Chapter 18
Detecting Linguistic Markers of Violent Extremism in Online Environments

Fredrik Johansson
Swedish Defence Research Agency (FOI), Sweden

Lisa Kaati
Uppsala University, Sweden

Magnus Sahlgren
Gavagai, Sweden

ABSTRACT

The ability to disseminate information instantaneously over vast geographical regions makes the Internet a key facilitator in the radicalisation process and preparations for terrorist attacks. This can be both an asset and a challenge for security agencies. One of the main challenges for security agencies is the sheer amount of information available on the Internet. It is impossible for human analysts to read through everything that is written online. In this chapter we will discuss the possibility of detecting violent extremism by identifying signs of warning behaviours in written text – what we call linguistic markers – using computers, or more specifically, natural language processing.

INTRODUCTION

In recent years, there have been many examples of various types of terrorist attacks taking place all over the world. There have also been several severe school shootings that resulted in many victims. When studying the reason behind why these attacks took place, Internet often has an important role to play. For example, the use of the Internet for terrorist recruitment and operations has increased significantly in recent years (Torok, 2013), not least due to an emergence of social media services such as Facebook, Twitter, Instagram, and YouTube.

DOI: 10.4018/978-1-5225-0156-5.ch018

There are several examples of terrorists and terrorist organisations that use or have used the Internet and social media in different ways. One example of a terrorist who used the Internet extensively is Jose Pimentel, who was arrested for planning attacks with home-made pipe bombs against police vehicles and postal facilities in the United States. Pimentel was very active on the Internet, where he maintained a website on Blogger and a YouTube channel containing radical works which connected him with like-minded individuals (Weimann, 2012). Another example of a terrorist that used the Internet is the Norwegian terrorist Anders Behring Breivik. He used the Internet to obtain the necessary knowledge on how to construct a large fertiliser bomb, and to express and discuss his critical view on Islam and socialism (Ravndal, 2013). The suicide bomber Taimour Abdulwahab al-Abdably that killed himself in the middle of Stockholm in 2010 is another example of a person that used the Internet for various reasons. Abdably was active on various forms of social media such as YouTube and Facebook and he also searched for a second wife on Islamic web pages.

Due to the nature of the Internet and social media, it is possible to communicate and express radical views and intentions as well as to connect to other persons with similar interests. This is also noted in Europol's annual terrorism situation and trend report of 2012, which states the following: "Online social media sites attract high numbers of users. Internet forums are an effective means to address targeted audiences, including supporters who have no off-line links to terrorist organisations" (Europol, 2012, p. 10).

The ability to disseminate information instantaneously over vast geographical regions makes the Internet a key facilitator in the radicalisation process and preparations for terror attacks. This can be both an asset and a challenge for intelligence and security agencies. While it is troublesome that radical and violent extremism content can be spread globally with very low costs, this fact also provides an opportunity for intelligence analysts and police to act preventively. By collecting, fusing and analysing 'weak signals' or 'digital traces' that are present on the Internet, there is a possibility to detect attackers before they strike.

One of the main challenges in doing such analysis is the sheer amount of information available on the Internet. For this reason, analysts need support from computerised tools to be able to perform analysis on a large scale. Although web data have become an important source of information for law enforcement agencies working to prevent terrorist attacks, it is impossible for human analysts to read everything that is written on the Internet.

As an example, a human analyst capable of speed-reading may be able to read up to 1,000 (or even more) words per minute (cf. the average adult reading speed is around 300 words per minute). If the analyst is able to read at the same speed consecutively for eight hours (which is perhaps not very likely), he/she will have read about 480,000 words, which given an average word length of six characters, equals 2.8 megabyte of data (cf. the average human would have read less than one megabyte of data in the same time period). It has been estimated that the Internet carries around 1.8 exabyte of data per day. Furthermore, approximately 80 percent of all available data is in unstructured form, and a large portion of this unstructured data is textual data. This means that it would require several hundred million incredibly focused speed-reading human analysts (and consequently more than a billion humans with normal reading capacity) to be able to read all the text data generated on the Internet in only one day. This example illustrates the need for computer support to be able to detect and analyse content that is of interest for law enforcement agencies if we do not know where to find the information of interest.

