BIRMINGHAM CITY University

Extremism Online - Analysis of extremist material on social media

Professor Imran Awan, Hollie Sutch and Dr Pelham Carter

Abstract

This paper examines the role of anonymity as a driver of extremist and Islamophobic language use on Twitter, as well as looking at the difference in language used and emotion expressed in response to both YouTube videos inspired by offline events (Event Specific) and those that are general online based discussions (General Discussion). Four approaches have been used in this paper; 1) a quantitative corpus linguistic analysis of extremist language on Twitter in relation to levels of user anonymity; 2a) a corpus linguistic analysis of conspiracy and media bias language posted in response to two categories of YouTube videos; 2b) a sentiment analysis of the emotional content of comments/language in response to two categories of YouTube videos; and 3) a qualitative thematic analysis/case study of comments in relation to two specific offline recent events. Over 100, 000 tweets, and over 100, 000 YouTube comments were collected for the quantitative analysis, and approximately 600 tweets were collected for the case study based analysis. Our findings suggest that increased anonymity is associated with an increase in extremist language, that conspiracy theory and media bias based language is more common in response to offline events than general online discussion, and emotional sentiment around fear, anger, disgust is more common for Event Specific videos but there is significantly more hatred and pride expressed in the General Discussion comments. This is explored further in the thematic case study analysis where the

data suggests these findings relate to how individuals seek to marginalise or lessen the impact of real offline events and how pride in identity and hatred can be used in the absence of a real world event to refute. Implications for policy such as soft verification, traffic light warnings and better reporting facilities are also discussed.

Introduction

This paper examines the role of extremism online and uses two primary studies to generate empirical evidence that examines the differences between general online discussion of extremism (General Discussion) and discussion inspired by offline events (Event Specific) through the analysis of tweets and YouTube comments. Though both sets of comments are situated online, only one set is in direct response to a public and identifiable offline event. We focus on two offline events (the Shamima Begum case and the New Zealand Christchurch terrorist attacks). The reason for choosing these social media platforms was mainly because of the role of anonymity and users numbers. This research will be adopting a more holistic definition where extremism is not limited to a single type, it is not restricted to violence and should be carefully distinguished from terrorism – although, it can include, but is not limited to actions and speech associated with terrorism. It can take place across different situations and platforms, due to globalisation and the ease of use of the internet, people can disseminate extremism in a variety of different ways both online and offline.

With the developments and ever-increasing use of the internet, particularly social media (Kemp, 2018), there has been an increase in extremism and its dissemination online (Awan, 2016). The internet is distinguished for its ability to allow information dissemination, provide conversation and support public debates, and now also as a tool that allows individuals to promote extremism and reinforce prejudices (O'Hara, & Stevens, 2015).

Incidences of online extremism and prejudice has contributed to the marginalisation of different communities (Barlett, 2010).

Research has suggested that the virtual nature of online extremism is a case of old crime in new bottles, assuming that it is in fact no different to terrestrial crime such as extremism and hate that occurs offline (Grabosky, 2001). The issue with this postulation is that cyber-mediated crime such as online extremism is not comparable to conditions of offline crime, the cyber nature of crime provides criminals with alternative methods of access and opportunities for harm and disorder (Brenner, 2004). The infrastructural properties of the internet particularly on social media sites can provide individuals with different opportunities when distributing extremism online, compared to disseminating extremism offline. Extremists may take advantage of how the internet provides the potential for greater reach and impact and an ability remain anonymous.

The infrastructure of social media networks can provide individuals with the ability to remain effectively anonymous online. The lack of a real name policy means users can disseminate and access information without being identified (Peddinti, Ross, & Cappos, 2017). It has been suggested that being behind a computer compared to face to face communication provides individuals with a sense of security and anonymity, which can cause them to act differently (Peebles, 2014). This has important implications as we know that this can lead to deindividuation which is when individuals use anonymity to cause group members to fail to acknowledge themselves as individuals (Zimbardo, 1969). Research has also demonstrated that the level of extreme narrative disseminated online can be a product of a user's levels of anonymity (Zhou, Qin, Lai & Chen, 2007).

The design of social media networks can create echo chamber environments. Social media sites can facilitate the dissemination and polarisation of extremism through the development of echo chamber environments (Awan, 2016). These echo chamber

environments can form as a result of how the internet has created an environment which prevents opinionated discussions, but rather encourages individuals to engage with like-minded individuals whose beliefs and views are aligned (Sunstein, 2001). Social media platforms produce an environment for like-minded individuals to cluster together and provide reaffirmation, these echo chamber environments are protected from alternative perspectives which therefore works to reinforce its beliefs and confirm any biases (Garrett, 2009; Sunstein, 2001). In this way, groups can be formed, and we can see sections of the online community become cyber mobs. Research has documented how these echo chamber environments can manifest differently on different social media sites. Facebook is regarded as being a platform that provides opportunities for individuals to make strong relationships with others, whereas Twitter is a stronger facilitator of information dissemination (Gruzd, Wellmen & Takhteyev, 2011). Research states that this difference highlights that Twitter provides a platform that allows strong echo chamber environments, that contain online extremism, to thrive (Gruzd, & Roy, 2014).

