Positing The Problem: Enhancing Classification of Extremist Web Content Through Textual Analysis

George R. S. Weir[1], Richard Frank[2], Barry Cartwright[2] and Emanuel Dos Santos[1]

[1]Computer and Information Sciences, University of Strathclyde,
Glasgow, Scotland, UK

[2]School of Criminology, Simon Fraser University,
Burnaby, BC, Canada

Abstract— **Webpages with terrorist and extremist content are key factors in the recruitment and radicalization of disaffected young adults who may then engage in terrorist activities at home or fight alongside terrorist groups abroad. This paper reports on advances in techniques for classifying data collected by the Terrorism and Extremism Network Extractor (TENE) web-crawler, a custom-written program that browses the World Wide Web, collecting vast amounts of data, retrieving the pages it visits, analyzing them, and recursively following the links out of those pages. The textual content is subjected to enhanced classification through software analysis, using the Posit textual analysis toolset, generating a detailed frequency analysis of the syntax, including multi-word units and associated part-of-speech components. Results are then deployed in a knowledge extraction process using knowledge extraction algorithms, e.g., from the WEKA system. Indications are that the use of the data enrichment through application of Posit analysis affords a greater degree of match between automatic and manual classification than previously attained. Furthermore, the incorporation and deployment of these technologies promises to provide public safety officials with techniques that can help to detect terrorist webpages, gauge the degree of extremism of their content, discriminate between webpages that do or do not require a concerted response, and take appropriate action where warranted.**

Keywords—sentiment analysis, web-crawling, classification

I. INTRODUCTION

Webpages with terrorist and extremist content are thought to be key factors in the recruitment and radicalization of disaffected individuals, who may then engage in terrorist activities at home or fight alongside terrorist groups abroad. That said, the sheer volume of online data makes virtually impossible the individual examination of each webpage, conversational thread or post that may or may not be related to terrorism or contain terrorist sympathies. Thus, computer-guided data collection and analysis have become increasingly vital.

This paper contrasts two different approaches to classifying data collected by the Terrorism and Extremism Network Extractor (TENE) web-crawler. TENE is a custom-written program developed by the International CyberCrime Research Centre (ICCRC) at Simon Fraser University in Burnaby, British Columbia. TENE browses the World Wide Web, where it collects vast amounts of data, retrieves the pages it visits, analyzes them, and recursively follows the links out of those pages. The raw data gathered by the TENE web-crawler are extracted to NetDraw/UICNet to create a network structure of the content of the terrorist and extremist webpages visited and other pages to which they are linked [1]. Two independent research groups, one located at the ICCRC, and the other at Computer and Information Sciences at the University of Strathclyde in Glasgow (UST), sought to develop an automated means of classifying the extremist web content extracted by TENE.

In the initial stage, the web content was subjected to a process of manual classification by the ICCRC, during which each webpage was classified as "pro-extremist," "neutral" or "anti-extremist." Examples of pro-extremist content included recognized extremist websites such as *Stormfront.org*, a white supremacist forum, *americannazipartyy.com*, an America-based neo-Nazi forum, *hizbuttahrir.org*, a pro-caliphate Islamic political party, and *www.ikhwanweb.com*, the website of the Muslim Brotherhood, to mention a few. Examples of "neutral" content were derived from media sources that could be expected to report routinely on terrorist events from what might arguably be a more impartial, journalistic perspective. Examples of "anti-extremist" content were collected from websites opposed to violence, such as those of public safety agencies, or groups like the Global Counterterrorism Forum.

The objective for each of the two research groups was to develop an automated means of classifying the raw web data. In each case, the initial manual classification would be regarded as a ground truth, against which the success of any approach would be gauged. Both approaches (ICCRC and UST) used a combination of language analysis and data mining techniques to generate their classification. ICCRC relied upon the use of keywords, part-of-speech (POS) tagging and sentiment analysis as a basis for characterizing the important features in the data. The resultant list of features and associated values was input to the WEKA analysis tool, where data mining algorithms were applied [2]. At UST, on the other hand, textual analysis focused on the quantitative syntactic features present in the webpages, including frequency data and type token ratios (the proportion of repeated words). These features and associated values were (as was the case with the ICCRC) input to the WEKA analysis tool, where data mining algorithms were applied. The following describes the differing approaches adopted by the two groups.