BACKGROUND

Tools for monitoring the Internet are not a new phenomenon. Large-scale surveillance systems and surveillance programs such as Echelon (Schmid, 2001), XKeyscore (Greenwald, 2013), and PRISM (Greenwald, 2013) have been heavily debated and criticised, mostly due to the threat towards personal integrity that monitoring of various types of communication pose. Another problem with such systems is that they often lack clear guidelines describing who and during what circumstances monitoring was allowed. Monitoring specific keywords or terms is the most common approach to obtain situational awareness using such systems. However, such approaches may generate a lot of irrelevant material due to the nature of our language – i.e., words can be used in many different contexts and have several different meanings.

Another problem with monitoring what people write (or what information they access) is the fact that even if some individuals show a great interest in radical content and participate in discussions, many of them will never become terrorists or even break the law, which makes it difficult for law enforcement agencies to identify actual terrorists (Bjelopera, 2013). For this reason, it is fundamentally important to realise that the output from such systems always have to be checked and verified by human analysts, and most often only can serve as a filtering mechanism to bring forward information that may be potentially interesting to the analysts.

MAIN FOCUS OF THE CHAPTER

In this chapter we discuss the possibility of automatically detecting signs of certain warning behaviours defined by Meloy, Hoffmann, Guldimann and James (2012). The warning behaviours might indicate that a person or a group has the intention to commit a violent attack. In the long run, our hope is that such warning behaviours can be used by systems to generate less false hits (i.e., false positives) than traditional surveillance systems.

This chapter is organised as follows. First we describe the typology of warning behaviours and then some related work and ideas for identifying a subset of such warning behaviours automatically in text using natural language processing techniques. Next, we describe the implementation of a prototype system for detecting warning behaviours in user-generated content posted on the Internet that is based on the triangulation of several linguistic markers. We also describe some preliminary results obtained when collecting data using the prototype system. It should be stressed that these are just preliminary results, but based on these we describe some operational and practical implications. There are many ethical considerations associated with these kinds of systems since there are many privacy-related issues to take into account when deciding on whether these kinds of systems should be used at all, and if so, how the intrusion of citizens' privacy can be minimised. These aspects are briefly discussed at the end of this chapter with some directions for future work.

WARNING BEHAVIOURS

It has been argued by various psychologists that violent attacks on public figures, mass murders, and acts of lone wolf terrorism, are often signalled by a set of more or less detectable warning behaviours.

Meloy and O'Toole (2011, p. 514) define warning behaviours as any behaviour that "precedes an act of targeted violence, is related to it, and may, in certain cases, predict it". Warning behaviours can in this context be viewed as indicators of increasing or accelerating risk of committing a violent attack.

Furthermore, in Meloy (2011), Meloy and O'Toole (2011), Meloy et al. (2012), and Meloy, Hoffmann, Roshdi, and Guldimann (2014), eight different warning behaviours are defined for targeted or intended violence. These warning behaviours are: (i) pathway warning behaviour, (ii) fixation warning behaviour, (iii) identification warning behaviour, (iv) novel aggression warning behaviour, (v) energy burst warning behaviour, (vi) leakage warning behaviour, (vii) last resort warning behaviour, and (viii) directly communicated threat warning behaviour. Recently, some validation research regarding the typology of warning behaviours have been done by Meloy et al. (2014). In their study, they show that some of the warning behaviours were also present in a sample of German school shooters.

In this work, we focus on a limited subset of those warning behaviours, namely the ones that have the highest potential to be discovered in textual content in social media. These warning behaviours are 'leakage', 'fixation' and 'identification', as suggested by Cohen, Johansson, Kaati, and Mork (2014). The interested reader is encouraged to read the works by Meloy et al. (2012) in order to get a full understanding of all eight warning behaviours from a psychological point of view, but for the purpose of this chapter, we will give a simplified description of leakage, fixation, and identification warning behaviours.