When examining the infrastructure of social media sites there are growing concerns for the material that is being accessed, circulated, promoted and supported within echo chambers. There is a continued reduction in the quality and source credibility of material/information transmitted on the internet and social media (Törnberg, 2018). A current trend in information sharing has seen high volumes of fake news, biased narratives and conspiracy theories (Törnberg, 2018). The problem as a result, highlighted by research, is how the propagation of fabricated material appears to disperse quicker and reach further audiences compared to legitimate news and sources (Vosoughi, Roy, & Aral, 2018). Researchers demonstrate how false news and misinformation on social media sites can encourage extremism (Lazer et al., 2018). For example, a story in 2017, emerged where a Russian bot falsely tweeted that a Muslim woman dressed in a Hijab (headscarf) had ignored

victims of the Westminster terrorist attack (Dixon, 2017). The image went viral after being sent by a Russian bot, who had history of disseminating fake news, which consequently fuelled a racist backlash against this woman and more wider the Muslim community (Dixon, 2017).

Extremist Language and Trigger Events

It is important to note that current events both in Britain and internationally have raised the issue of online extremism more prominently in the public eye. Quite often this is exacerbated when commentators and even politicians issue statements that refer to groups in a demonising manner. For example, when the former Foreign Secretary, Boris Johnson, described Muslim women wearing the burqa as being bank robbers and letterboxes, the incident caused a backlash and an increase in reporting of Islamophobic abuse. As the growth of extremism online continues, the need to tackle the root causes of where hate and extremism merge together has become critical. For example, research shows that trigger events can lead to reprisal attacks and an increase in the level of extremism and hate online. We argue that such incidents can provide a link between people's behaviour online and actual incidents reported offline.

From reviewing the literature, it is questionable whether the infrastructure of social media networks provide individuals with a sense of ease and ability to distribute extremist content which may be more harmful and radical – compared to extremism disseminated offline. It is clear from recent events and the literature that issues such as anonymity and online communication, the language used by extremists generally against Muslim communities, and the potential difference in comments between content inspired by offline trigger events and more general online discussion are all of importance. To address these issues our paper firstly applies corpus linguistics, the study of language in context and the comparison of different bodies of language (corpora or texts), to investigate the impact of

anonymity on extremist language frequency in general online Twitter comments. We then repeat this investigation of language used and frequency by comparing the use of language relating to conspiracy, fake news and media bias on YouTube, comparing the comments for videos that are inspired/related to offline trigger events like the Christchurch terrorist attack and the recent case of Shamima Begum to more general videos considering Islam that have not been posted in response to a clear and identifiable offline trigger event.. Using the same comment data in response to the offline trigger inspired videos and the General Discussion videos we have run further analysis of the emotional sentiment expressed within the comments (rated on affective dimensions such as fear, disgust, pride, hatred etc.), allowing exploration of emotional differences as well as language differences. To gain further detail and depth of understanding we have also conducted an in depth qualitative thematic analysis of 600 tweets around two cases studies (comments in response to either the Christchurch terrorist attack or the Shamima Begum case).

Study 1: Corpus linguistic analysis of Anonymity effects on Twitter

The following study was conducted to determine whether differing levels of online anonymity would impact on the type and level of extremist language used on Twitter.

Method

Design

This research utilised a cross-sectional design, comprising of a corpus linguistic analyses. This is an approach that quantitatively examines natural occurring language by comparing large bodies of text (corpora) against others to determine statistical differences in the frequency of language used, how keywords are paired and how they are paired differently between texts. A corpus linguistic analysis allows for the handling of large data sets and keeps corpora intact, whilst producing a high statistical reliability (Kennedy, 2014). A corpus

linguistic analysis was utilised to assess the association between anonymity and a Twitters users' level of extremism online. For the corpus linguistic analysis anonymity was operationalised by scoring usernames and profiles on the amount of personal and identifying information provided, with more anonymous users having lower scores and more identifiable users having higher scores (see appendix A for scoring). Visual binning was conducted on spss for the corpus linguistic analysis to create categorical variables for anonymity, this process effectively creates new variables by grouping original data sets into clear categories- producing categorical variables . In total, anonymity had three categorical variables; low anonymity (five to six identifiable items), moderate anonymity (three to four identifiable items) and high anonymity (one to two identifiable items). For example, users with low anonymity were classed as having high identifiability online, where as those with high anonymity were classed as having low identifiability online.

