SENTIMENT ANALYSIS OF WEB FORUMS: COMPARISON BETWEEN SENTIWORDNET AND SENTISTRENGTH

TAWUNRAT CHALOTHORN
Computing, Engineering & Information Sciences
University of Northumbria, Newcastle Upon Tyne
tawunrat.chalothorn@unn.ac.uk

JEREMY ELLMAN
Computing, Engineering & Information Sciences
University of Northumbria, Newcastle Upon Tyne
jeremy.ellman@unn.ac.uk

ABSTRACT
Internet has become a major tool for communication, training, fundraising, media operations, and recruitment, and these processes often use web forums. This paper intended to find suitable technique for analysing selected web forums that included radical content by presenting a comparison between SentiWordNet and SentiStrength. SentiWordNet is a lexical resource for supporting opinion mining by assigning a positivity score and a negativity score to each WordNet. SentiStrength is a technique that was developed from comments on MySpace. It uses human-designed lexical and emotional terms with a set of amplification, diminishing and negation rules. The results have been presented and discussed.

KEY WORDS
SentiWordNet, SentiStrength, sentiment, analysis, web forums, radical

1 INTRODUCTION

Web forums have become important places for social communication and discussion on the internet. Some radical groups also use them for communication and disseminating their ideologies to the public [1]. These kinds of forums can be referred to as part of the Dark Web. The Dark Web includes websites that are used by terrorists, radicals and extremist groups [2]. This paper presents the two system approach of two web forums in the area of sentiment and affects analysis. Their content is related to radicalization. The sections of this paper are structured as follows: section 2 provides some discussion on work related to sentiment analysis, SentiWordNet and SentiStrength. SentiWordNet is a lexical resource for supporting opinion mining by assigning a positivity score and a negativity score to each WordNet. SentiStrength is a sentiment analysis technique that was developed from comments on MySpace. It uses human-designed lexical and emotional terms [3]. Section 3 discusses the research question and this is followed by details of the data collection in section 4. System techniques were developed to assign and measure the effect of and sentiment found in the communication of web forums, as described in section 5. Finally, the analyses of the results are presented in section 6.

2 RELATED WORK

The term 'sentiment' was used by [4] and [5] in reference to the automatic analysis of evaluative text and tracking of predictive judgements, as well as analysing market sentiment [6]. Afterwards, the term 'opinion mining' was used at a WWW conference by [7]. They mentioned that the ideal opinion mining tools would press a set of search results for a given item, generating a list of product attributes and aggregation opinions about each of them [6]. Sentiment analysis has been used in many research fields, such as [8] who used sentiment analysis to analyse video comments and user profiles. [9] used the structure of lexical contextual sentences to classify sentiment from online customer reviews. Moreover, there are some researches that have used SentiWordNet and SentiStrength for classifying content, whether positive or negative. For instance, SentiWordNet was used for determining the polarity of reviews within the English and German languages by [10], and to classify movie reviews by [11]. [12] used SentiStrength to detect comments on MySpace. Also, SentiStrength was used for classifying emotions within reviews and analysing the content of Twitter by [3] and [13], respectively.

3 RESEARCH QUESTIONS

Web forums have become the main tool for communicating with others as they can be accessed anywhere. Sometimes they are used by a group of people who have radical ideologies for research, communication, training, fundraising, media operations, radicalisation and recruitment [14]. This paper presents our research on sentiment analysis and detection of radical content. In particular, this research attempts to answer the research question 'which technique of sentiment analysis can be used for classifying radical contents on web forums?'

4 DATA

Two forums have been selected for using in the research: Montada and Qawem. Both of them use the Arabic language. They have been selected by using research 21 people who are Arabic speaker by asking them that which websites they think that might have the contents related to radical Islamic ideologies. The results showed that Qawem and Montada are in the highest range.

5 METHODS

Data has been collected from the two web forums and classification of polarity has taken place using two techniques of sentiment analysis: SentiWordNet and SentiStrength. The model building phase when using SentiWordNet was started by splitting sentences into words and reducing the high-frequency text (stopwords) in sentences. Words were stored in a bag of words (BOW) and part of speech (POS) was used for tagging words and knowing the position of each word in the sentence. Lexicon, WordNet and SentiWordNet were used for assigning positive and negative scores for each synset in each word [8]. The formulas for calculating positive and negative scores were taken from [15]. The final scores of sentences were calculated using a formula taken from [9]. The scores of sentences were applied using the rule

that if the sentence had a positive score more than or equal to its negative score, then the sentence would be classified as positive. Otherwise, it would be negative. An objective (neutral) score was not used. The sum of positive, negative and objective was equal to 1.0. After that, the technique SentiStrength was applied for classifying the data on a scale from 1 to 5: 1 meant that there was no sentiment and 5 meant that there was a very strong positive or negative sentiment [12]. The overall results from SentiStrength were based on the formula shown in (1).

$$\left.\begin{array}{ll} if & positive > negative; \quad Positive\ Sentiment \\ if & positive < negative; \quad Negative\ Sentiment \\ if & positive = negative; \quad Neutrality\ Sentiment \end{array}\right\} \quad (1)$$

6 RESULTS

Model building of sentiment was applied to the web forums Montada and Qawem so as to analyse the results. After removing stopwords, the rest of the sentences were used for analysis using the technique SentiWordNet, while the full sentences were used for analyse using SentiStrength without removing any words. The results show that the Montada forum has less negative postings than the Qawem forum. In particular, the radical effect is quite strong in the communication found in the Qawem forum. Nearly 35% of the postings in Qawem were found to have a negative score between 0.050 and 0.100, while Montada had less than 15% of postings in the same score range when using SentiWordNet. When using SentiStrength it was found that nearly 50% of the postings in Qawem had a negative score at 2, while only 30% of the postings in Montada had the same score. On the other hand, using SentiWordNet it was found that the positive scores of postings in the Montada forum were higher than those in the Qawem forum, except in the range from 0.100 to 0.150. Using SentiStrength it was found that the positive scores of postings in Montada were higher than in Qawem in every range from 1 to 5. From the overall results it can be seen that both techniques seem to work well for classifying the content of web forums. However, there are some problems if checking the score of sentences one by one.

