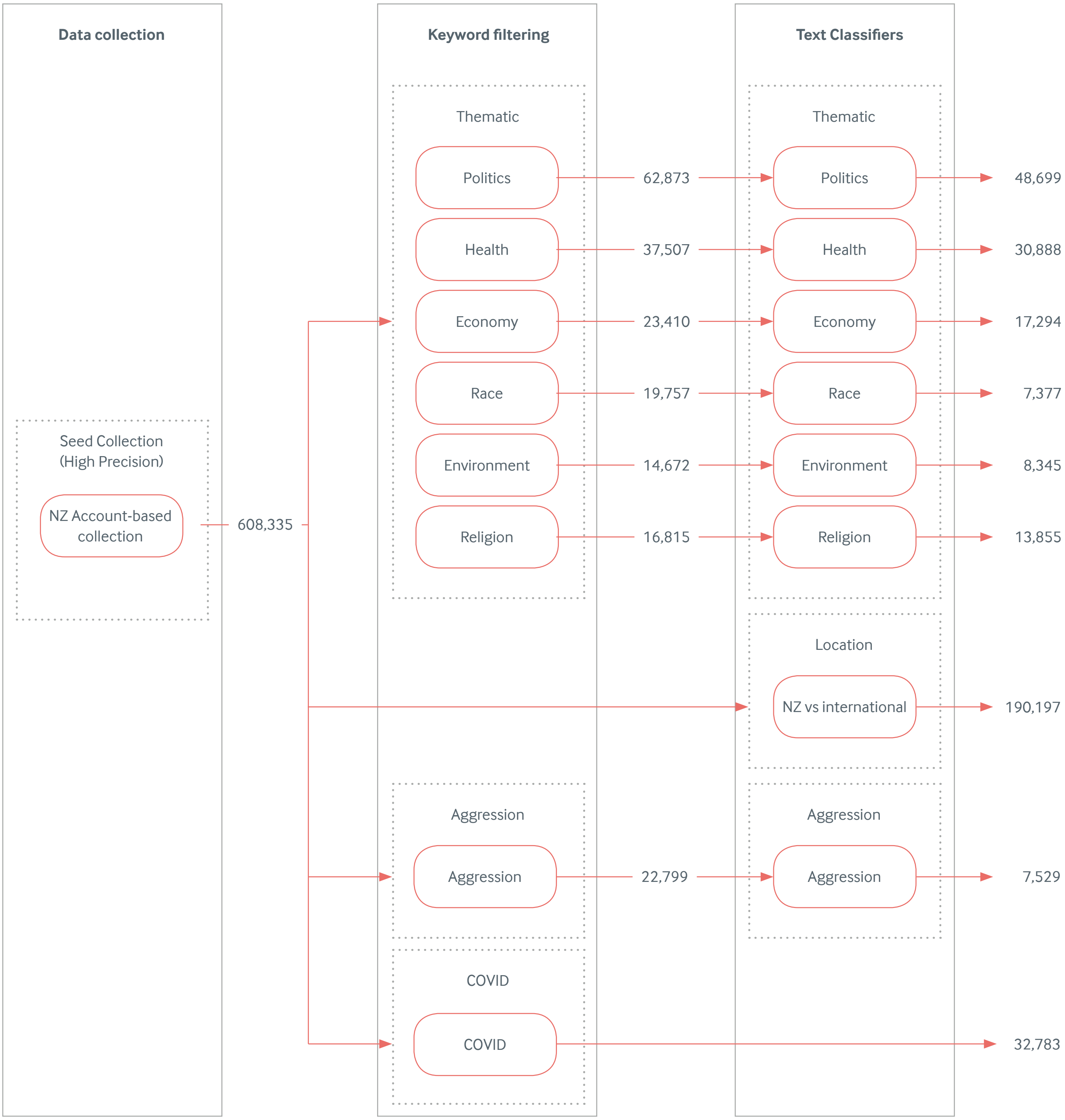
These augmented lists were manually reviewed again and filtered down into our final keyword lists. Records that did not contain any of these words were discarded, to remove records whose content were unlikely to be relevant to any of the themes in question.