Participants

When employing corpus linguistic techniques there is not an agreed upon corpus size, as this does not always guarantee its usefulness (Hiltunen, McVeigh, & Säily, 2017), although Haber (2015) suggests that with utilising these methods on Twitter, in order to gain a mean variability of data, 200 tweets per users is generally recommended. This requirement was exceeded, a total of 205 Twitter accounts and 102,290 tweets were examined for the corpus linguistic analysis.

Materials

For the purpose of the corpus linguistic analysis, using literature regarding online extremism and Islam (Awan, 2016; Sutch & Carter, submitted), a word list containing five extreme words/phrases such as Islamisevil and Islamiscancer was generated (appendix b). The current research utilised multiword-units for the word list as it is argued that it can reduce the risk of missing important findings (Johansson, Kaati, & Sahlgren, 2016). The

research also required the use of Antconc computer software (version 3.5.7) (Anthony, 2018), to perform keyness analyses for the corpus linguistic analysis. In order to collect the original tweet data, the research utilised FireAnt (version 1.1.4) (Anthony and Hardaker, 2017) which can be used to collate a large corpus of tweets. Twitter archiving google sheet (TAGS) was also used during the research as it performs automated collection for search results from Twitter.

Procedure

Utilising a non-traditional snow ball sample, search terms (words from the extreme word list) were used to identify an initial pool of Twitter users and through examining their account characteristics i.e. followers and communications, further Twitter users were identified and selected for this research. Twitter users were identified and rated based on their level of anonymity, by documenting the number of identifiable items their account contained. Using FireAnt, tweets were collected from users' accounts to create a corpus of data. Utilising this data, AntConc was used to run three keyness analyses on the three categorical variables of anonymity, this demonstrated which variables of anonymity were significantly associated with the extreme word list. A keyness analysis essentially identifies significant differences between two corpora of data, for the purpose of this research utilising keyness analyses displayed significant differences between the sets of corpora, which were distinguished based on their differing levels of anonymity. For the purpose of this research there were a total of three different levels of corpora meaning three keyness analyses were required to compare all the corpora against each other.

Data analysis strategy

A total of three keyness analyses was performed for the corpus linguistic analyses for study 1, this method was chosen as it can perform a log-likelihood test, this highlighted which words are significantly associated with the extreme word list and anonymity categories

(Anthony, 2004). Words that are more key for one corpus are words that occur significantly more frequently within it than the comparison corpus.

Results

A total of three keyness analyses were performed on the three categorical variables of anonymity. These keyness analyses highlighted whether the variables were significantly associated with the extreme words from the word list. Negative keyness values represent words that are unusually infrequent compared to words in a reference corpus (Anthony, 2004).

Table 1

Keyness analysis for Moderate Anonymity and suggested extreme words

	Moderate Anonymity comparisons					
	Moderate anonymity vs Low anonymity			Moderate anonymity vs High anonymity		
Key terms	Frequency value	Keyness value	Significance value	Frequency value	Keyness value	Significance value
Islamisevil	-	-	-	24	-628.11	<.001
Stopislam	-	-	-	8	-719.35	<.001
Banmuslims	-	-	-	12	-332.72	<.001
Islamiscancer	-	-	-	24	-264.07	<.001
Islamonazism	-	-	-	-	-	-
Islamistheproblem	-	-	-	123	-400.84	<.001

Table 2

Keyness analysis for High Anonymity and suggested extreme words

	High Anonymity comparisons	
	High anonymity vs Low anonymity	High anonymity vs Moderate anonymity

Key terms	Frequency value	Keyness value	Significance value	Frequency value	Keyness value	Significance value
Islamisevil	1136	+342.8	<.001	1136	+628.11	<.001
Stopislam	1124	+339.18	<.001	1124	+719.35	<.001
Banmuslims	596	+179.83	<.001	592	+332.72	<.001
Islamiscancer	576	+173.81	<.001	576	+264.07	<.001
Islamonazism	573	+172.91	<.001	573	+405.35	<.001
Islamistheproblem	1339	+164.66	<.001	1339	+400.84	<.001

The tables above represent the comparisons for the three levels of anonymity. Results illustrate that the high anonymity corpus contains more words which occur significantly more frequently when compared to both the low anonymity corpus and the moderate anonymity corpus. There is little difference when comparing the frequency of the Islamophobic terms used between low to moderate anonymity. All of the keywords above appear statistically more frequently in the tweets of high anonymity users than they do in either moderate or low anonymity users, clearly suggesting increased anonymity may be predictive of increased Islamophobic language use. Specifically, highly anonymous users (those with only two, maximum, identifiable details) are much more likely to use extremist terms than those with low to moderate anonymity (those with three or more identifiable details/items).

