Contextual Semantics for Radicalisation Detection on Twitter

Miriam Fernandez and Harith Alani

Knowledge Media Institute, The Open University, United Kingdom
{m.fernandez, h.alani}@open.ac.uk*

Abstract. Much research aims to detect online radical content mainly using radicalisation glossaries, i.e., by looking for terms and expressions associated with religion, war, offensive language, etc. However, such crude methods are highly inaccurate towards content that uses radicalisation terminology to simply report on current events, to share harmless religious rhetoric, or even to counter extremism. Language is complex and the context in which particular terms are used should not be disregarded. In this paper, we propose an approach for building a representation of the *semantic context* of the terms that are linked to radicalised rhetoric. We use this approach to analyse over 114K tweets that contain radicalisation-terms (around 17K posted by pro-ISIS users, and 97k posted by "general" Twitter users). We report on how the contextual information differs for the same radicalisation-terms in the two datasets, which indicate that contextual semantics can help to better discriminate radical content from content that only uses radical terminology. The classifiers we built to test this hypothesis outperform those that disregard contextual information.

Keywords: Radicalisation Detection, Semantics, Feature Engineering, Twitter

1 Introduction

In an increasingly digital world, radical individuals continue to successfully exploit social media platforms for their recruitment campaigns across the world. 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 funds, and radicalise and recruit individuals.

While law enforcement agencies, experts, and volunteers are actively working to target this problem (see for example Ctr-sec[1], where volunteers report the existence of ISIS propaganda in social media and they claim responsibility for closing more than 200,000 Twitter accounts), the vast amount of traffic generated in social media makes it nearly impossible to manually identify radical content and violent extremism online. Researchers and governments are therefore investing in the creation of advanced information technologies to identify and counter extremism through intelligent large-scale analysis of online data.[2]

However, existing methods to automatically identify radical content online mainly rely on the use of glossaries (i.e., lists of terms and expressions associated with religion,

[1] https://twitter.com/CtrlSec

[2] http://www.voxpol.eu/identifying-radical-content-online/

war, offensive language, etc.). These methods are not always effective and we continue to observe that many who use radicalisation terminology in their tweets are simply reporting current events (e.g., *"Islamic State hacks Swedish radio station'*, or sharing harmless religious rhetoric (e.g., *"If you want to talk to Allah, pray. If you want Allah to talk to you, read the Qur'an*'', or even countering extremism (*"armed jihad is for defence of muslim nation. Not for establishment of khilafah."*).

Reliability is indeed a well-known concern in automatic radicalisation detection. A popular example is the campaign launched by the hacker community Anonymous as a response to the Paris attacks, where they claimed taking down more than 20,000 Twitter accounts linked to ISIS. Those accounts included, among others, the ones of the U.S ex-president Barack Obama, the White House, the BBC news, to name just a few.[3] One important source of inaccuracy of automatic radicalisation detection approaches is their reliance on the appearance of terms and expressions without considering their context. In other words, although the content may contain terms that are indeed associated with common radicalised rhetoric ("Islamic State", "Allah", etc.), these words do not convey radicalisation meanings all the time. Instead, this meaning is only 'active' in particular situations or contexts. We hypothesise that, in order to provide effective mechanisms to identify radicalised content, and to develop intelligent algorithms to identify and counter extremism, it is vital to not only consider radicalised terminologies and expressions in an isolated manner, but to take into account the surrounding contextual information. Considering this, our work investigates the following two research questions:

- Are there significant variances between the semantic contexts of radicalisation terminology when this terminology is used to convey 'radicalised' meaning vs. when it is not?
- Can we improve the effectiveness of radicalisation detection by making it more context-relevant, thus increasing its accuracy and consistency?

The notion of context is not new and has been long acknowledged as being of high importance in a wide variety of fields, such as computational linguistics, automatic image analysis, information retrieval or personalisation, to name a few [16]. The research presented here focuses on the role of context for radicalisation detection in social media, and more particularly Twitter. Our approach for context modelling exploits semantic, ontology-based contextual information from DBpedia and Wikidata,[4] as well popular classification Taxonomies including IPTC media topics, IPTC newscodes and the Internet Advertising Bureau QAG segments.[5] Among the possible knowledge representation formalisms, ontologies and knowledge bases present a number of advantages as they provide a formal framework for supporting explicit, machine-processable semantic definitions, and facilitate inference and derivation of new knowledge based on already existing ones.

To answer the above research questions, we analysed over 114K tweets mentioning terms linked to radicalised rhetoric (17K originated from pro-ISIS users and 97K

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

[4] http://wiki.dbpedia.org/, https://www.wikidata.org

[5] http://cv.iptc.org/, http://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/

originated from "general" users). We study how contextual semantic information differs for the same terms in the two datasets. Aiming to use this notion of context to better discriminate radical content we proposed an approach to incorporate it into existing classification methods, obtaining a 4.5% improvement in precision when using contextual information. By conducting this work our paper provides the following contributions:

- A contextual representation for radicalisation detection on Twitter based on ontologies and knowledge bases.
- Demonstrate that contextual information differs between content originating from pro-ISIS accounts vs. content originating from general Twitter accounts, even though both contents use radicalised terminology.
- Show that contextual semantic information can help to enhance existing keyword-based radicalisation detection mechanisms.

