The Centre for the Study of Terrorism and Political Violence

Journal of Terrorism Research

Quantifying Salient Concepts Discussed in Social Media Content: An Analysis of Tweets Posted by ISIS Fangirls

by Shadi Ghajar-Khosravi, Peter Kwantes, Natalia Derbentseva and Laura Huey

Abstract

In this paper, we measure the extent to which we can accurately measure the salience of topics/concepts that might be of interest to an analyst tasked with analyzing the content posted on social media platforms. We also evaluate whether concepts like positive and negative sentiment can be meaningfully extracted from Social Media content. As a test case, we examined Twitter content generated by female users who are sympathetic to the Islamic State in Iraq and Syria (ISIS). Although the results were based on a small sample of users, we demonstrate that ISIS fangirls differ in the content of their tweets from other, non-radicalized, teenage girls, and that automated text analysis techniques can detect the differences. The basic technique proposed here is a promising step in devising techniques for quantifying the salient topics being discussed on social media platforms, and should be developed further to create more fine-grained exanimations of such content.

Keywords: *Social Media; Islamic State; Twitter; radicalization*

Introduction

Social Media has become an important source for information about people and important events around the world events. Its importance is driven largely by the enormous number of people generating and updating content in Social Media platforms on a constant, sometimes more than hourly, basis. Twitter represents one of several social media outlets that have become a widely used and popular technology for radicalized groups like ISIS proponents to spread their real time messages around the world. Lack of strict regulations, the ability to remain anonymous, easy access to a wide range of audience, and the fast flow of information within Twitter and similar Social Media platforms have made these tools a popular choice for broadcasting extremist beliefs and values to the more susceptible audience (Weimann, 2015). While many of radicalized community members have posted radicalized content on their social media pages before they committed a terrorist attack (for example, the shooting on Parliament Hill in 2014), there are limitations to the intelligence community ability to proactively and efficiently identify these individuals from their social media content. Part of the limitation comes from the volume and rate at which user data get posted in these environments and the low signal to noise ratio in social media content in terms of the amount of content suggesting pro-radical ideas versus anti-radicalization or neutral content.

The purpose of this paper is two-fold. First, we sought to explore some techniques and strategies for analyzing the content of twitter feeds. The proposed analytical strategy starts by defining a set of topics of relevance that might be relevant to an analysis, and then measures how often the concepts are used and the association of concepts to positive and negative sentiment. The second purpose of the paper is to apply the technique to see how the content of tweets posted by radicalized teenage girls who support the Islamic State in Iraq and Syria (ISIS) differs from a) content posted by an age-matched control group of non-radicalized users and b) content about ISIS from a random sample of twitter users. We refer to these radicalized teenaged girls as "ISIS Fangirls" (Huey & Witmer, 2016). Taken together, our analysis tries to answer the following questions:

1. Can we validly determine the salience of certain topics discussed in twitter content?
2. To what extent do the tweets posted by ISIS fangirls differ from control tweets in their expression of sentiment towards the content topics we have chosen?

Content Analysis

Broadly speaking, Content Analysis refers to a set of manual, automatic or semi-automatic techniques whereby the language used in a narrative is processed to objectively summarize the salient topics and extract the meaning being discussed. How the language is processed will vary depending on what the researcher wishes to include as the basic linguistic unit (e.g., words, phrases) for analysis. Once the linguistic units have been extracted from the text, they can be quantified to determine, for example, the salient topics being discussed, the frequency with which they are discussed, or what sentiments are associated to the topics. Various content analysis methods exist and the choice depends on the specific research objectives (Grimmer & Stewart, 2013). The analysis we conduct here is similar in nature to ones that use a lexicon-based technique (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Kouloumpis, Wilson, & Moore, 2011) rather than a corpus-based (Pak & Paroubek, 2010) or hybrid approach (Kumar & Sebastian, 2012). In it, concepts are formed by collecting the linguistic units mentioned above into a lexicon for a higher-level topic. For example, the concept of *friend* might include terms like, 'friend', 'bestie', and 'bff'. Analysis then proceeds by tabulating the presence of the topics in the text. In doing so, the researcher is able to characterize what can be a very large collection of text as a manageable collection of analysable concepts, and infers the salience of topics in a person's or group's generated content from the frequency with which they appear in text.