II. METHODS

Two independent research groups, one located at the International CyberCrime Research Center (ICCRC) at Simon Fraser University, and the other at Computer and Information Sciences at the University of Strathclyde in Glasgow (UST), sought to develop an automated means of classifying the extremist web content extracted by TENE. Both classification methods are described below.

A. The ICCRC Classification Method

The language analysis step followed by the ICCRC involved the identification of "keywords" (linguistic markers) that would represent the pro-extremist, neutral and anti-extremist categories. This was accomplished using Open NLP (a language processing tool) to develop a POS tagger [2, 3]. In turn, the POS tagger analyzed each page, extracting nouns to create a frequency distribution, in turn identifying the most commonly used nouns. The ten most commonly used nouns from the pro-extremist, neutral and anti-extremist data sets were used in the sentiment analysis phase.

The actual sentiment analysis was conducted using SentiStrength, which assigns positive or negative values to lexical units in the text [4]. In addition, the keyword function in SentiStrength was used to identify those terms considered to exhibit the strongest degree of sentiment. Terms in proximity to the keywords were assigned sentiment values derived from SentiStrength's General Inquirer dictionary.

The resultant data, comprised of the noun keywords for each Web page with the associated sentiment score and the manual classification for that page, were then input to WEKA for classification. This method employed WEKA's standard J48 tree classification method with 10-fold cross-validation. In this cross-validation, 10% of the data was hidden, and conditions were sought that would split the remaining 90% of the dataset in two, with each part having as many data-points as possible belonging to a single class. Accuracy of the tree was then considered relative to the hidden 10% of the data. This process was repeated 10 times, each time with a different hidden 10% subset. WEKA produced a measure of how many of the pages were correctly classified.

B. The UST Classification Method

The language analysis step followed by UST employed existing software in the form of the Posit Textual Analysis Toolset (Posit) [5]. Posit consists of several software modules working in combination to provide a broad range of textual analysis tools.

The Posit toolset uses a Unix-based scripting approach. When Posit was originally developed, many then-existing Windows tools proved unable to adequately accommodate large text corpora. Developing the Posit toolset as Unix scripts afforded a flexible and easily customizable set of analysis components. Large files can be processed with ease, since the text data is not loaded in its entirety into main memory at one time, but instead, is processed sequentially. The Posit scripts are readily modularized and easily combined with executables—off-the-shelf POS-taggers or lemmatizers—to facilitate relatively easy updating and maintenance.

The system provides a means of generating frequency data, as well as Part-of-Speech (POS) tagging. The data output from Posit includes values for total words (tokens), total unique words (types), type/token ratio, number of sentences, average sentence length, number of characters, average word length, noun types, verb types, adjective types, adverb types, preposition types, personal pronoun types, determiner types, possessive pronoun types, interjection types, particle types, nouns, verbs, prepositions, personal pronouns, determiners, adverbs, adjectives, possessive pronouns, interjections, and particles, or 27 features in all.

The rationale behind the Posit toolset is to provide extensive quantitative analysis of individual samples of textual data as a basis for comparisons across corpora. Such insights have proved useful in a variety of contexts, including analysis of English as a Foreign Language (EFL) textbooks used across several decades in Japan [6] and an in-depth study of a Scottish newspaper corpus [7].

Posit analysis was applied to each of the webpages manually classified by the ICCRC. The Posit output was then pre-processed, along with the manual classification, which provided 28 features (the 27 features from Posit plus the manual classification) for direct entry into WEKA. This generated a measure of how many of the pages were correctly classified using the Posit features. Several classification algorithms were explored, each with 10-fold cross-validation, including the J48 tree-based algorithm employed by the ICCRC research group. However, of the various algorithms considered, the best match in classification (using the Posit features) was produced by Random Forest. Random Forest is another decision tree-based classification algorithm that constructs classification or regression trees differently from standard decision trees, by splitting each node, "using the best among a subset of predictors randomly chosen at that node" [8, 9].

III. RESULTS

The results of the WEKA standard J48 decision-tree classification method employed by the ICCRC indicated that 80.51% of webpages were classified correctly. The anti-extremist and pro-extremist categories had the highest degree of correctly identified pages, with 92.7% of the pro-extremist cases and 88% of the anti-extremist cases correctly identified. This indicates that the decision tree worked well in classifying extremist content. Performance on the neutral category was less convincing, at 68% [1].