The following sections are structured as follows. Section 2 describes related work. Section 3 shows our proposed approach to automatically model and capture the semantic context of terms and expressions. Sections 4 and 5 discuss the evaluation and results of studying the semantic context divergence across terms and the results of incorporating contextual information for radicalisation detection. An in-depth discussion of our findings and the limitations of this study is reported in Section 6. Section 7 concludes.

2 Related Work

Researchers from the areas of counter-terrorism and cyber-security have begun to examine the radicalisation phenomenon and to analyse the social media presence and actions of extremist organisations [1] [9]. These works include computational approaches developed towards the **analysis** and **detection** of radicalisation.

Among the works developed towards **analysing** the online radicalisation phenomenon, we can highlight the works of Klausen [11], Carter [7], Chatfield [8], Vergani [17] and Rowe [13]. These works examine the type of language and communication strategies that ISIS members and supporters use to radicalise and recruit others. The works highlight: (i) the direction of the communication flow, from the terrorist accounts, to the fighters based in the insurgents' zones, to the followers in the west [11], (ii) the frequent use of certain terms in the conversations (islamic, Allah, fight, Mujahideen, ISIS, etc.) [7], (iii) the frequent mentioning of international media, regional Arabic media, IS sympathisers and IS fighters [8], (iv) the use of expressions related to achievement, affiliation and power with a focus on emotional language, (v) the frequent mentioning of death, female and religious related terms [6], and (vi) the high relevance of social-homophily on the diffusion of pro-ISIS terminology. These works are not focused on the automatic detection of online extremism, but rather on investigating and understanding the online radicalisation process in different channels, including Twitter [11, 7, 8, 13], Facebook [7], YouTube [6] and Dabiq; the official ISIS magazine [17].

Among the works concerned with the **detection** of extremist content and Tweets, we can highlight the works of Berger [4, 5], Agarwal [2], Ashcroft [3] and Saif [14]. These approaches tend to: (i) rely on the use of known radical users as seeds and expand through the network of followers to detect other potential radical users and content [4][5] or, (ii) build classifiers trained on data collected based on radicalisation glossaries, where users whose tweets contain certain words would be regarded as in the "radicalised" set

[1][3][14]. Sometimes these collected tweets are additionally filtered based on: (i) lists of known radicalised users [3][14], (ii) based on whether the user account has been suspended [5] or, (iii) validated by annotators, who, in many cases are non-experts on the topic [2]. It is also unclear how the manual annotations are performed and what criteria are followed by the annotators when categorising users and data [5].

The lack of grown truth datasets for radicalisation detection, and the difficulties for obtaining them (particularly when relying on glossaries of radicalisation terms) are well known problems in this field (see Section 6). To the best of our knowledge no previous studies have focused on understanding how the context of radical terminology semantically differs when used to convey a radical message vs. when not, which is one of the key contributions of this work.

3 Semantic Context

3.1 Modeling Semantic Context

Context is a difficult notion to grasp and capture in a software system. In this work, we focus our efforts on the notion of semantic context for radicalised content detection. The information objects in our model are therefore a set of posts P, and the units for which we want to extract the context are a set of terms $t \in T$, where T is a glossary/lexicon used for radicalisation detection (examples of these terms can be seen in Section 4.2).

We define the semantic context of a term t as the set of categories C_{t_p}, topics T_{t_p}, entities E_{t_p} and concepts (entity types) ET_{t_p} that emerge from the posts where the term appears $t_p \subset P$. Multiple semantic services have been built over time to enable the annotation of text with concepts, properties and values defined according to domain ontologies and stored in knowledges bases (KBs) [12]. For this work, we selected TextRazor[6] as the service to extract the semantic elements (categories, topics, entities and types) that we will use to compose the context for each term. For every annotation, TextRazor provides a confidence score. More details on how annotations are extracted and how this confidence score is calculated can be found on TextRazor's website. Our semantic context model is therefore composed by four dimensions:

1. **Categories**: The set of Categories C for the posts P is extracted considering three different taxonomies: IPTC media, IPTC news codes and the Internet Advertising Bureau (IAB) QAG segments. Examples of relevant IPTC categories include: {id: 16009001 label: unrest, conflicts and war>war>civil war}, {id:20000677 label: religion and belief>religious belief>Islam}. Examples of IAB taxonomy categories include: {id: IAB11 label: Law, Govt & Politics}.
2. **Topics**: High-level topics are assigned to posts P based on Wikipedia. These topics include their corresponding link to wikidata and wikipedia pages and categories. Examples of relevant topics include: {wikidataId: Q7283 label:Terrorism} and {wikidataId: Q42418 label: Taliban}.
3. **Entities**: The set of Named Entities E for the posts P is extracted based on DBpedia and Wikidata. Examples of relevant entities include: {wikidataId: Q41183 label: Aleppo types:[Place, PopulatedPlace, Settlement, City]}, {wikidataId: Q133207 label: Muslim Brotherhood types:[Agent, Organisation, PoliticalParty]} .

