Identifying Radical Social Media Posts using Machine Learning

Preprint · June 2017

DOI: 10.13140/RG.2.2.15311.53926

3 authors, including:

Prabhakar Gupta
Amazon
9 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Pulkit Varshney
Stony Brook University
1 PUBLICATION **4** CITATIONS
SEE PROFILE

Identifying Radical Social Media Posts using Machine Learning

Prabhakar Gupta[1], Pulkit Varshney[1], M. P. S. Bhatia[2]
1 Division of Information Technology, Netaji Subhas Institute of Technology, Delhi, India
prabhakar.gupta@nsitonline.in, pulkitvarshney186@gmail.com
2 Division of Computer Engineering, Netaji Subhas Institute of Technology, Delhi, India
bhatia.mps@gmail.com

Abstract—**Radicalization (like cyberterrorism) is one of the major concerns to all governments and law enforcement agencies to provide safety and security to their citizens. A lot of radical groups, extremists and insurgent organizations use social media platforms such as Facebook, Twitter, Reddit, YouTube etc. to post their ideology and propagate their message to their followers. Manual detection of these posts is nearly an impossible task. We propose an automated system for extracting data from Twitter employing investigative data mining technique using the hashtags used in the posts. The system preprocesses the data to clean it by tokenizing, stemming and lemmatization. Data is classified as radical or non-radical using supervised machine learning classification techniques (Naive Bayes, SVM, AdaBoost and Random Forest) with varying parameters. The idea is to classify posts by identifying the linguistic structure, their stylometry and detecting a time based pattern.**

Keywords—Radicalization; Cyberterrorism; Machine Learning; Data Mining; Short-text classification

I. Introduction

Since the late 1980s, the World Wide Web has become a highly powerful means of communication worldwide, reaching an ever-growing audience. According to the FBI, international radicalization has following characteristics: (1) it involves violent acts harming other human life that can violate government laws (2) it is intended to intimidate civilian masses (3) it affects the policy of any government by coercion (4) it influences the conduct of a government by kidnapping, mass destruction, or assassination. Radical groups have started using the Internet to disseminate information that aid their causes. The availability of terrorist related material on the Internet plays an important role in radicalization processes. Due to this increasing availability of content on social media websites (such as Twitter, Facebook and Reddit etc.) there is an urgent need to identify these radical tweets.

In the pro-radical networks, large amount of information is carried and commanded by a limited number of individuals [1]. Most of the radical groups have shifted their focus from mainstream media to digital and social media to broadcast the information to new or younger groups or individuals, who can get the information by searching through hashtags and further by following them [2].

Research Motivation Twitter is the most commonly used micro-blogging website. Due to the low dissemination barrier on Twitter, anonymity [3]; anybody can follow anyone (unless the account is a private account). It has also become one of the biggest platforms of cyberterrorism holding a lot of extremist users expressing their views through their posts (tweets). More than a million tweets are being posted on Twitter everyday, so analysis of each and every tweet cannot be done manually. Also, there is a maximum 140 character limit (including text and hashtags) and noise i.e. online internet slang, abbreviations, incorrect grammar and spelling errors makes the automatic classification of these posts quite challenging. Moreover, Twitter Search REST API [4] only provides the data for one week. The aim is to automatically identify if a tweet is radical or not regardless of all the limitations.

We broadly classify our research work and experiment into three major steps:

A. Data extraction

Extraction of tweets is done from Twitter. Attributes associated to a single tweet are URLs, text, user mentions, hashtags, media files (image, audio and video), timestamp, number of retweets, likes etc. and the user information. The main focus was on text mining as storing media files would, firstly, need a lot of space and secondly, analysing media files for radical content would require additional machine learning techniques. The texts were collected without knowing the sentiment. For example, when collecting tweets on hashtag #Syria (which is in the list of top 10 most frequent hashtags), it is not known initially whether:

- The tweet is posted by a person associated with terrorist organisation or not;
- Someone abominates the very idea of radicalization, showing disbelief and expressing against a pro-extremist person;
- Someone is discussing something general related to Syria, not in context with terrorism.

