Sentiment-based Classification of Radical Text on the Web

Ryan Scrivens and Richard Frank
International CyberCrime Research Centre
School of Criminology, Simon Fraser University
Burnaby, Canada
{rscriven, rfrank}@sfu.ca

***Abstract* — The total number of webpages has grown substantially since the birth of the Internet. So too have the number of webpages dedicated to radical yet subtle content. As these new circumstances have necessitated a guided data collection method, one that can sidestep the laborious manual methods that have been classically utilized, simple keyword analysis has not been sufficient to identify radical sites on Web 1.0 – pro-extremist, anti-extremist, and news sites, for example, may use the same keywords to discuss the same event but have a very different motivation. In an effort to explore this problem, we completed an exercise involving the use of a web-crawler to collect 20,000 webpages from five sentiment-based classes to assess their differences: (1) radical Right sites; (2) radical Islamic sites; (3) anti-extremist sites; (4) news source sites discussing extremism; and (5) sites that did not discuss extremism. Parts-of-Speech (POS) tagging was used to identify 198 of the most frequent keywords within the data, and the sentiment value for each of these keywords was calculated for each webpage using sentiment analysis. With these values, a decision tree was applied to three classification models. Results suggest that radical Islamic text can be classified at a much higher rate of success than radical Right text.**

Keywords - Sentiment Analysis; Decision Trees; Extremism

I. INTRODUCTION

Since the mid-1990s, we have seen a rapid growth in the number of radical websites acting as a hub of the movement they represent. In many respects they serve as a convergence setting for newcomers to familiarize themselves with radical teachings [2, 13], connecting users to other similar websites, web-forums, and pages on the Dark Web, to name but a few [4]. Counter-extremism organizations continue to look for ways to identify these sites [11], yet the Internet is constantly growing and information is being generated at a rapid pace [9]. This has led to a growing flood of data, resulting in manual data analysis becoming less efficient or entirely infeasible [5]. In response, researchers have developed data collection tools that allow decisions to be made about the thousands of webpages that are extracted and, by extension, classify radical text on the Web.

Researchers have explored this critical point of departure via automated computational and semi-automated tools, both to improve the methods of collection and identification of extremist content. Chen's *Dark Web Project*, for example, produced a number of computational and data centric studies on the content and structure of surface level and Dark Web sites containing extremist material [9]. In light of this comprehensive work, the entire process was automated and they did not assess the subtle language – in the English text – found on websites featuring extremist content. Arguably, a fully automated system should be avoided when human intelligence is required [5] (i.e., identifying subtle forms of radical text in English).

In the work of Mei and Frank, the authors developed a model that combined sentiment analysis and a web-crawler with a decision tree to differentiate pro-extremist webpages from anti-extremist pages, news pages, and pages that did not relate to extremism. Here, the overall goal of this semi-automated approach was to create a tool that made predictions about the content found on the sites it downloaded [10]. While this technique managed to classify extremist text at a high rate of accuracy (i.e., 92%), both extremist-based web-forums and websites of extremist organizations were analyzed within one model. Arguably, the sentiment found on both types of platforms are different in terms of tone and subject matter. Indeed, this problem requires further exploration.

II. METHODS

The purpose of this study was to add to the literature by using a sentiment analysis tool and a decision tree to differentiate pro-extremist webpages from anti-extremist pages, news pages, and pages that did not relate to extremism. This was done by building three text-based classification models (Chapter II.A). First, we used the Terrorism and Extremism Network Extractor (TENE) to collect data (Chapter 0) and OpenNLP's Parts-of-Speech (POS) analysis to develop a list of keywords (Chapter II.C). The sentiment expressed within the webpages was then calculated based on the POS keywords, detailing the relative sentiment scores for each page for each of the keywords (Chapter II.D). Finally, a decision-tree was built based on the sentiment scores (Chapter II.E) which produced the classification results for three classification models (Chapter II.F) (see Figure 1).

