Classification of Radical Web Text using a Composite-Based Method

Kolade Olawande Owoeye
Department of Computer and Information Sciences
University of Strathclyde, Glasgow
United Kingdom
kolade.owoeye@strath.ac.uk

George R. S. Weir
Department of Computer and Information Sciences
University of Strathclyde, Glasgow
United Kingdom
george.weir@strath.ac.uk

Abstract—**The spread of terrorism and extremism activities on the Internet has created the need for intelligence gathering via Web and real-time monitoring of potential websites for extremist activities. However, the manual classification for such contents is practically difficult and time-consuming. In response to this challenge, an automated classification system called Composite technique was developed. This is a computational framework that explores the combination of both semantics and syntactic features of textual contents of a Web page. We implemented the framework on a set of extremist Web pages - a dataset that has been subjected to a manual classification process. Thereby, we developed a classification model on the data using the J48 decision algorithm, to generate a measure of how well each page can be classified into their appropriate classes. The classification result obtained from our method when compared with other states of the art, indicated a 96% success rate overall in classifying Web pages when matched against the manual classification**.

Keywords—*Extremist, Posit, Classification, Sentiment, Web pages, Composite*

I. INTRODUCTION

The spread of extremist documents on the Internet is alarming and has become a major concern for government and security agencies. The potential dangers of online extremism cannot be overemphasized. For example, 3,000 people were killed in the 9/11 terrorist attacks in United States [1], while 4 people were killed and many were injured in extremist attack at Westminster, London [2] to mention a few. However, a survey from the National Consortium for the Study of Terrorism and Responses to Terrorism (START)[2], also reported 2,794 terrorist attacks that resulted in 3,659 deaths from 1970-2016 in the United States. The Global Terrorism Index, GTI [3] reported that Boko-Haram in Nigeria was the deadliest extremist group in 2014 with the record of 6,700 deaths. In 2016, this group was known as the third deadliest extremist group. Just a single terrorism attack in Nigeria was recorded among the 20 most deadly terrorist attacks worldwide in 2016. In 2014, nine similar attacks happened in that country.

The adoption of online presence such as YouTube, Facebook, online forums and Twitter gave extremist groups like, Boko-Haram and ISIS the opportunity to rise to thousands of members. Many of the extremist activities online involve radical discussions, fund raising, campaigns and recruitments etc. [4-5]. Examples of extremist websites are jihadi websites, far-right propaganda and bomb-making instructional websites.

However, one form of counter-terrorism measure is the classification of such extremist documents (Web pages) on the Internet. But manual classification of such content on the Internet is impractical due to billions of Web pages of diverse use. Faced with this challenge, we developed a computational framework, which is based on the combination of a data-mining algorithm, and the hybrid of both linguistic and syntactic features of the Web texts to build a model for the automatic classification of extremism Web pages.

This article details how we implemented our framework on a set of manually classified extremist Web pages. These Web pages were extracted from extremist websites through the aid of the Extremism Network Extractor (TENE) WebCrawler software designed at the International cyber Crime Research Centre (ICCRC), Simon Fraser University, Canada. This crawler follows links based upon keyword searches online, analyses and extracts each Web page visited [4]. However, the Web page data was subjected into the process of manual classification by ICCRC into three categories namely, "pro-extremist", "anti-extremist" and "neutral" based on the content each category exhibited. The main objective reported in this paper is to develop an automated means of classifying extremist Web pages. Therefore, the manual classification of the TENE-sourced Web pages serves as a threshold to measure the success of our automated method.

The linguistic features are obtained through the use linguistic markers to pinpoint sentiment in the Web page and a lexical approach, a Sentistrength resource then assigns the sentiment score to each Web page. While, syntactic features of textual contents of a Web are obtained through, a textual analytic tool called Posit. Posit is a Unix-Scripting program that is capable of generating frequency data, as well as Part-of-Speech (POS) tagging in unstructured textual data [reference to Posit here}.

In our research, we implemented a data-mining algorithm in the knowledge extraction software WEKA (Waikato Environment for Knowledge Analysis). WEKA is an open source software that has a collection of different algorithms and visualization tools utilized for different machine learning or data mining tasks [6].

