A Semantic Graph-based Approach for Radicalisation Detection on Social Media

Hassan Saif,[1] Thomas Dickinson,[1] Leon Kastler,[2] Miriam Fernandez,[1] and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, United Kingdom
{h.saif, thomas.dickinson, m.fernandez, h.alani}@open.ac.uk
[2] University of Koblenz Landau, Germany
lkastler@uni-koblenz.de

Abstract. From its start, the so-called Islamic State of Iraq and the Levant (ISIL/ISIS) has been successfully exploiting social media networks, most notoriously Twitter, to promote its propaganda and recruit new members, resulting in thousands of social media users adopting a pro-ISIS stance every year. Automatic identification of pro-ISIS users on social media has, thus, become the centre of interest for various governmental and research organisations. In this paper we propose a semantic graph-based approach for radicalisation detection on Twitter. Unlike previous works, which mainly rely on the lexical representation of the content published by Twitter users, our approach extracts and makes use of the underlying semantics of words exhibited by these users to identify their pro/anti-ISIS stances. Our results show that classifiers trained from semantic features outperform those trained from lexical, sentiment, topic and network features by 7.8% on average F1-measure.

Keywords: Radicalisation Detection, Semantics, Feature Engineering, Twitter

1 Introduction

Traditionally, the process of radicalisation has occurred directly, person to person. However, in the age of social media platforms and access to the Internet, this process has moved to a virtual sphere where terrorist organisations use 21st century technology to promote their ideology and recruit individuals. Particularly, the so-called Islamic State of Iraq and the Levant (ISIL/ISIS) is one of the leading terrorist organisations on the use of social media to share their propaganda, raise money and radicalise and recruit individuals. According to a 2015 U.S government report[3] this organisation has lured more than 25,000 foreigners to fight in Syria and Iraq, including 4,500 from Europe and North America.

Aiming to hinder ISIS recruiting efforts via social media, researchers, governments and organisations are actively working on identifying ISIS-linked or ISIS-supporting social media accounts. A popular example was the campaign launched by the hacker community Anonymous as a response to the Paris attacks,[4] where they claimed taking

[3] https://homeland.house.gov/wp-content/uploads/2015/09/TaskForceFinalReport.pdf
[4] https://en.wikipedia.org/wiki/November_2015_Paris_attacks

down more than 20,000 Twitter accounts linked to ISIS. A key criticism received by this initiative was the wrong categorisation as pro-ISIS of multiple Twitter accounts, including the ones of the U.S president Barack Obama, and the one of BBC news.[5] While it is unclear the strategy used by Anonymous to identify these accounts, this incident emphasises the difficulty and sensitivity of the problem at hand.

Current research works that have aimed to analyse radicalisation and pro-ISIS stances of social media users mainly rely on features extracted from the lexical representation of words (e.g., word n-grams, topics, sentiment), or from the online profile of users (e.g.,network features). While effective, these approaches provide limited capabilities to grasp and exploit the conceptualisations involved in content meanings. This involves limitations such as the inability to capture relations between terms (countries *attacking* ISIS vs. countries *attacked* by ISIS), or the weakness to properly capture contextual information by understanding which groups of terms co-occur together and how they relate to one another. The above limitations constitute a problem when trying to discriminate the stance expressed by users in social media. We therefore hypothesise that, by exploiting the latent semantics of words expressed in tweets, we could identify additional pro-ISIS and anti-ISIS signals that will complement and enhance the ones extracted by previous approaches.

Starting from this position, this paper investigates the use of ontologies and knowledge bases to support a graph-based analysis of tweets' content. Entities are extracted from the tweets of users' timelines (e.g. "*ISIS*", "*Syria*", "*United Nations*") and expanded with their corresponding semantic concepts (e.g. "*Jihadist_Group*", "*Country*", "*Organisation*") and relations (*e.g., Military_intervention_against_ISIL, place, Syria*) by using DBpedia[6]. Frequent sub-graph mining is applied over the extracted semantic graphs to capture patterns of semantic relations that help discriminating the radicalisation stances of users. These patterns are then used as features (so-called *semantic features* in our work) for detecting the radicalisation stances of users on Twitter.

The effectiveness of semantic features to identify pro-ISIS and anti-ISIS stances is compared against several baseline features, particularly unigram features, sentiment features, topic features and network features. This comparison is performed by creating classifiers, based on the different sets of features, from a training dataset of 1,132 European Twitter users equally divided in pro-ISIS and anti-ISIS. Our results show how classifiers trained with semantic features outperform the baselines by 7.8% on average F1-measure, showing a positive impact on the use of semantic information to identify pro and anti ISIS stances. An additional analysis is performed over the data to identify signals of radicalisation. Our results show that pro-ISIS users' discussions tend to mention entities and relations focused on religion, historical events and ethnicity, such as "*Allah*", "*Prophet*", "*(Mohamed, ethnicity, Arab)*", while anti-ISIS users' discussions tend to focus more around politics, geographical locations, and interventions against ISIS (*e.g., Military_intervention_against_ISIL, place, Syria*). Anti-ISIS users tend to mention the entity ISIS with a higher frequency than pro-ISIS users.

[5] `http://www.bbc.co.uk/newsbeat/article/34919781/anonymous-anti-islamic-state-list-features-obama-and-bbc-news`

[6] `http://dbpedia.org`

The rest of the paper is structured as follows: Section 2 assesses the related work in the areas of radicalisation studies. Section 3 describes our graph-based approach for radicalisation detection. Section 4 describes our experimental setup, including the dataset used for the analysis and the baseline features selected for comparison. Section 5 reports the results of comparing semantic features against the baselines. Section 6 discusses our identified pro- and anti-ISIS signals. Discussions and conclusions of this work are reported in Sections 7 and 8 respectively.