In what we report here, we apply our analysis to text generated by ISIS Fangirl on Twitter. Our analysis of the ISIS Fangirl content begins with an *a priori* set of topics to examine. We then find and count segments of text belonging to each of the topics. The topics are then ordered by their frequency of occurrence—a property that, we postulate, reflects the topics' relative importance to those generating the content. An aspect of the approach we take in this paper includes validating the salience of topics when it is based on the frequencies of the topics. The frequencies with which concepts are used in text *could* be determined randomly, making their order unrelated to their salience. We validate the order of topics by comparing it to that from text sampled from two control groups. Specifically, our analysis on twitter content generated by radicalized high school girls was compared to that of two collections of content that serves as controls: a sample of tweets written by non-radicalized high school girls, and a sample of tweets from multiple authors that reference the Islamic State. The first control collection allows us to compare how radicalized and non-radicalized teenaged girls differ with respect to how our concepts are discussed. The second control group allows us to compare how the concepts that radicalized teenagers discuss might differ from others who also generate content related to the Islamic State, but may or may not be supporters.

Once tabulated, the topics can also be assessed for expressions of positive and negative sentiment. The same technique we used to create concepts for the content discussed in the tweets is used to create concepts for positive and negative sentiment. That is, two collections of words were created: one comprising positive words and one containing negative words. The sentiment concepts allowed us to measure the relative extent to which the content concepts were expressed in a positive or negative manner.

ISIS fangirls

Following Huey and Witmer (2016) we refer to the radicalized teenaged girls as "ISIS Fangirls". In the terrorism context, 'fan girl' has been used to describe a girl or woman – usually a teenager – who joins a jihadist network in order to enjoy the notoriety that comes from participating in a group that is seen by some as 'rebellious' and therefore 'cool' (Huey & Witmer, 2016). The term, which is commonly used by both IS and AQ affiliated networks to refer to enthusiastic, but often naive joiners, comes from popular culture. It was spawned by the sometimes hysterical reactions generated among young fans to such pop culture phenomenon as the Twilight movies, Justin Bieber and/or One Direction fans (Herrmann, 2008). What marks IS fan girls as distinct from their One Direction loving counterparts, is that the object of their interest is the Islamic State. Although female adherents to IS ideology are supposed to refrain from open adoration of males, and certainly from contact with males, beyond the benefits of association with a group espousing a form of 'jihadi cool' (Huey, 2015; Picart 2015), fan girls are likely also attracted by a plethora of images of highly attractive young male jihadists, whose pictures, often in highly romanticized form, frequently dominate their twitter streams (Erelle, 2015).

To the extent that these young women are often recent joiners, and typically reveal an ignorance of IS ideology – often behaving in ways that contradict this ideology by, as an example, posting pictures of their faces – we could argue that they have not been completely radicalized yet. However, their association with regular IS members and, in many instances with notable recruiters and IS propagandists, does place them at significantly higher risk of radicalization than other young people in the general population. Increasing this risk is, as some commentators have noted, the presence of a "jihadi girl power subculture" that recruiters and propagandists use to increase the attractiveness of joining IS (Pandith & Havelicek, 2015). For these reasons, the content of their posts is of scientific value to the extent that it can tell us something about how to differentiate individuals who may be at different stages in the process of radicalization.

Method

Concepts

Content-Related Concepts: The first step in addressing the research questions was to prepare a list of topics that would be of interest to analysts studying radicalized individuals. Normally, such an analysis would be conducted by analysts who are subject-matter experts in the domain and would use it to decide upon concepts of interest. For the example we report here, we identified ten concepts relevant to radicalization: Islamic State (IS), Punishment, Unbeliever, Jihad, God, Islam, Marriage, Violence, Middle-East countries that are enemies of ISIS, and Western countries ('the West'). A concept was constructed by compiling a potentially exhaustive list of terms that described the concept.

In addition to the concepts related to radicalization, we included four control concepts that represent topics that would be of more general interest to teenage girls. These were: High School, Friends, Dating, and Sports. The complete list of concepts is shown in Table 1.