This paper is organized as follows: section 1 is the Introduction followed by related work in section 2. Section 3 describes the methods used. Section 4 reports the classification results. Our conclusion is drawn in section 5 with bibliography presented in section 6.

II. Related Work

This section describes existing classification techniques for the identification and classification of extremist Web contents.

Various types of textual classification techniques have been used in identifying and classifying radical documents on the Internet. Examples of such techniques are Topic Modeling and Sentiment Analysis to mention a few. Sentiment analysis tends to determine opinion or emotion in unstructured textual data. Methods used in sentiment analysis include, Machine learning and Semantic orientation. A sentiment classification method was presented in [4]. The authors designed a WebCrawler to make a decision on each Web page it downloaded whether the page is pro-extremist, anti-extremist or neutral. The process was achieved through the use of frequently used keywords as linguistic markers to pinpoint the sentiment in each page. In [7] Sentiment and social analysis were combined as a technique used to survey the agenda of a radical group in YouTube. The polarity for each topic discussed within the group was obtained and explored to model individual's behaviour. Eventually, it was spotted that extremism and intolerance were prominent among female users. Hierarchical clustering was applied to divide extremist Web pages in politics and religion categories [8]. Data retrieved from the Dark Web Portal Project was used to conduct the first proposed method to detect cyber recruitment effort [9].

A sentiment-based classification method was employed for Twitter analysis classification in [10]. Web Forums were used for opinion classification in [11]. Twenty eight (28) different extremist religion forum discussions translated from Arabic to English were compiled for annotation Thereafter, the authors used a set of textual features and Bayesian criteria to classify the corpus. An accurate result was obtained and the most predictive terms were highlighted [12].

Machine learning algorithms such as Naïve Bayes and Support Vector Machines, were used to classify positive and negative features in given data [13]. A machine learning framework that explores a mixture of network, metadata and temporal features to detect extremist users, predict content adopters and interaction reciprocity in social media was presented in [14].

Another method of sentiment analysis is Semantic orientation. This depends on exploring a corpus annotated for sentiment functions or a dictionary comprising words with unique sentiment values [15].

Posit is a textual analytic tool that enriches representation for text by giving counts on syntactic and quantitative values for texts which are useful for textual classification models. Posit textual analysis has been deployed for diachronic analysis of English textbooks used in Japan [16] and Posit was employed for the analysis and categorization of a Scottish newspaper corpus [17].

Two different techniques used for automatic classification of extremist Web pages were contrasted in [18]. The aim of the research was to determine the best automated classification system among approaches that can efficiently place each Web page into the appropriate classes. The two approaches are Posit-textual analysis [16] and a Sentiment classification rule-based technique [4]. These techniques were applied separately on the aforementioned extremist Web pages. A classification model was then developed on the features generated by each technique, using J48 decision tree as the classifier algorithm. Eventually, the results obtained indicated that Posit results out-performed the results obtained from the sentiment-based classification method.

While several methods have been developed in the literature, the proposed method presented in this study is an underpinning of existing work on sentiment analysis that uses keywords as a linguistic marker technique to pinpoint sentiment in a Web page and the Posit textual analytic tool that generates syntactic features of textual content from a Web page which are useful input for classification models. However, the use of the linguistic marker technique in some sentiment analysis (i.e. the use of frequently used keywords) can result in non-capture of some sentiment values from larger Web page data, thereby hindering the training process of useful sentiment features of the Web pages. In addition, the Web pages with no sentiment values in the aforementioned method are due to non-keyword presence in those pages, which may be misclassified into a class of highest probability due to application of the classifier generalization rule during the classification process. This situation could also produce false positive results in some classes. Based on the evidence, we propose a method that tends to offer a wider coverage of more useful features in the Web pages. The proposed method explores a hybrid of both semantics and syntactic features from the textual contents of Web pages to build a classification model for extremist Web content.

III. Method

In this section, we describe the Web dataset and the different features extraction techniques used.

