2015 IEEE 15th International Conference on Data Mining Workshops

Detecting Multipliers of Jihadism on Twitter

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

Enghin Omer
Uppsala University
Uppsala, Sweden
omer.enghin@yahoo.com

Nico Prucha
ICSR
London, UK
nico.prucha@univie.ac.at

Amendra Shrestha
Uppsala University
Uppsala, Sweden
amendra.shrestha@it.uu.se

***Abstract*—Detecting terrorist related content on social media is a problem for law enforcement agency due to the large amount of information that is available. This work is aiming at detecting tweeps that are involved in media mujahideen - the supporters of jihadist groups who disseminate propaganda content online. To do this we use a machine learning approach where we make use of two sets of features: data dependent features and data independent features. The data dependent features are features that are heavily influenced by the specific dataset while the data independent features are independent of the dataset and can be used on other datasets with similar result. By using this approach we hope that our method can be used as a baseline to classify violent extremist content from different kind of sources since data dependent features from various domains can be added.**

In our experiments we have used the AdaBoost classifier. The results shows that our approach works very well for classifying English tweeps and English tweets but the approach does not perform as well on Arabic data.

I. Introduction

Detecting and removing terrorist related content on the Internet is an important and difficult task for law enforcement agencies all over the world. Jihadist groups, and specifically ISIS, have been able to maintain a persistent online presence by sharing content through a broad network of ”media mujahideen”. 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” [1]. 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) [2]. Since 2011 the Syrian conflict, recognized as the most ”socially mediated” in history, has developed into the new focal point for jihadi media culture [3].

Facilitating the Internet as the prime and most effective (as well as cost-effective) communication facility to lure consumers into their specific interpretation or world perception is not restricted to the jihadi web. Militant and hate groups of all colors employ similar means to gain sympathy through modern and pop-cultural elements. However, the quantity as well as quality, not to neglect the multi-lingual capacity, of jihadi media departments is unmatched and unprecedented. The rhetoric is inseparable from the (audio-) visual content and enforces key elements while reaching out to the audience to get active, empower the ”Islamic community” (*umma*) by individual response.

Extremist content appeals to Arabs and non-Arabs, while the latter group is directly approached by militant media operatives who provide translations of Arabic language materials and reach out via social media. Content, Arabic and non-Arabic is not often removed, for example, the first video of the self-proclaimed ”Caliph” Abu Bakr al-Baghdadi, published in June 2014 is still available to download as of writing.[1] The speech, given in al-Baghdadi’s mother tongue Arabic was published in English, German, French and Russian. The English version of a video by the *Ansar Bayt al-Maqdis* group showing attacks against Egyptian soldiers published in February 2014 is still available to download on the links published by the group to highlight the operations on the Sinai Peninsula[2] - in the meantime the group has merged with ISIS while Sinai was already announced a ”province” (*walaya*) of a future Islamic State in the February video [4]. Take the last big al-Qaeda speech featuring Ayman al-Zawahiri declaring a franchise in India published September 2014 in multiple languages, including Bangladeshi, all links are still functioning with the only difference to ISIS videos the download- and view count.[3]

With abundant jihadist material freely available online and with an ever present crowd of sympathizers at hand, the nature of jihadist spheres on the Internet have changed, adapted, and increased with the technical developments of the World Wide Web over time. It is al-Qaeda who are the pioneers on the front lines and in attacking western capitals, while maintaining for decades ideologically consistent materials, published to indoctrinate and recruit potential new members. AQ has created the very foundation of Sunni jihadist ideology - and managed to ensure it’s persistent and take-down resilient presence on the Internet [5].

