Identifying potential criminals and terrorists from social media analysis (SNA)

Ranadip Chatterjee
University of Calgary
ranadip.chatterjee@ucalgary.ca

Dora Tan
University of Calgary
dora.tan@ucalgary.ca

Karmvir Singh Dhaliwal
University of Calgary
karmvir.dhaliwal@ucalgary.ca

ABSTRACT

With the increasing number of online social media platforms, the world has dwindled in terms of communication. As they are pretty much free, unsophisticated and widely adopted, these platforms and tools have given extremists a virtual gateway to communicate, collaborate and convince readers with their propaganda globally. This is especially dangerous given the ease of how anyone can be radicalized consuming endless amounts of content, often subconsciously.

The digital footprints left behind however can be exploited to identify spreaders of terror. Our research aims at studying, extending and applying various terrorism-related content analysis

frameworks with the primary focus of classifying and eventually determining potential high-risk individual (accounts).

General Terms

Algorithms, Measurement, Documentation, Performance, Human Factors.

Keywords

Social Media, Terrorism, Machine Learning, extremism, sentiment analysis, text classification, Twitter, Natural Language Processing

1. INTRODUCTION

[1] Global accessibility to the Internet has reshaped our perception of the world. One of the most prominent and widely prevalent examples being Social Media (SM), which is present in many forms such as, forums, online news services, and social networks. Different social networks aim at different objectives, for example opinion transmission is done through Twitter or Facebook and business contacts through LinkedIn. In this modern digitalized era, this gigantic volume of data can be used to obtain different insights to improve the experience of everyone using the Internet.

Social Media, according to Margetts et al. (2015: 5) [2] can be defined as an 'Internet-based platform that allows the creation and exchange of user-generated content, usually using either mobile or web-based technologies

Though the internet can be undoubtedly considered a blessing, it is not without some serious flaws. Problems related to the safety of people are identified and detected daily on different social networks, like the violation of privacy, dependency and risk of use of these networks by minors, harassment and insults, and finally the security of people who are exposed to violent content like the video of the terrorist attack in New Zealand on Friday, March 15, 2019.

Vulnerable people can be easily targeted by terrorist groups seeking support and future recruits, subject to being victims of propaganda. From [3], Online Extremism can be defined as advocating support of groups or causes that in any distribution of opinion would lie on one of the "tails" [4]. Although the methods and ends espoused by ISIS' online marketing campaign clearly meet the definition of extremism, the campaign's global reach has generated an operationally significant amount of online and offline support.

Terrorist organizations have highly benefited by the worldwide reach, speed, and growth of the internet. By using these social media platforms, mainly Twitter, terrorist organizations spread their foul views.

From [5], the Islamic State, also known as ISIS, has distinguished itself as a pioneer in the use of social media for recruitment. But, while ISIS continues to be one of the most influential terrorist groups in the material world, other extremists are closing the gap in the virtual realm. Twitter is the ISIS's most popular and used social media platform.

[6] With the latest criminal and terrorist attacks occurring in the world, analyzing criminals' behavior and their connections is the main priority for authorities. The major challenge faced by authorities monitoring criminal activities is to accurately and efficiently extract criminals' or terrorists' information from huge volumes of criminal data available. To this end, we propose a framework built in Python that makes use of sentiment analysis and machine learning techniques to help detect content which can be considered to be extremist.

2. RELATED WORK

Since this is an active area of research in social media analysis, we will briefly focus on and overview the most relevant research, focusing on extremism/terrorism detection on Twitter, and sentiment analysis. Research into online radicalisation has mainly focused on two areas of enquiry, that is, the role of social media and ISIS' activity online.

Through sentiment analysis, and more concretely ML classification algorithms, messages in a social network such as

Twitter can be analyzed and tagged to determine the writer's attitude on a topic. In our case, we design and implement a sentiment analysis model to identify tweets that contain specific terrorism related keywords such as Jihad, ISIS, etc.

M. Ashcroft et al. [7] devised a method so as to detect jihadist messages on Twitter. They performed sentiment analysis to detect if a message supported ISIS or not. They compiled their dataset by searching for selective keyword(s) to extract tweets from Twitter feeds. This work has three features(time based, Sentiment based and Stylometric) used to detect jihadist text. About 90% accuracy was achieved through their results.