A. Webpage Data

Researchers have suggested that the sentiment found on both radical Right [1, 3] and radical Islamic [13, 16] websites are elusive. As a means of exploring this claim on the pro-extremist content and evaluating whether the decision tree could classify content based on those subtle ideological motivations, five distinct and non-overlapping classes of pages were attained.

1) The *radical Right* class was selected through two methods: (1) a Google search of the keywords 'extremist websites', 'white supremacy websites', and other similar words, and (2) using an index of extremist sites generated by researchers and sources (for

978-1-5090-2857-3/16 $31.00 © 2016 IEEE
DOI 10.1109/EISIC.2016.41

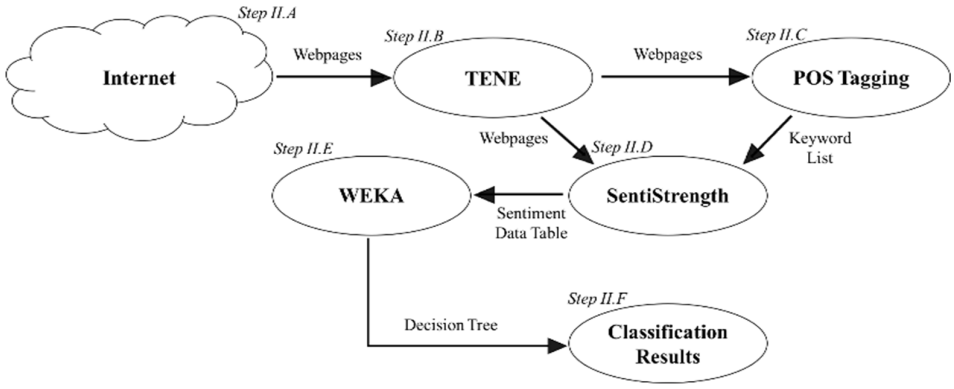

Figure 1 – Data Collection and Tree Generation Process

examples, see [3, 10]). This class contained 1,300 webpages, spread over 20 websites.

2) The *radical Islamic* class was selected through the keywords 'extremist websites', 'jihad organizations', and previously compiled indexes [10]. This class contained 3,700 webpages from 14 websites.
3) An *anti-extremist* class included intelligence agencies and organizations dedicated to countering extremism and were identified via an online search of such keywords as 'counter-extremism' and 'anti-extremism' groups, for example. The sample contained 5,000 webpages from 20 websites.
4) A *news* class included websites if they were sources that reported on extremist events. Here it was assumed that they presented a story in a more "impartial" manner than pro-extremist or anti-extremist sites. This sample contained 5,000 webpages from 17 websites.
5) A final class, *other*, served as the control group in the sample, where one would not expect to uncover material relating to extremism, and included topics relating to art and culture, education, entertainment, food, travel and lifestyle, social media, online blogs, automobiles, sports and business. For this class, 5,000 webpages from 31 websites were sampled.

Three models were then constructed for the purposes of comparing and contrasting their classification results:

- *Two-class model:* (1) *radical Right* and *Islamic* versus (2) *anti-extremist*, *news*, and *other* webpages;
- *Four-class model:* (1) *radical Right* and *radical Islamic* versus (2) *anti-extremist* versus (3) *news* versus (4) *other* webpages;
- *Five-class model:* (1) *radical Right* versus (2) *radical Islamic* versus (3) *anti-extremist* versus (4) *news* versus (5) *other* webpages.

B. Terrorism and Extremism Network Extractor (TENE)

The Terrorism and Extremism Network Extractor (TENE), a custom-written computer program that was designed to collect vast amounts of data online, automatically browsed and captured all of the content from the 102 websites used for classification between July 13, 2015 and July 26, 2015 (for a more detailed description of the crawler, see [10]).

