On Detecting Online Radicalization Using Natural Language Processing

Mourad Oussalah[1(✉)], F. Faroughian[2], and Panos Kostakos[1]

[1] Centre for Ubiquitous Computing, University of Oulu, Oulu, Finland
Mourad.Oussalah@oulu.fi
[2] Aston University, Aston, UK

Abstract. This paper suggests a new approach for radicalization detection using natural language processing techniques. Although, intuitively speaking, detection of radicalization from only language cues is not trivial and very debatable, the advances in computational linguistics together with the availability of large corpus that allows application of machine learning techniques opens us new horizons in the field. This paper advocates a two stage detection approach where in the first phase a radicalization score is obtained by analyzing mainly inherent characteristics of negative sentiment. In the second phase, a machine learning approach based on hybrid KNN-SVM and a variety of features, which include 1, 2 and 3-g, personality traits, emotions, as well as other linguistic and network related features were employed. The approach is validated using both Twitter and Tumblr dataset.

Keywords: Natural language processing · Radicalization · Machine learning

1 Introduction

The variety, easy access and popularity of social media user-friendly platforms have revolutionized the sharing of information and communications, facilitating an international web of virtual communities. Violent extremists and radical belief supporters have embraced this changing digital landscape with active presence in online discussion forums, creating numerous virtual communities that serve as basis for sympathizers and active users to discuss and promote their ideologies. As well as disseminating events and inspiration to gain new resources and the demonization of their enemies [2, 13, 16]. They have exploited the Internet's easy to use, quick, cheap and unregulated, relatively secure and anonymous platforms.

Within the extremist domain, online forums have also facilitated the 'leaderless resistance' movement, a decentralized and diffused tactic that has made it increasingly difficult for law enforcement officials to detect potentially violent extremists [5, 17].

It is becoming increasingly difficult – and near impossible – to manually search for violent extremists or users that may embarrass radicalization through Internet because of the overwhelming amount of information, and the inherent difficulty to distinguish self-curiosity, sympathizer and real doctrine supporter or genuine participation in violence acts.

Uncovering signs of extremism online has been one of the most significant policy issues faced by law enforcement agencies and security officials worldwide [6, 16], and the current focus of government-funded research has been on the development of advanced information technologies to identify and counter the threat of violent extremism on the Internet [13].

In light of these important contributions in digital extremism, an important question has been set aside: how can we uncover the digital indicators of 'extremist behavior' online, particularly for the 'most extreme individuals' based on their online activity? To some extent, criminologists have begun to explore this critical point of departure via a customized web-crawler, extracting large bodies of text from websites featuring extremist material and then using text-based analysis tools to assess the content [3, 9, 11]. Similarly, some computational-based research has been conducted on extremist content on Islamic-based discussion forums [1, 6]. Salem et al. [14] proposed a multimedia and content-based analysis approach to detect Jihadi extremist videos and the characteristics to identify the message given in the video. Wang et al. [15] presented a graph-based semi-supervised learning technique to classify intent tweets. They combined keyword based tagging (referred as an intent keyword) and graph regularization method for classifying tweets into six categories. Both Brynielsson et al. [4] and Cohen et al. [6] hypothesized a number of ways to detect online traces of lone wolf terrorists, although, no practical platform has been demonstrated and evaluated. Davidson et al. [7] annotated some 24,000 Tweets for 'hate speech', 'offensive language but not hate', and 'neither'. They began with filtering Tweets using a hate speech lexicon from Hatebase.org, and selected a random sample for annotation. The authors pointed out that distinguishing hate speech from nonhate offensive language was a challenging task, as hate speech does not always contain offensive words while offensive language does not always express hate. O'Callaghan et. al. [12] described an approach to identify extreme right communities on multiple social networking websites using Twitter as a possible gateway to locate these communities in a wider network and track its dynamic. They performed a case study using two different datasets to investigate English and German language communities and implemented a heterogeneous network employing Twitter accounts, Facebook profiles and YouTube channels, hypothesizing that extreme right entities can be mapped by investigating possible interactions among these accounts.

In this paper, we propose a multi-facet based approach for identifying hate speech and extremism from both Twitter and Tumblr dataset. Building on previous research (e.g., see [8, 10]), we use various n-gram based features such as the presence of religious words, war-related terms and several hashtags that are commonly used in extremist posts. Furthermore, other high-level linguistic cues like sentiment, personality change, emotion and emoticons as well as network related features are employed in order to grasp the rich and complexity of hate/extremism like text.