B. Preprocessing of data

Data preprocessing is a necessary step after data extraction

in order to make the data unique and non-redundant. As tweets which have been retweeted by other users occurred multiple times in the dataset, thus duplicates were removed using the unique tweet ID from the Twitter API response. After that, the text was strictly restricted to English language. But the problem arose that a lot of the tweets extracted were in other languages (like Arabic). Many of the Twitter accounts had their primary language as Arabic (or languages other than English) and even if they were radical, could not be considered. As the results are preliminary, so to write a unicode to work on tweets in different languages is an extensive task and out of the scope of this experiment. Then, to clean the tweets' texts, URLs were removed for the smooth processing of data and to reduce overhead. Stopwords were removed and tokenization was done.

C. Classification of tweets using machine learning techniques

Supervised machine learning classification is generally employed for processing a large quantity of data which cannot be done manually. Various algorithms can be implemented on these elucidations to make the classification of tweets as radical or not. A set of training data containing tweets which were 100% radical and tweets which were 100% non-radical. A feature vector for each tweet was generated on the basis of a unique feature-set. The feature vector contained the value as 1 or TRUE if that particular feature was present in the tweet, and 0 or FALSE if feature was absent. Using this training data, a system is made to learn using classification algorithms: (1) SVM (Support Vector Machine), (2) Naive Bayes, (3) Adaboost Classification and (4) Random Forest Classification using scikit-learn (Python library for machine learning) [5].

II. Related work

Machine learning is the most common approach for classification, regression and clustering. Classification involves identification among various categories a particular object belongs to. Regression predicts a continuous-valued attribute which is associated with an object. Clustering means grouping of objects which exhibit nearly similar properties.

A lot of research in clustering was done in [6], a topic entity relationship graph was made. The most discussed topics were the central nodes. All the users who were talking about it, or have tweeted earlier were linked to that central hub in the form of a star topology using k-means clustering. A burst rate was also taken into consideration i.e. if an account was linked to that topic in the past, now not posting anything or not in context with the topic was not considered. In [1] [7], a list consisting of 66 Twitter accounts which were identified as radical but were yet blocked by Twitter and the tweets from these accounts were taken as the training data. The system was made to learn accordingly and accuracies of each classifier (SVM, Naive Bayes and AdaBoost) were calculated. The problem with the list of Twitter accounts was, firstly, majority of tweets were in Arabic language and secondly, most of the accounts have been suspended by Twitter leaving not a single account good for this experiment's consideration making it unfit for using it as training dataset.

Many researchers are solicitous for depicting the phenomena of this social problem revolving around extremist propaganda.They use online data as a proxy to study the behavior of individuals and groups. In a 2016 study [8], the authors carried out three forecasting tasks: (1) Detection of extremist accounts, (2) To reckon normal user to adopt extremist content, (3) Predicting whether a normal user will retaliate contacts generated by the extremist account. Various machine learning tools are used to create a framework that generate features of multiple dimensions including network statistics, user metadata and temporal pattern of activity. Two scenarios were taken into account for this forecasting process: a time independent, post hoc prediction task on collected data, and a real-time simulated prediction task. Further concluded by determining the emanating signals that provide a thorough feature analysis by prediction in different scenarios.

A few researches worked on alternative data sources. Twitter data was used as a source archive [9], some studies were based on Arabic tweets and classified these as pro-ISIS and anti-ISIS. Provided with the Arab Twitter users (called tweeps in their research), the account history of users were known, a model was developed to examine the interest of both groups (pro-ISIS and anti-ISIS) before and after they started opposing or supporting ISIS. Trends of tweets or any literature were analysed to find out the motivations that made people to follow ISIS online (not considering whether they were willing to join ISIS or not), identifying ascertained injustices known as triggers. Their analysis was divided into two parts: (1) After collection of data, global trends were determined. This provided an insight into external affairs such as 'videos of beheadings' or 'sites of terrorists camps' (2) Individual Historic Analysis for the patterns of user history before supporting ISIS. Limitations for [9], were: their data was biased to less openly hateful and less offensive users. Something less vivid than suspension of accounts that users themselves can delete their tweets. If an ISIS supporter tweets about Coldplay, now he can go back in time and delete his tweet. But expecting this kind of behaviour was limited. Another restriction was that Twitter only provides with 3200 tweets for a user, so individual historical analysis would not provide good results as complete history of a user is not available.