Within the framework of the turmoil in the wake of the ”Arab Spring”, and especially with the conflict in Syria growing in intensity and scope, AQ has been able to re-emerge with two linked groups, *Jabhat al-Nusra* (JN) and *The Islamic State of Iraq and al-Sham* (ISIS) [6]. While JN has pledged allegiance (*bay’a*) to Ayman al-Zawahiri, ISIS under the rule of Abu Bakr al-Baghdadi has morphed into the greatest success of al-Qaeda iconography and doctrine by adhering to the ideology without being subjected to its formal leadership. As thus, the renaissance of al-Qaeda doctrine is largely based on its most powerful tool that allows the network to morph and spread in many directions: the professional and coherent decentralized use of the Internet and the continuous and tireless deployment of media-workers and media-dedicated brigades

[1] Khutba wa-salat al-jum’a fi l-jami’a al-kabir bi madinat Mosul, Mu’assasat al-Furqan. Avaliable at *https://archive.org/download/KhotbaJomaa/*

[2] Sawlat al-ansar - ’amaliyat al-mujahidin dudd jaysh al-ridda al-masriyya, Fursan al-Balagh. Available at *http://justpaste.it/fursan-trv-swlansar*

[3] I’lan insha’ far’a jadid l-jama’a qa’idat al-jihad fi shiba al-qara al-hindiyya, Mu’assasat al-Sahab. Available at *https://archive.org/details/Indiann*

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.9

embedded with fighting jihadi units in the real-life battle arenas [6].

Twitter has become an important way of communicating for jihadist groups, most likely since it is easy to use and provide rapid updates to a large amount of users. Twitter has also been used live at the battlefield to report about injuries or deaths of fighters and battle outcomes without any censorship. One of the most common approaches to stop terrorist groups from spreading propaganda and other forms of terrorist related content is to suspend accounts when they are discovered. Usually this approach requires that human analysts manually read and analyze an enormous amount of information on social media. In this work we use machine learning to automatically detect tweeps (user accounts on Twitter) that are supporters of jihadist groups and disseminate propaganda content online. We also use machine learning to detect individual tweets that contains propaganda.

A. Outline

This paper is outlined as follows. In Section II work related to ours is described. In Section III we described our approach to detect tweeps that are involved in media mujahideen. We also described the data dependent features and the data independent features that we use in our experiments. In Section IV the experimental setup is described, including the datasets that we have used. Our experimental results are reported in Section V and a discussion about the results is presented in Section VI. Finally, some conclusions and directions for future work is presented in SectionVII.

II. Related work

Analyzing terrorist related content on the Internet has previously done in many variations. The approach that we use is text classification, an approach that has been used in several cases to detect online radical content. The problem is usually transfered into a binary text classification problem where linguistic features in the content are used to train a machine learning classifier.

One approach to analyze extremist online content is describe in [7] where affect analysis is used to analyze the intensity of emotions in extremists discussion boards. To achieve this a manually created affect lexicon is used to measure the usage of violence and hate affects in U.S. and Middle Eastern extremist group forum postings. Another approach to analyzing affects using machine learning is described in [8] where classifiers are used to analyze affects in two jihadist discussion boards: Al Firdaws and Montada. Four different classifiers are used: violence, anger, hate and racism. To build the classifiers, a number of linguistic features like character n-grams, word n-grams, root n-grams and collocations are used. The features that are used relies heavily on the data.

In [9] machine learning and semantic-oriented techniques were used to identify radical opinions in hate group web forums. Semantic orientation is a sentiment classification that is based on the total sum of both positive and negative sentiment words and phrases contained in the text that is evaluated. Four different classes of features are used: syntactic, stylistic, content-specific, and lexicon features. The lexicon features are a semantic oriented technique. In their work two domain experts annotate the collected messages independently - this is one step that we have been able to avoid due to the nature of the dataset we use. Three classification techniques are used: SVM, Naive Bayes and Adaboost. In their result SVM outperformed the other classifiers and the results showed that by adding more feature sets (i.e., syntactic, stylistic, content-specific, and lexicon features) the effectiveness of the classifiers for radical opinion identification was improved significantly. A conclusion that can be drawn from this paper is that content-specific features are important features in opinion mining, something that can be seen in our experiments as well.