[6] https://www.textrazor.com

4. **Entity Types**: Entity types refer to the DBpedia concepts (types) of the previously extracted entities. These concepts are also linked to wikidata. E.g., the types for the entity: Muslim Brotherhood include: {dbo:Agent wikidataid: Q24229398}, {dbp:Organisation wikidataid:Q43229} and {dbo:PoliticalParty, wikidataId:Q7278}[7]

Once we have extracted the different semantic elements for every post $p \in P$ we use these elements to extract the semantic context of every radicalised term $t \in T$. In our work, we extract two different semantic contexts for each term: $SCR(t)$ and $SCNR(t)$. The first semantic context $SCR(t)$ refers to the context extracted from a subset of tweets posted by pro-ISIS users. The second semantic context $SCNR(t)$ refers to the context extracted from a subset of tweets posted by non pro-ISIS users.

Let $P_{t_r} \subset P$ be the set of tweets posted by pro-ISIS users where the term t appears. $SCR(t)$ is then defined as the set of categories C_{t_r}, topics T_{t_r}, entities E_{t_r}, and entity types ET_{t_r} that emerge from the semantic annotations of P_{t_r}. Note that the same category, topic, entity, or entity type may emerge multiple times in the tweets associated with the term. We therefore define the vectors of unique categories, topics, entities and entity types as:

$$\boldsymbol{Vc_{t_r}} = \{c_1, c_2, ...c_n\}, c_i \in C_{tr} \text{ and } val(c_i) = \frac{\sum_{p=1}^{p=|P_{t_r}|} confscore(c_{i_p})}{|P_{t_r}|}$$

$$\boldsymbol{Vt_{t_r}} = \{t_1, t_2, ...t_m\}, t_i \in T_{tr} \text{ and } val(t_i) = \frac{\sum_{p=1}^{p=|P_{t_r}|} confscore(t_{i_p})}{|P_{t_r}|}$$

$$\boldsymbol{Ve_{t_r}} = \{e_1, e_2, ...e_o\}, e_i \in E_{tr} \text{ and } val(e_i) = \frac{\sum_{p=1}^{p=|P_{t_r}|} confscore(e_{i_p})}{|P_{t_r}|}$$

$$\boldsymbol{Vet_{t_r}} = \{et_1, et_2, ...et_s\}, et_i \in ET_{tr} \text{ and } val(et_i) = \frac{\sum_{p=1}^{p=|P_{t_r}|} confscore(et_{i_p})}{|P_{t_r}|}$$

Note that the value for each contextual element is computed based on the confidence score of the annotations. For example, if we are computing the context for the term *Allah*, and this term appears in five posts, three of them annotated with the category *religious belief*, the value of the category *religious belief* in the context of the term *Allah* is computed as the sum of the confidence scores of the three annotations, divided by the total number of posts in which the term *Allah* appears, in this case five. By doing this, the value incorporates both, the frequency of the semantic element (i.e., in how many annotations it appears) as well as the confidence of those annotations. Note also that confidence scores values are between [0, 1]. By normalising by the number of posts we ensure that the value for every element (c_i, t_i, e_i, et_i) in the context of the term is always between [0,1] as well. We then define the two semantic contexts for each radicalised term $t \in T$ as:

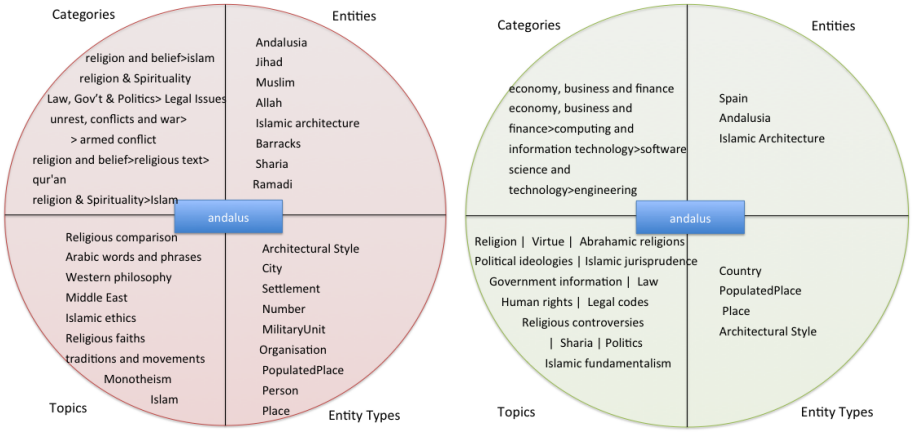
$SCR(t) = \{Vc_{t_r}, Vt_{t_r}, Ve_{t_r}, Vet_{t_r}\}$, which is built based on $P_{t_r} \subset P$, the set of radical tweets, or tweets posted by pro-ISIS users.

$SCNR(t) = \{Vc_{t_{nr}}, Vt_{t_{nr}}, Ve_{t_{nr}}, Vet_{t_{nr}}\}$, which is built based on $P_{t_{nr}} \subset P$, the set of non radical tweets, or tweets posted by non pro-ISIS users.

A reduced example of these contexts for the term *Andalus* is displayed in Table 1.