Sentiment Concepts: To compare and contrast the sentiment that Fangirls and control users have toward content-related concepts, we created two concepts related to sentiment. Sentiment concepts were created in three steps. First we started with the positive and negative sentiment word lists created by Hu and Liu (2004) for their lexically-based sentiment analysis approach. Second, we sorted the words in each list by their frequency of usage in everyday usage using the CELEX word frequency database from the Max Plank Institute for Psycholinguistics and selected only the 300 most frequent words as candidates for each sentiment concept. As a final step, we examined each word of the two lists and removed any word that did not have a clear association to a sentiment. For example, the terms 'wicked' and 'hot' can be associated with negative or positive sentiment. The resultant lists contained 210 keywords comprising the positive sentiment concept and 159 keywords representing the negative sentiment concept.

Datasets

Three datasets of tweets were collected from Twitter using Twitonomy.

ISIS Fangirls: The first dataset contained tweets from 14 different teenage girls. The dataset, referred to as the "Fangirls", contained approximately 5000 tweets and included expressions of appreciation or dedication to ISIS in their tweet content. The tweets were collected from January 2015 to April 2015. One fangirl (#6) was removed from the analyses due to the limited number of tweets available for her in the dataset.

Control Girls: A second dataset, referred to as "Control Girls", contained tweets collected from 14 randomly selected teenage girls who, at least from the content of their tweets, did not appear to have any affiliation with ISIS or radical Islam.

ISIS-control: A third dataset was collected to explore the differences in how ISIS Fangirls and other users who mention ISIS differ with respect to the concepts we had selected. These tweets were extracted using the Twitter API and by searching for tweets containing the keywords assigned to the IS concept. The tweets collected could be pro-ISIS, anti-ISIS, or neutral. Hence, this tweet collection was a second control collection and referred to as the "ISIS-control" dataset. After the dataset was cleaned (detailed in the next section), 3332 tweets remained in the dataset. All the tweets were posted on July 10, 2015 the date the online search was conducted.

Importantly, all of the tweets used in the analysis below were completely anonymized before being submitted for analysis. We therefore have no information about the nationalities or identities of the persons. Table 2 details the amount of content we were able to sample for each twitter user in the Fangirl and Control datasets.

Pre-processing of Data

Clean-up and substitution: A 3-step clean-up process was conducted on the three datasets: 1) any non-English words, characters or emoticons were removed from the tweets. 2) Empty or repeated tweets (including retweets from Fangirls and Control Girls) were removed from the tables and 3) example, different spellings of the transliterated Arabic terms *kuffar, kufar*, *kuffaar*, *kafirs*, *kafiroon*, etc. were all replaced with *kuffar*.

Tokenization and Preliminary Variables: A Java program was coded to parse the individual tweets and search for keywords related to each of the concepts. The list of keywords was fed into the program to return the following variables:

1. Number of tweets containing each keyword for each girl.
2. Number of tweets containing each concept for each girl. Put simply, for every tweet, we determined whether it contained words from any of the 14 concepts. If a tweet contained one or more keywords

from a concept, a counter for the number of tweets containing the concept was incremented by 1.

3. Number of tweets expressing positive or negative sentiment for each concept together for each girl.

Sentiment Expression in Tweets: The expression of sentiment was measured separately for each content concept by calculating the conditional probabilities that positive and negative concepts appeared with tweets containing terms from the content concept (see equation below). Done this way, we can measure, for each content concept, the extent to which a concept, or topic, accompanies expressions of positive and negative sentiment.

$$P(\mathrm{Concept}_{\mathrm{Sentiment}} | \mathrm{Concept}_C) = \frac{n(\mathrm{Concept}_{\mathrm{Sentiment}} \,\&\, \mathrm{Concept}_C)}{n(\mathrm{Concept}_C)}$$

where n(Concept$_C$) refers to the number of tweets in which the content-related Concept$_C$ (i.e., one or more of its associated terms) has occurred and n(Concept$_{\mathrm{Sentiment}}$ & Concept$_C$) refers to the number of tweets in which both Concept$_{\mathrm{Sentiment}}$ (i.e., a positive or negative sentiment conept) and Concept$_C$ have occurred together. Hence, P(Concept$_{\mathrm{Sentiment}}$ |Concept$_C$) refers to the probability that the terms from a sentiment concept appear with tweets containing terms from a content-related concept.