Machine learning on ISIS related tweets is also done in [10] where the authors used data from twitter to classifying users as potentially supporting or opposing ISIS before the user explicitly write a tweet identifying his/her stance. The dataset that they use consist of a collection of Arabic tweets referring to ISIS, the tweets are then classified into pro-ISIS and anti-ISIS. The classification into pro-ISIS and anti-ISIS is done automatically using the name variants that is used to refer to the organization: the full name and the description as ”state” (in Arabic: *dawla*) is associated with support, whereas abbreviations (*da'ish*) usually indicate opposition. In our work we use a network of known jihadists accounts whom we have identified and graded as such by reading and manually assessing their Arabic and non-Arabic tweets. The features used in [10] are similar to the features that we use in this work: bag-of-words features, including individual terms, hashtags and user mentions. What can be noted is that the results from [10] shows that supporters and opposers of ISIS can be separated with high accuracy using features that are data dependent.

In [11] experiments are done on classifying ISIS related individual tweets. The dataset that is used is a subset of the dataset that we have used in this paper and it consist of English tweets from a network of known jihadists accounts. The features that are used are stylometric features, time based features and sentiment based features. This work can be seen as an extension of the work done in [11].

III. Detecting tweeps involved in media mujahideen

This work is aiming at detecting tweeps that are involved in media mujahideen - the supporters of jihadist groups who disseminate propaganda content online [2]. Our approach towards this problem is to use machine learning and train a model that can recognize if a twitter user is supporting a jihadist groups and disseminate propaganda content online (also called a media mujahid). To do this we have transferred the problem of detecting tweeps involved in media mujahideen into a binary text classification problem and we use linguistic features in the content to train a machine learning classifier. The datasets that we have used are described in Section IV.

We use two different sets of features: data dependent features and data independent features. The data dependent features are features that are heavily influenced by the specific dataset, this kind of features can be very useful if a specific topic or domain is considered but the results can not be generalized to work on different datasets. Data independent features on the other hand are features that are independent of

Feature class	Description	Number of features
Word length	Relative frequency of words with 1-20 characters	20
Letters*	Relative frequency of a to z (ignoring case)	26
Digits	Relative frequency of 0 to 9	10
Punctuation	Relative frequency of characters . ? ! , ; : ( ) ” - ´	11
Arabic Function words	Relative frequency of various function words	160
English Function words	Relative frequency of various function words	293
Time	Features related to time	39
Emotion words	Relative frequency of various sentiments words	108
Hashtag	Relative frequency of hashtags	1

*Only for English tweets

TABLE I. DATA INDEPENDENT FEATURES

Feature class	Description	Top 10 English features
Hashtags	100 most common hashtags	#is, #iraq #islamic_state, #alleyesonisis, #syria, #islamicstate, #khilafarestored, #islam, #isis, #muslims
Word bigrams	100 most common word bigrams	of the, islamic state, to the, the islamic, in the, from the, for the, by the, on the, islamic states
Letter bigrams	100 most common word bigrams	th, he, in, er, an, is, re, st, es, at
Frequent words	100 words that are used most frequently	the, of, in, to, and, a, from, is, for, islamic

TABLE II. DATA DEPENDENT FEATURES

the dataset and that can be used on other datasets with similar result.

A. Data independent features

The classes of data independent features that we have used are described in Table I. The data independent features are stylistic features (as described in [12]), time features (e.g. what time or what day a tweet is posted, we use a subset of the time features described in [13]), and emotion words. The emotions words are used to capture emotions and sentiments. Most of the features are similar for both English and Arabic text.

B. Data dependent features

The data dependent features that we have used are influenced by the specific dataset. In this case, where we are interested in building a model for classifying twitter users that are communicating jihadi content, the data dependent features can be valuable. It might be the case that certain hashtags and frequently mentioned words change over time but many of the data dependent features remain the same and are representative for the group of users that is targeted in this work. The data dependent feature classes that we use are the most common hashtags, most common word bigrams, most common letter bigrams and the most frequent words. In Table II the set of data dependent features that we have used are listed. Not surprisingly, words and hashtags related to IS, Islam and Islamic state are in the top 10 most used features on English. Similar results can be seen for Arabic where the most frequently used words are *in, from, Allah, on* and *that*.