2 Related Work

Understanding how an individual becomes radicalised online has become a burgeoning, albeit relatively-new, topic of research, and has recently focused on the role that social media plays in radicalisation. Existing works in this space have spanned multiple research domains (social science, psychology, computer science) and have often sought to understand the *pathway* to radicalisation: (i) picking out key signifiers of increasingly radicalised behaviour [2] (e.g. distribution of jihad videos); (ii) defining pathway models of the stages towards radicalisation (e.g. isolation, disillusionment, anger, etc.) [9], and; (iii) the process used by those radicalised to recruit others [4, 8, 18]. Indeed, in our own prior work [14], by Rowe and Saif, we found that users adopted radicalised rhetoric from other users with whom they shared common interactions and connections - suggesting a potential *nascent* community of influence.

Moving away from more general studies of *how* radicalisation occurs, towards *predicting* who will become radicalised has been the focus of several recent works. For instance, although O'Callaghan et al. [11] did not necessary *label* Twitter users as being pro or anti-ISIS (as we do in this paper), the authors instead clustered Twitter users collected from Twitter lists related to the Syria conflict into high-modularity clusters. The authors subsequently identified a cluster of '*jihadist*' users, which contained those who support ISIS. Inspection of the videos shared by users in that cluster found that videos were often shared from YouTube channels related to ISIS, the Nusra Front, and Aleppo (a key city that has been under ISIS control). In a more direct approach, Berger and Morgan [3] collected 90K ISIS supporters, manually, from Twitter and then induced a machine learning model (it is not clear which) to differentiate between pro and anti-ISIS supporters. Using this approach, the authors found that pro-ISIS supporters could be accurately predicted ($\sim 94\%$ accuracy) from their profile descriptions' terms alone: with keywords such as succession, linger, Islamic State, Caliphate State or In Iraq being key indicators of ISIS-support. Similarly, Magdy et al. [10] were able to accurately (87% F1) differentiate between pro and anti-ISIS users, finding that ISIS supporters talked a lot more about the Arab Spring than ISIS opponents. Users were defined from a collection of Arabic Tweets as pro-ISIS if they used '*Islamic State*' more, and anti-ISIS is they used '*ISIS*' more.

Unlike the above reviewed works, this paper investigates the role of semantics in classifying users as pro or anti-ISIS - as opposed to using terms in users' profile descriptions [3] or terms alone [10]. In carrying out our work, our results empirically validate the utility of semantics (i.e. semantic concepts, entities and relations) over merely unigrams.

3 Semantic Graph-based Approach for Pro-ISIS Stance Detection

In this section we describe our semantic graph-based approach for detecting pro-ISIS stances on Twitter. As discussed before, the discriminative power of features used for radicalisation detection often relies on the latent semantic interdependencies that exist between certain words in tweets. As such, the proposed approach aims to extract and use such interdependencies and relations to learn patterns of radicalisation.

The proposed semantic graph-based approach breaks down into four main steps, as depicted in Figure 1: (1) extract named entities and their semantic concepts in tweets, (2) build a semantic graph per user representing the concepts and semantic relations extracted from her posted content, (3) apply frequent sub-graph mining on the semantic graphs to capture patterns of semantic relations that discriminatingly characterise the radicalisation stances of users, and lastly (4) use the extracted patterns as features for radicalisation classifier training. Our approach uses a dataset of 1,132 European Twitter users (together with their timelines) equally divided in pro-ISIS and anti-ISIS. This dataset is further described in Section 4.1.

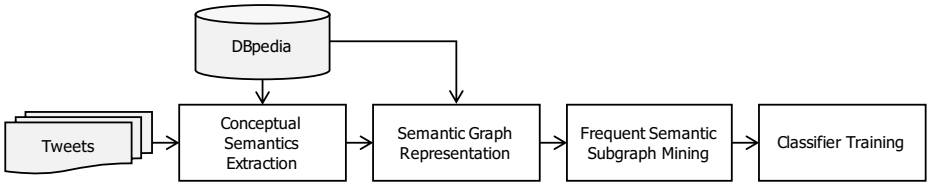


Fig. 1: Pipe of detecting pro-ISIS stances using semantic sub-graph mining-based feature extraction

Step 1. Conceptual Semantics Extraction Given a training set, consisting on labelled (pro-ISIS, anti-ISIS) users' timelines, this steps extracts named-entities from the tweets of the users' timelines (e.g. `ISIL`, `Syria`, `Al-Baghdadi`[7]) and expands them with their corresponding semantic concepts (e.g. `Jihadist_Group`, `Country`, `Leader`). The semantic extraction tool AlchemyAPI[8] is used for this purpose due to its accuracy and high coverage of semantic types and subtypes in comparison with other semantic extraction services [13, 15]. Table 1 lists the total number of unique entities and concepts and the top 10 frequent entities and concepts, extracted from our dataset, for both pro-ISIS and anti-ISIS user accounts. Visible differences can be observed within these top 10 entities and concepts.

