2015 European Intelligence and Security Informatics Conference

# Detecting Jihadist Messages on Twitter

Michael Ashcroft
Uppsala University
Dept. of Inf. tech.
Uppsala, Sweden
michael.ashcroft@it.uu.se

Ali Fisher
VORTEX
University of Vienna
Vienna, Austria
ali@humanshuddle.com

Lisa Kaati
FOI/Uppsala University
Stockholm, Sweden
lisa.kaati@foi.se

Enghin Omer
Uppsala University
Dept. of Inf. tech.
Uppsala, Sweden
omer.enghin@yahoo.com

Nico Prucha
ICSR
King's College
London, UK
nico.prucha@univie.ac.at

***Abstract*—Jihadist groups such as ISIS are spreading online propaganda using various forms of social media such as Twitter and YouTube. One of the most common approaches to stop these groups is to suspend accounts that spread propaganda when they are discovered. This approach requires that human analysts manually read and analyze an enormous amount of information on social media. In this work we make a first attempt to automatically detect messages released by jihadist groups on Twitter. We use a machine learning approach that classifies a tweet as containing material that is supporting jihadists groups or not. Even tough our results are preliminary and more tests needs to be carried out we believe that results indicate that an automated approach to aid analysts in their work with detecting radical content on social media is a promising way forward. It should be noted that an automatic approach to detect radical content should only be used as a support tool for human analysts in their work.**

## I. Introduction

Jihadist groups, and specifically ISIS, have been able to maintain a persistent online presence by sharing content through a broad network of "media mujahedeen" in one of the clearest incarnations of netwar since it was first envisaged. ISIS uses dispersed forms of network organization and strategy to disseminate rich audiovisual content from the battlefield in near-real time. Its interconnected network constantly reconfigures itself, much like the way a swarm of bees or flock of birds constantly reorganizes in mid-flight. It marks a shift from the broadcast models of communication during conflict to a new dispersed and resilient form, the user curated swarmcast [1]. This makes ISIS a challenge for traditionally hierarchical organizations to counter.

The internet has been identified by senior Sunni extremists as a "battlefield for jihad, a place for missionary work, a field of confronting the enemies of God" [2]. This was further encouraged by a "Twitter Guide" (dalil Twitter) posted on the Shumukh al-Islam forum which outlined reasons for using Twitter as an important arena of the electronic front (ribat) [3]. Since 2011 the Syrian conflict, recognized as the most "socially mediated" in history, has developed into the new focal point for jihadi media culture [4]. Within this online battlefield Twitter has become the Beacon for Jihadist Activity Online [2].

In particular the self-proclaimed "Islamic State" that is in the process of consolidating territory in Syria and Iraq has effectively built up and maintains a persistent presence online. One may argue, this online presence reflects the real-world consolidation of territory and one may be inclined to term this ideologically coherent and technical resilient presence as online territories of terror [5].

Jihadist content is removed every now and then and while Twitter actively shuts down extremist handles [6], the jihadist networks and clusters are able to cope with shut downs and removal. Content cannot be removed from all platforms at once and media workers and sympathizers are highly dedicated and committed, reopen accounts and oversee the constant flow of extremist materials across platforms, often using a number of languages. It is difficult to remove material for good from the Internet, especially not when committed individuals ensure specific videos can be found on YouTube or Facebook and elsewhere. With dedicated media operations, part of militant movements, the inclusive public relationship of terrorist groups as a tactic seeks to explain and justify militancy while inviting the consumers to join the cause.

In this work we will focus on Twitter and the distribution of jihadist messages on Twitter. We use a machine learning approach to classify tweets as supporting ISIS or not. We sometimes refer to this as classifying tweets into radical or non-radical even tough the problem that we are considering can not be generalized into solving the problem of detecting radical content in general.