[7] dbo refers to DBpedia Ontology

Table 1: Semantic contexts for the term *andalus*. $SCR(andalus)$ vs. $SCNR(andalus)$. Scores for each element are not added to the image for better visibility



3.2 Calculating Semantic Context Divergence

To assess the divergence of semantic contexts among radicalisation terminology, we compute for each term $t \in T$ the similarity between $SCR(t)$, and $SCNR(t)$. We maintain the four defined contextual dimensions (categories, topics, entities and entity types), and use the cosine similarity metric (cos)[8] to compute the similarity between the two semantic contexts for each dimension. The semantic context divergence for each term is then defined as: $SCD(t) = \{1 - cos(Vc_{t_r}, Vc_{t_nr}), 1 - cos(Vt_{t_r}, t_{Vt_nr}), 1 - cos(Ve_{t_r}, Ve_{t_nr}), 1 - cos(Vet_{t_r}, Vet_{t_nr})\}$

3.3 Exploiting Semantic Context for Radicalisation Detection

There are multiple ways in which the proposed semantic context model could be incorporated into existing radicalisation detection approaches. It could particularly be helpful for defining rules associated to current radicalisation lexicons, so that posts and users are not automatically categorised as 'radical' if they contain or use certain terms. Instead, the use of those terms could be weighted based on their contextual information, similarly to existing contextually enhanced lexicon-based sentiment analysis approaches [15].

In this work however, our purpose is not to generate a novel method for radicalisation detection, neither to improve current state of the art approaches for radicalisation detection, but to study whether semantic information could help to enhance keyword-based methods by making them more context relevant. To test this hypothesis we have build classifiers based on n-grams and test them against classifiers that incorporate as additional features the extracted contextual information. To train these classifiers we make use of a dataset of posts P, along with their class labels C =(pro-ISIS, non pro-ISIS). Features are extracted in the following way. For each posts $p \in P$ we create two vectors: the unigram vector and the semantically enhanced vector.

[8] `https://en.wikipedia.org/wiki/Cosine_similarity`

$\boldsymbol{V_{unigram_p}} = (w_1, w_2, ...w_n)$, where w_i is the $tf * idf$ value of the unigram.[9]

$\boldsymbol{V_{semantic_p}} = (w_1, w_2, ...w_n, c_1, c_2, ...c_m, t_1, t_2, ...t_o, e_1e_2, ...e_p, et_1, et_2, ...et_q)$, where $(c_1, c_2, ...c_m)$, $(t_1, t_2, ...t_o)$, $(e_1e_2, ...e_p)$, $(et_1, et_2, ...et_q)$ are the semantic context elements (categories, topics, entities and entity types) extracted for p. To generate these vectors we consider only semantic elements for which the confidence score of their annotations is higher than 0.5.

4 Experimental Setup

4.1 Dataset

We use two publicly available datasets to study radicalisation, from Kaggle datascience community. The first dataset contains 17,350 tweets from 112 distinct pro-ISIS accounts.[10] collected based on a three-month period study.[11] To ensure that this dataset contains only users that are **pro-ISIS**, we checked the profiles of the 112 accounts using the Twitter API. We assume that if the account is no longer available (i.e., has been blocked), then most likely the account did indeed belong to a radicalised individual. Only two of those accounts were still active at the time of writing, one belongs to a journalist, and the other one to a researcher who focuses on Jihadi groups. We deleted these two profiles and their tweets from the dataset, thus obtaining a final dataset of **110 users and 16,949 tweets**.

The second dataset was created as a counterpoise of the previous dataset. It contains 122K tweets from 95,725 distinct users collected on two separate days 7/4/2016 and 7/11/2016. Tweets were collected based on the following keywords (isis, isil, daesh, islamicstate, raqqa, Mosul, 'islamic state').[12] Many of these accounts have now been blocked. To ensure that this dataset contains only users that are **not pro-ISIS** (they could be anti-ISIS or neutral), we collected the profiles of the 95,725 users by making use of the Twitter API. We found that 76,819 of them were still active at the time of writing. We assume that, if the account is still active even though two years have passed since it was shared on Kaggle, then the profile does not belong to a radicalised individual. To assess this hypothesis, a random subset of 40 profiles was selected and manually assess by two annotators (authors), who agreed, for all the 40 profiles, that these accounts do not show signs of support for ISIS. The final dataset after removing the content of the closed accounts contains **97,287 tweets from 79,819 active users**.

Note that in the cases of both datasets, tweets contain radicalised terminology, but, while in the first dataset that terminology was used by radical individuals to spread their message, in the second dataset, 'common' Twitter users are the ones using that terminology to, for example, report or discuss related news and events.

While most tweets in these datasets are written in English, few of them are written in arabic. TextRazor recognises a variety of languages.[13] However, the annotator could not extract information for 3,249 tweets from the pro-ISIS dataset and 2,573 from the non

[9] https://en.wikipedia.org/wiki/Tf-idf

[10] https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[11] http://blog.kaggle.com/2016/06/03/dataset-spotlight-how-isis-uses-twitter/

[12] https://www.kaggle.com/activegalaxy/isis-related-tweets

[13] https://www.textrazor.com/languages

pro-ISIS dataset. This was due to the use of unknown languages (e.g., arabic script) or because the tweets contained very few pieces of information (for example, just a URL).