A. Web Dataset

The dataset comprises 7500 Web pages manually classified into "pro-extremist," "neutral" and "anti-extremist". The Web pages were categorized according to what their content revealed. For example, the anti-extremist category revealed contents that are against violence and intelligent agencies such as the Global Counterterrorism Forum. This category contains 2500 Web pages that were harvested from 10 different Websites. The neutral category contained material that reports violence and terrorist matters but from a journalistic perspective. This category consists of 2500 Web pages were harvested from 15 websites such as news websites. The pro-extremist category comprises terrorist and Jihad society contents from a white supremacist forum, an America-based neo-Nazi forum, a pro-caliphate Islamic political party, and the website of the Muslim Brotherhood, to mention a few. In this category, 2500 Web pages were obtained from 11 different Websites.

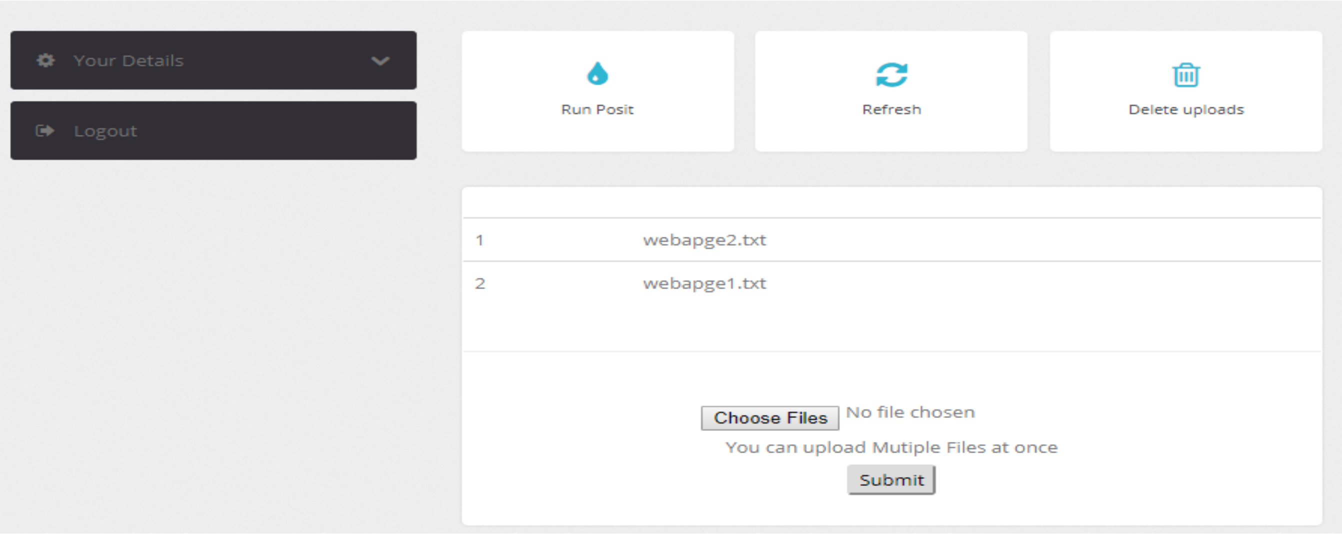
B. *Posit-Textual Analysis*

The Posit textual analysis tool-set is a program written mainly in UNIX script and is capable of generating a detailed syntactic and frequency analysis of a textual corpus [16]. Posit outputs quantitative data from any text, including, word count, number of characters and sentences, number of token and types, n-gram frequencies and finally, part-of-speech tagging (POS) [17]. By default, the Posit produces data on 27 features. The features include, noun types, possessive pronoun, personal pronouns, average sentence length, determiners, adverbs values for total words (tokens), total unique words (types), type/token ratio, number of sentences, number of characters, average word length, verb types, adjective types, adverb types, preposition types, personal pronoun types, determiner types, types, interjection types, particle types, nouns, verbs, prepositions, adjectives and interjections.

Recently, Posit has been implemented in an integrated full-featured Cloud-based version [20]. This system provides the full scope of the Posit application in the analysis of text data sets. The Cloud-Posit system was developed with two modes, the interactive Cloud-based and the Posit-API version. In the interactive Cloud-based mode, third-parties can access a Posit facility which enables the upload of several data files in a set. After uploading the file, the 'run Posit' option is selected and each file in the dataset is analysed in a logical order. The output result for each file is created in a separate folder and the complete set of analysis folders is compressed into a single file, which is downloaded to the remote Web client. Through the Posit-API version, remote users can process multiple files for analysis and fetch the result files directly for further processing.