IV. EXPERIMENTAL SETUP

We have conducted a set of experiments to get an understanding of how well tweeps that are involved in media mujahideen can be identified on Twitter using data dependent, data independent features and a combination of both. All experiments are done on Arabic and English datasets.

A. Classifier

In our experiments we have used an AdaBoost classifier. The AdaBoost algorithm was introduced by [14]. AdaBoost models were created using the ada R package [4] [15]. Before choosing AdaBoost, other classifier[5] was evaluated but since AdaBoost outperformed the other classifier the experiments were done using AdaBoost. AdaBoost is a machine learning algorithm based on boosting that combines moderately inaccurate rules of thumb or simple classifiers to create a very accurate classifier. The boosting algorithm calls the simple classifiers repeatedly. When learning each classifier in the sequence, the data is weighted so that weak classifiers are tweaked in favor of those instances misclassified by previous classifiers. Final classifications are combined into a weighted sum that represents the final output of the boosted classifier. We have used classification trees as the base classifiers.

B. Datasets

We use two datasets with tweeps that are involved in media mujahideen: one with tweeps communicating on English and one with tweeps communicating on Arabic. We call the set of tweeps in English TWEEP-PRO-E and the Arabic TWEEP-PRO-A. We also use two sets of ”random” tweeps. These tweeps are not really random but we use them to represent regular tweeps (in this case, regular in the sense that they are not supporters of jihadist groups). We call the set containing tweeps on English TWEEP-RAND-E and the set of randomly collected tweeps on Arabic TWEEP-RAND-A.

In our experiments we also classify individual tweets. To do this we have similar datasets: two sets containing tweets containing jihadist propaganda, one on English called TW-PRO-E and one on Arabic called TW-PRO-A. We also have two sets of ”random” tweets one English called TW-RAND-E and one on Arabic called TW-RAND-A. The datasets, a short description of them and how the datasets where used in the classification are presented in Table III.

Media mujahideen: To obtain a dataset containing tweeps that can be seen as multipliers of jihadism we have collected data in two different ways. As described in [2] a posting on the Shumukh al-Islam forum provided a ”Twitter Guide”. In this guide 66 users are highlighted as ”The most important jihadi and support sites for jihad and the mujahideen on Twitter”. We have used 30 tweeps from the guide and downloaded the latest 3400 tweets from each of the 30 users

[4]The package is freely available from http://CRAN.R-project.org/

[5]Support Vector Machine (SVM) was used with default parameters.

Dataset	Description	Size	Training	Testing	Validation
TWEEP-PRO-E	English tweeps involved in media mujahideen.	93	41	31	21
TWEEP-RAND-E	Randomly collected tweeps on English.	742	377	220	145
TWEEP-PRO-A	Arabic tweeps involved in media mujahideen.	81	33	28	20
TWEEP-RAND-A	Randomly collected tweeps on arabic.	256	136	74	46
TW-PRO-E	English tweets containing jihadist propaganda based on hashtags and network of known jihadists.	27753	13994	18027	5459
TW-RAND-E	Randomly collected tweets discussing various topics on English.	60000	29882	8299	12091
TW-PRO-A	Arabic tweets containing jihadist propaganda. based on hashtags and network of known jihadists.	16000	7973	4836	3191
TW-RAND-A	Randomly collected tweets discussing various topics on arabic.	45013	22534	13468	9011

TABLE III. THE DIFFERENT DATASETS USED IN THE EXPERIMENTS.

(the other accounts were suspended by Twitter). We have also used a set of 45 tweeps that where manually identified to be multipliers of jihadism, all these users are followers of the 66 users and are spreading jihadist propaganda.

We have also collected a set of tweets containing hashtags that were related to jihadists, and in particular ISIS. The hashtags we have used to collect data are the following: #IS, #ISLAMICSTATE, #ILoveISIS, #AllEyesOnISIS, #Calamaity-WillBeFallUS, #KhalifaRestored and #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. 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 clusters of known Jihadist sympathizers [2]. The list we used consisted of 6729 usernames and the tweeps (and tweets) that we use are all from this list.