4.2 Radicalisation Terminology

Radicalisation lexicons have been developed by experts, and have also been created from content generated by ISIS, such as the Dabiq [14] and Inspire[15] magazines. In this work we make use of a lexicon containing 305 entries, including 556 terms, expressions and variances created in a previous work [10]. An example can be seen in Figure 1

	Term	Translation and definition	Variants
1.	Abu Mus'ab az-Zarqawi	ISIS's spiritual founder & a former leader of al-Qaeda in Iraq	Abu Musab az-Zarqawi

Fig. 1: Examples of entries in the radicalisation glossary

4.3 Semantic Context Extraction

To extract the categories, topics, entities, and entity types that define our semantic context (see Section 3.1) we have made use of the TextRazor semantic annotator. To compute/build our context we are only considering annotations for which the confidence score is higher than 0.5. Categories, topics, entities, and entity types that emerge from the text but with a confidence score lower than 0.5 are discarded. Context is only built for radicalised terms that appear in a minimum of five tweets per dataset.

5 Evaluation Results

5.1 Semantic Context Divergence

The first step when conducing our analysis was to distinguish between: (i) the terms that do not appear in any of the datasets, (ii) the terms that appear in one dataset but not in the other and, (iii) the terms that appear in both datasets, which will be the ones for which we will extract and compare contextual information.

- **Pro-ISIS**: 48 terms, (15.7%) have been found only in the Pro-ISIS dataset. Examples of the terms appearing with higher frequency include: *Sahwat* - Awakening Councils, derogatory term denoting militias/groups who are allegedly supported by the West - (e.g. "Sahwat spent three months battling IS in #Aleppo with the support of #US planes to create a "Safe zone")" , or *taghut* - terminology denoting a focus of worship other than Allah (e.g., "After the recent killings of Sheikh faris Zahrani and other Sheyukh by the taghut of al saud AQAP has warned this [url]"[16]).
- **Non Pro-ISIS**: 17 terms (5.6%) have been found only in the Non Pro-ISIS dataset. Examples among the terms appearing with higher frequency include: *Fundamentalist* -a person who believes in the strict, literal interpretation of scripture in a religion-

[14] https://en.wikipedia.org/wiki/Dabiq_(magazine)

[15] https://en.wikipedia.org/wiki/Inspire_(magazine)

[16] Sheikh Faris bin Ahmed Jamaan al-Showeel al-Zahrani was on Saudi Arabia's list of 26 'most-wanted' suspected terrorists. On 2 January 2016 Sheikh Zahrani was executed by the Saudi state along with 46 others convicted of terrorism

(e.g., “IF ISIS IS NOT ISLAMIC TERRORIST OR ISLAMIC EXTREMIST OR ISLAMIC FUNDAMENTALIST OR NOT ISLAMIC AT ALL,WHY SHARIA LAW NEEDED??”), or “*Zindiq* -Muslim individuals who are considered to hold views or follow practices that are contrary to central Islamic dogmas- (e.g., “#ISIS are takfiri, khawarij, zindiq & the real enemy of #Islam.. They did a frontal attack on #Islam. [username]”).

– **Pro and Non Pro-ISIS**: 146 terms (47.9%) appear in both datasets. Common terms include *Allah* -God -, *Amaq* - A news agency linked to ISIS -, I*slamic State*, *Khilafa*, etc. See for example the term *Allah* used by the pro-ISIS group “Our Prophet, has ordered us to fight you till you worship Allah Alone or give Jizya” vs. a use by the non pro-ISIS group “May ALLAH the Almighty destroy ISIS and rid humanity of their disgusting actions. #Istanbul [username]”.
– **Not found**: 94 terms (30.8%) do not appear in any of the datasets. Examples of these terms are: *Millah* - which refers to any variety of religion, except the “true” religion of Islam, or *Tatarrus* - the practice of using civilians as human shields.

We have further analysed the use of the 146 terms appearing in both datasets to better understand the semantic context divergence when the terms are used by the pro-ISIS vs. the non pro-ISIS group. As reported in section 3.1 we consider four different dimensions of semantic context: (i) categories, (ii) topics, (iii) entities and (iv) entity types. Each of these contextual dimensions is extracted for each of the 146 terms under analysis by considering the tweets in which the terms appear within the non-pro-ISIS and pro-ISIS datasets.

Our experiments show that, while there exist a semantic contextual divergence among terms, this divergence is not equally distinctive across the different contextual dimensions (see Figure 3). The more discriminative contextual dimension is Entities. Out of the terms analysed, 95 (65%), displayed contextual divergence > 0.25 based on entities. This is lower for categories (29 terms, 20%), topics (12 terms, 8.2%) and entity types (8 terms, 12.33%). Entities are more fine-grained/specific descriptors of the context, while topics, categories and types provide higher-level conceptualisations. The Venn diagram showing these intersections is displayed in Figure 2. In total, out of the 146 terms analysed, 100 (68.5%) showed a context divergence $>= 0.25$ for at least one of the semantic context dimensions. Those terms that exhibit low context divergence for all dimensions refer to locations (such as Afghanistan or Iraq), as well as historic and religious-related terms (e.g., aqdah, khawarij). Interestingly, this group of terms also includes proper nouns used for the radical groups including ISIL, Taliban, or Al-Qaeda. These terms appear in a very high number of tweets and are therefore associated with a high number of topics, categories, entities and entity types in both groups. An indication that, while they may be useful to detect social media conversations around jihadism, they may not be that helpful when discriminating radical content among those conversations.