Along the same trend, interesting examples proposed by various researchers [10] - [13] on different machine learning strategies aimed at detecting hate promotion, extremist support and cyber recruitment on various social media platforms like Tumblr, Twitter and YouTube [10]. Their training data was obtained by semi-supervised learning methods and their framework mainly constituted features of content and metadata. Their research was divided into many stages primarily consists of six steps: data extraction, creation of training data, pre-processing of data, feature set creation and

extraction, data classification and evaluating performance. In first stage, all the tweets that were in English language were extracted and combined to form a single unique dataset. Second stage consisted of the creation of training data based on hashtags, as hashtags are the best indicator of the sentiment of tweets. Tweets were labelled manually and recursively extended to find new hashtags based on some seed hashtags. Stage three involves data pre-processing to remove the hashtags and @username. Stage four includes extraction of different features on the basis of data extracted. In stage five, on the basis of two independent classifiers, tweets were classified as supporting ISIS or not. To complete the process, the last stage involves creation of confusion matrix to judge the accuracy of above classifiers. The problem with this approach was that only the hashtags were considered for the classification parameters without considering the text or any other linguistic structure of the tweets since there are a lot of tweets without any hashtags and there are many tweets with more hashtags than just one.

III. Approach for Identification

Twitter Search REST API [4] was used for extracting the public tweets for the hashtags which are associated with the radical groups. A few seed hashtags (#ISIS, #IslamicState) were selected manually and all the tweets for those hashtags were extracted. The API only provides the data for 7 days so the process of extraction was repeated for 4 weeks, giving the data for roughly 1 month (mid February 2017 to mid March 2017). For the tweets extracted, the frequencies of all the hashtags were calculated and recursively the most frequent and unique hashtag was used as the new search query. This process was repeated for 18 more hashtags (apart from seed hashtags) giving around 57,698 tweets of which 48,644 tweets were unique. The tweets which didn't have primary language as English were not considered regardless of their content. This dataset is called TT-FEATURE [19]. After extracting tweets, the most popular hashtags are shown in Table - I.

This data was used to extract the linguistic features and most popular hashtags. For extraction of these features, a tool called NLTK (Natural Language ToolKit) [14] was used. It is an open-source suite of various libraries for statistical Natural Language Processing (NLP) for English. It is written in Python language.

The texts of these tweets were cleaned by removing the URLs, user mentions (@), hashtags (#) and the term 'RT' (Retweet). The clean text was, then, tokenized using nltk.tokenize package in the NLTK library giving us the POS (Part-of-Speech) tags for every word. Only the words with noun POS tags (NN, NNS, NNP, NNPS) [15] were considered. Many of the words had same meaning but occurred in different form like jihad, jihadi, jihadist and jihadology. The derivationally related forms and inflectional forms of a word were reduced to a common base word ("jihad" in this example). This technique is called stemming.

TABLE I. Most Popular Hashtags

Hashtag	Occurrence in Tweets (Out of 48,644)	Occurrence Percentage
#ISIS	16535	33.992%
#IslamicState	14890	30.610%
#Taliban	7373	15.157%
#IS	7217	14.836%
#Wahhabism	6934	14.254%
#AlQaeda	6738	13.852%

For all the noun words from tweets were stemmed and lemmatized using nltk.stem.wordnet (WordNet stemmer) package of NLTK library. The frequency for the base word was calculated in a Python dictionary. For the abbreviations like "Islamic State" and "IS" both represent the same thing but were considered as separate entities during the experiment. To consider only English noun words, a spell checker tool was used called PyEnchant. PyEnchant [16] is a Python library for spell checking. With correct parameters, it helps in identifying if a word is an English word or not.

Test data for experiment was extracted using a few of the most popular hashtags (#IslamicState, #ISIS, #AlQaeda, #Wahhabism, #Taliban and #Daesh) as seed hashtags. All the tweets with at least one of the seed hashtags were extracted using the Twitter API for one week. 10,282 unique tweets were extracted (TT-TRAIN-PRO [19]). A random hashtag (which was safely assumed to be not about any radical discussion) was taken (#IPL). 15,200 unique tweets were extracted from this (TT-TRAIN-CON [19]). The dataset of these 25,482 tweets is called TT-TRAIN [19].

The tweets in TT-TRAIN dataset which were extracted from the seed hashtags were random in nature as they could be talking about these radical groups in a supportive way or opposing them. In order to classify them as radical or not-radical, many of the tweets were manually classified and the remainder of the tweets were classified on the basis of the occurrence of other hashtags. The tweets extracted from the random hashtags were all safely considered to be not radical in nature. The tweets were then cleaned in a similar way as the tweets in TT-FEATURE. Each of them was converted to a feature vector of size 613 in length having boolean values. They were assigned a value corresponding to their manual or assumed classification as TRUE for radical tweets and FALSE for non-radical tweets.