An example of a English tweet from the dataset is:

"Who wants the truth about ISIS??? Well here it is from Sheikh AlAdnani in his recent speech #AllEyesOnISIS http://t.co/LFT790b5bo"

This particular tweet is referring to Sheikh Abu Muhammad al-Adnani, one of the top ISIS ideologues who frequently produces audio speeches that are issued in Arabic and mostly published in English, German, French and Russian.

Tweeps discussing various topics: A set of tweeps discussing various topics is also used to train our classification model. To get tweeps discussing various topics we used two approaches. To collect English tweets we collected a set of tweets during a certain time period and use some of the tweeps that had written these tweets. For the Arabic tweets we collected tweeps from a list of Twitter influential in Arabia and a list of the 100 most influential Arabic female Twitter users.

An example of a tweet from the English dataset is:

"Toby Keith Tickets http://t.co/oV6US49t at Blossom Music Center in Cuyahoga Falls OH on July 13 #tobykeith"

V. EXPERIMENTAL RESULTS

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 IV.

TABLE IV. CONFUSION MATRIX

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

We have done two sets of experiments where we use tweeps and tweets on English and Arabic. For each experiment we use three different sets of features.

A. Experiment 1: Classifying tweeps

In the first experiment we use two different datasets (English and Arabic) with two different sets of features (data independent and data dependent). While classifying tweeps the AdaBoost model has been performed with 500 boosting iterations and rest of the parameters were set default. How the datasets were used for training, testing and validation is shown in Table III.

1) Arabic tweeps: The results for using data dependent and data independent feature are shown in Table V. While using data independent features the accuracy is 0.9242 and the precision 0.8. With data dependent features the accuracy is 0.9848 and the precision 0.8260. Using both data independent and data dependent features the accuracy is 0.9697 and the precision 0.9. As could be expected precision, accuracy and recall are significantly lower when using data independent features compared to data dependent features. In this small sample dataset, we can not tell if a combination of data dependent and data independent features improves the result.

2) English tweeps: The results for using data dependent and data independent feature are shown in Table VI. The results using the different features are almost the same and when data dependent features are used the result is perfect. The reason could be that the tweeps spreading jihadist propaganda and the "random" tweeps are totally different from each other; a much larger dataset would be needed to investigate this further. Another reason for the results could be that the tweets are downloaded during different time periods and there might be a difference in what topics that are discussed.

B. Experiment 2: Classifying individual tweets

In the second experiment we use the same setup as in the first experiment but instead of classifying tweeps we classify individual tweets. Since the dataset is much larger than in the previous experiment and due to time constraint the AdaBoost model has been performed with 300 boosting iterations in these experiments. The default parameters were used for the rest of the AdaBoost parameter settings.

TABLE V. RESULTS FOR CLASSIFYING ARABIC TWEEPS

Features	Confusion Matrix	Precision	Accuracy	Recall
Data independent	45 4 1 16	0.8	0.9242	0.9411
Data dependent	46 1 0 19	0.8260	0.9848	1.0
Data dependent + Data independent	46 2 0 18	0.9	0.9697	1.0

TABLE VI. RESULTS FOR CLASSIFYING ENGLISH TWEEPS

Features	Confusion Matrix	Precision	Accuracy	Recall
Data independent	145 1 0 20	0.9528	0.994	1.0
Data dependent	145 0 0 21	1.0	1.0	1.0
Data dependent + Data independent	145 0 0 21	1.0	1.0	1.0

TABLE VII. RESULTS FOR CLASSIFYING ARABIC TWEETS

Features	Confusion Matrix	Precision	Accuracy	Recall
Data independent	8542 1679 469 1512	0.4738	0.824	0.7633
Data dependent	8699 1560 312 1631	0.5111	0.8466	0.8394
Data dependent + Data independent	8724 1375 287 1816	0.5691	0.8638	0.8635