Step 2. Semantic Graph Representation The second step in our approach aims to extract the sets of semantic relations for every pair of named-entities (e.g., `Syria`, `ISIL`) co-occurring together in tweets, and represent these relations as graph structures. For the purpose of our study we extract semantic relations using the approach proposed by Pirro [12] over DBpedia, since DBpedia is a large generic knowledge graph which captures a high variety of relations between terms. To extract the set of relations between

[7] Abu Bakr Al-Baghdadi is the leader of the Islamic State in Iraq and Syria `http://dbpedia.org/page/Abu_Bakr_al-Baghdadi`

[8] `http://www.alchemyapi.com/`

	pro-ISIS		anti-ISIS	
No. of Unique Entities	32,406		30,206	
No. of Unique Concepts	35		36	
Top 10 Frequent Entities & their Concepts	**Entity**	**Concept**	**Entity**	**Concept**
	MSNBC	Company	BBC	Company
	Iraq	Country	UK	Country
	Allah	Person	Kobane	City
	America	Continent	London	City
	Muslim	Person	ISIS	Organisation
	Officer	JobTitle	Syria	Country
	Wounds	HealthCondition	Europe	Continent
	Syria	Country	Iran	Country
	WAPO	PrintMedia	Kurdish	Person
	Israel	Country	Police	Organisation

Table 1: Total number and top 10 frequent entities and their associated semantic concepts extracted from our dataset.

two named entities this approach takes as input the identifiers (i.e., URIs) of the source entity e_s, the target entity e_t and an integer value K that determines the maximum path length of the relations between the two named entities. The output is a set of SPARQL queries that enable the retrieval of paths of length at most K connecting e_s and e_t. Note that in order to extract all the paths, all the combinations of ingoing/outgoing edges must be considered. For example, if we were interested in finding paths of length $K <= 2$ connecting $e_s = Syria$ and $e_t = ISIL$ our approach will consider the following set of SPARQL queries:

SELECT * WHERE {:Syria ?p1 :ISIL}
SELECT * WHERE {:ISIL ?p1 :Syria}
SELECT * WHERE {:Syria ?p1 ?n1. ?n1 ?p2 :ISIL}
SELECT * WHERE {:Syria ?p1 ?n1. :ISIL ?p2 ?n1}
SELECT * WHERE {?n1 ?p1 :Syria. :ISIL ?p2 ?n1}
SELECT * WHERE {?n1 ?p1 :Syria. ?n1 ?p2 :ISIL}

As it can be observed, the first two queries consider paths of length one. Since a path may exist in two directions, two queries are required. The retrieval of paths of length 2 requires 4 queries. In general, given a value K, to retrieve paths of length K, 2^k queries are required. Figure 2 shows an example of the semantic relations for the entities `Syria` and `ISIL`. As can be noted, these two entities are either connected via a direct semantic relation (e.g., $ISIL < headquarters > Syria$) or via linking nodes (e.g., $ISIL < ideology > Pan - Islam < ideology > MuslimsBrotherhood < location > Syria$)

Once we have the entities' pairwise semantic relations (i.e., paths) extracted from the users' timelines, we represent these relations for each user as a directed graph $G = (V, E)$ comprising a set V vertices of co-occurring entities with a set of edges E denoting the semantic relations between these entities.

Step 3. Frequent Patterns Mining In this step we apply frequent pattern mining to the users' semantic graphs which we extracted in step 2. As mentioned earlier, the goal behind this approach is to find patterns of similar semantics among pro- and anti-ISIS users, which can help characterise their stances. To this end, we apply frequent pattern

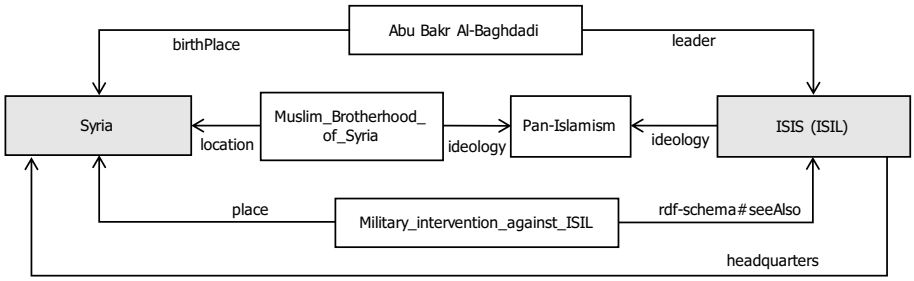


Fig. 2: Example for semantic relations between the entities Syria and ISIS with a path length of 3

mining to the users' semantic graphs and extract the sub-graphs appearing more than n times, where n is set to 2 in our experiments. This allows us to identify as many sub-graphs as possible, without returning a user's entire graph.

To mine these frequent sub-graphs we use CloseGraph [20], which performs an exhaustive sub-graph search, returning all frequent closed graphs within our dataset. We use the Parallel and Sequential Mining Suites (ParSeMiS[9]) implementation of this algorithm. Applying the aforementioned algorithm to the users' semantic graphs results in 187 unique sub-graphs for pro-ISIS users and 723 unique sub-graphs for anti-ISIS users in total, with the top frequent sub-graph appearing more than 500 times.

Figure 6 depicts three of the top most discriminative semantic sub-graphs mined from pro-ISIS and anti-ISIS users. These sub-graphs differ in the underlying semantics represented by the entities and the semantic relations. While the pro-ISIS sub-graphs (Figure 6:a) denote entities and relations around historical and religious topics, e.g., "`Muhammad <religion> Islam`", the anti-ISIS sub-graphs (Figure 6:b) denote entities and relations around key military interventions and geographical locations, e.g., "`Military_intervention_against_ISIL <place> Syria`". Further details on the analysis of these sub-graphs are provided in Section 6.