C. Parts-of-Speech (POS) Tagging

The initial step in analyzing the content on webpages (automatically) was to determine the topics of discussion. To do this we wanted to take a data-driven approach and isolate the particular nouns with the highest rate of occurrence within the data collected. We assumed that the most frequently discussed topics would most likely be the ones in which extremist content was likely to be detected, based on the work of [12]. Further, nouns were chosen because they are the words most likely to be surrounded by relevant sentiment terms [8]; names and places are often described or denoted by the adjectives linked to them, and adjectives are often the words that have sentiment values attached to them [15]. By specifying adjectives as keywords, the sentiment of the word itself would be lost.

Parts-of-Speech (POS) tagging was used to perform this analysis. POS is a data analysis method that collects and divides supplied text into word groupings, such as nouns and verbs, based on their usage and position within the sentence (for more information, see [12]).

From each class, we selected 80 of the most frequent nouns and aggregated them into one list, which, after removal of duplicates and mis-categorized words (symbols and non-words, for example), we ended with 198 keywords. This final list formed the keyword list for the sentiment analysis. Domain experts may replace or extend the list of keywords with relevant domain-specific words. However, the exploratory and data-driven nature of the study suggested that we rely on the data to identify the relevant keywords.

D. Sentiment Analysis

After the keyword list was developed, it was necessary to identify and evaluate the context surrounding the keywords. To allow a proper analysis of the webpages, sentiment analysis was used to highlight relevant text.

Sentiment analysis is a data collection and analysis method that allows for the application of subjective labels and classifications [6]. It can evaluate the opinions of individuals by organizing data into distinct classes and sections, and assigning an individual's sentiment with a negative or positive polarity value [6]. It also provides a more targeted view of a dataset by

allowing for the demarcation between cases that are sought after and those without any notable relevance. Since the purpose of the current study was not to push the boundaries of sentiment analysis algorithm, SentiStrength [14] was utilized.

SentiStrength is Java-based software that uses a specific algorithm to run through large volumes of text and create sentiment scores for the supplied documents. While there are several configurations of the software, we utilized a keyword-focused method, as a central feature of SentiStrength is its ability to evaluate sentiment around any given keyword [14]. This process involves the utilization of a dictionary of catalogued terms and Harvard's general inquirer database to determine sentiment values [14]. It locates words that correspond with its dictionary and database, and then it employs a stemming method to evaluate a text by assigning polarity values of either positive or negative to the words. Values are augmented by characters that can influence the values assigned to the text, such as booster words, negative words, repeated letters, repeated negative terms, antagonistic words, punctuation, and other distinctive characters suited for studying an online context [14] (for more information on SentiStrength, see [14]).

E. Decision Tree

A tool in machine learning and data mining that is widely accepted amongst academics, Waikato Environment for Knowledge Analysis (WEKA), was run on the sentiment data for each of the three classification models. Its standard J48 tree classification method was used because it contains an algorithm for text classification that allows for a rule-building process [7]. WEKA was applied using 10-fold classification on the outputted sentiment tables, searching for differences in sentiment values between the classes of sites. As a result, this algorithm produced a decision tree with recursive leaves and branches, with each leaf representing a specific set of sentiment thresholds.

It is with these thresholds that the decision tree established whether a certain webpage was organized into the *radical right*, *radical Islamic*, *pro-extremist*, *anti-extremist*, *news*, or *other* class. For example, if the keyword 'marriage' was included in the text with a sentiment value of less than or equal to 2 and the term 'price' was also in the text with a sentiment value of less than or equal to a value of -5, then the J48 pruned tree indicated that 2 of the 20,000 pages fell within the *news* classification.

F. Classification Results

After the J48 algorithm was run on the sentiment data for each of the three models, WEKA created a measure that specified the number of accurate pages that fit into each of the correct classes. In turn, a confusion matrix displayed the number of pages that were correctly and incorrectly identified (i.e., false positives and false negatives) in each model, measured with precision and recall. Precision is a measure which represents the exactness of the rules. It can be thought of as the number of true positive entities identified by the classifier out of the set of all entities identified as positive. The larger the variability in the results, the smaller the precision will be. Recall is a measure of completeness and is the number of positive entities *identified* out of *all* true positive entities. A recall score of 1, perfect, means that all true positives are identified without any false negatives.