Among the contextual divergences we have also observed that some contextual elements (either categories, topics, entities or entity types) appear in one context but not in the other one, and that certain semantic elements, although they appear in both contexts, they show more ’strength’ in one of them. Note that value for a semantic element in the context of a term is computed based on the frequency and confidence of the semantic element in the annotations (see Section 3.1)

Let's take as an example the term *fatwa* (a ruling on a point of Islamic law given by a recognized authority). Among the entities that appear in $SCR(fatwa)$ but not in $SCNR(fatwa)$ we found: Jihad, Al-Qaeda, Syria, Ayatollah, Prophets and messengers in Islam, Mujahideen, etc. Among the entities that appear in $SCNR(fatwa)$ but not in $SCR(fatwa)$ we found: Terrorism, Sacrilege, Kurt Westergaard, and Charlie Hebdo (who published controversial cartoons of figures and elements of Islam) and were the target of terrorist acts. Among the entities that appear in both, but stronger in one context than the other, the entity Muslim is used with more strength in the non pro-ISISl context, while the entity Sunni Islam appears with higher strength in the pro-ISIS one. All these semantic variances can therefore help us to better understand when these terms may be conveying a radical connotation rather than descriptive meanings.

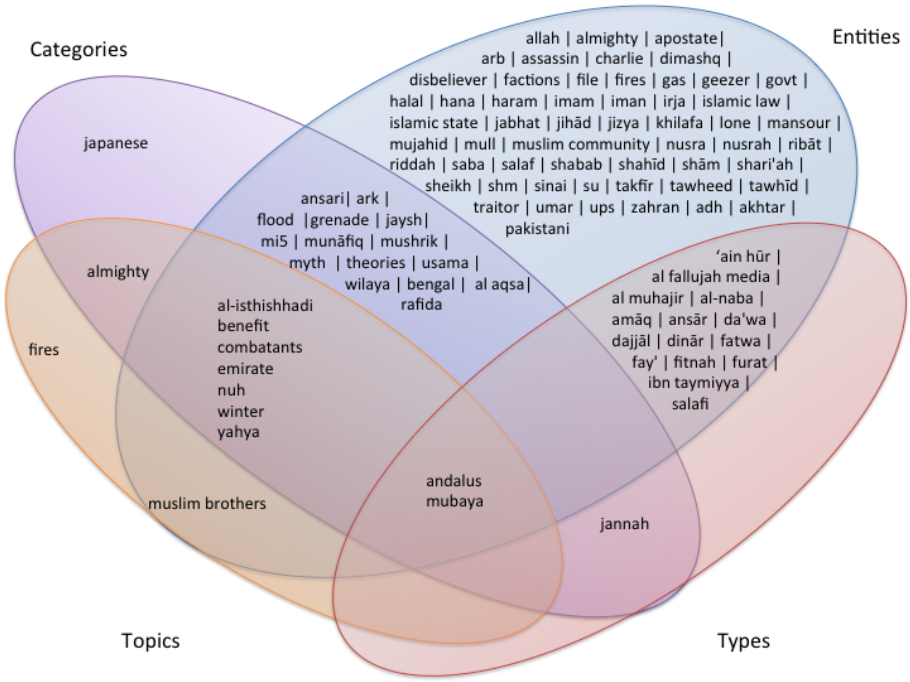


Fig. 2: Venn Diagram showing the terms whose context divergence is > 0.25 for any of the four semantic context dimensions

5.2 Radicalisation Detection

As mentioned in Section 3.3, while semantic contextual information can be incorporated in a variety of ways to enhance current radicalisation detection approaches, in this work we are testing its effectiveness against the use of unigrams for radicalisation classification. Our dataset for training is composed the 13,700 tweets of the pro-ISIS dataset for which we could obtain semantic annotations (see Section 5.1) (labelled as pro-ISIS) and a random sample of 13,700 tweets from the non pro-ISIS dataset (labelled as non pro-ISIS).