The interactive Cloud-based Posit is suitable for analysis of small data sets, while the API access needs no user interaction. The advantage of the interactive mode is that, researchers attempting to 'train' their classification model could initially use it for an appropriate classification algorithm. Once the appropriate model is developed, the larger volume of data would then be processed through the non-interactive API of Cloud-Posit. In this study, the API of Cloud-Posit was employed. Figure 1 shows the Cloud-Posit interactive facility.

When Posit is applied on the extremist Web pages through the API access, it outputs statistical details of the text content of the Web in terms of individual words (tokens) and word types. The frequency data is generated for particular parts of speech, including frequency ordered accounts of each specific word in the analyzed text. The output produces three different levels of detail, a summary level, the intermediate (aggregate) part-of-speech analysis and the finely detailed word types together withn the part-of-speech analysis. The summary level includes the total number of verbs, nouns, adverb, etc. (Figure 2). In addition, frequency data is produced in the intermediate level for the contents of the text analyzed in terms of particular parts-of-speech. For example, it generates analysis of different forms of verb such as, the base type of verbs, the gerund, the past tense, the past participle, the 3rd person present, the present tense (non-3rd person) form and the 3 modal auxiliary form as shown in (Figure 3). In the fine detail level, frequency data for each word in terms of part-of-speech type is provided, such as, the number of occurrences of every word that are in the past participle form, etc. An illustration of this level is given in Figure 4. Using the summary level of detail from the Posit analysis, the resultant feature data generated by Posit together with the manual classification, produces 28 features across the 7500 Web pages.