TABLE VIII. RESULTS FOR CLASSIFYING ENGLISH TWEETS

Features	Confusion Matrix	Precision	Accuracy	Recall
Data independent	12006 122 85 5337	0.9776	0.9882	0.9843
Data dependent	12087 159 4 5300	0.9708	0.9907	0.9992
Data dependent + Data independent	12085 80 6 5379	0.9853	0.9951	0.9988

1) Arabic Tweets: When classifying individual tweets on Arabic the results are not as good as when classifying tweeps (that had written at least 60 messages). The accuracy is 0.824 for data independent features, 0.8466 for data dependent features and 0.8638 for a combination of both set of features. The precision and the accuracy has decreased significantly compared to the experiments done on Arabic tweeps. The precision had dropped to 0.4738 for data independent features and 0.5691 for the combination data dependent and data independent features. Figure 1 shows the precision, accuracy and recall for classifying Arabic tweets using the data independent features (Indep), data dependent features (Dep) and both data dependent and independent features (All).

2) English Tweets: The results for both data dependent and data independent feature are shown in Table VIII. The accuracy is over 0.98 for all sets of features and using data dependent features improve the overall result. Figure 2 shows the precision, accuracy and recall for classifying English tweets using the different sets of features. The results are very promising - with high accuracy, precision and recall.

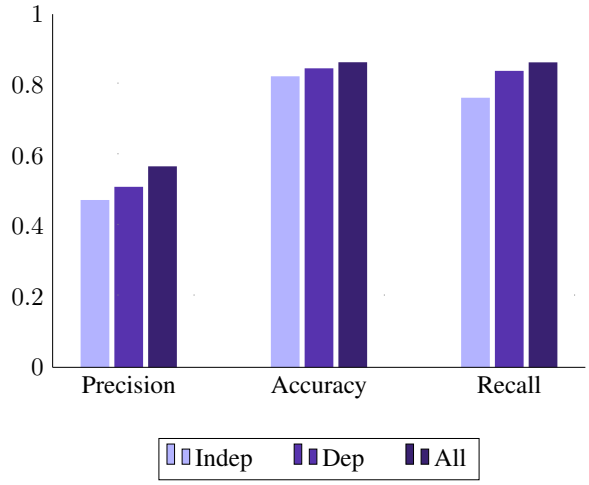

Fig. 1. Results for classifying Arabic tweets

VI. DISCUSSION

In our experiments we use both data dependent features and data independent features to get an understanding of how well our approach work and would work in a real setting. Most previous work has used data dependent features while we in this work investigate to what extent data independent features can be used.

Our models work very well for classifying English tweeps and English tweets. As can be seen in our experimental results, our models performs significantly worse on Arabic data. This is the case for both tweeps and individual tweets but for the individual tweets the results are increasing dramatically.

When analyzing the most important features for classifying English tweeps we can see that when data dependent fea-

Data dep. features	Feature class	Data ind. features	Feature class	Data ind. + dep. features	Feature class
islamic	Frequent words	word length 17	Word length	islamic	Frequent words
islamic state	Word bigrams	a	Letters	#is	Hashtags
#is	Hashtags	i	Letters	islamic state	Word bigrams
#syria	Hashtags	-	Punctuation	-	Punctuation
#islam	Hashtags	n	Letters	#syria	Hashtags
soldiers	Frequent words	word length 14	Word length	al	Letter bigrams
#iraq	Hashtags	s	Letters	i	Letters
al	Letter bigrams	the	Function words	#iraq	Hashtags
the	Frequent words	of	Function words	#islam	Hashtags
#islamic_state	Hashtags	as	Function words	soldiers	Frequent words