Step 4. Classifier Training This step takes as input a training set $\mathcal{T}^{train} = \{(\mathbf{U}_n; c_n) \in \mathcal{U} \times \mathcal{C} : 1 \leq n \leq N^{train}\}$ of users U in our dataset along with their class labels $\mathcal{C}$ = {pro-ISIS, anti-ISIS}). After that, it constructs for each user $U \in \mathbf{U}$ a semantic vector $\mathbf{v}_{us} = (e_1, e_2, ..., e_l, s_1, s_2, ..., s_m, b_1, b_2, ..., b_j)$ as the joined vector of entities $\mathbf{e} = (e_1, e_2, ..., e_l)$, concepts $\mathbf{s} = (s_1, s_2, ..., s_m)$, and $\mathbf{b} = (b_1, b_2, ..., b_j)$ semantic sub-graphs (patterns) extracted from the user's timelines as explained in the previous steps. The generated semantic vectors are then used to train multiple machine learning classifiers. SVM was selected in our experiments as the best performing one. Further details on the creation of this classifier are provided in Section 4.3.

4 Experimental Setup

Our proposed approach, as shown in the previous section, extracts frequent patterns of semantics commonly expressed by users of a certain radicalised stance. We assess the extracted patterns by using them as features to train supervised classifiers for user-level

[9] `https://www2.informatik.uni-erlangen.de/EN/research/zold/ParSeMiS/index.html`

radicalisation classification, i.e., classifying users in our dataset according to their stance as pro-ISIS or anti-ISIS. Hence, our experimental setup requires the selection of (i) an annotated dataset of Twitter users (pro-ISIS and anti-ISIS) together with their timelines, (i) baseline features for cross-comparison and (ii) a supervised classification method. These elements are explained in the following subsections.

4.1 Dataset of pro-ISIS and anti-ISIS Twitter users

Our approach relies on a training dataset of $1,132$ European Twitter users (together with their timelines) collected in our previous work [14]. In this work the pro-ISIS stance of 727 Twitter users was determined based on their sharing of incitement material from known pro-ISIS accounts and on their use of extremist language. By the time of conducting this research, 161 of these Twitter accounts were suspended or changed the privacy to *protected*, preventing us from accessing their profile information. As such, we resorted to remove them from the original set, resulting in 566 pro-ISIS users in total. To balance our dataset, we added 566 anti-ISIS users, whose stance is determined by the use of anti-ISIS rhetoric. Table 2 shows the total number, and distribution of tweets and words for each user group. As we can observe, both the number of tweets and words for anti-ISIS users are significantly higher than the ones for pro-ISIS users. We refer the reader to the body of our work [14] for more details about the construction and annotation of this dataset.

	pro-ISIS Users	**anti-ISIS Users**
Total number of Tweets	602,511	1,368,827
Average Number of Tweets per User	1,065	2,418
Total number of Words	3,945,815	9,375,841
Average Number of Words per User	6,971	16,570

Table 2: Statistics of the Twitter dataset used for evaluation

4.2 Baseline Features

Unigrams Features: Word unigrams are features traditionally used for various classification tasks of tweets data. For example, in the context of a sentiment analysis task, models trained from word unigrams were shown to outperform random classifiers by 20%. [1] We generate the user's unigram vector $\mathbf{t}_{uunig}$ as the vector $\boldsymbol{t}_{uunig} = (w_1, w_2, ..., w_m)$ of the words in his timeline. Note that stopwords, non-English words and special characters are removed from the timeline prior to building $\boldsymbol{t}_{uunig}$ in order to reduce its dimensionality.

Sentiment Features: Sentiment features denote the sentiment orientation (positive, negative, neutral) of users in our dataset. The rational behind using these features is that the sentiment conveyed by the users' posts may help discriminating between pro- and anti-ISIS stances. To extract these features for a given user u, we first extracted the sentiment orientation of each tweet in the user's timeline. To this end, we used SentiStrength [17], a lexicon-based sentiment detection method for the social web. To construct the sentiment vector $\boldsymbol{t}_{usentiment}$ for user u, we augment the unigrams feature vector $\boldsymbol{t}_{uunig}$ with the extracted sentiment orientation of tweets as: $\boldsymbol{t}_{usentiment} = (w_1, w_2, ..., w_m, p_{pos}, p_{neg}, p_{neu})$, where p_{pos}, p_{neg} and p_{neu} are the

numbers of positive, negative and neutral posts in the user's timeline. Note that, due to the low dimensionality of sentiment features, the sentiment vector for each user is constructed as a combination of ngrams and sentiment attributes.

Topic Features: Topic features denote the latent topics extracted from tweets using the probabilistic generative model, LDA [6]. LDA assumes that a document is a mixture of topics and that each topic is a mixture of probabilities of words that are more likely to co-occur together under the topic. For example the topic "ISIS" is more likely to generate words like "behead" and "terrorism". Therefore, LDA topics represent groups of words that are contextually related. To extract these latent topics from our dataset we use an implementation of LDA provided by Mallet.[10] The topic feature vector for user u is constructed as $\boldsymbol{t}_{utopic} = (t_1, t_2, ..., t_k)$, where $t_i \in \boldsymbol{t}_{utopic}$ represents a topic extracted from the user's timeline. It is worth noting that LDA requires defining the number of topics to extract before applying it on the data. To this end, we ran LDA with different choices of numbers of topics between 5 and 10,000. We trained a SVM classifier from the features extracted by each of these choices and measured the classification performance in F1-measure. The best performance 82.9% F1 is reached with 5,000 topics, which is the number of topics used in our analysis.

Network Features: Network features refer to the profile information/attributes of Twitter users.[11] This includes: *number of followers*, *number of followee*, *number of hashtags*, *number of mentions* (i.e., @user), *favourites count*, *status count*, *profile description*, and *geographic location*. The notion behind using these features for radicalisation detection is that users of a certain radicalisation stance are more likely to interact with other users of the same stance than with users from a different stance, as we discussed in our previous work [14]. The network feature vector for user u is constructed as $\boldsymbol{t}_{unetwork} = (n_1, n_2, ..., n_l)$, where $n_i \in \boldsymbol{t}_{unetwork}$ represents an attribute derived from the user's profile.