From Y. Noguchi and E. Kholmann [8], about 90% of the activities of terrorists on the internet are made on social networks. Several studies have been done which have focused on this topic/

Walid M. et al. [9] attempted to predict future support, or opposition, for ISIS from tweet textual analysis. In their study, the authors used Twitter data to study the ISIS support of users. They used the bag of words model as a feature vector which included individual terms, user mentions and hashtags. For their model, they used SVM with a linear kernel to train a classifier to predict the support or opposition of ISIS. Through this, they reached about 87% accuracy.

Furthermore, Saif et al. [10] found that semantic features based models out-performs other lexical, topic and sentiment based models in detecting ISIS supporting accounts. Berger et al. [11] discovered that these accounts can also be identified through their profile description. Agarwal et al. [12] expressed the presence of offensive, war and hate speech terms in ISIS propaganda material.

3. METHODOLOGY

We broadly define the stages involved in our project as:

I. Data gathering

II. Data cleaning

III. Sentiment analysis and Modelling

IV. Result visualization

3.1 Corpus creation

In this work, we restricted our dataset generation to be from Twitter only. The reasoning behind choosing Twitter to be our primary source of data was simply because Twitter is a massively successful microblogging social network, which has gained tremendous popularity over the last few years. Twitter users are restricted to writing messages of no more than 140 characters (short messages known as 'tweets'). Twitter has about 463 millions of daily users and 500 millions of tweets transmitted every single day [13].

However, this also means that a large portion of the transmitted content is hateful and extremist in nature. So much so that recently [14], Twitter banned/penalized about 1,126,990 different accounts between July and December 2020 for infringing its hateful conduct policy, a roughly 77% increase over the prior six-month period. Actions range from removing a tweet to banning an account.

A total of 110 tweets were extracted using *snscrape*, which is a scraper tool for social networking services (SNS). It is designed to scrape user profiles, hashtags, or searches and return the discovered items and results. We constructed our dataset corpus by filling in 10 tweets from a list of terrorism related keywords from a report [15], which contained words such as ISIS, Jihad, etc. and then tweets containing these words were extracted into a singular comma-separated values (csv) dataset. So, the dataset consists of 110 tweets (10 tweets for each keyword in terrorism keyword list) and includes the tweeter_id, tweet (text) and the username of the person who tweeted the message. We limited ourselves to this figure for reasons of performance of the machines used for the project and other similar limitations.
Since we were also interested in performing sentiment analysis, we manually labeled each of the 110 tweets as being either 0 (Not harmful or non-extremist) or 1(Terrorist/Extremist).

This was done manually since we believed that having a relatively diverse group of authors with sufficient knowledge of identifying extremist content will restrict the overall bias as opposed to using a predefined sentiment analysis model such as TextBlob[16] or VADER[17], which are often inaccurate, especially on highly sensitive topics such as terrorism.

We separated the task of data (keyword) extraction and data analysis into separate files for higher readability and documentation.

3.2 Data preprocessing and cleaning

[18]A tweet, for the most part, contains a lot of opinions about the data it represents in a very compact 140-character limited space. These additional features do not affect the sentiment of the tweets and are largely unnecessary as well as redundant in our analysis. Tweets also contain some language-based features including informality and is overall not a structured use of language. To overcome this, preprocessing of tweets is performed by taking multiple steps. As suggested in [19], the following tasks were performed on the raw tweet texts:

- **Letter casing**: Converting all letters to lowercase
- **Noise removal**: Eliminating unwanted characters, such as HTML tags, punctuation marks, special characters, white spaces etc.
- **Tokenizing**: Turning the tweets into tokens. Tokens are defined as words separated by spaces in a text.
- **Vectorizing Data**: Vectorizing is the process of converting tokens to numbers. It is an integral step as the machine learning algorithm works with numerical values and not text.

To go about implementing this, we made use of a popular tweet preprocessing library in Python (tweet-preprocessor)

[20] and some simple regex (regular expressions) to handle the cleaning, tokenizing and parsing of tweets.

3.3 Tweet sentiment classification

As previously mentioned in Related Works, there exists a considerable amount of research literature in the area of sentiment analysis. Since the goal of our research is to detect posts supporting extremist views, a tweet with positive sentiment towards ISIS(or terrorism-specific keywords) is labeled positive (1), while a tweet with negative OR neutral sentiment towards extremism is labeled negative(0).

We consider four different binary classifiers for our dataset: Multinomial Naive Bayes (NB), Logistic Regression (LR), Support Vector Classification(SVC) and Random Forest(RF).