TABLE IX. MOST IMPORTANT FEATURES FOR CLASSIFYING ENGLISH TWEEPS

Arabic			English		
Data dep. features	Data ind. features	Data ind. + dep. features	Data dep. features	Data ind. features	Data ind. + dep. features
which	and	on	islamic state	its	islamic state
ha	(	about	support	need	isis
years	what	today	for the	so	the islamic
eboumn	today	that	of the	it	#iraq
sheikh	)	)	their	their	they
God	after	’	in the	an	”
it	it was	with	to the	about	for
no	he	after	#isis	me	#islam

TABLE X. MOST IMPORTANT FEATURES FOR CLASSIFYING ENGLISH AND ARABIC TWEETS

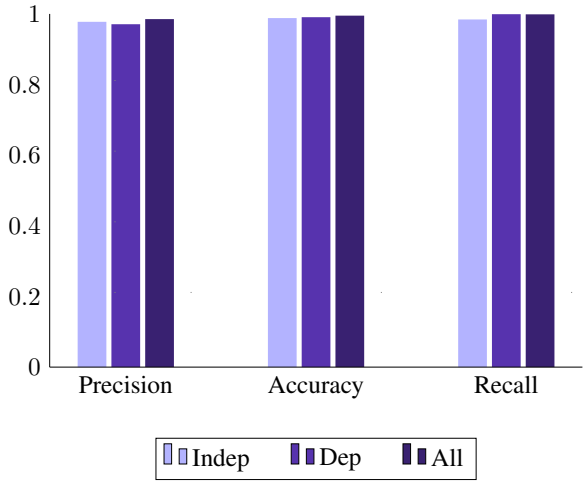

Fig. 2. Results for classifying English tweets

tures are used words like *islamic, soldiers, islam and islamic state* together with hashtags like *#is, #syria #islam, #iraq, #islamic_state* are important features. For data independent features like word length and letters a and s together with some function words are among the top 10 most important features. The top 10 most important features for classifying English tweeps are shown in Table IX. The most important feature classes for the Arabic tweeps when using data independent features are function words (*on, after and from*), word length (9, 11, and 1) and letters. When using a combination of data dependent and independent features letter bigrams, a hashtag (referring to the Caliphate state), letters and most frequent words (*where, visit, except*) are among the top 12 features. What is noticeable is that the data dependent features does not include any words relating to islam, islamic state or soldiers like the most important features on English.

A similar observation can be seen on the most important features for classifying tweets. For English tweets the most important (data dependent) features includes words and hashtags like *islamic state, isis, the islamic, #iraq, #islam, #isis* and *abu*. For the Arabic tweets, some of the most important features are *on, about, that, today, with* and *after*. These features are all prepositions that appear frequently within Arabic texts. In Table X the most important features for both Arabic and English tweets are listed.

One possible explanation for much worse results on Arabic is the complexity of the Arabic language. Arabic is a very specific language, in particular in an orthodox-conservative Sunni Islamic environment, of which groups like ISIS and al-Qa’ida claim to be part of. The ideology of Sunni extremism that we refer to as jihadist or jihadism is dominated by Arabic and native Arabs who have crafted the very foundation (in Arabic: *al-Qa’ida*) of today’s framework employed by ISIS. Users and sympathizers in general of ISIS as much as al-Qa’ida need to be initiated, by understanding and getting to know the ideology, that is in great parts based on interpretations of specific parts of religious scripture and thus employs elements of - for example - Islamic law (*fiqh*). Thus, the readers and consumers of audio-visual content must be rooted within the mainly written ideology to fully understand how the implementation of ideology functions, what the theological references are and how actions are sanctioned - even when these are not further explained in the videos, micro and macro codes are conveyed that are clearly perceived by the more initiated sympathizers and online followers. Tweeps can just refer to one word in Arabic to reference a complex theological concept, that is enriched by providing a picture of a mujahid or subject of the ”Islamic State” to underline the true re-enactment of divine scripture within the self-proclaimed Caliphate. The visual layer cannot be assessed without being rooted in the Arabic texts of jihadism, which originate in the 1980s in Afghanistan and have since likewise rapidly grown and reached out. To fully understand such individual key words that imply greater and in-depth nodes of the ideology, the key words have to be read

and examined hermeneutically - within the respective tweet and it's framework as well as within the overall corpus of jihadist ideology.