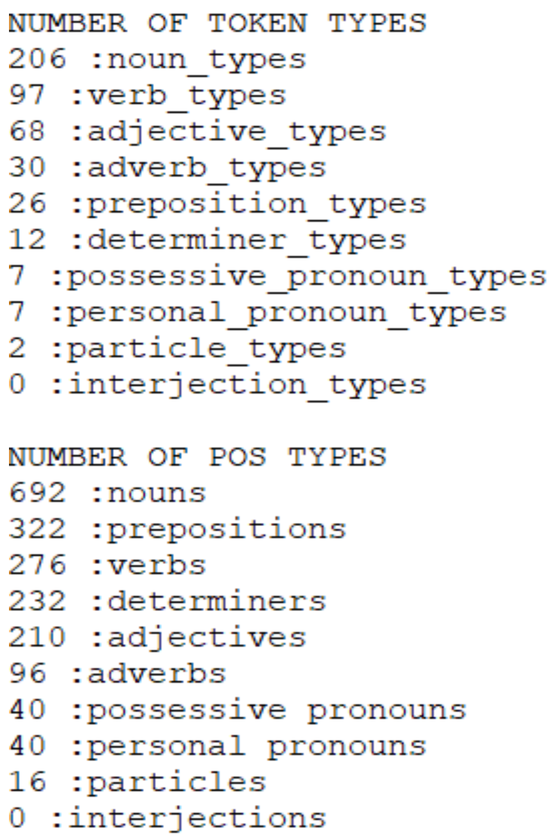


1 webapge2.txt

2 webapge1.txt

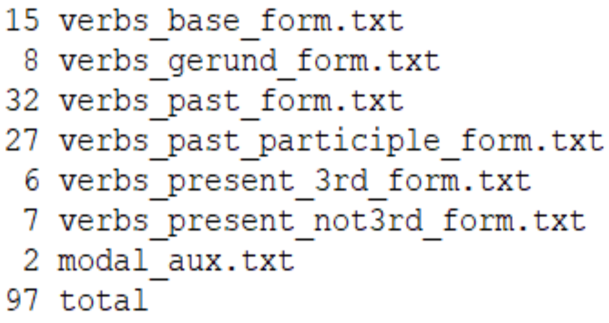
You can upload Mutiple Files at once

Figure 1: Cloud-Posit interactive facility



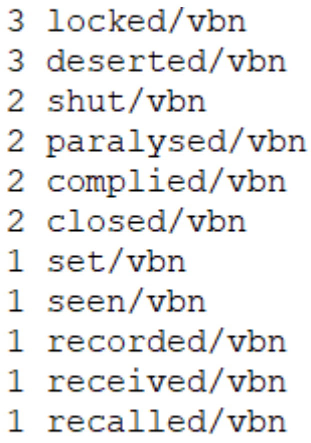
```
NUMBER OF TOKEN TYPES
206 :noun_types
97 :verb_types
68 :adjective_types
30 :adverb_types
26 :preposition_types
12 :determiner_types
7 :possessive_pronoun_types
7 :personal_pronoun_types
2 :particle_types
0 :interjection_types

NUMBER OF POS TYPES
692 :nouns
322 :prepositions
276 :verbs
232 :determiners
210 :adjectives
96 :adverbs
40 :possessive pronouns
40 :personal pronouns
16 :particles
0 :interjections
```

Figure 2: Example of Posit Summary Output



```
15 verbs_base_form.txt
 8 verbs_gerund_form.txt
32 verbs_past_form.txt
27 verbs_past_participle_form.txt
 6 verbs_present_3rd_form.txt
 7 verbs_present_not3rd_form.txt
 2 modal_aux.txt
97 total
```

Figure 3: Example of Posit Aggregate Output of Verbs



```
3 locked/vbn
3 deserted/vbn
2 shut/vbn
2 paralysed/vbn
2 complied/vbn
2 closed/vbn
1 set/vbn
1 seen/vbn
1 recorded/vbn
1 received/vbn
1 recalled/vbn
```

Figure 4: Example of Posit Detail for Past Participle Form

C. *Sentiment Analysis*

The aim of this section is to report on the generation of sentiment features. The sentiment features of the text document in each Web page of the Web data are obtained by subjecting the Web data to part-of-speech (POS) tagging. Therein, the top ten frequently used noun keywords that have significant meanings were used to pinpoint terms that show a high degree of sentiment in each page. In fact, 26 noun keywords were obtained due to the overlapping of some keywords. The keywords used are Syria, Counter terrorism, program, affairs, Court, Ebola, Facebook, Islam, Jihad, Military, Muslim, News, Policy, Politics, President, Press, Rights, Safeguards, Syria, Trial, Twitter, CNN, Crime, victims, war and security. Five word terms range around each of this specific keyword were selected in each page and deployed into a lexical resource (Sentistrength) to generate the sentiment value for each page, which is derived from Sentistrength's General Inquirer dictionary. Details of the sentiment analysis are explained in [19].

D. *Composite Method*

The Composite Method operates through a custom written script that merges together semantic features derived from sentiment analysis and the frequency of syntactic features obtained from Posit. The rationale behind the hybrid features in the composite approach is to explore the richer feature set that feeds into building a classification model. Sentiment analysis tends to offer prospective method on unstructured data because it contains opinionated contents while Posit provides the quantitative syntactic features that 'enrich' the information given by the text corpus. The output features generated from Posit and the sentiment-rule base have proven to be significant in developing a classification model [4][18][19].

However, the output generated by the composite technique across the three categories produces 54 features.

E. Evaluation Metrics

Precision, Recall, F-measure, and Accuracy were employed as metrics used for the performance evaluation of our system. TN means True Negative, False Positive (FP), False Negative (FN) and True Positive (TP).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{F-measure} = 2\times\frac{\text{Precision X Recall}}{\text{Precision+Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (4)$$

$$\text{Error Rate (E)} = \text{1-Acc} \quad (5)$$

F. WEKA Implementation

This section describes the classification results using the J48 decision tree algorithm.

We implemented our classifier, J48 decision tree algorithm in WEKA (Waikato Environment for Knowledge Analysis) to create a rule-building process for the automated classification system. The aim of this process is to generate measures that show how the system assigns each page into their appropriate classes. The J48 decision tree algorithm is employed because it is an efficient algorithm for text classification which supports a rule building process and variable screening performance. The J48 algorithm is applied to the features produced from the composite approach with 10-fold cross validation. That is, the dataset was split in a way that 90% of the dataset was used for training while the remaining 10% were used for testing the accuracy, this process was done 10 times and the average accuracy was taken. The J48 decision tree algorithm splits each level of the data in a manner corresponding to different attributes. The non-leaf nodes are denoted by attributes while the leaf nodes indicate the predicted variable. Eventually, the J48 algorithm generated a measure of how the pages were correctly classified into their respective class in each automated method.