4.3 Classification Method

We tested several machine learning classifiers with our semantic features including Naive Bayes, Maximum Entropy, and SVMs with linear, polynomial, sigmoid, and RBF kernels. The classifiers were tested using 10-fold cross validation over 30 runs. Results showed that SVM with RBF kernel produced the highest and most consistent performance in accuracy and F1-measure among all the other classification methods. Section 5 reports performance using this classifier. Note that, to generate these classifiers, we perform a feature selection process on the 4 baseline feature sets, as well as our semantic features, by excluding features with low discrimination power. To this end, we use Information Gain (IG) [7] to compute the discriminative score of features in each feature set and filter out those with low scores (IG $\approx$ 0) from the feature space.[12]

Figure 3 shows the original number of features under each feature set and the impact of feature selection using the IG method on them. Here we notice a reduction rate of 88% for all feature sets on average. The highest reduction rate of 97% is achieved on the semantic features, reducing the number of features from 346,512 (considering entities,

[10] http://mallet.cs.umass.edu/

[11] https://dev.twitter.com/overview/api/users

[12] IG measures the decrease in entropy when the feature is given vs. absent [21]

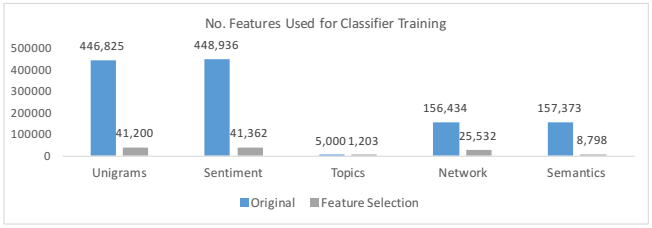


Fig. 3: Number of features used for classification with and without feature selection

concepts and semantic sub-graphs) to 8,429 only. This indicates that pro- and anti-ISIS users share a high degree of terminology and semantics when posting in Twitter.

5 Evaluation Results

In this section, we report the results obtained from using the proposed semantic features for user-level radicalisation classification. Our baselines of comparison are SVM classifiers trained from the 4 sets of features described in Section 4.2. Results in all experiments are computed using 10-fold cross validation over 10 runs of different random splits of the data to test their significance. Statistical significance is done using *Wilcoxon signed-rank test* [16]. Note that all the results in average Precision, Recall and F1-measure reported in this section are statistically significant with $\rho < 0.001$.

Table 3 shows the results of our binary stance classification (pro-ISIS vs. anti-ISIS) using *Unigrams*, *Sentiment*, *Topic*, and *Semantic* features after feature selection, applied over the 1,132 users in our dataset. The table reports three sets of precision (P), recall (R), and F1-measure (F1), one for anti-ISIS stance identification, one for pro-ISIS stance identification, and the third shows the averages of the two. The table also reports the total number of features used for classification under each feature set.

		anti-ISIS			pro-ISIS			Average		
	No. of Features	P	R	$F1$	P	R	$F1$	P	R	$F1$
UNIGRAMS	41,200	0.814	0.919	0.863	0.907	0.79	0.844	0.86	0.854	0.854
SENTIMENT	41,362	0.814	0.919	0.863	0.907	0.79	0.844	0.86	0.854	0.854
TOPICS	992	0.771	0.943	0.848	0.927	0.719	0.81	0.849	0.831	0.829
NETWORK	25,532	0.897	0.827	0.86	0.839	0.905	0.871	0.868	0.866	0.866
SEMANTICS	8,798	0.994	0.852	0.917	0.87	0.995	0.928	0.932	0.923	0.923

Table 3: Classification performance of the five feature sets with IG feature selection. The values highlighted in grey correspond to the best results obtained for each feature. Results in average P, R and F1 are statistically significant with $\rho < 0.001$.

According to the results presented in Table 3, the proposed Semantic features outperform the 4 baseline feature sets in all average measures by a large margin. In particular, classifiers trained from Semantic features produce 7.8% higher Recall, 7.7% higher precision, and 7.82% higher F1 than all baselines on average. Network features come next, followed by Unigrams features, with approximately 87% and 85% in average F1 respectively. On the other hand, Topic features produce the lowest classification

performance with 82.9% in average F1. We also notice that sentiment features, which consist of both, word unigrams and their sentiment (Section 4.2), have no impact on the classification performance compared to using Unigrams only.

As for per-stance classification performance, we observe that Unigrams, Sentiment and Topic features produce higher performances on detecting anti-ISIS stance than pro-ISIS stance. For example, the F1 produced by Unigrams when identifying anti-ISIS stance is 2.2% higher than the F1 produced when identifying pro-ISIS stance. This might be due to the imbalanced distribution of words and tweets in both classes. As described in Section 4.1, the number word unigrams in anti-ISIS users' timelines is ≈ 2.5 the number of those in pro-ISIS users' timelines.

On the other hand, classifiers trained from either Network or Semantic features seem to be more tolerant to the imbalanced distribution of words in our dataset. Specifically, the performance of both, anti-ISIS and pro-ISIS classification becomes more consistent, with $\approx 1\%$ difference in F1 only when using Network or Semantic features.

The above results show the effectiveness of using semantic features for radicalisation classification of users on Twitter, substantially improving per-stance, as well as overall, classification performance. It is worth noting that our results here are directly comparable to prior work of Magdy et al. [10]. However, while in this work the authors used unigrams alone to identify pro and anti-ISIS users, our work shows the enhancement obtained by using semantics for this task.