**Outline** This paper is outlined as follows. In Section II we describe how Sunni extremists, to whom we refer to as jihadists, use social media and in particular Twitter. Some related work is also described in this section. In Section III we describe how we use classification to identify tweets that are supporting Jihadist grops and in Section IV we describe the experiments that we have done and their results. Section V some concluding remarks and directions for future work is presented.

## II. Radical Content on Social Media

Jihadist supporters, in particular ISIS supporters, use Twitter to distribute radical content through a dispersed network of Twitter accounts. The structure of the network gives ISIS a stable and persistent online presence even if key players are detected and suspended.

### A. Related work

Machine learning is the most common approach to sentiment classification previous work in this area includes [7], [8] and [9]. In [10] tweets are classified into five topic categories and in [11] trending topics on Twitter are classified into 18 general categories using bag-of-words and network-based classification. Results shows that accuracy is up to 65% and 70% for the different classification methods respectively. The

978-1-4799-8657-6/15 $31.00 © 2015 IEEE
DOI 10.1109/EISIC.2015.27

most common approach to classification is done using features such as keywords, entities, synonyms, and parts of speech [12], [13].

In [14] techniques based on machine learning and semantic-oriented approaches was used to identify radical opinions in hate group web forums. Four different types of text features: syntactic, stylistic, content-specific, and lexicon features where used and three classification techniques (SVM, Naive Bayes, and AdaBoost).

An approach using ISIS related tweets to predict future support or opposition for ISIS was done in [15] where the authors used twitter data to study the antecedents of ISIS support of users. Predictions about future support or opposition for ISIS could be done with with 87% accuracy using a SVM classifier.

## III. Classification of Jihadist Messages

Automatically analyzing messages on Twitter is an important task for law enforcement agencies. It is impossible for human analysts to manually read all information that is available. Not only is the available information made up of various forms of media such as texts, pictures and video content, the information is also made available by different extremists media groups in multiple languages - with Arabic dominating this field. With media groups operating alongside fighting elements, and with the influx of non-Arab foreign fighters, content is created across a range of languages appealing to different audiences while providing a consistent message and ideology. Using computers to classify content automatically as radical or not would significantly speed up the analysis so that radical content can be removed earlier. However, it is important to emphasize that computers should only be used to support analysts - in this case an automatic classification of tweets can be presented to the analysts that can make a decision whether the tweet is radical or not.

We use a machine learning approach as a first step towards determining if a tweet is supportive of Jihadist groups or not. To build a classifier that can do this we need to have a suitable dataset to select features from that are useful in determining if a tweet is supporting Jihadist groups or not. The most common approach is to use humans that manually classify tweets as either supporting or non-supporting but in this work we have used another approach. We have collected a set of tweets containing hashtags related to jihadists, and in particular ISIS, from the English language spectrum of pro-ISIS clusters on Twitter. All of the hashtags we used have a corresponding Arabic hashtag and are often used within Arabic and non-Arabic tweets to widen the availability of ISIS material in general. In this work we have focused on the English hashtags and the hashtags we have used to collect data are the following: #IS, #ISLAMICSTATE, #ILoveISIS, #AllEyesOnISIS, #CalamaityWillBeFallUS, #KhalifaRestored, #Islamicstate. The tweets where collected between 25th of June 2014 and 29th of August 2014. Some of the messages that were collected containing the hashtags mentioned above were not related to ISIS and they had no violent/radical message. For example, in some cases the #IS hashtag was not referring to the Islamic state but to the verb "is" (to be). In other cases, some of the hashtags were used since the tweets contained messages that were against ISIS. To tackle this issue we used a list of user accounts describing clusters of known Jihadist sympathizers [3]. The list we used consisted of 6729 user names. We use only tweets that were posted by these users. Apart from the dataset containing tweets supporting Jihadist groups we also collected 2000 random tweets discussing various topics and a set of tweets that was from accounts that are against ISIS. The assumption that the identified accounts were against ISIS and are not posting messages supporting ISIS was made based on the user name and manual verification of the content.