Results

Dominant Content-Related Concepts

Table 3 shows, in order from highest to lowest frequency, the content-related concepts mentioned in the datasets. The content-related concepts are comprised of varying numbers of keywords. For example, the concept, Dating has three terms, while Punishment has 13. These differences raise a possibility that a concept's frequency in the text is driven by the number of words that constitute the concept. In other words, the frequency of a concept may increase with the number of terms comprising the concept because the more terms in a concept, the more opportunity there is to have a match between the tweet and the concept. As a check, we correlated the number of keywords in the concepts and their frequency in the tweets for each dataset. The correlations were not significant (Table 4). That is, the frequency of concepts' presence in tweets was independent of how many terms comprised the concept.

The Order of Concepts' Frequencies

Table 3 orders the 14 content-related concepts from highest frequency to lowest. We can interpret the order as reflecting the salience of the concepts to the authors providing content. One way to validate the order of the concepts is to compare the order to that of one or more control groups. In our case, we can compare the order of concepts in Fangirls tweets with Control Girls and ISIS-Control tweets. Because Fangirls and ISIS-Control tweets may be written by twitter users that are similar is some respects, we hypothesised that the prioritization of the concepts they discuss might be similar. To test the notion, we calculated Kendall's τ, a correlation measure that is applied to ordinal data, on the order of the topics in Table 4. Specifically, we measured the extent to which the order of topics in Fangirls tweets matched that of Control Girls and ISIS-Control tweets. We found, as hypothesized, that the order of Fangirls and Control Girls concepts was in far less agreement ($\tau = .18$) than it was for ISIS-control tweets ($\tau = .67$).

The key finding from this section is that, we were able to measure and prioritize the concepts discussed in tweets. We validated the order in two ways: first, by establishing that the position of a concept in the list was independent of the number of terms comprising the concept, and second by comparing the order of concepts in Fangirls tweets to that of two control groups of tweets. We demonstrated that the order reflected in the Fangirls' concepts was far more similar to that of a random collection of tweets about ISIS than a collection of tweets from non-radicalized teenaged girls.

Sentiment Towards Concepts

The next analysis concerned the expression of sentiment associated with our 14 concepts. Specifically, for each concept, we calculated the likelihood that it appears with at least one term from the positive and negative concepts described. We did the analysis for the tweets in each of the three tweet samples. So, for each concept, we have a measure of the extent to which the concept is associated with positive sentiment and negative sentiment. Table 5 shows the results of the analysis.

Using the table, we can inspect the extent to which Fangirls differ or are aligned with the pattern of sentiment expression in the control groups. In other words, to what extent are the positive and negative expressions of sentiment about the concepts similar to, or different from, those of controls?

Four examples of where the groups differ stand out and are highlighted with x's in the table. First, whereas the expression of sentiment around unbelievers is neutral for Control Girls, it is decidedly more negative than positive for both ISIS-Control and Fangirls tweets. Second, whereas Control Girls express more positive than negative sentiment associated with The West, Fangirls (not surprisingly) associate The West with more negative than positive sentiment. Sports are associated with more negative sentiment than positive sentiment for Fangirls; a pattern that reverses for Control Girls. Finally, whereas jihad is associated with positive sentiment for Fangirls, it is associated with negative sentiment for Control Girls.

As a second analysis on the sentiment data, we can compare the pattern with which positive and negative sentiment are assigned to concepts, and measure how similar the pattern is across different groups. For example, in our dataset here, we can look at the pattern of Positive and Negative associations across the concepts and measure the extent to which it is shared with our control groups. We predict, for example, that the pattern of positive and negative expressions of sentiment for Fangirls will be more similar to that of ISIS-Control tweets than Control Girls' tweets because the ISIS-Control tweets are more likely to express pro-ISIS sentiment like the Fangirls. To test the hypothesis, we calculated the difference between positive and negative associations for each concept separately for each group. We then used Pearson's correlation coefficient on the differences between Fangirls and Control Girls and between Fangirls and ISIS-Controls. As predicted, ISIS-Control tweets contained a pattern of sentiment expression that was more similar to the Fangirls' pattern ($r = .53$) than it was for Control Girls ($r = .34$). Although suggestive and consistent with our hypothesis, the difference between the two correlations was not statistically significant ($p = .29$).