IV. CLASSIFICATION RESULTS.

This section details the classification results obtained in the Composite method.

The results interpreted in this section are largely focused on the overall classification and the pro-extremist class. The result obtained from the Composite method indicated that overall 96% of the Web pages from the three classes were accurately classified into their appropriate classes when matched against manual classification. From the confusion matrix, it was noted that pro-extremist and anti-extremist classes had the highest proportion of correctly identified pages, at 97% and 95% respectively. The performance in the neutral class was lower compared to the other two classes. Considering the pro-extremist class in the model, we observed higher precision and recall in the pro-extremist class than the other two classes, at 97.9% and 97.8% respectively. Also, the F-measure from the model indicated 97.8% performance. This indicates that the algorithm is efficient in classifying Web pages with extremism contents. The classification result is presented in Table 1 below.

A. Comparison with Sentiment Rule-Based classification

In the overall Web page classification, the Composite method showed the highest proportion of correctly identified pages across all the classes at 96% success rate when matched against the manual classification, unlike the sentiment–rules method that achieved 93% in [19]. The error rate is minimal in the Composite approach at 0.023 compared to the sentiment-rule based method. However, the time to build a model in the sentiment-rule based method is less than the Composite approach at 0.55sec and 1.09sec respectively. In addition, high recall and precision rates were observed in the Composite approach at 97% and 97% respectively, better than the Sentiment rule-based result. Considering the pro-extremist class in the Composite method, the model indicated a lower false positive result of 1%. This indicates that there was a lower level of erroneous classification of Web pages in the pro-extremist class compared to the results obtained in the sentiment rule-based classification method reported in [19]. Taken into account the overall parameters in the analysis, the parameters show that Composite method is more efficient in discerning contexts that contain extremist content than the sentiment rule-based classification method

Table 1-J48 algorithm Classification Results from Composite Method

Correctly Classified Instances	7204	96.0133 %
Incorrectly Classified Instances	296	3.9467 %
Kappa statistic	0.9408	
Mean absolute error	0.0229	
Root mean squared error	0.1353	
Relative absolute error	6.8717 %	
Root relative squared error	33.1378 %	
Total Number of Instances	7500	

=== Detailed Accuracy by Class ===

TP	FP	Precision	Recall	F-Score	Class
0.957	0.028	0.944	0.957	0.950	Anti-Extremist
0.978	**0.010**	**0.979**	0.978	0.978	**Pro-Extremist**
0.947	0.020	0.959	0.947	0.953	Neutral
0.961	0.020	0.961	0.961	0.961	

=== Confusion Matrix ===

```
a    b    c   <-- classified as
 2392  27  81      a = ANTI-EXTREMIST
  35 2444  21      b = PRO-EXTREMIST
  107  25 2368      c = NEUTRAL
```

T

Table 2-J48 algorithm Classification Results of Sentiment-Rule Based Method [19]

Correctly Classified Instances	7040	93 %
Incorrectly Classified Instances	460	6.1%
Kappa statistic	0.908	
Mean absolute error	0.0466	
Root mean squared error	0.1691	
Relative absolute error	0.4804 %	
Root relative squared error	35.8756 %	
Total Number of Instances	7500	

=== Detailed Accuracy by Class ===

TP	FP	Precision	Recall	F-Score	Class
0.987	0.042	0.921	0.987	0.953	Pro-Extremis
0.942	0.029	0.942	0.942	0.942	Anti-Extremis
0.887	0.020	0.956	0.887	0.920	Neutral
0.939	0.031	0.940	0.939	0.938	

=== Confusion Matrix ===

```
a    b    c   <-- classified as
2468   16   16 |    a = PRO-EXTREMIST
59 2355   86 |    b = ANTI-EXTREMIST
153  130 2217 |    c = NEUTRAL
```

The rapid increase of extremism documents online has created the need for efficient automated systems for the classification and identification of Web pages with extremism contents. This will assist in triage and further investigation on particular Web pages that are likely to relate to terrorism or extremism. This will also aid in countering extremist activities such as recruitment and radicalization on the Internet. The composite classification method developed in this research has demonstrated a high degree of robustness and efficiency in building an automatic classification system for a representative set of extremist and terrorist Web documents. The results presented out-performed the existing method of sentiment rule based classification.