IV. Features

In the experiment, a total of 613 features were considered which belonged to 2 different types of feature classes:

- stylometric features (SF)
- time pattern features (TF)

A. Stylometric Features (SF)

There were a total of 582 stylometric features (Table - II). We used the frequency of all English noun words occurring in the tweets in TT-FEATURE dataset after cleaning the tweets and preprocessing them.

TABLE II. STYLOMETRIC FEATURES (SF)

frequent words	occurrence of most frequent noun words	443
hashtags	occurrence of most frequent hashtags	139

The most frequent 443 words were used as the features. The 10 most frequent words from these 443 were, "isis", "monitor", "imam", "thanks", "world", "woman", "group", "saudi",, "attack", "mischief", "fighter", "car". Among the most frequent words used, we can notice the words that are related to radical activities ("isis", "attack"), words that can or cannot relate to radical activities ("world", "woman", "car") and also the word imam. The word imam refers to the person who leads prayers in a mosque. It is a word originated from Arabic language.

The 139 most frequent hashtags were used for the experiment. The 10 most frequent hashtags were, #ISIS, #IslamicState, #Taliban, #IS, #Wahhabism, #AlQaeda, #SaudiArabia, #ISIL, #Syria, #Daesh. It can be observed that all of the most frequent hashtags are related to radical groups and their activities. #ISIS, #IslamicState, #IS, #ISIL, #Daesh all signify the Sunni militant group. #AlQaeda and #Taliban also are militant groups, whereas #SaudiArabia and #Syria are the countries which are affected by these groups.

B. Time Pattern Features (TF)

The date and time at which the tweet was posted for recognizing a pattern was used. The day of the week and the hour of day were used since other attributes like month were not relevant in our experiment since the data from 4 weeks. The considered attributes are:

- Hour of day (Hour 0, Hour 1, Hour 2, …, Hour 23)
- Day of week (Monday, Tuesday, …, Sunday)

In total 31 time pattern features were used, 24 for hours and 7 for days.

V. EXPERIMENTAL RESULTS

A tool called scikit-learn was used for various classifications. It is an open source Python library for machine learning built using NumPy, SciPy, and matplotlib libraries. A dataset with 5,297 tweets was extracted using the same extraction method as TT-FEATURE dataset using a seed hashtag (#GameOfThrones). This dataset was called TT-TEST [19] and the ratio between training data and testing was approximately 1:5. Different classification algorithms used were: Naive Bayes, SVM (Support Vector Machines), AdaBoost and Random Forest Classification. SVM was tested over different values of its "kernel" parameter to get the optimum results for various kernels.

TABLE III. RESULTS USING ALL FEATURES ON TT-TEST

Classifier		Correctly Classified	Incorrectly Classified	Accuracy
Naive Bayes		4933	364	93.1282
Random Forest Classifier		5214	83	98.4331
AdaBoost Classifier		**5245**	**52**	**99.0183**
SVM	Linear	5161	136	97.4325
	RBF	5199	98	98.1499
	Sigmoidal	5193	104	98.0366

It can be seen that AdaBoost Classifier works slightly better than Random Forest Classifier. Naive Bayes provides the worst results out of all the classifiers for this experiment. For SVM, choosing the correct kernel improves the results. Radial basis function (Gaussian) (RBF) kernel performs better than Sigmoidal and Linear kernels.

TT-TRAIN-PRO dataset was analysed for understanding the time based pattern of when people generally tend to post pro-radical tweets. It was observed that the maximum radical tweets were tweeted between 1:00 AM GMT and 2:00 AM GMT while the minimum radical tweets were tweeted between 10:00 PM GMT and 11:00 PM GMT. The results are visualised in graphical format in Figure I.