6 Signals of Radicalisation (pro-ISIS vs. anti-ISIS)

In the previous sections we showed how to detect radicalisation stances of users on Twitter and investigated which features help achieving higher performance levels. In this section, we reflect on the semantics expressed by both, pro- and anti-ISIS users, aiming at finding signals of radicalisation or anti-radicalisation in their timelines.

To this purpose we look for possible variations in the types of semantics in both, pro-ISIS and anti-ISIS users and study whether such variations indicate radicalisation or anti-radicalisation stances. We compute the frequency distribution of the entities, concepts and semantic sub-graphs that pro-ISIS and anti-ISIS users adopt in their tweets and we discard entities with zero information gain score (IG ≈ 0), since they have no discrimination power for identifying pro- and anti-ISIS stances, as discussed in Section 4.3. We also compute the sentiment of each entity by taking the average sentiment of the tweets where the entity is mentioned in. Looking at the entities and concepts used by pro-ISIS users (Figure 4(a)) we observe that the majority of discriminative entities found within pro-ISIS users' discussions focus on religion, with many positively mentioned entities such as "`Allah`", "`Prophet`", and "`Khilafah`" (Khilafat). On the other hand, the most discriminative entities and concepts found within anti-ISIS users' discussions focus on locations and politics, with entities like "`Kurdistan`", "`Europe`" and "`Putin`". We can also observe several common entities between pro- and anti-ISIS users. Some of these entities are perceived positively by both groups (e.g., "`Hillary`", "`Gulf`", "`China`") although the vast majority are commonly mentioned with negative sentiment (e.g., "`Assad`", "`Syria`", "`Israel`")

Figure 5 shows the per-user distribution of six highly frequent and commonly mentioned entities by pro-ISIS and anti-ISIS users. We can notice that, although these words receive similar sentiment by users in both groups (see Figure 4), they have

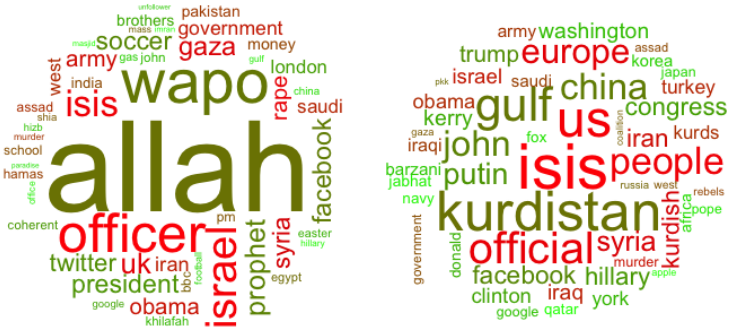


Fig. 4: Word clouds of the top-50 named-entities published by pro-ISIS and anti-ISIS users, the colour indicates the sentiment attached to the entity - with red being negative, and green being positive.

different frequency distributions. For example, an entity like "`Allah`" is used more frequently by pro-ISIS users (mean = 42.64) than anti-ISIS users (mean = 1.28). On the other hand anti-ISIS users tend to use the word "`ISIS`" more frequently than pro-ISIS users, with a mean frequency of 46.88 for the former and 18.13 for the latter. While the sentiment of "`ISIS`" is negative for both groups, a manual analysis of the tweets containing "`ISIS`" reveals that pro-ISIS users generally refer positively to the term, although the context in which it is mention might be negative. For example, the tweet, "`If your goal is killing Abu Sayyaf then our goal is killing Obama and the worshipers of the cross. #ISIS`", shows support to "`ISIS`" but expresses negativeness towards Americans and Christians, and therefore, the tweet is categorised as negative. This comes inline with the work of Magdy et al. [10] (see Section 2), which reported that "`ISIS`', when mentioned in a negative context, often indicates anti-ISIS stance whereas it indicates pro-ISIS stance when it is mentioned in a positive context.

Figure 6 shows the three most discriminative sub-graphs for pro-ISIS (a) and anti-ISIS (b) users. As we can see in this figure pro-ISIS sub-graphs are formed of concepts and relations related to religion and ethnicity (Muhammad, Islam, Arabs), to relevant historical events (the invasion of Badr[13]) and to the United States and some of its relevant political figures, particularly Barack Obama (ex-president of the United States), and Cyrus Amir-Mokri, (Iranian-American, ex-Assistant Secretary for Financial Institutions at the U.S. Treasury Department, and reporter on the situation of Iran under Foreign Affairs[14]). On the other hand, sub-graphs for anti-ISIS users are particularly focused on the military interventions against ISIS, on key locations, such as Iraq and Syria, and also on relevant United States political figures. It is worth noting that while both, Barack Obama and Amir-Mokri, appear within the sub-graphs of pro- and anti-ISIS users, they do it in very different semantic contexts. In the sub-graphs extracted for pro-ISIS users