The composite-based method might also be well suited to a wider variety of textual classification tasks in other Web content domains due to its richer context for automated classification. This future work will further test the approach's robustness and versatility.

VI. References

[1] L. Eric, "Bin Laden Chose 9/11 Targets, Al Qaeda Leader Says" in the New York Times, 2003

[2] G. Birchall, W. Chrismas and P.Harper, "TERROR IN THE CAPITAL," in the Sun News Paper, Retrieved from: https://www.thesun.co.uk/news/3151868/london-westminster-terror-attack-bridge-victims-about/

[3] J. S. Rivinius, "START Background Report," published in Department of Homeland Security Center of Excellence, University of Maryland. Retrieved from: http://www.start.umd.edu/news/proportion-terrorist-attacks-religious-and-right-wing-extremists-rise-united-states, November 2017.

[4] 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.

[5] Q. Schiermeier, "Terrorism: Terror Prediction Hits Limits." Nature, vol. 517, no. 7535, p. 419, 2015.

[6] 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.

[7] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton, "Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation," in 2009 International Conference on Advances in Social Network Analysis and Mining (ASONAM). IEEE, 2009, pp. 231–236.

[8] X. Qi, K. Christensen, R. Duval, E. Fuller, A. Spahiu, Q. Wu, and C.-Q. Zhang, "A Hierarchical Algorithm for Clustering Extremist Web Pages," in 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2010, pp. 458–463.

[9] J. R. Scanlon and M. S. Gerber, "Automatic Detection of Cyber Recruitment by Violent Extremists," Security Informatics, vol. 3, no. 1, pp. 1–10, 2014.

[10] Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," Inproceedings of the Workshop on Language in Social Media (LSM 2011), Portland, Oregon, 23 June 2011, pages 30–38.

[11] A. Abbasi, H. Chen and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," ACM Transactions on Information Systems, 26(3), 2008, pp 1-34.

[12] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann, "Uncovering the Dark Web: A Case Study of Jihad on the Web," Journal of the American Society for Information Science and Technology, vol. 59, no. 8, pp. 1347–1359, 2008.

[13] Kechaou, Z., Ammar, M., & Alimi, A. "A Mutli-Agent Based System for Sentiment Analysis of User-Generated Content,". International Journal on Artificial Intelligence Tools, 22(2), (2013). 1350004-28.

[14] E. Ferrara, W.-Q. Wang,O. Varol,A. Flam- mini, and A. Galstyan,, "Predicting Online Extremism, Content Adopters, and Interaction Reciprocity," InSocial Informatics: 8th Intl. Conf., SocInfo 2016, Bellevue, WA, USA, 2016, pp.22–39.

[15] Feldman, R., "Techniques and Applications for Sentiment Analysis" in Communications of the ACM, 56(4), (2013), pp. 82-88.

[16] 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, 2010, pp. 56-66.

[17] 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.

[18] G. R. S. Weir, E. D Santos, B. Cartwright and R.Frank, "Positing The Problem: Enhancing Classification of Extremist Web Content Through Textual Analysis," in Proceedings of the 4th International Conference on Cybercrime and Computer Forensics (ICCCF), Simon Fraser University, Vancouver, Canada.2016

[19] K.O. Owoeye and G. R. S. Weir, "Classification of Extremist Text on the Web using Sentiment Analysis Approach", in Proceedings of the 5th Annual Conf. on Computational Science & Computational Intelligence (CSCI'18), Las Vegas, USA, 2018

[20] George R S Weir, Kolade Owoeye, Alice Oberacker and Haya Alshahrani, “Cloud-Based Textual Analysis as a Basis for Document Classification” in the 7th International Workshop on Security, Privacy and Performance in Cloud Computing (SPCLOUD 2018)