We only use tweets that were written in English. The selection of tweets that were written in English was done using Shuyo's language detection library for java [16].

## IV. Experiments

### A. Datasets

As mentioned in the previous section, we used three different datasets. We will label these datasets TW-PRO, TW-RAND and TW-CON (Table I).

| Dataset | Description |
|---|---|
| TW-PRO | Tweets that are pro ISIS, based on hashtags and network of known jihadists. |
| TW-RAND | Randomly collected tweets discussing various topics. |
| TW-CON | Tweets from accounts that are against ISIS. |

TABLE I: The datasets used for the experiments

The data was preprocessed and re-tweets, annotations and urls were removed. The data was tokenized and lemmatized to obtain the base form of the words. In the process of lemmatizing the toolkit [17] was used.

### B. Features

In our experiments we have used three different classes of features:

- stylometric features (S)
- time based features (T)
- sentiment based features (SB)

**Stylometric features** The stylometric features that we have used in this work are shown in Table II. The stylometric features also contain a set of the words that are the most frequent in the dataset. We have used 173 frequently used words where the 10 most frequently used words in the dataset are: $state, islamic, not, do, kill, support, abu, allah, people,$ and $al$. Among the most common words we can notice words that are related to ISIS ($state$ and $islamic$), verbs like $kill$ and $support$ as well as the word $abu$. The word $abu$ refers to individuals who usually assume a nom de guerre starting with abu (father of) [name] by which various foreign fighters, martyrdom operatives, ideologues etc are introduced and made popular within the respective clusters; $allah$ simply refers to "God" and the Arabic appellation is preferred by pro-ISIS clusters who generally adhere to Arabic dominated language sets; "al" is the Arabic article "al" ("the").

| | | |
|---|---|---|
| function words | frequency of various function words | 293 |
| frequent words | frequency of most frequent words | 173 |
| punctuation | frequency of characters . , , ; : , ’ ,- , [ , ] , { , } , !,?,& | 13 |
| hashtags | frequency of most frequent hastags | 100 |
| letter bigrams | frequency of most frequent letter bigrams | 133 |
| word bigrams | frequency of most frequent word bigrams | 99 |

TABLE II: The stylometric features. The list of function words that we have used can be found in [18].

Stylometric features also includes punctuation, letter bigrams, word bigrams and the most frequently used hashtags. In the dataset that we use, the 10 most frequently used hashtags in the dataset are: #IS, #AllEyesOnISIS , #Iraq, #Syria, #Islam, #ISIS, #Muslims, #Mosul, #Caliphate, and #Khilafa.

**Time based features** Time based features contain information about when a tweet is posted. The attributes that we have used are similar to what was used in [19] except for the month feature. Including month as a feature would not be feasible since we only have data from two months. The attributes are:

- **Hour Of Day:** Hour1, Hour2, . . . , Hour24,
- **Period Of Day:** Morning, MidDay, Evening, Night,
- **Day:** Sunday, Monday, . . . , Saturday
- **Type Of Day:** WeekDay, WeekEnd

**Sentiment based features** Sentiment analysis seeks to determine the attitude of text towards a specific topic or the general contextual polarity of the text. It has extensively been used to classify the sentiment of film, book and product reviews as well as the sentiments of tweets [8], [20]. The analysis of the sentiment was done using [17], the values the sentiment can take are: very negative, negative, neutral, positive, very positive.

Most tweets in the dataset had a negative sentiment. This is most likely since the tweets contains words that are considered to be negative such as killing.

**Feature vectors** Each tweet is transformed into a feature vector where position in the feature vector corresponds to the absolute frequency of how many occurrences each individual feature has. For each set of stylometric features: function words, frequent words, punctuation, hashtags, letter bigrams and word bigrams an internal normalization is done. This means that the sum of each set of stylometric features is equal to 1.