[13] `https://www.britannica.com/event/Battle\-of\-Badr`

[14] `https://www.foreignaffairs.com/articles/iran/2015-10-20/windfall-iran`

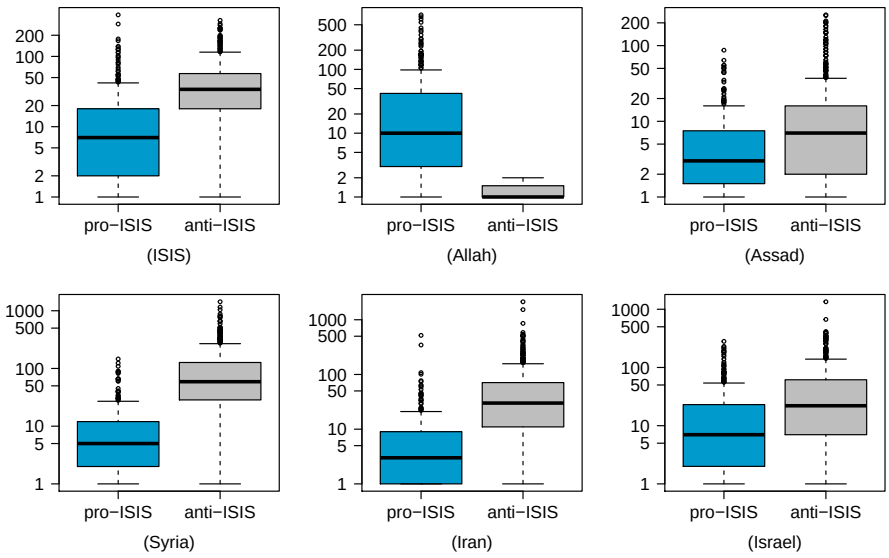


Fig. 5: Per-user distribution of the named-entities: *ISIS, Allah, Assad, Syria, Iran and Israel* for pro-ISIS and anti-ISIS users.

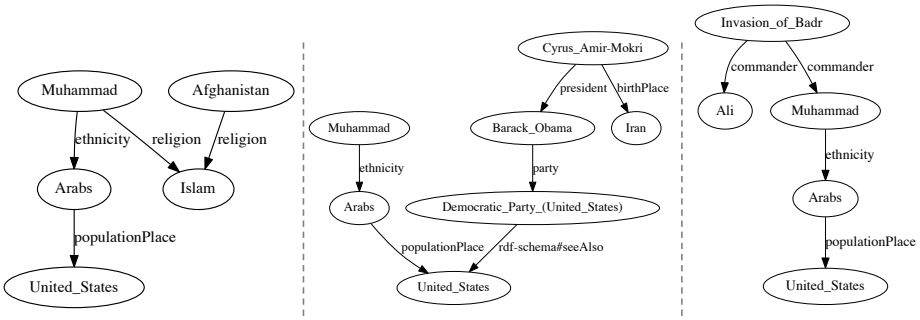
they are semantically related to religion and ethnicity (Muhamad, Arabs), while in the sub-graphs extracted for anti-ISIS users they are related with military interventions and media organisations (GEM_TV).

7 Discussion and Future Work

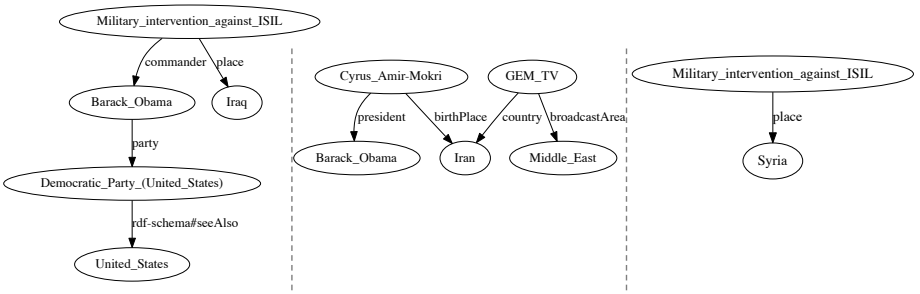
In this paper we demonstrated the value of using the semantics as features for identifying pro-ISIS and anti-ISIS stances of users on Twitter. This section discusses the limitations of our study as well as our directions for future work.

We experimented with a dataset of $1,132$ Twitter users. These users were annotated as pro-ISIS or anti-ISIS based on their sharing of content from known pro-ISIS accounts and on their use of pro-ISIS or anti-ISIS rhetoric [14]. Although the ratio of pro-ISIS users (727 pro-ISIS accounts were identified from a an original set of 154K accounts) is similar to the one identified in other works [3], the dataset is relatively sparse, which leads to question the *generality* of the obtained results. Our future work will therefore replicate the process described in this paper with similar datasets. It is also relevant to observe that word distribution in this dataset was found to be skewed towards anti-ISIS users, with approximately 58% more words found in anti-ISIS users' timelines (see Table 2). Although we reduced the impact of such bias by performing feature selection, our future work will study whether such differences in content are similar for other datasets. Since this dataset focuses on users based in Europe, anti-ISIS users may express their views more freely and actively than pro-ISIS users, which may lead to the variation in the amount and diversity of generated content.

While previous works focus on identifying pro-ISIS users mainly in Middle East [3] [10], we are interested in studying European radicalisation stances. In addition, as



(a) pro-ISIS sub-graphs



(b) anti-ISIS sub-graphs

Fig. 6: Example of 3 of the most discriminative semantic sub-graphs for (a) pro-ISIS and (b) anti-ISIS users in our Twitter dataset

opposed to these works, which include Arabic tweets in their analyses, we only processed English tweets. Note that extracting conceptual semantics for the Arabic language is a challenging task, with little research being done in this regard. As future work, we plan to explore different ways for extracting semantics from Arabic tweets and include them in our analysis.

As described in Section 3, our proposed semantic features include entities, concepts and semantic sub-graphs. We consider these three elements for completeness, since semantic sub-graphs might not be found in the timelines of all Twitter users. The generation of these sub-graphs depends on the richness of semantics existing within the users' content, the accuracy of AlchemyAPI to extract these semantics, and the coverage of the knowledge base, in this case DBpedia, to capture the semantic relations between the extracted entities.