2a. Event Specific versus General Discussion YouTube Comments: Conspiracy Theories

The second part of the study was conducted to determine whether there is a relationship between the type of discussion around videos (that are either Inspired by Offline events, or those that are more general General Discussion videos) and the use of language around conspiracy theories.

Method

Design

This research utilised a cross-sectional design, comprising of a corpus linguistic analysis. A corpus linguistic analysis was utilised to assess the association between video type and a commentator's level of conspiracy related language. For the corpus linguistic analysis, the operationalisation of video type differentiated between videos that discussed specific offline events such as the New Zealand terror attack and the Shamima Begum case (Event Specific) and videos that discussed Islam in general without reference to one specific event (General Discussion).

Participants

All video searches were filtered by view count and selected in order of view count. Videos were collected until the overall comment threshold of 50,000 for each category was met. Initially 229, 182 comments to appropriate YouTube videos were collected. Due to the nature of our analysis and YouTube's comment system any replies to comments were removed, as were any comments that were blank, contained only numbers or non-English characters. After filtering a total sample of 107, 582 YouTube comments remained with 62, 374 being Event Specific, and 45, 208 being General Discussion. All of these comments were replies to or comments on the video.

Materials

As in study 1 the research required the use of Antconc computer software (version 3.5.7) (Anthony, 2018), to perform keyness analyses for the corpus linguistic analysis.

Procedure

YouTube videos relating to or inspired by recent offline events were selected by using associated search terms (such as Christchurch terrorist attack or Shamima Begum). For YouTube videos relating to general discussion about Islam, and not related to a specific offline trigger event were selected. The comments were collected using a YouTube comment

scraper, and converted into individual text files using a custom visual basic script, and were then combined into larger text documents to form the corpora.

Results

Table 3

Keyness analysis for conspiracy terms between Event Specific and General Discussion groups

Key terms	Frequency value	Keyness value	Significance value
Fake	693	287.822	<.001
Flag	229	67.950	<.001
MSM	150	171.835	<.001
Left	1483	347.852	<.001
News	1581	679.147	<.001
Media	1900	684.013	<.001

Using a range of terms associated with the discussion of conspiracy theories (conspiracy, fake news, Mainstream Media/MSM, false flag) the corpus for Event Specific comments and General Discussion were compared. As can be seen in table 3, terms associated with the discussion of conspiracy theory appeared significantly more in response to videos that were Event Specific than in those about general online discussion. A positive keyness value in table 3 indicates a word that is significantly more frequent in the Event Specific comments, a negative value would instead indicate a word that was significant more frequent in the General Discussion comments. Phrases like ‘fake news’, ‘fake media’, ‘false flag/s’ and ‘left news/media’ were significantly more common. Though not included in table 3 tt is also worth noting that ‘media’ was significantly frequently paired with either ‘western’ ($p < .001$) or ‘fake’, indicating a potential duality in the discussion of media bias due to a

similar approach by users from either end of the political spectrum to discredit media sources. The pairing of ‘fake’ or ‘western’ both serve the purposes of assigning bias to a source or reducing the reliability of a source.However this relationship was significantly stronger for ‘fake’ than it was ‘western’. This has clear implications for how terminology surrounding conspiracy theories, media bias and false flags is used in response to specific offline events rather than being a common part of negative online discussion. This may represent an attempt to lessen the impact of events that may garner sympathy towards the Muslim community, and lessen the negative impact on those holding an Islamophobic ideology. The attacking and undermining of media sources would be a potentially viable strategy in general discussion comments also, but as evidenced here it does not appear to be occurring frequently, with users in a general online discussion favouring other discursive strategies (see 2b).

2b. Event Specific Versus General Discussion YouTube Comments: Sentiment Analysis

In this part of the study the same YouTube data set is used but was analysed via Sentiment Analysis. This allows for individual comments to be rated according to the emotional sentiment expressed within them and then further analysed using traditional statistical measures to determine any group based differences between the Event Specific and General Discussion video comments.

Method

Design

The online data for the sentiment analysis comprised of the data set of YouTube comments from videos of offline events such as Christchurch and Shamima Begum, and more general comments on videos found using search terms presented in study 1. This was an independent design with video type acting as the independent variable (Event Specific Vs. General Discussion) and the Geneva Affect Label Coder (GALC)(Scherer, 2005) sentiment

scores (for affective dimensions such as fear, anger, disgust, hatred etc.) acting as the dependant variables.

Participants

The same data for 2a was used.

Materials

In order to perform a sentiment analysis, the Sentiment Analysis and Cognition Engine (SEANCE) (Crossley, Kyle & McNamara, 2017) was used.

