Classification of Extremist Text on the Web using Sentiment Analysis Approach

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 high volume of extremist materials online makes manual classification impractical. However, there is a need for automated classification techniques. One set of extremist web pages obtained by the TENE Web-crawler was initially subjected to manual classification. A sentiment-based classification model was then developed to automate the classification of such extremist Websites. The classification model measures how well the pages could be automatically matched against their appropriate classes. The method also identifies particular data items that differ in manual classification from their automated classification. The results from our method showed that overall web pages were correctly matched against the manual classification with a 93% success rate. In addition, a feature selection algorithm was able to reduce the original 26-feature set by one feature to attain a better overall performance of 94% in classifying the Web data.

Keywords—Extremism, Sentistrength, Classification, Sentiment, Web pages

1. Introduction

While technology has advanced the use of Internet information, text data is the most common content type in the Web. However, terrorist and extremist groups are also adopting Web technologies for different functions including dissemination of information, propaganda, fundraising, recruitment and assignment of deadly missions [1-2]. In such contexts, the Internet poses a threat to national security. Many law enforcement and intelligence agencies are interested in countering the use of the Internet for extremism or terrorism. One of the ways to accomplish this task is through classification and identification of Web text with radical contents. Yet the huge volume of information on the Internet makes it a heavy burden on humans to classify all contents, therefore machines are required to assist with automated classification. Automated methods for overall review and classification of Web documents are necessary to give insights and aid law-enforcement agencies in decision-making.

In research reported in this paper, we developed an automated classification system for web pages obtained from extremist Websites through the TENE-WebCrawler. The TENE-WebCrawler is software developed at the International Cyber Crime Research Centre (ICCRC), Simon Fraser University, Canada. This crawler follows links based upon keyword searches through the Internet, extracts web pages and analyses each page visited [1]. One set of such web pages was initially subjected to manual classification by ICCRC personnel, whereby each Web page was grouped as "pro-extremist", "neutral" or "anti-extremist" based on their contents. This initial manual classification then served as a threshold to measure the success of any automated classification approach. Based on the initial manual classification of the web pages, we developed a computational approach capable of effectively classifying extremist web pages, and able to identify which particular data items (pages) differed in manual classification from the automatic classification and, finally, able to determine which features are most relevant for the desired classification.

The aim of this work is to develop a computational framework that explores both a lexical approach and a data-mining algorithm toward automating the classification of web pages that contain radical contents. In our method, we implemented a data-mining algorithm in knowledge extraction software (WEKA). WEKA is an open source software that comprises a collection of different algorithms and visualization tools used for different machine learning tasks.

To this end, we extracted the linguistic features of the Web content in each page and assigned sentiment value to each page. These details were fed, together with their manual class, into WEKA, where a data-mining algorithm was applied in order to build a classification system. We conclude our analysis by confirming the broad-brush approach to manual classification using the automated classification system. In addition, we explored different feature selection algorithms to reduce the feature space of the dataset and obtain better classification accuracy. This paper describes the approaches used and the results achieved.

.

II. Related Work

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

Recently, much attention and efforts have focused on issues relating to online terrorism, propaganda, radicalization, campaigns and extremism [2]. Among various techniques used in identifying online radicalization or extremism are topic modeling and sentiment analysis to mention a few. However, the present research focuses more on sentiment analysis. Sentiment analysis is widely used in recent research due to its traditional means of analyzing customer reviews and understanding the public's view (negative, positive or neutral) towards specific products [6][3].

Sentiment analysis uses a computational approach to obtain opinionated contents and classifies the overall review of the topic into positive, negative and neutral. It also reveals user's intentions, emotions and opinion hidden in unstructured text [4]. The techniques used by sentiment analysis for classification include machine learning and lexicon-based approaches. Machine learning algorithms such as support vector machine [5] and Naïve Bayes classifier [6] were used in classifying text document to rate positive and negative contents based on given data.