Then, a series of natural language processing algorithms were trained to identify whether any social media post (containing, at this stage, one of the keywords) was relevant to any of the themes. To train each classifier, a human analyst marked up a series of social media posts into one of the six categories of interest, or indeed into none of these themes. The classifier then attempts to statistically understand the distribution of words and phrases between these categories, and on the basis of that understanding, will then place additional social media posts into these categories. The training of the Natural Language Processing algorithms adhered to the following process:

- **Definition of categories.** The formal criteria explaining how each record should be annotated is developed. Practically, this means that 'the economy', 'the environment', 'health', 'politics', 'race' and 'religion' were each defined, and a number of edge cases were surfaced and discussed. In this way, the exact definition of the categories developed throughout an early interaction with the data. This process ensures that the categories reflect the evidence, rather than the preconceptions or expectations of the analyst.
- **Creation of a 'gold standard' test set.** This phase provides a source of truth against which the classifier performance is tested. 100 records were randomly selected to form a gold standard test set. These are manually coded into the categories defined during the prior stage. The records comprising this gold standard are then removed from the main dataset, and are not used to train the classifier.
- **Training.** This phase describes the process wherein training data is introduced into the statistical model, called 'mark up'. Through a process called 'active learning', each unlabelled record in the dataset is assessed by the classifier for the level of confidence it has that the record is in the correct category. We then select the records with the lowest confidence scores, and these are presented to the human analyst. The analyst reads each Tweet, and decides which of the pre-assigned categories (see Phase 1) that it should belong to. A small group of these (usually around 10) are submitted as training data, and the NLP model is recalculated. The NLP algorithm then looks for statistical correlations between the language used and the meaning expressed to arrive at a series of rules-based criteria, and presents the researcher with a new set of records which, under the recalculated model, it has low levels of confidence for.
- **Performance Review and Modification.** The updated classifier is then used to classify each record within the gold standard test set. The decisions made by the classifier are compared with the decisions made (in stage two) by the human analyst. On the basis of this comparison, classifier performance statistics – 'recall', 'precision', and 'overall' (see 'assessment of classifiers', below) - are created and appraised by a human analyst.
- **Retraining.** The steps above are iterated until classifier performance ceases to increase. This state is called 'plateau', and, when reached, is considered the practical optimum performance that a classifier can reasonably reach. Plateau occurred once a number (ranging from 104 at the lowest to 225 at the highest) of annotated records had been marked up as training data.

These steps, together, created a pipeline that can be logically drawn as overleaf:

Diagram of Pipeline 1 Architecture



4. Measuring Performance of Natural Language Processing

The use of machine learning for natural language processing is inherently probabilistic, and inevitably entails an error margin. To calculate the performance of each of the classifiers used for this project, as stated above, 100 social media posts were sampled at random and annotated by an expert researcher according to relevance to any of the themes. This is called the 'gold standard test set'.

The decisions of the trained classifier for each of these records were then compared to those of the human: where they agree, it is understood that the classifier was correct; where they disagree, it is understood that the classifier was wrong.

On this basis, the performance of the classifiers was measured in three ways:

- Precision: Of the records marked as relevant to the theme, how many will actually be relevant?
- Recall: of the records actually relevant to the theme, how many did the classifier find?
- F-score: a floating harmonic mean of precision and recall.

Theme	Precision	Recall	Fscore
Politics	72.2%	92.9%	81.3%
Health	74.4%	89.7%	81.3%
Economy	67.1%	79.7%	72.9%
Race	63.6%	67.7%	65.6%
Environment	79.4%	76.9%	78.1%
Religion	76.5%	87.3%	81.6%

Two 'cross cutting' analyses were also created, applied across each of the pipelines, so they could create a series of more universal measurements across the data. The performance of this classification was as follows:

- 'COVID' - a keyword-based annotator which sought to identify whether any record contained one of a number of words strongly associated with COVID-19 and the pandemic.
- Aggression and calls to action – a further classifier was built to detect aggressive or violent language use by extremists, alongside concrete calls to action against a perceived existential threat from an out-group. Examples of relevant linguistic markers for aggression include calls by extremist New Zealanders to 'fight back', against 'traitors' in a coming 'civil war', with irrelevant results filtered out by a trained classifier. Its performance is detailed in the table below.

Classifier	Precision	Recall	Fscore
Aggression/Call to action	71.40%	56.80%	63.3%

International Pipeline

1. Account-based Collection

The second pipeline was intended to understand the extent to which international extremist mobilisations online conducted behaviour which was about New Zealand and the people, places or organisations within it.

Drawing on previous research conducted either by CASM or ISD across a range of Anglophone country contexts, a total of 5,661 accounts which had previously been qualitatively gathered through ethnographic analysis across a range of online platforms, were added to the data system. The key inclusion criterion was that the account had to be identified as an expert researcher as extremist, but did not, of course, have to be based in New Zealand.