Discussion

In this paper, we explored a method for conducting a quantitative analysis to measure the salience of concepts discussed in tweets generated by teenaged girls who support ISIS, and used tweets from two different populations to serve as controls. The results of the analysis were promising. We were able, with some certainty, to accurately characterize the salience of content-related concepts. We were also able to measure the extent to which positive and negative sentiments were expressed in tweets by measuring the extent to which terms in a sentiment lexicon were present in tweets mentioning our 14 topics of interest. For the most part, the pattern of tweets seems reasonable. For example, Fangirls associated positive sentiment with jihad, whereas for Control Girls it was negative. Conversely, Control Girls had a more positive than negative association to the West, where the pattern reversed for Fangirls. Although the results were based on a small sample of users, they demonstrated that users with different ideological orientations differed in the content of their tweets, and that automated text analysis techniques can detect the differences. We would caution however, that such automated techniques should always be interpreted with care and be supplemented with a domain expert's validation.

Although we do not claim that the methods we describe in this paper provide a complete treatment of the analysis of the content in tweets, it is a promising step in devising techniques for quantifying the salient topics being discussed in social media platforms. We see the techniques as being complementary to social media analytic techniques performed on massive collections of content. Whereas social media analytics focuses on understanding the concepts expressed in massive aggregated collections of say, Twitter content, understanding the more nuanced content of individual users can be accomplished techniques similar to the one we used here. We therefore recommend further work to develop quantitative methods for analysing social media content at the lower, user-based level.

There are two aspects of the work here that, in our opinion, represent opportunities for improvement to the analysis capability. First, in what we report above, the analyst must decide *a priori* what concepts will be examined in the analysis. While such a strategy is useful, it is somewhat limited in that there may be other salient concepts in the text that will remain undiscovered because they are not present in the set of concepts under examination. There are computational models that are capable of automatically extracting topics from text (e.g., Blei, 2012). These so-called, Topic Models are typically applied to documents that are longer than tweets, so it is unclear how well they can be applied in the social media context where texts are very short. The Topic Model's applicability to extracting topics automatically from tweets will be explored in upcoming work.

The second opportunity for improvement could be developed using the lexicon-based approach we used to tabulate the frequency of concepts. Recall that each concept was comprised of a set of words, and that if a tweet contained one or more words of the concept, the frequency of the concept was incremented. Clearly then, how well the system is able to characterise the presence of a concept will depend on how comprehensively the concept is represented by the words in the lexicon. The technique worked well in the results we reported here, but could be improved by algorithms that are able to match terms on their meaning rather than a strict match on spelling. For example, our concept for marriage did not include the terms, 'bride' and 'groom', when clearly they could (or perhaps should) have been. Not having them in the concept means that tweets containing either 'bride' or 'groom' would not be identified as being relevant to the concept. The solution would be an algorithm that knows when words are related to a concept even if the word is not present in the list of words that comprises it. There are semantic models that can, in an unsupervised fashion, generate 'meaning' representations for words. Perhaps the most popular example is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA performs statistics on a matrix describing the frequency with which terms appear in each document of a large collection of text. In the end, each word is represented as a vector which behaves much like a "meaning" in that the vectors for two semantically related words like dog and puppy will be similar as measured by their cosine (akin to a correlation coefficient wherein a value of 1 means they are identical, and a 0 means that the two are completely dissimilar). The power of LSA as a semantic model lies in its ability to deduce that terms are related even if they never occur together in the same document.

Unfortunately, models like LSA do not represent the form of the semantic relationship between words. Specifically, they cannot differentiate among the various forms of semantic association. So, a term like married is as similar to its synonym, wed as it is to its antonym, divorced. As they currently exist then, unsupervised models of semantics do not provide the required precision to support the technique and would require further development.

The most promising way to advance the technique is to exploit knowledge stored in ontologies to identify terms that are relevant to a concept, even if they are not contained in the lexicon describing it. Ontologies formally represent the relationships between words/concepts in a domain as a network of connected nodes where the links between them describe the nature of the relationship. So, for example, an ontological representation of English words would identify divorce as an antonym of marriage, and wed as a synonym. Perhaps the most extensive lexical ontology for identifying synonyms is Princeton University's WordNet (Fellbaum, 2005) which if it could be incorporated into the analysis we conducted here, would increase its precision (https://wordnet.princeton.edu/).