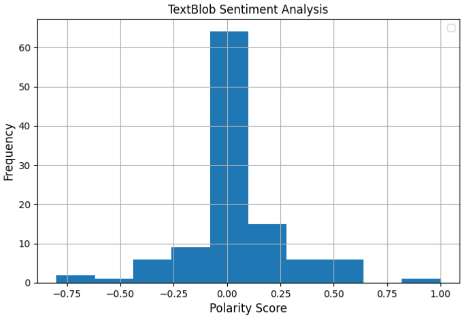

Figure 1. Preliminary sentiment analysis using TextBlob

3.4 Models:

Supervised classification is comprised of predicting the value of a qualitative response variable. This response variable is commonly known as a category or a class. There are many different classifiers, and their performance depends on the problem. However, given the fact that there is limited literature specific to detecting extremism/terrorism related conversations in social media. Chatfield et al. [21] analyze the content and sentiment of tweets published by a verified information disseminator of ISIS.

Amongst several tweet sentiment classifications, the dominant ones are those with supervised algorithms. The set of models used for evaluation in this project are:

i) *Multinomial Naïve Bayes classifier* [82][85]:

Naive Bayes models are a group of extremely fast and simple classification algorithms that are often suitable for very high-dimensional datasets. The features are assumed to be generated from a simple multinomial distribution and is very frequently used in text classification.

ii) Logistic regression (LR)[24]:

uses features(predictors) for building a linear model that estimates the probability that an observation belongs to a class.

iii) Support Vector Classifier (SVC) [25]:

classifiers whose result is based on a decision boundary generated by support vectors. The shape of the boundary is determined by a kernel function.

SVC in particular uses a linear kernel so in theory is faster and has better scalability. It is widely used for binary classifications and multi-class classifications.

iv) Random Forest (RF) [26]:

fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

The dataset was shuffled and split into a (8:2) ratio with 80% as the training dataset and 20% as the testing dataset. To keep the code clean, each of these classifiers were modeled in separate python programs with the same split ratio and seed.

The different classification and deep learning models used for profiling extremists that dynamically learn the continuous representation of tweets and then pick features from them extracted using count vectorizer and tf-idf vectorizer.

4. Experimental Results

Table I

Original/Raw Tweet	**Processed Tweet**
Wake Up Muslims Ummah Its The Time To Fight Its Time For jihad Its Time To listen To Oppressed Its Time To liberate Muslims of #Palestine #Syria #Myanmar #Kashmir https://t.co/lJhE48BpMm	wake muslims ummah time fight time jihadits time listen oppressedits time liberate muslims

4.1 Classification Metrics

For evaluating our models, we used precision, recall, F1 Score, support, and accuracy as classification metrics. We obtained them by using classification reports and confusion matrixes, as these metrics are widely used for evaluating supervised machine learning models for classification.

From (Afra 2019), Confusion matrix shows the performance of a classifier compared to a test data by highlighting the potential sources of error. Columns of the matrix represent prediction results while the rows represent the actual classes. Four cases are possible in a two-class classification problem.

True positive (tp) means a post was manually annotated

as extremist content and it was successfully predicted as so.

False negative (fp) means the latter post was predicted as non-terrorist related.

True negative (tn) means a post was manually annotated as non-extremist and it was successfully predicted.

False positive (fp) means the latter post was predicted as crime

related.

Accuracy is the ratio of correctly classified results, that is, the fraction of tp and tn.

Precision is the number of positive predictions divided by the total number of positive class values predicted.

Recall is the number of positive predictions divided by the number of positive class values in the test data

F-measure is the harmonic mean of precision and recall values.

The range of all 4 of these values vary between 0 and 1 (inclusive), with 1 being the best value and 0 being the worst value.

5. Conclusions

In this paper, we explore several features that can be useful for detection of extremism content on Twitter through the use of several different classification algorithms and models. Table 2 shows the accuracy of the model predictions.

We can observe that the Logistic Regression classifier had the highest accuracy out of the four other models with a 77% accuracy rate of detecting extremist content on Twitter.

From our experiments, the classification is limited to collect and analyze tweets that are only written in the English language.

It is indeed possible to extend upon this to support other languages. The accuracy of the proposed system can also be enhanced by analyzing images or videos (media content) on Twitter. Lastly, we restricted our study on the Twitter social media platform, which, despite being very popular, is dwarfed in userbase by other platforms such as Facebook, YouTube, Instagram and TikTok, thus the possibility remains to perform similar user classifications on these platforms so as to reach a wider audience and help detect the spread of extremism.