The total number of features we have used are 853. To sort out the features that contributed to the classification, we applied information gain, which is an entropy-based feature selection method [21]. The features that contributed most to the classification were:

1) Friday
2) ey (letter bigram)
3) #AllEyesOnISIS
4) ye (letter bigram)
5) hour: 4 (time feature)

Friday is the most important feature, with most radical tweets (and re-tweets, though they were removed from the dataset) being sent on Fridays. This seems to be a strategy that is used to make the tweets live longer (over the weekend) and avoid having twitter accounts are blocked and the content removed. This strategy is also backed by the finding that ISIS high quality videos are often released over the weekend, as a deliberate strategy to pick the days when those employed to challenge or disrupt the propagation of information about the video were least likely to be available [22]. Additionally, Muslims traditionally hold a congregational prayer (in standard Arabic called jum’a) every Friday where a sermon (khutba) is held. The khutba is given by an Islamic cleric and is thus received as a person of authority. The sermon oftentimes touches on religious, moral and contemporary subjects. Another reason for the high number of posts on Friday might be the fact that jihadist media supporters and sympathizers seek to contribute on this important day by using twitter to spread extremists’ messages that are published around the clock.

After removing the features that did not effect the classification we used 579 stylometric features, 36 time based features and 4 sentiment based features.

### C. Experimental results

All our experiments were conducted using a machine learning tool called Weka which is a suite of machine learning software written in Java. For the experiments we have used between 4000 and 7500 tweets in total and the ratio between testing data set and training data set was approximately 1:4. We have used three different classification algorithms: SVM, Naive Bayes and Adaboost.

The results from the experiments using the different feature sets are reported using confusion matrices in which we present the number of true positives, false negatives, true negatives, and false positives as illustrated in Table III.

| | Predicted class | |
|---|---|---|
| Actual class | **True Pos. (TP)** | **False Pos. (FP)** |
| | **False Neg. (FN)** | **True Neg. (TN)** |

TABLE III: Confusion matrix

Results differed depending on what datasets we used. Using TW-PRO and TW-RAND led to better results than if TW-PRO and TW-CON were used. The results for TW-PRO and TW-RAND and the features (S + T + SB) are shown in Table IV. As can be noted AdaBoost performs very well with 100 % accuracy on the test set.

The results for using the datasets TW-PRO and TW-CON are shown in Table V, the accuracy when using the AdaBoost classifier is still high (99.5%). Since the datasets that are used for the experiments in Table IV and Table V differs the results are expected. TW-RAND contain randomly selected tweets

| | Non Radical | Radical | Correctly Classified Instances |
|---|---|---|---|
| SVM | 1974 | 24 | 99.1 % |
| | 11 | 1990 | |
| Naive Bayes | 1997 | 1 | 99.9 % |
| | 1 | 1990 | |
| AdaBoost | 1998 | 0 | 100 % |
| | 0 | 1991 | |

TABLE IV: Results when using features (S + T + SB) on the datasets TW-RAND and TW-PRO.

while TW-CON are tweets that are against ISIS. The tweets that are against ISIS contain similar hashtags and topics as the TW-PRO dataset and is therefore harder to separate than the randomly collected tweets.

| | Non Radical | Radical | Correctly Classified Instances |
|---|---|---|---|
| SVM | 1155 | 38 | 98.5 % |
| | 24 | 2783 | |
| Naive Bayes | 1178 | 15 | 96.8 % |
| | 114 | 2693 | |
| AdaBoost | 1182 | 11 | 99.5 % |
| | 8 | 2799 | |

TABLE V: Results when using all features (S + T + SB) on TW-CON and TW-PRO

| | Non Radical | Radical | Correctly Classified Instances |
|---|---|---|---|
| SVM | 3099 | 92 | 97.9 % |
| | 74 | 4813 | |
| Naive Bayes | 2877 | 314 | 89.0 % |
| | 574 | 4313 | |
| AdaBoost | 1600 | 0 | 100 % |
| | 0 | 2905 | |

TABLE VI: Results when using all features (S + T + SB) on the full dataset

Table VI shows the results for three different classifiers using all features on all the datasets TW-PRO, TW-RAND and TW-CON. As can be seen in the table AdaBoost performs slightly better than both Naive Bayes and SVM.