Data Analysis

The sentiment data gathered via SÉANCE was analysed using a series of independent t-tests to determine group differences in the emotions expressed in the comments to the Event Specific and General Discussion videos. Due to the sample size parametric assumptions regarding normality and variance were assumed to have been met.

Results

Using SENACE the sentiment scores for the GALC emotional/affective indices were obtained for the Event Specific and General Discussion video comments. A relevant selection of these were then analysed using a series of independent t-tests. As can be seen from table 5 generally higher scores for fear, anger, anxiety, disgust, sadness and shame can be observed for the Event Specific comments.

Table 4

Selected Mean GALC scores for Event Specific and General Discussion Comments

Mean GALC scores				
	Event Specific		General Discussion	
GALC	Mean	Standard Deviation	Mean	Standard Deviation
Admiration/Awe	0.82	1.47	1.60	2.40

Amusement	1.05	1.29	1.38	1.25
Anger	1.76	1.50	1.08	0.93
Anxiety	0.10	0.27	0.09	0.23
Compassion	0.31	0.65	0.16	0.44
Contempt	0.07	0.27	0.06	0.15
Disappointment	0.17	0.86	0.29	0.81
Disgust	1.42	2.16	0.65	1.25
Dissatisfaction	0.00	0.02	0.01	0.10
Fear	13.43	5.80	2.68	2.00
Happiness	1.75	1.43	0.85	0.81
Hatred	1.68	1.47	2.06	1.48
Hope	1.34	1.14	0.86	0.74
Pride	0.12	0.35	0.20	0.62
Relief	0.02	0.41	0.01	0.05
Sadness	2.04	2.50	0.50	1.08
Shame	0.82	1.24	0.38	0.74
Tension/Stress	0.12	0.33	0.10	0.21

The independent t-test results in table 5, demonstrate several significant differences in sentiment score between the two groups. With scores for Anger, Compassion, Disgust, Fear, Sadness and Shame being significantly higher in response to the Event Specific than the General Discussion. The higher compassion and sadness potentially aimed at the victims of the events. The higher anger and fear may potentially relate to the threat of further attacks, reprisals or in the case of Shamima Begum the fear of allowing individuals to return. Levels of hatred, disappointment and pride for the General Discussion comments may reflect Islamophobia comments that do not have to contend with comments surrounding a specific event (and the related emotional response), increased targeting of outgroups. The increase in pride may also reflect statements around being proud of an identity that is then used in arguments within the General Discussion comments. Identity could be used to create a sense of threat to a group's identity (the in-group) to allow for anger to be directed at the out-group or the threats to that established identity Whilst those responding to specific offline events have an event to either rally against/for, or undermine the validity of through claims of

media bias and untruth as we have seen in 2a as a strategy, the more general discussion online may require the more abstract invocation of identity and pride as an alternative strategy. The anger and fear that is readily available in response to an offline event or threat may be replaced by the use of pride and identity, and by proxy a suggested threat to that identity.

Table 5

t-test results for GALC scores between Event Specific and General Discussion Comments, with a positive t value indicating higher scores for the Offline

Independent t-test		
GALC	t	Significance value
Admiration/Awe	-6.6	<.001
Amusement	-4.24	<.001
Anger	8.53	<.001
Anxiety	0.016	>.05
Compassion	4.21	<.001
Contempt	0.132	>.05
Disappointment	-2.36	.018
Disgust	6.75	<.001
Dissatisfaction	-2.78	.005
Fear	37.79	<.001
Happiness	12	<.001
Hatred	-4.17	<.001
Hope	7.85	<.001
Pride	-2.41	.016
Relief	0.682	>.05
Sadness	12.33	<.001
Shame	6.67	<.001
Tension/Stress	1.15	>.05

When breaking down the comparisons within the Event Specific comments about the Christchurch terrorist attack and the Shamima Begum with further independent t-tests one particular significant difference stands out. Levels of anger in response to the Shamima Begum case are significantly higher than in response to the Christchurch terrorist attack.

3. Thematic Case Study Analysis

Utilising a selection of the previously collected data a more in-depth qualitative analysis was performed to further explore some of the reoccurring and importance themes within the sample. This allowed for further exploration and investigation of the language analysed in studies 1 and 2 in its original context.

Method

Design

A small-scale case study was employed to address important themes associated with the evidence of online extremism associated within Islam. The case study was comprised of tweets or YouTube comments that represented contemporary events that was associated with extremism towards the Christchurch terrorist attack, Shamima Begum and general Islamic hate.

Participants

When completing case study techniques, the number of participants is dependent on data saturation, this is where new data fails to highlight any distinguishable new data (Sargeant, 2012). The total number of Tweets utilised for the case study was 600, at this point it was evident that data saturation had occurred. These comments were drawn from those collected for study 1 and 2, from both Twitter and YouTube.