In a process similar to the data collection step of Pipeline 1 social media data was collected that was associated with those accounts. This resulted in over 26,000,000 records being collected, including over 16,000,000 records from the /pol/ board of 4Chan.

Platform	Records
4Chan /pol/	16,077,002
YouTube comments	5,080,038
Twitter	2,439,551
Facebook	2,130,808
Telegram	1,231,971
Reddit	256,969
Gab	78,690

2. Geoparsing

A workflow was then built to identify as many references to New Zealand as possible across all of the posts sent by international extremists and collected as detailed above. We call this a geo-parsing architecture, and it has a number of sub-steps:

- **Ingestion.** All the posts collected from international extremists are turned into a single data structure and fed into the geo-parsing architecture.
- **Named Entity recognition (NER).** This is the process of identifying sequences of text in a document which correspond with entities that exist in the world, along with classifying them into categories such as whether they are likely to be a 'Person', 'Organisation', or 'Location'. To do this, we applied the Stanford Named-Entity Recognition (NER) tool to recognise entities within the messages.
- **Geo-parsing.** Geoparsing is the joint process of (a) recognising place names mentioned within documents and (b) resolving these place names to their corresponding geographical coordinates. CLIFF, an open-source tool that performs both of these steps was employed for this process. To disambiguate which place-on-earth a given location extraction belongs to, CLIFF utilises the GeoNames geographic database. Geoparsing is still very much an open area of research, and this process is by no means guaranteed to either identify all location mentions, nor resolve them to the correct place-on-earth. To help improve the precision (the number of places identified to be within New Zealand that are within New Zealand), and recall (the number of New Zealand locations picked up by the geoparser) for New Zealand-based geoparsing, a set of additional rules were developed around the output of CLIFF. These rules include single high-precision terms (e.g. "NZers", "Zealanders"), multi-term heuristics ("Napier" in the same passage of text as one or more high-precision terms), and exclusionary rules to remove places that have been resolved to New Zealand which should not be (most notably: "East Coast", "Pearl Harbour", and "Mankind"). Rules were developed in an iterative manner, using a random sample of project data to identify and validate improvements. From a random sample of 100

international posts, the geoparser identifies posts mentioning New Zealand locations with 96% precision.

- **Resolving Regional-Level Mentions in New Zealand.** Using the geo-parsing architecture, all posts were annotated with country-level mentions for all countries. CLIFF also uses the GeoNames database to annotate each extracted place with intra-country level information, if applicable. For New Zealand places, these are the 16 Local Government Regions. This was used to annotate all posts where this information existed.

3. Classification

All posts that the process above had determined to be about New Zealand were then passed through the cross-cutting classification architecture described above, seeking to identify COVID-related vocabulary and aggressive or violent language. The results, as with the other pipeline, were then outputted for visualisation and interpretation.

Part Two: Research Ethics and Design

It was essential that this work be conducted according to the highest research ethical standards. With a focus on extremism, this research takes as its main theme a social phenomenon which itself is a matter of passionate and often controversial public dispute and debate. It combines this with the use of technologies and analytical methods which may be unfamiliar to many people, both in how they work and the nature of what they produce. This makes it extraordinarily important both for ethical challenges to be explicitly identified and navigated, and for them to be done so in a way that is transparent.

An ethical framework was constructed at the outset that tried to balance two different public goods: on the one hand, there was the public good of privacy and autonomy online, and on the other the public good of social cohesion, public security and safety. The aim of the framework was to help shape a project that could provide the clearest, most accurate picture possible for the New Zealand government about extremism online, whilst doing so in a way that would not constitute in any way an intrusion on any individual given the data being collected, and where the research would not create any risk to any individual

This framework constituted the following foundational principles and values:

Only completely public spaces were studied at scale

Issues of privacy online are complex, and an important part of online research ethics is the recognition of the many different kinds of online space which are used. Sometimes within the same platform, there are actions (such as direct messages) which most users consider private, and others (such as Tweets) which are widely understood to be entirely public.

In addition to this, public perceptions of what social media spaces are public and which are private can vary significantly from the legal reality or terms of service. Because these discrepancies between reality and perception relate to issues of autonomy, where information that is not public or information that might reasonably be perceived as private is sought within a research project, the acquisition and recording of this data must be well considered, justified and documented.