2 Method

Our general approach for radicalization identification undergoes a two-step strategy. First, a radicalization score is obtained by exploring mainly the characteristics of negative sentiments. Second, a machine learning strategy is explored to separate radical post from non-radical one using a wider and diverse set of features involving both linguistic and network features together with previously estimated radicalization score.

2.1 Radicalization Score

Similarly to alternative works in [7], we first explore the sentiment analysis of user's posts. The rationale behind this reasoning is to hypothesize that an extremist user is characterized by the dominance of negative materials over a certain period of time, suggesting that such user espouses an extremist view. Typically sentiment score enables quantifying such trend. Indeed, sentiment analysis is a well-known data collection and analysis method that allows for the application of subjective labels and classifications, by assigning an individual' sentiment with a negative, positive or neutral polarity value. We employed the established Java-based software SentiStrength, which allows for a keyword-focused method of determining sentiment near a specified keyword.

In line with Scrivens et al. [19], the radical score accounts for:

- Average sentiment score percentile (AS), it is calculated by accounting for the average sentiment score for all posts in a given forum. The scores for each individual were converted into percentiles scores, and percentile scores were divided by 10 to obtain a score out of 10 points.
- Volume of negative posts (VN). This is calculated in two parts: (1) the number of negative posts for a given member, and (2) the proportion of posts for a given member that were negative. To calculate the number of negative posts for a given member, we counted the number of negative posts for a given member and converted these scores into percentiles scores.
- Severity of negative posts (SN). This is calculated in two steps: (1) the number of very negative posts for a given member and (2) the proportion of posts for a given member that were very negative. 'Very negative' was calculated by standardizing the count variable; all posts with a standardized value greater than three were considered to be 'very' negative.
- Duration of negative posts (DN). An author who posted extreme messages over an extensive period of time should be classified as more extreme than an author who posted equally extreme messages over a shorter period of time. It is calculated by determining the first and last dates on which individual members made negative posts.

The radical score is therefore calculated as an aggregation of the four previous elements, see Fig. 1. Unlike Scrivens et al. [19] where a simple arithmetic operation was employed, we advocate a non-linear combination of these attributes:

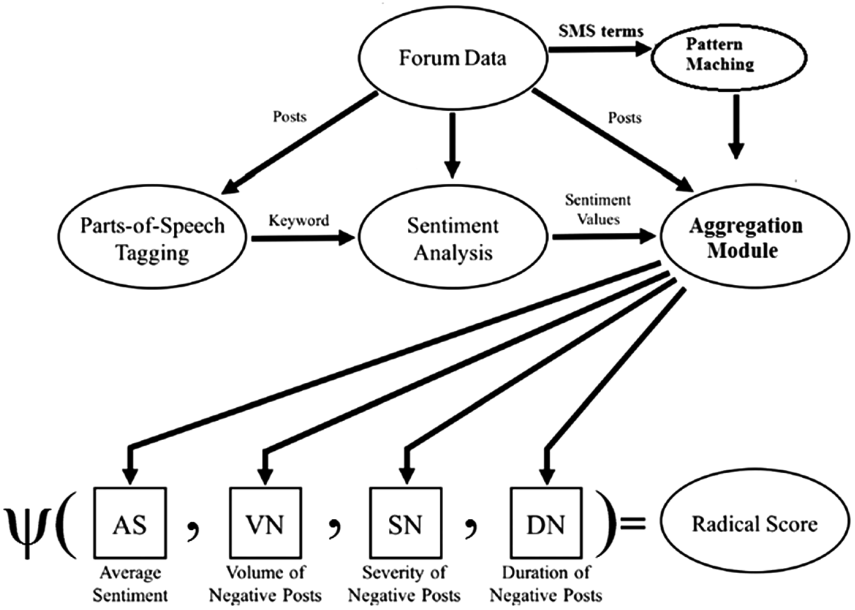

Fig. 1. Extremism radicalization/extremism estimation

$$Radical\,Score = \Psi(AS, VN, SN, DN) \tag{1}$$

Especially, Radical Score is intuitively increasing with respect to VN, SN and DN. Therefore, we argue that the combination operator Ψ linear but rather multiplicative where AS plays only a normalization like role yielding (for some constant factor K):

$$Radical\,Score = \left(K/AS^3\right)(VN.SN.DN) \tag{2}$$

2.2 Machine Learning Based Classification

The second phase in the radicalization identification is to use a one-class classification framework involving advanced machine learning techniques where SVM, K-NN, Random Forest were implemented. More specifically, the approach uses the following:

- Extensive preprocessing stage is employed at the beginning in order to filter out stop list words, unknown characters, links, and
- A hybrid SVM-KNN in the same spirit as Zhou et al. [18] is adopted.
- Three types of features are considered. The first one is related to the use of N-gram, especially, 1-g, 2-g and 3-g features were employed as primarily input to the classifier.
- The second type of features relate to personality traits (using the five personality model), emotion and writing cues. The implementation of personality trait identification is performed using the MRC Psycholinguistic database, Linguistic

Inquiry and Word Count (LIWC) feature and Random Forest classifier. Emotion recognition is performed using WordNet Affect, an extension of WordNet domains that concerns a subset of synsets suitable to represent affective concepts correlated with affective words together with Bayes classifier. Finally, writing cues were only considered from its basic content with respect to psychological process as quantified using LWIC features.

– The third set of features are related to various semantic and network related measures. This includes, the length of the post, emoticons, personal pronouns, interrogation and exclamation marks, offensive words, swear words, war words, religious words. We use both LIWC categorization as well as wordnet taxonomy in order to identify war and religious related words. Next, social network related features concern mainly the frequency of messages of the user, average number of posts by the user as well as centrality value whenever possible. Furthermore, the radicalization score computed in previous step is also employed as part of input to the two-class classifier (presence or absence of radicalization case) based on hybrid KNN-SVM.
– We utilize some of existing corpus gathered from DarkWeb project and repository (https://data.mendeley.com/datasets/hd3b6v659v/2) in order to enhance the training of the hybrid KNN-SVM classifier. The overall architecture of this classification scheme is highlighted in Fig. 2.

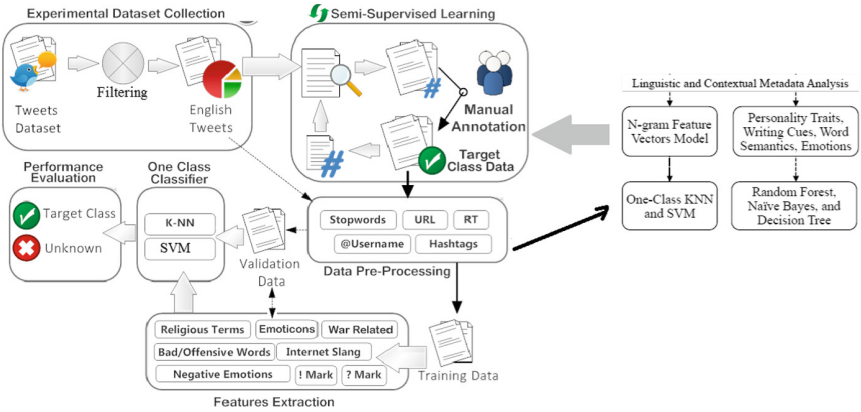

Fig. 2. One-class classification approach

3 Method

3.1 Dataset

Two types of dataset were employed: Twitter dataset and Tumblr dataset.

The initial attempt to collect related tweets is to crawl the hashtags that contains terms "#islamophobia", "#bombing", "#terrorist", "#extremist", "#radicalist". For each set of identified hashtags, Twitter Search API was employed to collect up to one hundred tweets per identified hashtag. A total of 12,202 tweets were collected. Eight thousands of these tweets were sent to Amazon Mechanical Turk in order to perform manual labelling. For each distinct user with a set of tweets, three independent annotators were employed to test whether the user is classified radical or not.

Similarly, we use close hashtags in order to collect data from Tumblr, especially, we employed keywords #islamophobia, #islam is evil, #supremacy, #blacklivesmatter, #white racism, #jihad, #isis and #white genocide. A total of 8000 posts were collected. We deliberately attempt to choose scenarios where a user is associated with several posts in order to provide tangible framework for application of our methodology. Likewise Twitter-dataset, close to 6000 of these posts are sent to Amazon Mechanical Turk in order to manually annotate the post whether the underlying user is considered radical/extremist or not. The results of this analysis are summarized in Table 1, which highlight the usefulness of the approach and its capabilities.