Studies in [7] presented a machine-learning framework that combined a mixture of metadata, network and a temporal feature that were used to discover the followings: radical users, if common users will use radical content and if there will be an exchange of contacts initiated by extremists. The work explored millions of tweets from more than 25 thousand common users that were manually tagged, reported and suspended as a result of their involvement with extremist movements by twitter and another sample of tweets obtained randomly from 25 thousand common users who are open to the extremist contents. All the information was used for the forecasting tasks. Eventually, the performance of the framework revealed 93% success rate for extremist user detection and 80% rate for predicting content adopter.

Another method used in sentiment analysis for classification of a text document is semantic orientation. This operates by depending on a method that utilized a corpus annotated for sentiment or a sentiment value derived from a dictionary of words to classify text document [8]. Many studies have explored a hybrid of both data mining algorithm and semantic orientation (a lexical approach) in classifying or identifying extremism web pages, such as [1][9][10]. An authorship analysis framework was implemented on the linguistic features extracted from online messages in [10]. The result was evaluated to determine the stylistic features of terrorist communications. A multilingual model comprising a set of algorithms and related features was used to detect Arabic messages and their language's unique peculiarities on an Arabic and English Web forum associated with radical groups. Two classifiers namely, C4.5 and Support Vector Machine were used on the features. The results from their model indicated that SVM out-performed C4.5, and a high degree of success in identifying the communication pattern was produced.

In [9] 20,000 web pages were collected with the aid of a WebCrawler to assess differences in five sentiment classes namely: anti-extremist sites, radical Islamic sites, radical right sites, sites that did not discuss extremism and news source sites discussing extremism. That is, pages that relate to extremism or not. 198 frequently used keywords were identified through the aid of POS tagging. These keywords were used to calculate sentiment values for each page through sentiment analysis. The result obtained showed that the radical Islamic text class was classified at a much higher rate of success than the radical right text class. A WebCrawler called TENE-WebCrawler was designed to make a decision on each Web page it downloaded whether the page is pro-extremist, anti-extremist or neutral in [1]. The process was achieved through the combination of semantic orientation and data mining techniques to produce their classification.

Classification tools are WEKA [11] RapidMiner5 and LIBLINEAR [12] to mention a few. Among the few classification tool sets mentioned, WEKA (Waikato Environment for Knowledge Analysis) is widely used because it is an open source software. WEKA was explored for classification tasks in [1, 13, 14, 15, 16, and 17].

In the literature, many studies developed an automated process to categorize and label a large amount of Web data. However, in our classification approach we further the exploration on reducing the feature space of the Web dataset for desired web page classification accuracy. In addition, we carry out an identification process to identify the data items that differ during the manual and automated classification process.

III. Method

In this section, we describe the linguistic features extraction process for the text in each Web page.

A. *Sentistrength*

Sentistrength is a lexical program written in java that contains some specific algorithms that run on a set of texts [14]. Sentistrength explores the dictionary of categorized terms to determine the sentiment of a document by analyzing a text and assigning polarity values of either negative or positive to each word within the text based on rules such as idiom list, word list emoticon, spelling, boosting words (e.g., very) and negation words (e.g., not) [18]. Sentistrength uses several dictionaries, for example, Harvard's general inquirer database, to determine the sentiment values of some terms [18]. The output from Sentistrength is a single scale (-4 to +4), binary (negative/positive), or trinary (neutral/positive/negative) results.

B. *Feature Extraction Process*

In our approach, linguistic features of the text content of each Web page in the dataset are obtained through the aid of part-of-speech (POS) tagging where the top ten noun keywords that accrued significant meanings were explored to spot terms that show a strong degree of sentiment in each page. However, due to the overlapping of some keywords, we obtained 26 keywords. The keywords 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. Noun keywords are employed because the context around them contains more sentiment. Similar approaches that use Noun keywords in the sentiment analysis include [1, 19, and 20]. In addition, five words range at either side around each of the specific keywords were selected in each page and fed into Sentistrength to produce each page with their sentiment value that is derived from Sentistrength's General Inquirer dictionary. The scope of five words is selected because Sentistrength has a good accuracy level for short non-political Web texts in English [18]. However, in the process of obtaining the sentiment values, there are situations where a Web page from the sample contains more than one occurrence of a particular keyword and each occurrence has different scores in a page based on the context. In our approach, we obtain a single sentiment value for such pages by finding the mean of the sentiment scores.