Conclusion

Social Media represents an increasingly important source of information about people and events. This article explored how well a lexicon-based technique characterises the salience of concepts and the sentiment associated with them. We see our techniques as being complementary to, so called, 'big data' techniques for analysing Social Media content. Specifically, while big data analytic tools are adept at extracting themes and networks from very large repositories of content, they are not designed to explore content at the individual user-level. The analysis we conducted here is designed to be done after analysis on a large repository has been done, and key content generators of interest have been identified for further analysis. The results were promising, and justify further exploration in order to increase the precision of the technique.

About the authors

Shadi Ghajar-Khosravi *is a scientist at DRDC Toronto Research Centre. She has a PhD in Mechanical and Industrial Engineering (human factors) from the University of Toronto, and a Masters of Information Systems from the Faculty of Information at the University of Toronto. Her research interests include social network analysis, information systems design and analysis, and the user-centered design and evaluation of interfaces.*

Peter Kwantes *is a scientist at DRDC Toronto Research Centre. He has a PhD in Cognitive Psychology from Queen's University at Kingston. He has worked as a scientist for DRDC since 2002 after completing a two-year post-doctoral fellowship in the ARC Key Centre for Human Factors and Applied Cognitive Psychology at the University of Queensland in Australia.*

Natalia Derbenstseva *is a scientist at DRDC Toronto Research Centre. She has a PhD in Management Sciences from the University of Waterloo. She conducts research on information representation, development of collaborative understanding, individual sense making and behavioural aspects of cyber security.*

Laura Huey *is the author of several articles on issues related to policing, cyber-security and terrorism. She is currently conducting research (with Johnny Nhan, TCU) on the role of gender in online radicalizing milieus and exploring women's participation in the creation and dissemination of pro-jihadist propaganda. Other current* *research is in the areas of cyber-security (as a member of the SERENE-RISC network) and alternate forms of police reporting.*

References

Blei, D.M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55* (4), 77-84.

Erelle, A. (2015). *In the Skin of a Jihadist*. New York: Harper Collins.

Fellbaum, C. (2005). WordNet and wordnets. In: K. Brown, et al. (Eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Herrmann, M. (2008). Not one of Those Screaming Girls: Representing Female Fans in 1990s' Germany. *Women's Studies in Communication*, 31(1): 79-103.

Hu, M., & Liu, N. (2004). Mining and Summarizing Customer Reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 22-25, 2004, Seattle, Washington, USA.

Huey, L., & Witmer, E. (2016). #IS_Fangirl: Exploring a New Role for Women in Terrorism. *Journal of Terrorism Research, 7*(1), 1-10.

Huey, L. (2015). This is Not Your Mother's Terrorism: Social Media, Online Radicalization and the Practice of Political Jamming. *Journal of Terrorism Research*, *6*(2).

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, *11*, 538-541.

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, *9*(3), 372-378.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM*, *11*(122-129), 1-2.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC, 10*, 1320-1326.

Pandith, F. & Havelicek, S. (2015). The Female Face of Terror. *The Telegraph*. Available online at: http://www.telegraph.co.uk/news/uknews/terrorism-in-the-uk/11374026/The-female-face-of-terror.html

Picart, C. (2015). "Jihad Cool/Jihad Chic": The Roles of the Internet and Imagined Relations in the Self-Radicalization of Colleen LaRose (Jihad Jane). *Societies, 5* (2):354-383.

Weimann, G. (2015). *Terrorism in Cyberspace: The Next Generation*. New York: Columbia University Press.

Tables

Table 1: The 14 content-related concepts along with their keywords.