Table 1. Twitter and Tumblr classification scores for various feature sets

Dataset	Features	Classification score
Twitter	1, 2, 3-g	46%
	1, 2, 3-g + personality + emotion	57%
	All features + radicalization score	68%
Tumblr	1, 2, 3-g	54%
Tumblr	1, 2, 3-g + personality + emotion	63%
Tumblr	All features + radicalization score	72%

References

1. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. Intell. Syst. **20**(5), 67–75 (2005)
2. Bowman-Grieve, L.: Exploring "stormfront:" a virtual community of the radical right. Stud. Confl. Terror. **32**(11), 989–1007 (2009)
3. Bouchard, M., Joffres, K., Frank, R.: Preliminary analytical considerations in designing a terrorism and extremism online network extractor. In: Mago, V., Dabbaghian, V. (eds.) Computational Models of Complex Systems, pp. 171–184. Springer, New York (2014). https://doi.org/10.1007/978-3-319-01285-8_11
4. Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Martenson, C., Svenson, P.: Analysis of weak signals for detecting lone wolf terrorist. In: Proceedings of the European Intelligence and Security Informatics Conference, Odense, Denmark, pp. 197–204 (2012)

5. Chen, H.: Dark Web: Exploring and Data Mining the Dark Side of the Web. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-1557-2
6. Cohen, K., Johansson, F., Kaati, L., Mork, J.: Detecting linguistic markers for radical violence in social media. Terror. Polit. Violence **26**(1), 246–256 (2014)
7. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of ICWSM (2017)
8. Foong, J.J., Oussalah, M.: Cyberbullying system detection and analysis. In: European Conference in Intelligence Security Informatics, Athens (2017)
9. Frank, R., Bouchard, M., Davies, G., Mei, J.: Spreading the message digitally: a look into extremist content on the internet. In: Smith, R.G., Cheung, R.C.-C., Lau, L.Y.-C. (eds.) Cybercrime Risks and Responses: Eastern and Western Perspectives, pp. 130–145. Palgrave Macmillian, London (2015). https://doi.org/10.1057/9781137474162_9
10. Kostakos, P., Oussalah, M.: Meta-terrorism: identifying linguistic patterns in public discourse after an attack. In: SNAST 2018 Web Conference (2018)
11. Mei, J., Frank, R.: Sentiment crawling: extremist content collection through a sentiment analysis guided web-crawler. In: Proceedings of the International Symposium on Foundations of Open Source Intelligence and Security Informatics, Paris, France, pp. 1024–1027 (2015)
12. O'Callaghan, D., et al.: Uncovering the wider structure of extreme right communities spanning popular online networks (2013). https://arxiv.org/pdf/1302.1726.pdf
13. Sageman, M.: Leaderless jihad: Terror networks in the twenty-first century. University of Pennsylvania Press, Philadelphia (2008)
14. Salem, A., Reid, E., Chen, H.: Multimedia content coding and analyzis: unraveling the content of Jihadi extremist groups's video. Stud. Confl. Terror. **31**(7), 605–626 (2008)
15. Wang, J., Cong, G., Zhao, W.X., Li, X.: Mining user intents in Twitter: a semi-supervised approach to inferring intent categories for tweets. In: AAAI (2015)
16. Weimann, G.: The psychology of mass-mediated terrorism. Am. Behav. Sci. **52**(1), 69–86 (2008)
17. Weimann, G.: Terror on Facebook, Twitter, and YouTube. Brown J. World Affairs **16**(2), 45–54 (2010)
18. Zhou, H., Wang, J., Wu, J., Zhang, L., Lei, P., Chen, X.: Application of the hybrid SVM-KNN model for credit scoring. In: 2013 Ninth International Conference on Computational Intelligence and Security (2013). https://doi.org/10.1109/cis.2013.43
19. Scrivens, R., Davies, G., Frank, R.: Searching for signs of extremism on the web: an introduction to sentiment-based identification of radical authors. Behav. Sci. Terror. Polit. Aggress. **10**(1), 39–59 (2017)