6. Graphs and Figures

TABLE 2: Comparison overview of accuracy of classifiers in classifying tweet sentiment

Classifier	Accuracy (in %)
Naïve Bayes	72.73
Logistic Regression	77.27
Support Vector Classification	72.72
Random Forest	68.18

7. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

8. Conflicts of Interest

The authors declare no conflicts of interest. The information and views set out in this paper are those of the author(s) and do not necessarily reflect the official opinion of the affiliation institutions.

9. BIBLIOGRAPHY

[1] Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. 2019. Detecting and monitoring hate speech in Twitter. Sensors (Basel, Switzerland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6864473/.

[2] Hopkins, J. 2017. How to define social media – an academic summary. Julian Hopkins PhD. http://julianhopkins.com/how-to-define-social-media-an-academic-summary/.

[3] Benigni, M.C., Joseph, K., and Carley, K.M. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0181405.

[4] Lake, D.A. Rational extremism: Understanding terrorism in the twenty first century. https://quote.ucsd.edu/lake/files/2014/06/Rational-Extremism.pdf.

[5] Oh, J. 2016. Isis vs. Nazi's on Twitter. *Homeland Security Digital Library*. https://www.hsdl.org/c/isis-vs-nazis-on-twitter/.

[6] Afra, S. and Alhajj, R. 2019. Integrated Framework for Criminal Network Extraction from Web - Salim Afra, Reda Alhajj, 2021. *SAGE Journals*. https://journals.sagepub.com/doi/10.1177/0165551519888606.

[7] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, "Detecting jihadist messages on twitter," in Intelligence and Security Informatics Conference (EISIC) European, Sept 2015, pp. 161-164

[8] Bedjou, K., Azouaou, F., and Aloui, A. 2019. Detection of terrorist threats on Twitter using SVM. *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems*.

[9] Magdy, W., Darwish, K., and Weber, I. 2015. #failedrevolutions: Using Twitter to study the antecedents of Isis Support. *arXiv.org*. https://arxiv.org/abs/1503.02401.

[10] H. Saif, Hassan and Dickinson, Thomas and Kastler, Leon and Fernandez, Miriam and Alani, "A semantic graph-based approach for radicalisation detection on social media," in European semantic web conference, 2017, pp. 571–587.

[11] Berger, J.M. and Morgan, J. 2016. The isis twitter census: Defining and describing the population of Isis supporters on Twitter. Brookings. https://www.brookings.edu/research/the-isis-twitter-

census-defining-and-describing-the-population-of-isis-supporters-on-twitter/.

[12] Agarwal, A. 2015. Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. springerprofessional.de. https://www.springerprofessional.de/de/using-knn-and-svm-based-one-class-classifier-for-detecting-onlin/2368512.

[13] Most used social media 2021. 2021. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[14] WAGNER, K. 2021. Twitter penalizes record number of accounts for hate speech. *Time*. https://time.com/6080324/twitter-hate-speech-penalties/.

[15] AHMED FRED LLOYD GEORGE, M. 2017. A war of keywords: How extremists are exploiting the internet and what to do about it. *Institute for Global Change*. https://institute.global/policy/war-keywords-how-extremists-are-exploiting-internet-and-what-do-about-it.

[16] SIMPLIFIED TEXT PROCESSING¶. *TextBlob*. https://textblob.readthedocs.io/en/dev/.

[17] CJHUTTO. CJHUTTO/Vadersentiment: Vader sentiment analysis. *GitHub*. https://github.com/cjhutto/vaderSentiment.

[18] SHAH, P. 2020. Basic tweet preprocessing in Python. *Medium*. https://towardsdatascience.com/basic-tweet-preprocessing-in-python-efd8360d529e.

[19] SINGHAL, G. 2020. Building a Twitter Sentiment Analysis in Python. *Pluralsight*. https://www.pluralsight.com/guides/building-a-twitter-sentiment-analysis-in-python.

[20] TWEET-PREPROCESSOR. *PyPI*. https://pypi.org/project/tweet-preprocessor/.

[21] CHATFIELD, A.T., REDDICK, C.G., AND BRAJAWIDAGDA, U. 2015. Tweeting propaganda, radicalization and recruitment: Islamic State supporters multi-sided Twitter Networks. *Tweeting propaganda, radicalization and recruitment | Proceedings of the 16th Annual International Conference on Digital Government Research*. https://dl.acm.org/doi/10.1145/2757401.2757408.

[22] A

[23] B

[24] SKLEARN.LINEAR_MODEL.LOGISTICREGRESSION. *scikit*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[25] SKLEARN.SVM.LINEARSVC. *scikit*. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.

[26] SKLEARN.ENSEMBLE.RANDOMFORESTCLASSIFIER. *scikit*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[27]