C. Dataset

In the following, we describe the data set used, the evaluation metrics and the classification model using the J48 decision tree classifier.

The dataset comprises 7500 web pages manually classified into "pro-extremist," "neutral" and "anti-extremist." the web pages were classified based on the contents they exhibited. For example, the neutral group reflects contents from the media/news that report impartially on terrorist events. In the neutral class, 2500 web pages were obtained from 30 Websites. The anti-extremist class contains Web content that reports the countering of terrorism and operations of intelligence agencies. The anti-extremist class consists of 2500 web pages from 10 Websites. Pro-extremist pages express extremism content "extremist and jihad organization Websites". Examples of such Web sources are a white supremacist forum, America-based neo-Nazi forum. In this class, 2500 web pages were obtained from 11 different Websites. The dataset was setup in line with the configuration used in the existing work where the classification method used was based on sentiment-rule [1]. The dataset was chosen to establish a basis for comparison with our existing study with a view to improvement for the classification method

D. Decision Tree

The J48 decision tree is a supervised data-mining algorithm that develops a classification or regression model in a tree-like form. It operates by determining the dependent variable, that is, the target value of new sample using the various attribute values in a given data set. The branches between the nodes of the decision tree show the possible values of the attributes in a given sample; the internal nodes indicate the different attributes and the terminal nodes produce the final value, that is, the classification of the dependent variable [21]. The decision tree is built top down in a recursive manner and uses information gain heuristic to choose the attribute to split on. The 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. Each of the leaves indicates a certain set of sentiment thresholds. It is with this threshold that the decision tree confirmed whether a particular Web page is classified in the pro-extremist, anti-extremist or neutral class. The different types of decision tree include ID3, (CART) and C4.5 [21]. However, in this research, we explored the C4.5 algorithm that is implemented as J48 in WEKA. Other algorithms are Support Vector Machine, Naïve Bayes and Neural Networks. The J48 decision tree algorithm is adopted because it gives a better understanding of how the algorithm makes decisions. In addition, it contains an algorithm for text classification, which allows a rule-building process [16].

E. Implementation

The data, comprising each Web page with their associated sentiment score and manual classification were deployed into WEKA, where the J48 algorithm was applied with 10-fold cross-validation. That is, the dataset was split in such a way that 90% of the dataset was used for training and the remaining 10% was used for testing, this process was repeated 10 times and the mean accuracy was taken.

Precision, Recall, F-measure, and Accuracy were employed as the metrics used for performance evaluation of the system.

IV. WEKA Classification Results

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

The results from the J48 decision tree algorithm indicated that of the 7500 web pages processed, 93.8% of the pages were correctly classified into their respective classes. The pro-extremist and anti-extremist classes had the most correctly identified pages, with 98.7% of pro-extremist cases and 94.2% of anti-extremist cases. However, performance on the neutral class was low at the rate of 88.7%. Table 1 shows the classification result.