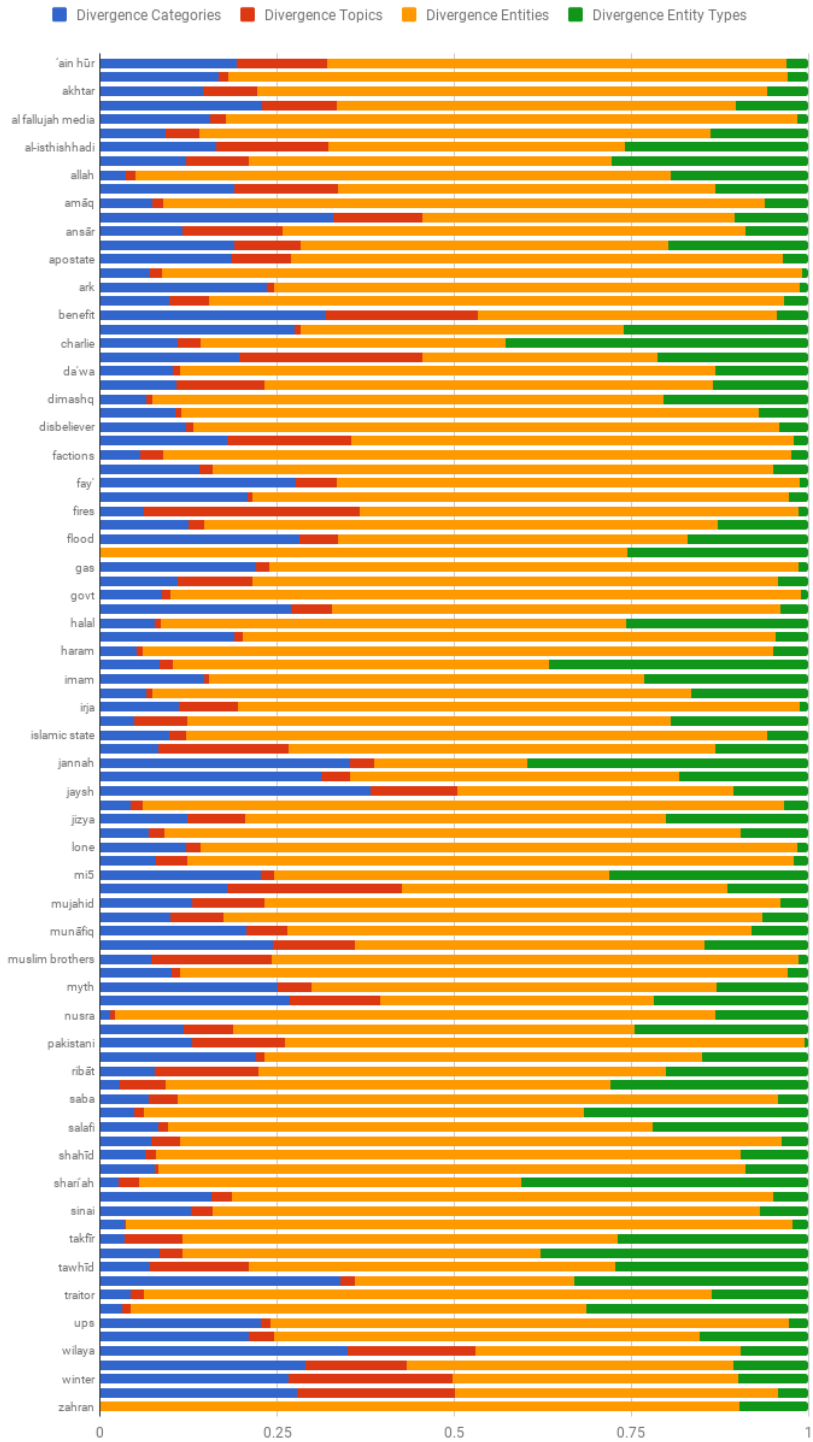


Fig. 3: Divergence across the four different context dimensions for those terms showing a context divergence > 0.25 for at least one of the dimensions.

We tested multiple classifiers including Naive Bayes (NB), SVM and decision trees (J48) and compared the use of n-gram only features vs. the use of n-grams enriched with contextual semantics. The top results were obtained with SVM in both cases (see Table 2). Results are reported based on 10-fold cross validation. As we can see in the table, the semantically enhance classier outperforms the unigrams-only classifier in precision, recall and F1-measure. This indicates that the use of semantics can indeed help providing additional meaning to terms in order to enhance radicalisation detection.

Table 2: Keyword-based vs. Semantic-based radicalisation detection

Features	**Precision**	**Recall**	**F-measure**
Uni-grams Only	0.816	0.801	0.822
Semantic Context Enrichment	0.859	0.843	0.851

6 Discussion and Future Work

Accurate detection of radicalised content and users online is a major problem faced by current governments and policing organisations. This is due to the fact that media agencies, journalists, political figures, and religious non-radical individuals may be using the same terminology that radical individuals use to spread their ideology. The problem is not only technically challenging but also ethically, since the wrong association of an individual to a radical movement can have serious implications.

In this work, we explored the use of semantic information to better understand the contextual variances in which radicalisation terms are used when conveying 'radicalised meaning' vs. when not. We also looked at whether keyword-only radicalisation detection could be enhanced by incorporating contextual semantic information. While our work indeed shows that semantics can provide a more fine-grained understanding of online radicalisation, this research has various limitations.

The most difficult challenge confronted by anyone who studies online radicalisation is the lack of gold standard datasets. Existing datasets are rarely verified by experts to ensure that false positives (i.e., content or user profiles that have been wrongly associated to radicalisation) are identified. Annotating this data is also not trivial, since annotators should have a certain degree of background knowledge on religion and politics to correctly annotate such content. We are currently working with law enforcement agencies and experts to be able to achieve such gold standards.

As explained in Section 5.1, we checked to see whether a Twitter account was blocked or not to consider the account to be a pro-ISIS or not. However, Twitter could wrongly block an account, or block it for reasons other than radicalisation. Therefore, a better approach is to manually annotate these accounts. To be on the safe side, instead of including blocked accounts as part of the pro-ISIS dataset we decided to discard them.

Another issue is the imbalanced number of users and tweets in the pro-ISIS and the non pro-ISIS datasets. While we have balanced the dataset to create our classifiers (see Section 5.2), we have decided to maintain all tweets to study the semantic context associated to radicalisation terms. Our assumption here is that, the more tweets we have, the more complete can be the semantic context we extract. Terms appear in average more than 100 times in each dataset and, as mention in Section 3 we have discarded those terms for which we have less than 5 tweets in each dataset to build their context.

The pro-ISIS and the non pro-ISIS datasets were collected in different time periods. Global information (i.e., news and events around the world) do vary over time, and leaders and locations that are relevant now may not be in a few months. This could particularly affect the contextual divergence obtained by looking at entities (persons, organisations, locations), which is indeed the highest for all terms across the four considered contextual dimensions. Ideally one should analyse two datasets collected in the same time period to avoid the potential time influence.