III. Results

The results were interpreted in two ways. First, the overall classification results were evaluated for each classification model, followed by the quality of the entire rule set for each. Here the analysis was largely focused on the pro-extremist class.

A. Two-Class Model: Results and Quality of Rule-Set

The two-class tree analysis indicated that 93.25% of webpages were successfully classified across classes. The combined *radical Right* and *radical Islamic* class was successfully classified at a considerably lower rate than the combined *anti-extremist*, *news*, and *other* class (i.e., 76.76% of pages were accurately identified). Furthermore, the recall of the entire rule-set indicated that actual two radical classes were classified at a rate of 77% in comparison to the non-radical classes' recall measure of 99% (see precision and recall in Table 1).

B. Four-Class Model: Results and Quality of Rule-Set

The four-class tree analysis revealed that 76.83% of webpages were accurately classified across classes. Interestingly, *news* pages had the highest number of correctly classified pages (i.e., 88.84%) followed by the combined two radical classes (i.e., 77%), which again showed moderately high precision rates (i.e., 94.4%) and moderately low recall rates (i.e., 77%). *News* class sites were also misclassified across the remaining three classes at the highest frequency (see Table 2).

C. Five-Class Model: Results and Quality of Rule-Set

The five-class tree analysis indicated that 76.22% of the pages were accurately classified. Of particular interest here were the *radical Right*- and *radical Islamic* pages. First, the confusion matrix indicated that a mere 53.92% of *radical Right* pages were successfully classified. Within this particular class, *news* pages were misidentified at the highest rate in the model; 28.54% of its pages were misclassified as radical right-wing. Second, 85.22% of pages that featured radical Islamic content were accurately classified. *News* pages were also misclassified at the highest rate (i.e., 9.60%) within the *radical Islamic* class. *News* pages again had the highest number of correctly classified pages across the entire sample (84.10%), similar to the classification results in the four-class model. Measuring the quality of the rule for the five-class model, the actual *radical Right* pages were classified at a low recall rate of 54%. At the other end of the spectrum, both the precision and recall measures indicated that *radical Islamic* pages were classified at a higher rate of success than the *radical Right* class; 96% of the *radical Islamic* pages were successfully classified, and actual pages within this class were successfully classified at a rate of 85.2% (See Table 3).

IV. Conclusions

We sought to build on a sentiment-guided web-crawler that could accurately classify the subtle yet radical-based text on radical Right and radical Islamic webpages. A few notable findings were produced. First, classification results from the two- and four-class models suggested that the sentiment-guided web-crawler lacked the overall ability to differentiate between the sentiment found on the selected websites that did and did not promote radical ideologies. This finding contrasts that of [10], where pro-extremist web-forum and webpage data were

		Predicted Pages		Precision	Recall
		Radical Right* and *Radical Islamic	***Anti-extremist, News* and *Other***		
Actual Pages	***Radical Right* and *Radical Islamic***	3,838	1,162	0.95	0.77
	Anti-extremist, News* and *Other	189	14,811	0.93	0.99

Table 1 – Confusion Matrix from the J48 Tree Analysis for the Two-Class Model

		Predicted Pages				Precision	Recall
		Radical Right* and *Radical Islamic	***Anti-extremist***	***News***	***Other***		
Actual Pages	***Radical Right* and *Radical Islamic***	3,850	184	906	60	0.94	0.77
	Anti-extremist	96	3,479	1,364	61	0.82	0.70
	News	57	336	4,442	165	0.57	0.89
	Other	75	244	1,085	3,596	0.93	0.72

Table 2 – Confusion Matrix from the J48 Tree Analysis for the Four-Class Model

		Predicted Pages					Precision	Recall
		Radical Right	***Radical Islamic***	***Anti-extremist***	***News***	***Other***		
Actual Pages	***Radical Right***	701	20	158	371	50	0.87	0.54
	Radical Islamic	4	3,153	168	355	20	0.96	0.852
	Anti-extremist	25	55	3,589	1,251	80	0.74	0.72
	News	40	35	561	4,205	159	0.59	0.84
	Other	34	36	399	936	3,595	0.92	0.72