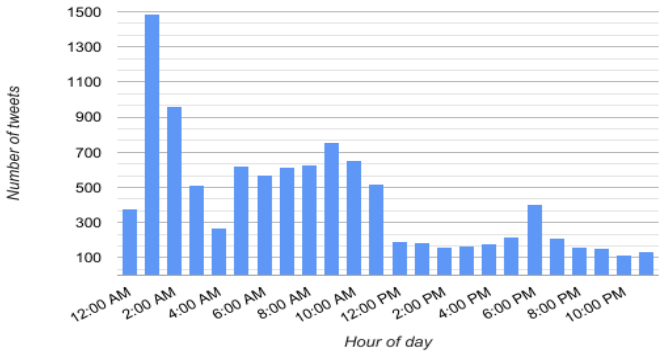

Fig. 1. Tweets tweeted at various hours of day

VI. CONCLUSION AND FUTURE WORK

In this work, supervised machine learning was used to identify radical social media posts. The major problem with this classification was manually labelling the test data as radical or not radical. The dataset extracted was dependent on the seed hashtags which were selected initially. In future more independence to dataset should be there. The texts, images and videos issued by various radical and extremists groups which are collected by intelligence agencies can also be used to

improve the performance of the system. Moreover, this classification should be seen as a support to the manual checking of the tweets since the accuracy of no classifier is absolute 100% and there is a chance of wrong tweets being identified as radical. This is due to the dynamic nature of the tweets and the noise in tweets due to abbreviations, Internet slang and 140 character limit.

The untagged datasets used in the experiment (TT-FEATURE, TT-TRAIN and TT-TEST) have been made public for future use which can be found on repository hosting website, GitHub [19]. The data is directly downloaded from Twitter using the Twitter API and is currently in JSON format.

The obtained results are from a limited dataset, that too only from one social media website, Twitter. In future the experiment can be extended to a diverse and large dataset from multiple websites (like Facebook, YouTube, Reddit, etc.) and platforms. The URLs in the tweets which were ignored here can also be analysed to identify the nature of tweets. Also, considering non-English languages tweets (like Arabic) can increase the reach of the classifier providing better results.

VII. References

[1] Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer,Nico Prucha, "Detecting Jihadist Messages on Twitter", Intelligence and Security Informatics Conference (EISIC), 2015 European

[2] "Dataset spotlight: How ISIS uses Twitter - interview with Khuram Zaman"(http://www.voxpol.eu/dataset-spotlight-isis-uses-twitter-interview-khuram-zaman), 2016

[3] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, Sanda M. Harabagiu D. Correa and A. Sureka, "EmpaTweet: Annotating and Detecting Emotions on Twitter", LREC 2012

[4] Twitter, Inc.. Twitter Developer Documentation (https://dev.twitter.com/rest/public/search), 2017

[5] Google, scikit-learn - Machine Learning in Python, (http://scikit-learn.org), 2017

[6] Pooja Wadhwa and Dr M.P.S Bhatia, "Tracking on-line radicalization using investigative data mining", Communications (NCC), 2013 National Conference

[7] A.Fisher and N.Prucha, "The call-up: The roots of a resilient and persistent Jihadist presence on Twitter", CTX Vol.4 No.3, 2014

[8] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini and Aram Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity", arXiv:1605.00659v1, [cs.SI] 2 May 2016

[9] Walid Magdy, Kareem Darwish, and Ingmar Weber, "#FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support", arXiv:1503.02401v1 [cs.SI] 9 Mar 2015

[10] Swati Agarwal, Ashish Sureka, "Learning to Classify Hate and Extremism Promoting Tweets", 2014 IEEE Joint Intelligence and Security Informatics Conference

[11] Denzil Correa and Ashish Sureka, "Solutions to Detect and Analyze Online Radicalization : A Survey," arXiv preprint arXiv:1301.4916, 2013

[12] S. Agarwal and A. Sureka, "A focused crawler for mining hate and extremism promoting videos on YouTube.", Proceedings of the 25th ACM conference on Hypertext and social media, 2014, pp. 294–296

[13] A. Sureka and S. Agarwal, "Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter," ,Distributed Computing and Internet Technology. Springer, 2015, pp. 431–442.

[14] NLTK Project, Natural Language Toolkit, (http://www.nltk.org/), 2017

[15] Penn Treebank P.O.S. (Part-of-speech) Tags, (https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

[16] Ryan Kelly, PyEnchant - A Spell checking Library for Python, (https://pythonhosted.org/pyenchant/), 2014

[17] Adam Badawy and Emilio Ferrara, "The Rise of Jihadist Propaganda on Social Networks", ArXiv 2017

[18] J.M. Berger and Jonathon Morgan, "The ISIS Twitter Census Defining and describing the population of ISIS supporters on Twitter", The Brookings Project on U.S. Relations with the Islamic World Analysis Paper | No. 20, March 2015

[19] Radicalization Twitter Dataset, (https://git.io/vHTUP), 2017