On the Role of Semantics for Detecting pro-ISIS Stances on Social Media

Hassan Saif,[1] Miriam Fernandez,[1] Matthew Rowe,[2] and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, United Kingdom
{h.saif, m.fernandez, h.alani}@open.ac.uk
[2] School of Computing and Communications, Lancaster University, United Kingdom
m.rowe@lancaster.ac.uk

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 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-based approach for radicalisation detection on Twitter. Unlike most previous works, which mainly rely on the lexical and contextual 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 words' semantics outperform those trained from lexical and network features by 2% on average F1-measure.

Keywords: Radicalisation Detection, Semantics, Feature Engineering, Twitter

1 Introduction

The so-called Islamic State of Iraq and the Levant (ISIL/ISIS) is one of the leading terrorists organisation 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. 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 and the contextual representation of words [1, 4] (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 conceptualizations involved in content meanings. This includes, for example, the weakness to properly cope with linguistic phenomena such as polisemy (e.g., ”ISIS” as Islamic State of Iraq and Syria vs. ”Isis” as the goddess from the polytheistic pantheon of Egypt). The

[3] https://homeland.house.gov/wp-content/uploads/2015/09/TaskForceFinalReport.pdf

aforementioned limitation constitutes 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 conceptual-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*"), by using ontologies like DBpedia. The extracted conceptual semantics of words 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 two baseline features, particularly unigram 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 2% on average F1-measure, showing a positive impact on the use of semantic information to identify pro and anti ISIS stances.

2 Dataset

Radicalisation detection of Twitter users can be considered as a text classification problem where features extracted from the users' timelines are used to train and build radicalisation classifiers using machine learning methods. In this work we use a dataset of 1,132 European Twitter users, equally divided into pro-ISIS and anti-ISIS, along with their timelines. Users in this dataset are collected and labelled with their radicalised stance in our previous work [4]. Table 1 shows the total number, and distribution of tweets and words for each user group.

	pro-ISIS Users	**anti-ISIS Users**
Total number of Users	566	566
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 1: Statistics of the Twitter dataset used for evaluation

3 Semantic Features for Radicalisation Detection

The process of extracting and using semantic features for detecting radicalisation stances consists of the following steps: Firstly, a training set, consisting on labelled (pro-ISIS, anti-ISIS) users' timelines needs to be provided. To this end, we use the dataset described in the previous section, which we formalise as: $\mathcal{T}^{train} = \{(\mathbf{W}_n; c_n) \in \mathcal{W} \times \mathcal{C} : 1 \leq n \leq N^{train}\}$ where $\mathcal{W}$ is the input space and $\mathcal{C}$ is a finite set of class labels (In our case $\mathcal{C}$ = {pro-ISIS, anti-ISIS}). Secondly, the training set is processed with AlchemyAPI.[4] In particular, named-entities are extracted from the tweets of the users' timelines (e.g.

[4] http://www.alchemyapi.com

“ISIS”, “Syria”, “United Nations”) and expanded with their corresponding semantic concepts (e.g. “Jihadist_Group”, “Country”, “Organisation”), by using ontologies and knowledge bases like DBpedia, YAGO, OpenCyc, Freebase, and others.[5] The semantic extraction tool AlchemyAPI is used for this purpose due to its accuracy and high coverage of semantic types and subtypes in comparison with other semantic extraction services [3, 5]. Table 2 lists the total number of unique entities and concepts and the top 5 frequent entities and concepts, extracted from our training dataset, for both pro-ISIS and anti-ISIS user accounts. Thirdly, a semantic vector $\mathbf{t}_{us} = (e_1, e_2, ..., e_l, s_1, s_2, ..., s_l)$ is then constructed for each user as the joined vector of entities $\mathbf{e}_n = (e_1, e_2, ..., e_l)$ and concepts $\mathbf{s}_n = (s_1, s_2, ..., s_l)$ extracted from the user's timeline. After that, a training set is constructed from the semantic vectors of all users, and used to train Naive Bayes classifiers.

	pro-ISIS		**anti-ISIS**	
No. of Unique Entities	32,406		30,206	
No. of Unique Concepts	35		36	
	Entity	**Concept**	**Entity**	**Concept**
Top 5 Frequent Entities & their Concepts	MSNBC	Company	BBC	Company
	Iraq	Country	UK	Country
	Allah	Person	Kobane	City
	America	Continent	London	City
	Muslim	Person	ISIS	Organisation

Table 2: Total number and top 5 frequent entities and their associated semantic concepts extracted from our dataset.

4 Evaluation and Preliminary Results

In this section, we report the results obtained from using the proposed semantic features for user-level radicalisation classification, that is classifying users in our dataset according to their stance as pro-ISIS or anti-ISIS. To this end, we use Naïve Bayes classifiers (NB). Our baselines of comparison are NB classifiers trained from: (i) word unigrams (Bag-of-Words) and (ii) network features, which denote the profile information/attributes of Twitter users. This includes: *number of followers*, *number of followee*, *number of hashtags*, *number of mentions* (i.e., @user), *favourites count*, *status count*, *profile description (Unigrams)*, and *geographic location (Unigrams)*.

Note that we perform a feature selection process on all the feature sets to reduce the size of the classifiers' feature space. To this end, we use Information Gain (IG) [2] to compute the discriminative score of features in each feature set and filter out those with low scores from the feature space.

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* [6]. 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 stance classification using *Unigrams*, *Network*, and *Semantic* features, 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 average of the two. The table also reports the total number of features used for classification under each feature

[5] The list of ontologies and knowledge bases used by AlchemyApi is listed under http://www.alchemyapi.com/api/entity/textc.html

set. From the results presented in Table 3, we notice that semantic features have the highest impact on the classification performance among all other features. Specifically, semantic features outperform unigrams features by 2.7% and 2.8% in accuracy and average F1 respectively. Also, semantic features improve classification performance by 2.5% in accuracy and by 1.18% in F1 in comparison with the network features. Overall, semantic features increase the classification performance by 2% in F1 in comparison with the average performance of all the baseline features (F1 = 90.7%).

		anti-ISIS			pro-ISIS			Average		
	No. of Features	P	R	$F1$	P	R	$F1$	P	R	$F1$
UNIGRAMS	41,200	0.885	0.92	0.902	0.917	0.88	0.898	0.901	0.9	0.9
NETWORK	25,532	0.887	0.952	0.918	0.948	0.878	0.912	0.917	0.915	0.915
SEMANTICS	8,429	0.91	0.945	0.927	0.943	0.906	0.924	0.926	0.926	0.926

Table 3: Classification performance of the three 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 $\rho < 0.001$.

The above results show the effectiveness of using semantic features for radicalisation classification of users on Twitter.

5 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 how named-entities in tweets can be extracted and used, together with their corresponding semantic concepts, as features to train machine learning classifiers for stance detection of Twitter users.

We experimented with semantic features on a Twitter dataset of 1132 pro-ISIS and anti-ISIS users and compared the performance of a NB classifier trained from semantic features against classifiers trained from unigrams, and network features. Results showed that using the semantic features in radicalisation classification improves performance by 2% in F1 over the average performance of all baselines.

Acknowledgment

This work was supported by the EU-FP7 project SENSE4US (grant no. 611242).

References

1. 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)
2. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of machine learning research 3, 1289–1305 (2003)
3. 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)
4. 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)
5. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proc. 11th Int. Semantic Web Conf. (ISWC). Boston, MA (2012)
6. Siegel, S.: Nonparametric statistics for the behavioral sciences. (1956)