Table 1-J48 algorithm Classification Results

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-Measure	Class
0.987	0.042	0.921	0.987	0.953	Pro-Extremist
0.942	0.029	0.942	0.942	0.942	Anti-Extremist
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
```

A. Feature Selection

This section details the evaluation of different feature selection techniques used to select relevant features for the classification model.

Feature selection operates by selecting a subset of relevant features for use in model construction to improve accuracy and run-time most especially in model construction where there are numerous features and comparatively few samples (or data points). The objective of feature selection is to improve the prediction performance of the predictors, providing faster and more cost-effective predictors. The three different classes of feature selection are wrappers, filters and embedded methods. Filter algorithms use independent criterion to evaluate the efficiency or performance of feature subset without involving the machine-learning algorithm. The independent criteria refer to information measures, distance measures and consistency measures [22]. The search method used by filter method is ranker. Ranker ranks features by their corresponding individual evaluations. Few examples of filter methods include the information gain, Chi-squared test and correlation coefficient scores. The Wrapper technique uses a dependency criterion. That is, it involves machine-learning algorithm and explores the performance of the algorithm on the selected subset to identify which features are selected [23]. That is, a predictive model is used to evaluate a combination of features and assign a score based on model accuracy. The wrapper algorithm produces better performance but it is computationally expensive. The search process in wrapper method could be systematic, it might explore a random hill-climbing algorithm, best-first search or forward selection and backward elimination to add and remove features. Embedded method operates by learning which features most contribute to the accuracy of the model when the model was created [23]. The most common type of embedded feature selection methods are regularization methods (introduce additional constraints into the optimization of a predictive algorithm). Elastic Net, Ridge Regression and LASSO algorithms are the examples of regularization methods [23]. However, this study employs three different feature selection algorithms such as CorrelationAttributeEval, Information Gain and Gain Ratio algorithms to reduce the feature space of the dataset.

B. CorrelationAttributeEval

CorrelationAttributeEval operates by determining the value for an attribute, the process is achieved by calculating the correlation, (Pearson's) between the class and attribute [24]. Each value in the nominal attributes is considered as an indicator by the evaluator. The total correlation for a nominal attribute is obtained by calculating a weighted mean

C. Information Gain

To comprehend the gain ratio and information gain metrics, it is important to define entropy H(C) that evaluates the information content of the class, C. [25]. Information gain (Info Gain) estimates the quality of the feature by measuring information weight of each feature, but taking into account the class features. Information gain finds how much information is added when a feature is included. The information gain is calculated as follows:

$$\text{Information Gain}(\text{Class}, \text{Attribute}) = \text{H}(\text{Class}) - \text{H}(\text{Class}|\text{Attribute})$$

Where, H is the information entropy. However, the information gain is biased towards multi-valued attributes in the test results. Gain Ratio technique was developed to obtain the ratio so as to overcome the limitation of information gain. The gain ratio is calculated between a feature and the category as the information gain divided by the information value of the attribute:

$$\text{GainR}(\text{Class}, \text{Atrribute}) = \text{H}(\text{Class}) - \text{H}(\text{Class}|\text{Attribute})/\text{H}(\text{Attribute}$$

D. Feature Selection Implementation

CorrelationAttributeEval, Information Gain and Gain Ratio are the three different feature selection algorithms employed. The algorithms are adopted because they are computationally fast and scalable, unlike wrapper technique that is computationally expensive. The algorithms were applied each on the dataset deployed into WEKA, which comprises 26 features that were used in extracting the sentiment scores across the three categories (7500 web pages). The objective of this process is to improve the classification performance of the predictors. The results from each algorithm showed that none of the algorithms was able to select relevant features with a better classification accuracy until the features were best ranked on the first selected 25 features by the CorrelationAttributeEval algorithm. The experiment was successful in the feature selection process as the removal of a feature "program" led to the improvement and accuracy of the result. CorrelationAttributeEval was able to produce the best algorithm in terms of accuracy and error rate compared to other algorithms due to the way its function handles the nature of the dataset most importantly, the specific problem the method is to solve. Table 2 shows the result of CorrelationAttributeEval algorithm

Table 2-CorrelationAttributeEval Result

Correctly Classified Instances	7051	94 %
Incorrectly Classified Instances	449	5.9 %
Kappa statistic	0.9102	
Mean absolute error	0.0465	
Root mean squared error	0.1675	
Relative absolute error	10.4543 %	
Root relative squared error	35.5425 %	
Total Number of Instances	7500	

= Detailed Accuracy by Class ===

TP	FP	Precision	Recall	F-Measure	Class
0.988	0.046	0.915	0.988	0.950	Pro-Extremist
0.944	0.027	0.946	0.944	0.945	Anti-Extremist
0.888	0.017	0.963	0.888	0.924	Neutral
0.940	0.030	0.941	0.940	0.940	