Materials

The comments for the case study were selected at random from the data collected by the software from study's 1 and 2.

Procedure

Tweets that related to Christchurch mosque attack, Shamima Begum and general extremist tweets toward Islam were collected to form a small corpus of data, along with a random selection of comments collected from YouTube (from 2a and b). This data was then utilised to perform a content analysis using thematic analysis to demonstrate important themes that were present in the data.

Data Analysis

The thematic analysis used the Braun and Clarke's (2006; 2014) guidelines and stages. A thematic analysis was chosen as it has the ability to be used across a range of research questions (Nowell, Norris, White, & Moules, 2017), it can also produce trustworthy and insightful findings (Braun & Clarke, 2006).

Findings and Discussion

Whilst there were a mix of positive and negative comments the analysis revealed largely negative themes; False Flags and Take-Overs, But what about us, and, Patriotic Resistance. These themes were distinct but some had clear links and relationships between them.

Theme 1: False Flags and Take-Overs

A frequently expressed view by many of the commenters was that of being supressed and silenced by Muslims. Even when reacting to attacks against Muslims, commenters would often refer to this as either being a justified reaction to their perceived loss of rights and marginalisation in the face of Islam:

> *"I'm quite surprised this hasn't happened long before this. When you consider the outright atrocities done by Muslim immigrants in the US and Europe, and the authorities trying to minimize rape, murder, child molestation etc, and concentrate on prosecuting any speech*

negative against Islam rather than the actual criminal acts, it was bound to trigger some unstable person to act in what they see as retaliation."

Related to this there is some reference to Islam as a death cult that only has a purpose to cause harm, it is regarded as an illness like a plaque that is spreading uncontrollably.

"Maybe its got something to do with the obvious and demonstrable harm islam continues to cause. a Cancer survivor doesn't suddenly sing the praises of cancer because it has escaped its grasp.. and thats a perfect analogy for ex muslims"

Which is similar to the some of the keywords that were found to differ by anonymity in study 1, and the general fear response found increasingly for the Event Specific comments in study 2b.

Or perhaps interestingly the impact of attacks of Muslims would be undermined by the suggestion that such attacks were exaggerated, less common than attacks on those of other faiths, or that the attacks were 'false flags' and part of a wider Islamic or leftist conspiracy theory.

"Anyone notice the propoganda? […] Many of the so called 'hate crimes' have been false flag attacks to attempt to stir up sympathy. And the media loves to play up the ""white supremicist bullshit at every opportunity."

As noted earlier in the paper conspiracy was not often used and preferred terminology centred around the media, fake news, media agendas. Arguably this is part due to the

negative connotations associated with modern conspiracy theories as well as the increasingly populist use of fake news or 'main stream media' as an outgroup to aim accusations towards.

Theme 2: But What About Us?

In response to an attack like Christchurch rather than being sympathetic, many messages are focused of reminding others of what's happened to their religion and attacks that have affected them.

"A few dead muslims compared to millions of slaughtered innocents at the hands of islamic barbarians. #islamisevil #NewZealandTerroristAttack"

There are many clear attempts to reframe and contextualise attacks against Muslims as either being comparatively rare, exaggerated or being focussed on at the expense of individuals who feel they themselves are marginalised by or under threat from Islam.

"Let us not forget the thousands upon thousands of victims killed by the real 'terrorists', propagating the Islamic ideology. #AntiIslamic #IslamIsEvil #EndIslam #Muslims"

This attempt to reframe and lessen the impact of attacks against Muslims may represent an attempt, much like the use of references to conspiracies and media bias to gain control of the narrative around deeply emotional events. This could be used to convince or refute those without Islamophobic viewpoints in such online discussion. There may be less of a need to do this during general online discussions as many videos (such as those in our sample) are already negatively framed and may be found by those already using Islamophobic search terms/key words.

Sub-Theme: Patriotic Resistance

A commonly stated idea was that of holding onto a western identity (often New Zealand or British) in the face a threat, being patriotic, and fighting for that identity.

"This is what open borders give you. This is what extreme Muslim terrorists attacking people all over the world gives you. This is what happens when you try to steal the identity of a Nation with too much immigration. Pray for all the victims but never forget there is collateral damage on both sides."

This aggressive promotion of one's own culture and beliefs was often framed as being a response to increased threat and marginalisation from Islam, with violence in response being supported as justified or expected.

"I'm pretty sure the shooter did this in retaliation to all the violence done by muslim extremist in other countries such as the U.K and other countries where the locals were affected by some sort of violence done by muslim immigrants. I don't feel no sympathy for any loss from this shooting attack because what goes around may eventually come around. Don't expect things will be all peaceful if there are those of your own kind doing violence to others in other areas."