The WEKA standard J48 decision-tree classification method employed by UST (using Posit) resulted in the correct classification of 91.4% of the webpages. The pro-extremist and neutral categories had the highest degree of correctly identified pages, with 94.4% of the pro-extremist cases and 90.8% of the neutral cases correctly identified. Performance on the anti-extremist category was slightly lower, at 88.3%.

When the Random Forest classification algorithm was applied by UST (again using Posit), accuracy of classification improved noticeably. With Random Forest, the overall correct classification reached 95.3%. Once again, the pro-extremist category had the highest classification success, at 96.7%, while

the neutral category and the anti-extremist category were at 94.9% and 94% respectively.

When contrasting the degree of success for any approach in matching the manual classification of the target Web content, a result that is closer to a 100% match may at first blush appear preferable. While this view may be broadly correct, however, it may be misguided, since the initial manual classification is likely to be less than completely accurate in classifying the content as pro-extremist, anti-extremist or neutral. Thus, if an automated classification attempts to match a less than entirely consistent manual classification, a match of less than 100% should be the expected result. A future action in our collaboration is to revisit the manual classification process and consider its degree of credibility. In turn, this will shed light on the plausible degree of variance for the automated classification from the manual judgment.

IV. Discussion

The language analysis approach adopted by the ICCRC was largely premised upon sentiment analysis. In contrast, the UST approach to language analysis relied entirely upon frequency of syntactic features. On the face of it, sentiment analysis might be expected to offer a more appropriate perspective on the source data than clinical frequency analysis, since the former considers the emotive content, whereas the latter does not. Nevertheless, the latter approach consistently afforded a higher degree of classification accuracy. This contrary result may reflect the limited impact of the sentiment attribute in the ICCRC analysis, since this was applied only to noun keywords. In contrast, the frequency-based analysis derived from Posit may yield sufficiently rich data that the data mining process can establish a high degree of correlation with the manual classification.

There may be scope for refinement of the semantics-oriented ICCRC approach, for example, through in-depth qualitative analysis of a large but randomized sample of the webpage content extracted by TENE, perhaps generating a new set of linguistic markers that could then be fed back into the WEKA classification phase as part of an iterative cycle. The Posit approach might be well suited to a wider variety of textual classification tasks, because the source data are considerably enhanced as a result of the Posit process, thereby providing a richer context for automated classification.

The incorporation and deployment of technologies such as those described above could well provide public safety officials with a new software toolkit that could help to detect terrorist webpages, gauge the intensity of their content, discriminate between those webpages that either do or do not require a concerted response, and allow them to take appropriate action where warranted. Also, in-depth textual analysis of the data collected by the TENE web-crawler could shed light on why certain individuals are more open to radicalization than others, why certain webpages and their content have greater appeal than others, and how those particular webpages might contribute to the recruitment and radicalization process.

V. References

[1] J. Mei and R. Frank, "Sentiment crawling: Extremist content collection through a sentiment analysis guided web-crawler," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015,* Calgary,Alberta, 2015, pp. 1024-1027.

[2] M. Hall, E. Frank, H. Geoffrey, B. Pfahringer, P. Reutemann and I. Witten, "The Weka data mining software: an update," *SIGKDD Explorations,* vol. 11, pp. 10-18, 2009.

[3] L. Chambers, E. Tromp, M. Pechenizkiy and M. Gaber, "Mobile sentiment analysis," in 2012.

[4] M. Thelwall and K. Buckley, "Topic-based sentiment analysis for the social web: The role of mood and issue-related words," *Journal of the American Society for Information Science and Technology,* vol. 64, pp. 1608-1617, 2013.

[5] G. R. S. Weir, "The Posit text profiling toolset." in *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguisitics,* Pattaya, Thailand, 2007, pp. 106-109.

[6] G. R. S. Weir and T. Ozasa, "Learning from analysis of Japanese EFL texts." *Educational Perspectives, Journal of the College of Education/University of Hawaii at Manoa,* vol. 43, pp. 56-66, 2010.

[7] G. R. S. Weir and N. K. Anagnostou, "Exploring newspapers: A case study in corpus analysis," in *ICTATLL Workshop 2007,* International Education Centre, Hiroshima International University, Japan, 2007.

[8] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[9] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News,* vol. 2, pp. 18-22, 2002.