In a real world setting, the ratio between tweeps that are involved in media mujahideen compared to tweeps that has nothing to do with media mujahideen would be much lower than in our experiments. This fact is something that should be considered in future work.

VII. Conclusions and Future work

In this paper we have used machine learning to classify tweeps and tweets as being multipliers of jihadism. When using machine learning there is always the risk that the models that are built only are applicable on the specific dataset. This is something that we have tried to consider by using both data dependent and data independent features in our experiments. To get an understanding of how well this kind of classification would work in a real scenario more experiments on different datasets need to be done. In our experiments we have used a very limited set of tweeps and therefore it is hard to say anything about how the results would work in a realistic scenario. However, the results are promising and shows that further investigations should be done.

There are many possible directions for future work. One direction is to try our approach on more complex datasets and more realistic scenarios. Another direction is to improve the classification results for Arabic tweeps and tweets. As can be noted in our experiments, the results are much worse for Arabic and there is most likely room for improving the results. Since Arabic is such a deep rooted and rich language, with the need of hermeneutical analysis, the human component to understand, grade and priorities key words and features that should be used for classification is probably a way forward towards improving the results.

One feature that could be included in our models is the ageing factor (AF). It has been observed that radical twitter posts have a very low AF. The ageing factor measures how fast a tweet was re-tweeted in a period of time and is defined as follows [16]:

$$AF = \sqrt[i]{\frac{k}{k+l}}$$

where i is some cut-off time in hours, k is the number of re-tweets originating at least i hours after the original tweet and l is the number of re-tweets originating less than i hours after the original tweet.

A low AF value suggests by [16] that the topic is a short-term trending topic while a high value of the AF indicates that the topic is a sustainable topic since people have re-tweeted and discussed the tweet over a longer duration. The one hour ageing factor ($1hAF$) is the ratio of re-tweets in a sample set that originated more than one hour after the original creation time over the total number of re-tweets in the sample set.

The ageing factor plays an important role in our work due to the strategies that jihadist groups use to promote and promulgate messages. When Twitter is used to distribute radical content the network quickly reacts to a tweet and re-tweets the message within short time. In some of the dataset that we have used in our experiments, the average $1hAF$ factor for a tweets is 0.06. This indicates that messages are re-tweeted quickly. We believe that it will be particularly interesting to incorporate this known idiosyncrasy into our feature sets - especially for Arabic tweets and tweeps.

Finally, a direction for future work is also to extend our work to other kinds of social media and other violent extremist ideologies.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] A. Zelin, "The state of global jihad online. a qualitative, quantitative, and cross-lingual analysis," in *New America Foundation*, 2013.

[6] N. Prucha, *Jihadist Innovation and Learning by Adapting to the New and Social Media Zeitgeist*. Taylor & Francis, 2015. [Online]. Available: https://books.google.se/books?id=Q2ShCAAAQBAJ

[7] A. Abbasi, "Affect intensity analysis of dark web forums," in *Intelligence and Security Informatics, 2007 IEEE*, May 2007, pp. 282–288.

[8] H. Chen, "Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet." in *ISI*. IEEE, 2008, pp. 104–109.

[9] 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.

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

[11] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, "Detecting jihadist messages on twitter," in *European Intelligence and Security Informatics Conference (EISIC)*, 2015.

[12] 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 Symposium on Security and Privacy (SP)*, may 2012, pp. 300 –314.

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

[14] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *In Proceedings of the thirteenth International Conference on Machine Learning.*, 1996, pp. 148–156.

[15] M. Culp, K. Johnson, and G. Michailides, "ada: An r package for stochastic boosting," *Journal of Statistical Software*, vol. 17, no. 2, pp. 1–27, 2007.

[16] V. Uren and A.-S. Dadzie, "Nerding out on twitter: Fun, patriotism and #curiosity," in *Proc. of the 22Nd Int. Conf. on World Wide Web Companion*, 2013, pp. 605–612.