To extract semantic relations from DBpedia our approach considers a maximum path length of two (see Section 3). While a larger maximum path length could increase the likelyhood of a semantic relation existing between two entities, as well as the amount of existing semantic relations, higher values of maximum path length come close to the diameter of the DBpedia graph itself, and may lead to an explosion in the number of extracted relationships.[15] Additionally, extracting semantic relations between

[15] The effective estimated diameter of DBpedia is 6.5082 edges. See `http://konect.uni-koblenz.de/networks/dbpedia-all`

a high number of entities via a SPARQL endpoint is a high-cost process [12]. Our implementation uses multithreading to enhance the performance (i.e., queries are sent in parallel), and relations are extracted once per Twitter dataset. These relations are stored to be reused for future experiments.

When performing sub-graph mining we need to consider scalability as well as redundancy issues (similar extracted sub-graphs). In our experiments, the number of training instances was limited, but a greater number of users, as well as larger graphs per user, may be difficult to scale. To deal with scalability our plan includes the use of the parallel processing [5]. Regarding redundancy, although our work already filters sub-graphs with low discrimination power based on IG, our plan for minimising redundancy also includes using compression techniques to cluster together sub-graphs with similar information [19].

When performing feature selection (see Section 4.3) we observed a reduction rate of 88% for all feature sets on average and 97% on the semantic features. This indicates that pro- and anti-ISIS users share a high degree of terminology and semantics when posting in Twitter. Our analysis (see Section 6) has therefore focused on analysing those key entities, concepts and semantic sub-graphs with higher discrimination power, i.e., those that are different among the two groups. Our future work aims to complement this analysis by investigating whether entities associated to particular semantic concepts (e.g., countries, organisations) have more discriminative power than other ones.

Although we use Twitter as a case study in our analysis, our approach is not tied to Twitter data. Room for future work is investigating the applicability and performance of our semantic features on other social media platforms such as Facebook and Instagram.

8 Conclusions

In this paper we proposed the use of the conceptual semantics of words for detecting pro-ISIS and anti-ISIS stances of users on social media. We used Twitter as case study of social media platforms, and investigated: (i) how semantic graphs can be created by extracting entities in tweets, together with their corresponding semantic concepts and relations, (ii) how frequent semantic sub-graphs can be mined from these graphs and, (iii) how entities, concepts and semantic sub-graphs can be used as features to train machine learning classifiers for stance detection of Twitter users.

We experimented with our semantic features on a Twitter dataset of $1,132$ pro-ISIS and anti-ISIS users and compared the performance of a SVM classifier trained from semantic features against classifiers trained from Unigrams, Sentiment, Topics, and Network features. We also studied the impact of feature selection on the performance of our classifiers and showed that, using the most discriminative semantic features in radicalisation classification improves performance by 7.8% F1 over the average performance of all baselines.

We performed an exploratory analysis on the variations of semantics and sentiment used by pro-ISIS and anti-ISIS users in our dataset and showed that pro-ISIS users tend to discuss about religion, historical events and ethnicity while anti-ISIS users focus more on politics, geographical locations and interventions against ISIS.

Acknowledgment

This work was supported by the EU H2020 projects COMRADES (grant no. 687847) and TRIVALENT (grant no. 740934)

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38 (2011)
2. Bartlett, J., Miller, C.: The edge of violence: Towards telling the difference between violent and non-violent radicalization. Terrorism and Political Violence 24(1), 1–21 (2012)
3. Berger, J., Morgan, J.: The isis twitter census: Defining and describing the population of isis supporters on twitter. The Brookings Project on US Relations with the Islamic World 3, 20 (2015)
4. Berger, J.M.: Tailored online interventions: The islamic state's recruitment strategy. Combatting Terrorism Center (2015)
5. Bhuiyan, M.A., Al Hasan, M.: Fsm-h: Frequent subgraph mining algorithm in hadoop. In: 2014 IEEE International Congress on Big Data. pp. 9–16. IEEE (2014)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)
7. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of machine learning research 3, 1289–1305 (2003)
8. Hall, J.: Canadian foreign fighters and isis (2015)
9. King, M., Taylor, D.M.: The radicalization of homegrown jihadists: A review of theoretical models and social psychological evidence. Terrorism and Political Violence 23(4), 602–622 (2011)
10. Magdy, W., Darwish, K., Weber, I.: # failedrevolutions: Using twitter to study the antecedents of isis support. First Monday 21(2) (2016)
11. O'Callaghan, D., Prucha, N., Greene, D., Conway, M., Carthy, J., Cunningham, P.: Online Social Media in the Syria Conflict: Encompassing the Extremes and the In-Betweens. In: Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'14) (2014)
12. Pirrò, G.: Explaining and suggesting relatedness in knowledge graphs. In: The Semantic Web-ISWC 2015, pp. 622–639. Springer (2015)
13. Rizzo, G., Troncy, R.: Nerd: Evaluating named entity recognition tools in the web of data. In: Workshop on Web Scale Knowledge Extraction (WEKEX11). vol. 21 (2011)
14. Rowe, M., Saif, H.: Mining pro-isis radicalisation signals from social media users. In: Proceeedings of the International Conference on Weblogs and Social Media (2016)
15. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proc. 11th Int. Semantic Web Conf. (ISWC). Boston, MA (2012)
16. Siegel, S.: Nonparametric statistics for the behavioral sciences. (1956)
17. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. American Society for Information Science and Technology 63(1), 163–173 (2012)
18. Winter, C.: Documenting the virtual 'caliphate'. Quilliam Foundation (2015)
19. Xin, D., Han, J., Yan, X., Cheng, H.: On compressing frequent patterns. Data & Knowledge Engineering 60(1), 5–29 (2007)
20. Yan, X., Han, J.: Closegraph: mining closed frequent graph patterns. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 286–295. ACM (2003)
21. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML. vol. 97, pp. 412–420 (1997)