## V. Conclusions and Future work

In this work we have used machine learning to automatically identify jihadist messages. One of the major problems with classification is that in most cases the data is manually labeled by analysts as either jihadist on non-jihadist (we use the terms radical or non-radical). We avoided this and related issues such as analyst disagreement by working with data labeled from incorporated hashtags and using networks of known jihadists to assure radical content. One of the drawbacks with our method is that many of the features are dependent of the dataset. For future work it would be interesting to use both data dependent and data independent features and evaluate the results.

Detecting radical content in order to react on it or to work with partners to remove it is an important task for law enforcement agencies. Due to the enormous amount of official high quality jihadist material issued by media operatives as well as user generated content available on social media there is a need for tools that can aid analysts in their work of detecting radical content online semi-automatically. We say 'semi-'automatically, because it is important to note that automatic detection of radical content as described in this paper can never replace human analysts. Instead it should be seen as a complementary way to detect radical content and present it to an analyst for further actions.

The results we have obtained using this limited dataset indicates that classification is a viable way forward in the work of detecting radical content on social media, and in particular on Twitter. We look forward to trying to replicate these results on more diverse and or complex data.

## References


[1] A. Fisher, "Last gang in town: How jihadist networks maintain a persistent presence online," in *Perspectives on Terrorism (to appear)*, 2015.

[2] A. Fisher and N. Prucha, "Tweeting for the caliphate: Twitter as the new frontier for jihadist propaganda." in *CTC Sentinel 6.6 19-23*, 2013.

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

[4] M. Lynch, F. Deen, and S. Aday, "Syrias socially mediated civil war." in *United States Institute Of Peace 91.1 1-35*, 2014.

[5] N. Prucha, "Online territories of terror: how jihadist movements project influence on the internet and why it matters offline," in *Dissertation, University of Vienna*, 2015.

[6] S. Wright, N. Dorman, and C. Cortbus, "Twitter shuts down isis supporters and jihadists as mi5 launch anti−terror social media crackdown," in *Mirror UK News*, 2015.

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *ACL-02*, 2002.

[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *LSM 2011*, 2011, pp. 30–38.

[9] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *COLING 10*, 2010, pp. 36–44.

[10] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *ACM SIGIR*, ser. SIGIR '10, 2010, pp. 841–842.

[11] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *ICDMW*, 2011, pp. 251–258.

[12] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, "Precise tweet classification and sentiment analysis," in *ICIS*. IEEE, 2013.

[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, 2009.

[14] M. Yang, M. Kiang, Y. Ku, C. Chiu, and Y. Li, "Social media analytics for radical opinion mining in hate group web forums," *Journal of homeland security and emergency management*, vol. 8, no. 1, 2011.

[15] W. I. Magdy Walid, "#failedrevolutions: Using twitter to study the antecedents of isis support," *arXiv preprint arXiv:1503.02401*, 2005.

[16] N. Shuyo, "Language detection library for java," http://code.google.com/p/language-detection/, 2010.

[17] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. of 52nd Annual Meeting of the Ass. for Comp. Ling.: System Demonstrations*, 2014, pp. 55–60.

[18] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *2012 IEEE SP*, 2012, pp. 300 –314.

[19] F. Johansson, L. Kaati, and A. Shrestha, "Time profiles for identifying users in online environments," in *IEEE JISIC*, 2014, pp. 83–90.

[20] T. T. Thet, J.-C. Na, and C. S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *Jour. of Inf. Sci.*, 2010.

[21] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Formal Methods for Information Retrieval*, vol. 42, pp. 155–165, 2006.

[22] A. Fisher and N. Prucha, "Is this the most successful release of a jihadist video ever? part 2," in *Jihadica.com*, 2014.