A governing principle of this project was that no space would be studied where users would have a reasonable expectation of privacy of any kind. In many cases, study of these spaces is also technically impossible, because data from them are not made available by the platforms themselves. But regardless of any technical possibility, no spaces were researched where action has been taken to restrict viewership of any piece of content. This would include:

- Of websites, any explicit action to preclude data scraping (either technically, or stated as policy on the website itself);
- Any online space that is password protected;
- Any online space requiring especial membership of a given group
- Any non-public part of any social media platform, including private Facebook groups, pages and so on;
- Any space or post that is geo-gated to restrict its viewership.

Aggregate and Anonymous Outputs Only

All research was conducted on the foundational principle of respect for the persons present in the online spaces which we will be studying, and for the privacy of these persons. In order to mitigate the possibility of any individual harm to any person as the result of the research, the only outputs that were produced were aggregated and anonymous. The research was exclusively focussed on the understanding of broad, strategic trends and patterns over time and across platforms, and this meant that no individual was named during the research, and no individual-level behaviour was described, unless the individual in question could be assumed to be highly visible publically already. In some cases, quotations were used in order to illustrate a particular point, however these were bowdlerised to prevent the retroactive identification of the original post through, for example, an online search. As standard, all efforts comply with UK data regulation and GDPR requirements, as well as New Zealand's 2020 Privacy Act.

A Clear Framework for Managing Risk in Digital Ethnographic Work

As explained above, ISD conducted a range of qualitative digital ethnography as part of the mixed-methodological approach to this project. Here, it was recognised that digital ethnographic approach may use online data without the consent of research subjects, and so it was governed by a separate framework developed by the ISD that included:

- Anonymity: The anonymity of all research subjects must be guaranteed through research method (including the use of permanent de-identification where possible, the maintenance of a separate and secured coded name register where this is required by the research, and the limitation of access to identifiable data within ISD).
- Identifiability: Any case studies, quotations, examples or any individual level data of any kind must be non-identifiable, including where a third party might undertake research to identify ISD research participants.
- The Minimisation of Deception: No deception was used during the course of research to access closed groups or influence research participants. The principle guiding research that includes deception must be that the minimum level of deception must be practiced, and where the importance of the research and its public benefit cannot justify deception or a level of deception, the research method must be changed or the research should not take place.

No Intent to Identify Criminality or Inform Law Enforcement Investigation

This work was undertaken as a piece of research to identify broad trends and patterns, conducted by researchers entirely outside of the warranted powers of a law enforcement organisation. It was not intended, nor designed, as an investigatory project to uncover criminality of any kind. In addition to the fact that the research did not reveal the identity of individual involved, except for the caveats outlined above, two other important aspects of research design were explicitly chosen:

- No part of the research architecture (and especially the automated classification, see above) was designed to identify behaviour that passed any kind of criminal or legal threshold. The research was not trying to find or measure criminality.
- The research outputs were not shaped in any way to guide or inform any law enforcement investigation or organisation.

A Clear Referral Process

This project was not designed or directed to identifying criminality. However, it was important that researchers clearly understood what to do if they encountered behaviour online that implied the presence of a credible real and immediate threat to a loss of life, threat to cause serious harm or threat of injury to another. This includes serious sexual assault or rape, specifically targeted towards individuals, groups, events or places. ISD and New Zealand government partners agreed a referral process whereby researchers reported to relevant authorities any documents encountered during the course of this research, which might be identified as representing a real and present threat as described above.

Honesty and Transparency in Intent and Method

An important part of this project has been to be as clear as possible about the research itself: that it was carried out, the intent in doing so, how the research was conducted, and the various principles and values that guided the research team when they did so. Indeed, that is an important reason why this document has been written.

Alongside transparency, it was also considered especially important that researchers be as honest as they could possibly be regarding the research methodologies that they employed. In part, this meant that researchers would make efforts wherever possible to objectively measure the accuracy and performance of the methods that were being used, especially the natural language processing classifiers described above. In addition, however, researchers needed to make explicit the the limitations and weaknesses that any given method entails, and the caveats, therefore, that findings produced by such a method must be read against. It was only on this basis that the research could constructively contribute to appropriate thinking and decision-making that it might inform or add value to.

The Ethics of Inaction

As stated at the beginning of Part Two, the ethical framework was predicated on the balancing of two public goods: privacy and social cohesion and public safety. As such, researchers needed to also give consideration to the argument that in some cases a failure to conduct online research might itself be ethically unsound, particularly where such research might inform activities to improve social cohesion, public safety and to reduce hate crime and violence.

As such, the principles above were used to shape work that, we believe, minimises the individual harms that research can impose whilst maximising the capacity of the research to contribute to public goods. We regarded this to be the most ethical course of action to take.

ISD | Powering solutions
to extremism
and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org