Arguably this subtheme underpins both theme 1 and 2 as it represents the manipulation or use of an identity to promote a lack of empathy and increased violent resistance through impending (and manufactured) threat.

Policy Recommendations

Soft verification of identity

Our initial findings from study 1 suggest that the increase in Islamophobic language is associated strongly, and more specifically to anonymous users with fewer than three

identifiable details in their user name/profile. Significantly less extremist language is present in those with three or more identifiable details in their user name/profile. Therefore, one suggestion for social media and online platforms could be to encourage or insist in some or a minimum amount of identifiable information to reduce extremist language use. This would not have to be the same as full verification that removes all anonymity but instead would act as a form of 'soft' verification that could reduce the disinhibiting effect of online anonymity.

Better reporting tools for trigger events?

With a clear indication from our research of an increase in conspiracy based language and negative emotional responses such as disgust, fear and anger to offline trigger events compared to the general non-specific content there may be a need to improve to the reporting system for inappropriate comments in the wake of such events.

More emphasis on counter-narratives and speech.

The significant difference in conspiracy based language used to undermine real world attacks and Islamophobia suggest a real need for the presence of counter narratives. The sentiment expressed in response to Event Specific real world events requires a challenging in a different manner, these Islamophobic comments would likely require the challenging of such conspiracy theories, and claims of media bias. However, for General Discussion comments where general Islamophobia is present the construct of a patriotic identity and 'victimhood' in the face of Islam would instead need challenging.

Using a traffic light system to warn and remove users

Whilst there are some issues to be addressed corpus linguistics and sentiment analysis could be used to create a crude traffic light or early warning system for platforms or users, identifying videos with more toxic or concerning comments. If the comments for a video or platform pass a certain threshold, users could be warned about the potential content, or reminders to fact check could be presented by the platform. In some case the comment

section could become moderated or removed if thresholds for negative emotional responses such as hatred or disgust are passed.

Conclusion

Increased anonymity is associated with increased extremist and Islamophobic language use of Twitter, which has clear implications for the use of such social media platforms. Such findings could have a potential impact on how user anonymity is dealt with on social media to mitigate extremism. An examination of YouTube comments has also revealed differences in conspiracy based language used between YouTube videos that are responding to offline events and those that are general online discussion. This potentially reveals how individuals construct and approach discussion about Event Specific content differently from more general General Discussioncontent, in an attempt to actively reduce the impact of events like the Christchurch mosque terrorist attacks and regain control of the narrative from an anti-Islamic stance. Sentiment analysis also revealed clear differences between the two types of YouTube video, again suggesting a difference in how individuals react emotionally to such content, in terms of fear, anger, disgust, hatred and pride in particular. In both cases, language and sentiment, we can see implications for how different types of comments and individuals could be approached by those running such platforms as well as other bodies.

References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL,* 7-13. Retrieved from https://www.researchgate.net/profile/Laurence_Anthony/publication/267631346_Proceedings_of_IWLeL_2004_An_Interactive_Workshop_on_Language_E-learning_2004/links/5458cd870cf26d5090acf212/Proceedings-of-IWLeL-2004-An-Interactive-Workshop-on-Language-E-learning-2004.pdf#page=7

Anthony, L. (2018). *AntConc (Version 3.5.6) [Computer Software]*. Retrieved from Tokyo, Japan: Waseda University, Laurence Anthony

Anthony, L. & Hardaker, C. (2017). *FireAnt (Version 1.1.4) [Computer Software]*. Retrieved from Tokyo, Japan: Waseda University, Laurence

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1 doi:10.5281/zenodo.58517

Bartlett, J. (2010). *From suspects to citizens: preventing violent extremism in a big society.* Retrieved from http://cve-kenya.org:8080/jspui/bitstream/123456789/164/1/Bartkett%20%26%20Birdwell%20-%20FROM%20SUSPECTS%20TO%20CITIZENS_%20PREVENTING%20VIOLENT%20EXTREMISM%20IN%20A%20BIG%20SOCIETY.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Brenner, S. W. (2004). Cybercrime Metrics: Old Wine, New Bottles?. *VA. JL & TECH., 9*, 13.

Clarke, V., & Braun, V. (2014). Thematic analysis. In *Encyclopedia of critical psychology* (pp. 1947-1952). Springer New York.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods 49*(3), pp. 803-821. doi:10.3758/s13428-016-0743-z.

Dixon, H. (2017) Russian bot behind false claim Muslim woman ignored victims of Westminster terror attack. Telegraphy [online]. Retrieved from https://www.telegraph.co.uk/news/2017/11/13/russian-bot-behind-false-claim-muslim-woman-ignored-victims/

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, *14*(2), 265-285.