Finally it is also important to highlight that, while the tweets contained in the pro-ISIS dataset have been written by pro-ISIS users, that does not necessarily mean that all these tweets have radical content. Similarly the non pro-ISIS dataset may also contain some tweets of radical nature. Therefore, while our results show that pro-ISIS and non pro-ISIS users present divergences in their use of radicalisation terminology, we cannot claim that these terms are used differently in radical vs. non-radical content. The same remark applies to our classifier. Our classifier shows that semantic information can help us to better differentiate content coming from pro-ISIS accounts, but we can not claim that our classifier improves the detection of radical content. Regarding the creation of this classifier, it is also important to highlight that our aim here was not to enhance current state of the art approaches for content-based radicalisation detection, but to investigate whether the use of semantics could help to enhance current methods by providing additional contextual information.

Regarding our proposed approach for semantic context representation, we consider that if an entity and a radicalisation term appear in the same tweet they are context-related. However a more fine-grained definition of context could be considered by taking into account the specific relation between the term and the entity as expressed in the text. For simplicity, we have not considered relations in this study. Regarding the computed contextual values, although the performance of semantic annotators have significantly improved over the years, annotations are not always accurate, specially when it comes to social media text (which is generally short and ill-formed). Similarly, Knowledge Bases may also contain incomplete or non-updated data, or even mistakes. While our model already takes into account the notion of 'confidence' within the annotations, this notion could be expanded by considering the quality (accuracy, completeness) of the KBs used to generate the annotations.

Despite all the above mentioned limitations, every step towards the effective detection of radicalised rhetoric may have important societal implications. We hope that the presented study will serve as basis for future work within and across the Semantic Web, the Social Web, and the policing research communities.

7 Conclusions

In this paper we propose an approach for building a representation of the semantic context of the terms that are linked to radicalised rhetoric. We report on how the contextual information differs for the same radicalisation-terms in two datasets generated by pro-ISIS and not pro-ISIS users. We also build classifiers to test whether contextual semantics can help to better discriminate radical content. Our results show that the classifiers used to build this hypothesis outperform those that disregard contextual information.

Acknowledgment

This work was supported by the EU H2020 project TRIVALENT (grant no. 740934)

References

1. Agarwal, S., Sureka, A.: Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. arXiv preprint arXiv:1511.06858 (2015)
2. Agarwal, S., Sureka, A.: Using knn and svm based one-class classifier for detecting online radicalization on twitter. In: International Conference on Distributed Computing and Internet Technology. pp. 431–442. Springer (2015)
3. Ashcroft, M., Fisher, A., Kaati, L., Omer, E., Prucha, N.: Detecting jihadist messages on twitter. In: Intelligence and Security Informatics Conference (EISIC), 2015 European. pp. 161–164. IEEE (2015)
4. Berger, J., Strathearn, B.: Who matters online: Measuring influence, evaluating content and countering violent extremism in online social networks. International Centre for the Study of Radicalisation and Political Violence (2013)
5. Berger, J.M., 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), 4–1 (2015)
6. Bermingham, A., Conway, M., McInerney, L., O'Hare, N., Smeaton, A.F.: Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In: Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'09) (2009)
7. Carter, J.A., Maher, S., Neumann, P.R.: # greenbirds: Measuring importance and influence in syrian foreign fighter networks (2014)
8. Chatfield, A.T., Reddick, C.G., Brajawidagda, U.: Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In: Proceedings of the 16th Annual International Conference on Digital Government Research. pp. 239–249. ACM (2015)
9. Correa, D., Sureka, A.: Solutions to detect and analyze online radicalization: a survey. arXiv preprint arXiv:1301.4916 (2013)
10. Fernandez, M., Asif, M., Alani, H.: Understanding the roots of radicalisation on twitter. In: Proceedings of the 10th ACM Conference on Web Science. ACM (2018)
11. Klausen, J.: Tweeting the jihad: Social media networks of western foreign fighters in syria and iraq. Studies in Conflict & Terrorism 38(1) (2015)
12. Rizzo, G., Troncy, R.: NERD: evaluating named entity recognition tools in the web of data. In: ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany. Bonn, GERMANY (10 2011)
13. Rowe, M., Saif, H.: Mining pro-isis radicalisation signals from social media users. In: Int. Conf. Weblogs and Social Media (ICWSM). Cologne, Germany (2016)
14. Saif, H., Dickinson, T., Kastler, L., Fernandez, M., Alani, H.: A semantic graph-based approach for radicalisation detection on social media. In: European Semantic Web Conference. pp. 571–587. Springer (2017)
15. Saif, H., He, Y., Fernandez, M., Alani, H.: Contextual semantics for sentiment analysis of twitter. Information Processing & Management 52(1), 5–19 (2016)
16. Vallet, D., Castells, P., Fernández, M., Mylonas, P., Avrithis, Y.: Personalized content retrieval in context using ontological knowledge. IEEE Transactions on circuits and systems for video technology 17(3), 336–346 (2007)
17. Vergani, M., Bliuc, A.M.: The evolution of the isislanguage: a quantitative analysis of the language of the first year of dabiq magazine. Sicurezza, Terrorismo e Società= Security, Terrorism and Society 2(2), 7–20 (2015)