7 CONCLUSION

In this paper we have presented an analysis of two web forums, Montada and Qawem. They were chosen because their content relates to radicalization. The results of a comparison between SentiWordNet and SentiStrength were presented. The overall results of both techniques showed that Qawem had a higher percentage of postings with negative sentences than Montada. This said, both techniques could be used for classifying the content of web forums. However, when checking scores of each sentence, there were incorrect scores in some sentences. The reason might be that, when using SentiWordNet, stopwords were removed from sentences and some words with negative meanings did not have a strong negative score, such as traitor and kill. This might have affected the meaning and the score of the sentence. For example, “My avenger” gives a different meaning to the sentence than “Avenger”, removing the stopword

"My". "My avenger" would have had a higher negative score than "Avenger". Another sentence, "God cleans Syria from the traitors", should have had a negative score instead of a positive score. The sentence aims to encourage people to fight in Syria, which is obviously radical in nature. Therefore, it would be better if SentiWordNet were to score stopwords and it should review the scores of some words that are negative. On the other hand, some incorrect scores occur when using SentiStrength, such as "Shiites". "Shiites" should get a positive score instead of a negative score because the word refers to a group of people who believe in Islam and Ali [16,17 and 18]. This could be the reason why the sentence "God blesses Shiites everywhere" got a negative score instead of a positive score. Also, in the second sentence "armed" and "liberate" are negative in sentiment but SentiStrength showed that they were neutral. The reason for them getting a neutrality score might be that the words are not in their database. The methodology of SentiStrength was developed from comments on MySpace and such words may not have appeared in the comments. Therefore, there is a possibility of using SentiStrength as a model for developing another methodology for classifying and detecting the content of web forums, which will be part of our future work.

REFERENCES

[1] Jack Glaser; Jay Dixit; Donald P .Green. Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence? Journal of Social Issues 58 (1), 2002, PP: 177-193.

[2] Hsinchun Chen. Intelligence and Security Informatics For International Security: Information Sharing and Data Mining. Springer, 2006.

[3] David Garcia; Frank Schweitzer. Emotions in Product Reviews--Empirics and Models. Paper presented at the Privacy, security, risk and trust (passat), IEEE 3rd international conference on social computing (socialcom), 2011, PP: 483-488.

[4] Sanjiv Das; Mike Chen. Yahoo! for Amazon: Extracting market sentiment from stock message boards. Paper presented at the Proceedings of the Asia Pacific Finance Association Annual Conference APFA, 2001, PP: 37-56.

[5] Richard Tong. An operational system for detecting and tracking opinions in on-line. Paper presented at the In Proc of SIGR Workshop on Operational Text Classification, New Orleans, Louisianna, 2001, PP: 1-6.

[6] Bo Pang; Lillian Lee. Opinion Mining and Sentiment Analysis. Found Trends Inf Retr 2 (1-2), 2008, PP: 1-135.

[7] Kushal Dave; Steve Lawrence; David M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. Paper presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, 2003, PP: 519-528.

[8] Adam Bermingham; Maura Conway; Lisa McInerney; Neil O'Hare; Alan F. Smeaton. Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. Paper presented at the Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, 2009.

[9] Aurangzeb Khan; Baharum Baharudin. Sentence Level Semantic Orientation of Online Reviews and Blogs using SentiWordNet for Effective Sentiment Classification. International Journal of New Computer Architectures and their Applications (IJNCAA) 1 (2), 2011, PP: 627-643.

[10] Kerstin Denecke. Using SentiWordNet for multilingual sentiment analysis. Paper presented at the Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference, 2008, PP: 507-512.

[11] Bruno Ohana; Brendan Tierney. Sentiment classification of reviews using SentiWordNet. Paper presented at the 9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, 2009.
[12] Mike Thelwall; Kevan Buckley; Georgios Paltoglou. Sentiment in Twitter events. Journal of the American Society for Information Science and Technology 62 (2), 2011, PP: 406-418.
[13] Rene Pfitzner; Antonios Garas; Frank Schweitzer. Emotional Divergence Influences Information Spreading in Twitter. Paper presented at the The 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2012.
[14] Barbara Mantel. Terrorism and the Internet : should Web sites that promote terrorism be shut down? In. Washington, D.C. : CQ Press, 2009, PP: 129-155.
[15] Alena Neviarouskaya; Helmut Prendinger; Mitsuru Ishizuka. Textual Affect Sensing for Sociable and Expressive Online Communication. Paper presented at the Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 2007, PP: 218-229.
[16] George W. Braswell. What You Need to Know About Islam and Muslims. B&H Publishing Group, 2000.
[17] L.R. Reddy. Inside Afghanistan: End of the Taliban Era? APH Publishing, 2002.
[18] Heinz Halm. The Shiites: A Short History. Markus Wiener Publishers, 2007.