```
=== Confusion Matrix ===

 a    b    c   <-- classified as
 2470   14   16 |   a = PRO-EXTREMIST
   69 2361   70 |   b = ANTI-EXTRMIST
 159  121 2220 |   c = NEUTRAL
```

E. Manual Class Verification

The WEKA standard J48 decision-tree classification output produced broad-brush results between the manual and automated classification in the 3 categories. In the pro-extremist class, the result obtained indicated that 2470 pro-extremist pages (98%) were correctly classified, while 14 and 16 pages were incorrectly classified as anti-extremist pages and neutral pages respectively. This was due to the sentiment the pages exhibited and calculated by our automated method. The situation in manual classification might be that any web pages harvested from extremist web domains were assigned to the pro-extremist class but in fact, not all the pages from such domains exhibit radicalization. Our automated method was able to identify also, the web pages whose contents exhibit neutral and opposition to violence in pro-extremist class.

From the experiment, the result shows that the pro-extremist class is the class with the most correctly identified pages while the neutral class is the class with most misclassified pages as pro-extremist pages in the manual classification. Table 3 shows a sample of pages that differ in their manual classification being identified by our automated system.

Table 3- A Sample of Neutral pages identified in Pro-extremist Class

```
Inst#       Actual                Predicted error prediction
   2266 1:PRO-EXTREMIST 1:PRO-EXTREMIST       0.452
   2267 1:PRO-EXTREMIST 1:PRO-EXTREMIST       0.452
   2268 1:PRO-EXTREMIST 1:PRO-EXTREMIST       0.452
   2269 1:PRO-EXTREMIST  3:NEUTRAL  +   0.75
   2270 1:PRO-EXTREMIST 1:PRO-EXTREMIST       0.452
   2271 1:PRO-EXTREMIST 1:PRO-EXTREMIST       0.75
```

F. Comparison with Existing Method

The results of the WEKA standard J48 decision-tree classification method employed in [1] showed that out of 7500 webpages processed, 80.51% were classified correctly, while the pro-extremist and the anti-extremist classes had the highest degree of correctly identified pages, with 92.7% and 88% respectively. The performance on the neutral class was lower, at 68%. However, the results produced in table 1 of this study and [1] when compared indicated an improvement in the performance of the sentiment-rule based method when feature selection techniques were used.

G. Model Evaluation

Our model is measured using precision and recall and F-measure in table 2. Precision reveals a number of true positive entities recognized by the classifier out of all entities identified as positive while recall shows the exactness that the algorithm returns most of relevant. However, our system indicated high recall and precision rate of 98% and 91% respectively in pro-extremist class while F-measure is the harmonic or balanced mean of the recall and precision. The model indicated 95% success rate in the F-measure, while the time taken to build the model is 1.05 seconds. However, measuring the quality of the rule for the class model, the pro-extremist class was classified at a higher rate of success than anti-extremist and neutral.

V. Conclusion

The sentiment analysis-based classification method detailed in this paper has proven to be an effective technique for the automatic classification of extremist web pages and identification of particular pages that differ in their manual class from the automatic class. The results from our method showed that overall web pages were classified at 93% success rate while pro-extremist pages were classified at a higher rate. In addition, we reduced the feature space of the dataset by using three different feature selection algorithms. We evaluated the algorithms to determine the best algorithm and most relevant features for classification accuracy. CorrelationAttributeEval produced the best algorithm by reducing the feature space of 26 features by one feature to give an overall better performance of 94% in classifying the overall Web.