Table 3 – Confusion Matrix from the J48 Tree Analysis for the Five-Class Model

classified at a high rate of accuracy. Second, the five-class model showed that radical Islamic pages were classified at a much higher rate of success than radical right-wing pages. Previous research suggests that the sentiment on radical Right websites (e.g. [1, 3]) and radical Islamic websites (i.e., [13]) is presented in a subtle manner, both to appeal to a wider audience and recruit new members, for example. However, our results lend support for [16]'s assertion that while radical Islamic sites attempt to legitimize their efforts by presenting themselves as news source websites, the sentiment is almost always related to radical topics (i.e., discussions of violence). That said, a more in-depth analysis is needed to assess whether there is a clear distinction between an extremist website featuring elusive messages and a news site. This problem could be explored using other tools of classification, such as random forests, or Bayesian methods to support vector machines and neural networks. Future studies should also integrate a qualitative understanding of how machine learning tools make decisions about the webpages that are visited. Doing so may increase the reliability of the results and increase the likelihood of identifying radical text online.

V. REFERENCES

1) L. Back, "Aryans Reading Adorno: Cyber-Culture and Twenty-First Century Racism," in *Ethnic and Racial Studies, 25*(4), (2002), pp.628-651.
2) L. Bowman-Grieve, "Anti-abortion Extremism Online," in *First Monday, 14*(11), (2009).
3) M. Caiani, D. della Porta, & C. Wagemann, *Mobilizing on the Extreme Right: Germany, Italy, and the United States*. Oxford: Oxford University Press, (2012).
4) H. Chen, *Dark Web: Exploring and Data Mining the Dark Side of the Web*. New York: Springe, (2012).
5) K. Cohen, F. Johansson, L. Kaati, & J. C. Mork, "Detecting Linguistic Markers for Radical Violence in Social Media," in *Terrorism and Political Violence, 26*(1), (2014), pp. 246-256.
6) R. Feldman, "Techniques and Applications for Sentiment Analysis," in *Communications of the ACM*, 56(4), (2013), pp. 82-88.
7) M. Hall, E. Frank, H. Geoffrey, B. Pfahringer, P. Reutemann, & I. Witten, "The WEKA Data Mining Software: An Update," in *SIGKDD Explorations, 11*(1), (2009), pp. 10-18.
8) M. Hu, & L. Bing, "Mining and Summarizing Customer Reviews," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2004).
9) Internet World Stats, *Internet Growth Statistics*, (2016). Retrieved from http://www.internetworldstats.com/emarketing.htm
10) J. Mei, & R. Frank, "Sentiment Crawling: Extremist Content Collection through a Sentiment Analysis Guided Web-Crawler," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, (2015).
11) M. Sageman, "The Stagnation in Terrorism Research," in *Terrorism and Political Violence*, 26(4), (2014), pp. 565-580.
12) R. Scrivens, G. Davies, R. Frank, & J. Mei, "Sentiment-based Identification of Radical Authors (SIRA)," in *Proceedings of the 2015 IEEE ICDM Workshop on Intelligence and Security Informatics*, (2015).
13) P. Seib, & D. M. Janbek, *Global Terrorism and New Media: The Post-Al Qaeda Generation*. London: Routledge, (2011).
14) M. Thelwall, & K. Buckley, "Topic-based Sentiment Analysis for the Social Web: The Role of Mood and Issue-related Words," in *Journal of the American Society for Information Science and Technology*, 64(8), (2013), pp. 1608-1617.
15) T. Thet, J. Na, & C. Khoo, "Aspect-based Sentiment Analysis of Movie Reviews on Discussion Boards," in *Journal of Information Science*, 36(6), (2010), pp. 823-848.
16) Y. Tsfati, & G. Weimann, "www.terrorism.com: Terror on the Internet," in *Studies in Conflict & Terrorism, 25*(5), (2002), pp. 317-332.