Grabosky, P. N. (2001). Virtual criminality: Old wine in new bottles?. *Social & Legal Studies*, *10*(2), 243-249.

Gruzd, A., & Roy, J. (2014). Investigating political polarization on Twitter: A Canadian perspective. *Policy & Internet*, *6*(1), 28-45. **doi:** 10.1002/1944-2866.poi354

Gruzd, A., Wellman, B., and Takhteyev, Y. (2011). Imagining Twitter as an imagined community. American Behavioral Scientist, 55(10), 1294-1318. doi:10.1177/0002764211409378

Haber, E. M. (2015). On the Stability of Online Language Features: How Much Text do you Need to know a Person?. *arXiv preprint arXiv:1504.06391*. Retrieved from https://www.researchgate.net/profile/Eben_Haber/publication/275527256_On_the_Stability_of_Online_Language_Features_How_Much_Text_do_you_Need_to_know_a_Person/links/557a600808aeacff2003d375/On-the-Stability-of-Online-Language-Features-How-Much-Text-do-you-Need-to-know-a-Person.pdf

Hiltunen, T., McVeigh, J., & Säily, T. (2017). How to turn linguistic data into evidence? *Studies in Variation, Contacts and Change in English, 19*. Retrieved from http://www.helsinki.fi/varieng/series/volumes/19/introduction.html

Johansson, F., Kaati, L., & Sahlgren, M. (2016). Detecting linguistic markers of violent extremism in online environments. In *Combating Violent Extremism and Radicalization in the Digital Era* (pp. 374-390). IGI Global. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=qu0ODAAAQBAJ&oi=fnd&pg=PA374&dq=Johansson,+F.,+Kaati,+L.,+%26+Sahlgren,+M.+(2016).+Detecting+linguistic+markers+of+violent+extremism+in+online+environments.+In+Combating+Violent+Extremism+and+Radicalization+in+the+Digital+Era+(pp.+374-390).+IGI+Global.&ots=E15PiPt2hS&sig=ZqrbBauEwv7oycUBidcAw6ffcCo#v=onepage&q&f=false

Kemp, S. (2018) Digital in 2018: World's internet users pass the 4 billion mark. Retrieved from https://wearesocial.com/blog/2018/01/global-digital-report-2018

Kennedy, G. (2014). *An introductio*n to corpus linguistics. Oxford and New York: Routledge.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Schudson, M. (2018). The science of fake news. *Science, 359*(6380), 1094-1096.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*. https://doi.org/10.1177/1609406917733847

O'Hara, K., & Stevens, D. (2015). Echo chambers and online radicalism: Assessing the Internet's complicity in violent extremism. *Policy & Internet*, *7*(4), 401-422.

Peddinti, S. T., Ross, K. W., & Cappos, J. (2017). User Anonymity on Twitter. *IEEE Security & Privacy*, *15*(3), 84-87. doi:10.1109/MSP.2017.74

Peebles, E. (2014). Cyberbullying: Hiding behind the screen. *Paediatrics & child health*, *19*(10), 527-528. doi:10.1093/pch/19.10.527

Sargeant J. (2012). Qualitative Research Part II: Participants, Analysis, and Quality Assurance. *Journal of graduate medical education*, *4*(1), 1–3. doi:10.4300/JGME-D-11-00307.1

Scherer, K. R. (2005). What are emotions? And how can they be measured?. *Social science information*, *44*(4), 695-729.

Sunstein, C. R. (2001). *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton, NJ: Princeton University Press.

Sutch, H & Carter, P (Submitted). *Exposing Online Extremism: Anonymity, Membership-Length and Postage Frequency as Predictors of Extremism.* (Submitted dissertation). Birmingham City University.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PloS one*, *13*(9), e0203958

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Zhou, Y., Qin, J., Lai, G., & Chen, H. (2007, January). Collection of us extremist online forums: A web mining approach. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, 70-70, IEEE. Retrieved from https://pdfs.semanticscholar.org/d7d2/70aeccc12a41bdefbc9b3642ca52a066e6da.pdf

Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation*. University of Nebraska press. Retrieved from http://psycnet.apa.org/record/1971-08069-001

Appendix

The scoring system for anonymity

- Full name = 2 pts
- First or last name = 1 pt.
- Specified location = 2 pts
- General location = 1 pt.
- Potential identifiable profile picture = 1 pt.
- Personal website = 1 pt.

Any additional information such as a D.O.B, a snapchat profile, an Instagram profile is worth 1pt.

Appendix b

Wordlist: Extremist words referring to Islam

Islamisevil
Stopislam
Banmuslims
Islamiscancer
Islamonazism
Islamistheproblem