We confirm the broad-brush approach to manual classification using our automated classification system. The result showed that the pro-extremist class is the class with most correctly classified pages compared to the other two classes while the neutral class has the most pages incorrectly classified as pro-extremist. However, the linguistic marker, that is, the top 10 keywords technique used to pinpoint sentiment across all web pages, might not be suitable to capture all the sentiments in larger web pages, thereby hindering the training process of the useful sentiment of some pages. In addition, the pages with no sentiments due to non-keyword presence may be misclassified into a class of highest probability due to the J48 generalization rule.

Future work will also focus on how the system could scale to handle a massive number of web pages. In addition, we intend to develop a hybrid approach that will merge the combination of both semantic and syntactic features, which is generated by a textual analysis technique such as Posit in building an automated classification system for extremist web pages. We intend to test the proposed method for its robustness and versatility.

VI. 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] Q. Schiermeier, "Terrorism: Terror prediction hits limits." Nature, vol. 517, no. 7535, p. 419, 2015.

[3] R. Feldman, "Techniques and applications for sentiment analysis" in Communications of the ACM, 56(4), (2013), pp. 82-88.

[4] T. A. Rana1 and Y. Cheah,"Aspect extraction in sentiment analysis: comparative analysis and survey," in Artif Intell Rev, Springer vol. 46 pp459–483. 2016. DOI 10.1007/s10462-016-9472-z

[5] A. Abbasi and H. Chen, "Applying Authorship Analysis to Extremist-Group Web Forum Messages," IEEE Intelligent Systems, 20(5), 2005, pp 67-75 [doi>10.1109/MIS.2005.81]

[6] A. S. Patil and B.V. Pawar,"Automated Classification of Web Sites using Naive Bayesian Algorithm in Proceeding international Multi Conference Engineers and Computer scientist 2012 Vol 1, IMEC 2012 March 2012, Hong Kong.

[7] E. Ferrara, W.-Q. Wang, O. Varol, A. Flam- mini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," In Social Informatics: 8th Intl. Conf., SocInfo 2016, Bellevue, WA, USA, 2016, pp.22–39.

[8] R. Feldman, "Techniques and applications for sentiment analysis" in Communications of the ACM, 56(4), (2013), pp. 82-88.

[9] R. Scrivens and R. Frank 2016, "Sentiment-Base Classification of Radical Texts on the Web," in Proceeding of the European Intelligence and Security Informatics Conference, 2016, pp 104–107.

[10] A. Abbasi and H. Chen, "Applying authorship analysis to extremist group Web forum messages," in Intelligent Systems, 20(5), (2005), pp. 67-75

[11] 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.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research. Vol.9, 2008, pp1871–1874.

[13] G. Fiol-Roig, M. Miró-Julià, and E. Herraiz," Data Mining Techniques for Web Page Classification," in Proceedings of the 9th Conference of Practical Application of Agents and Multiagents System, AISC 89, Berlin Heidelberg, 2011, pp. 61–68.

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

[15] 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.

[16] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," In Proceedings of the ACL-02 conference on empirical methods in natural language processing-volume 10. Association for computational linguistics, pp 79–86. 2002

[17] 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

[18] 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.

[19] K. Bafna and D. Toshniwal," Feature-based summarization of customers' reviews of online products," ProcComputSci, vol.22, 142–151.2013

[20] X. Meng, H. Wang, "Mining user reviews: from specification to summarization," In Proceedings of the ACL-IJCNLP 2009 conference short papers," Association for computational linguistics, pp 177–180. 2009

[21] J. R. Quinlan," C4.5 Programs for Machine Learning," in Morgan Kaufmann, San Mateo, 1993.

[22] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining", Boston: Kluwer Academic Publishers, 1998.

[23] Liu, H. and Yu, L. (2005).Toward Integrating Feature Selection Algorithms for Classification and Clustering, Department of Computer Science and Engineering, University of Arizona

[24] M. A. Hall, "Correlation-Based Feature Selection for Machine Learning," The University of Waikato, 1999

[25] D. Oreski and T. Novosel, "Comparison of Feature Selection Techniques in Knowledge Discovery Process," TEMJOURNAL, pp. 285-290, 2014.