Concept	Keywords	Number of keywords
IS*	caliphate, dawla, ummah, sharia, Islamicstate, IS, ISIL, ISIS	8
Punishment	punish, punished, punishing, prisoner, execute, execution, revenge, behead, beheading, torture, tortured, kill, killed	13
Unbeliever	Kafir, kuffar, pagan, pagans, atheist, apostate, apostasy, unbeliever	8
Jihad	Jihad, jihadi, mujahid, mujahideen, mujahidin, mujahadein, mujahadeen, mujahidin, mujahedeen, mujahedin, martyr, martyrdom, salvation	13
God	God, Allah, Alla, Jehovah	4
Islam	Muhammad, Koran, Qu'ran, Islam, Muslim, Muslims, Islam, Mecca	8
Marriage	marriage, marry, married, husband, wife, wives, wedding, nikah, zawj	9
Violence	attack, attacking, attacked, exploded, explosion, bomb, bombing, invasion, invaded, destroyed, destroy, obliterated, obliterate, annihilate, annihilated, battlefield, battle, war	18
ME local*	Jordan, Saudi, Iran, peshmerga, PKK, Bahrain, Qatar, UAE, Turkey	9
The West	USA, America, France, UK, Britain, Australia, Canada, Japan	8
High School	essay, homework, teacher, principal, classroom, textbook, highschool, prom, classes	9
Friends	buddy, bestie, clique, friends, friend, bff	6
Dating	boyfriend, girlfriend, dating	3
Sports	soccer, baseball, basketball, hockey, sports, athlete, athletics, volleyball, wrestling, football	10

*IS = Islamic State; ME Local = Middle East local enemies

Table 2: Number of tweets collected from each user in Control and Fangirl datasets.

Fangirls		Control Girls	
User ID	Number of Tweet Posts	User ID	Number of Tweet Posts
1	1708	1	3156
2	1324	2	3170
3	199	3	3121
4	212	4	3041
5	254	5	1561
6	3	6	3158
7	795	7	3142
8	205	8	2114
9	146	9	918
10	54	10	432
11	250	11	3180
12	264	12	3064
13	236	13	2689
14	121	14	1752

Table 3. Frequency of content-related concepts in the three tweet datasets. Frequency is expressed as a percentage of tweets containing the concept.

Fangirls		Control Girls		ISIS-control	
Concepts	Frequency (%)	Concepts	Frequency (%)	Concepts	Frequency (%)
God	9.81	God	1.46	IS*	54.47
Islam	9.10	Punish	1.34	Islam	15.49
IS*	8.72	Friends	0.99	Violence	7.44
Violence	3.33	Violence	0.93	West	5.07
Punish	3.19	Islam	0.84	Punish	3.96
West	2.53	High School	0.75	ME local*	2.73
Unbeliever	2.50	West	0.75	Jihad	2.40
Jihad	1.61	Marriage	0.48	God	0.60
ME local*	1.30	ME local*	0.35	Unbeliever	0.30
Marriage	0.81	Dating	0.35	Sport	0.21
Friends	0.54	Sport	0.34	Marriage	0.15
High School	0.16	Unbeliever	0.03	Friends	0.12
Sport	0.10	IS*	0.01	High School	0.06
Dating	0.07	Jihad	0.01	Dating	0.03

* IS = Islamic State, ME Local = Middle East Local Enemies

Table 4. Pearson correlation of number of words in concepts with frequency of concepts in Fangirl and control tweet collections (N=14).

	Fangirl	Control	ISIS-control
r	*-.143*	*.001*	*.019*
Sig.	*ns.*	*ns.*	*ns.*

Table 5. The Proportion of Tweets Expressing Positive and Negative Sentiment for Tweets Mentioning Each of The 14 Concepts (P(Concept $_{Sentiment}$ |Concept$_{c}$).

	Fangirls		Control Girls			ISIS-Control	
Concepts	Negative	Positive	Negative	Positive		Negative	Positive
dating	0.00	0.25	0.17	0.28		0.00	0.00
friends	0.16	0.45	0.14	0.41		0.00	0.00
God	0.16	0.25	0.10	0.29		0.05	0.25
High School	0.00	0.33	0.09	0.18		0.00	0.00
IS*	0.16	0.16	0.00	0.00		0.15	0.12
Islam	0.18	0.18	0.14	0.18		0.09	0.09
jihad	0.09	0.25	0.33	0.00	x	0.04	0.08
marriage	0.15	0.26	0.13	0.17		0.20	0.40
ME Local*	0.16	0.11	0.14	0.17		0.21	0.05
punish	0.45	0.12	0.40	0.13		0.45	0.10
sport	0.33	0.17	0.04	0.28	x	0.00	0.00
unbeliever	0.23	0.17	0.17	0.17	x	0.40	0.00
violence	0.21	0.14	0.29	0.14		0.14	0.06
west	0.25	0.12	0.15	0.20	x	0.14	0.14

* IS = Islamic State, ME Local = Middle East Local Enemies