Leakage can basically be defined as the communication of intent to do harm to a third party. Leakage usually signals research, planning and/or implementation of an attack. It is also common that a preoccupation with the target is present. Leakage has been shown to occur in many different cases of targeted violence – everything from school shootings to attacks on public figures.

Fixation can loosely be defined as any behaviour which indicates an increasing pathological preoccupation with a person or a cause. According to Meloy, Mohandie, Knoll, and Hoffmann (2015), fixation can be measured by:

- Increasing perseveration on the person or cause;
- Increasingly strident opinion;
- Increasingly negative characterisation of the object of fixation;
- Impact on the family or other associates of the object of fixation, if present and aware; and
- Angry emotional undertone. Furthermore, fixation is typically accompanied by social or occupational deterioration.

Fixation is a behaviour that can be noted in many subjects' daily life. It could be fixation of hobbies, sports, and admiration of public figures, or in an early stage of a love affair. Fixation is considered to be pathological when it disturbs the social functions (Meloy et al., 2015). While fixation is a pathological preoccupation on an external person or cause, the identification warning behaviour is a cause or action that is within the person.

Identification can be defined as a behaviour which indicates a desire to be a 'pseudo-commando' – i.e., have a warrior mentality, closely associate with weapons or other military or law enforcement paraphernalia, identify with previous attackers or assassins, or identify oneself as an agent to advance a particular cause (Meloy et al., 2015). There are many examples of terrorist and mass murderers that showed signs of the identification warning behaviours. Meloy and colleagues (2015) use Timothy McVeigh who bombed the Alfred P. Murrah Federal Building in Oklahoma City on April 19, 1995 as an example. Due to some circumstances, McVeigh became a soldier without an army. He had developed

a rigid and disciplined 'warrior mentality' and collected military paraphernalia, including weapons. Some other signs that supported McVeigh's identification warning behaviour was the fact that during the bombing he wore a T-shirt with the text "The tree of liberty must be refreshed from time to time with the blood of patriots and tyrants", and that he communicated his desire to be the 'first hero' to his sister in writing (Meloy et al., 2012).

DETECTING WARNING BEHAVIOURS

In Cohen et al. (2014), tools and techniques for detecting some of the warning behaviours described by Meloy et al. (2012) using linguistic markers were outlined. Although the article is focused on the identification of digital traces related to potential lone wolf terrorism, it can easily be extended to also cover other kinds of violence – since warning behaviours may indicate dynamic and accelerating risk of targeted violence across a variety of domains, including school shootings, mass murder, public figure attacks and assassinations, and terrorist acts (Meloy et al., 2014).

The basic idea outlined in our previous work (Brynielsson et al., 2013; Cohen et al., 2014) is to first make use of web crawlers to extract and download relevant data from websites and social media that can be readily accessed through the Internet. Next, it is suggested that various natural language processing algorithms are applied to the data in order to detect warning behaviours from digital traces that might indicate signs of violent extremism. These kinds of techniques are argued to be of great potential use by intelligence analysts for monitoring and searching for relevant information in large amounts of available data. However, it is important to stress that automated techniques can never replace a human analyst, but merely aid or support his/her work.

In Cohen et al. (2014), it is argued that the warning behaviours that are likely to be the most easily detectable in the subject's written communication in social media (e.g., extremist discussion boards) are leakage, fixation, and identification. For each of these three warning behaviours, a set of linguistic markers is proposed. The linguistic markers are intended to be used as inputs to computer algorithms so that they may be able to recognise signs of radical violence. As already mentioned, such algorithms are not supposed to make any kinds of automated decisions, but merely are intended to help the analyst navigate through the massive amounts of data and help him/her to focus on potentially relevant information.

Starting out with linguistic markers for leakage, it is observed that the leakage of one's intention to take any kind of violent action is likely to contain auxiliary verbs signalling intent (i.e., 'I will …', '… am going to …') together with words expressing violent action, either overtly or through euphemisms. As a starting point for being able to detect such linguistic markers signalling a violent intention, it is proposed that predefined word lists of violent actions are used, and to extend such a predefined list of words using lexical databases such as WordNet. Now, by lemmatising posts and tagging them with their part of speech, it would be possible to match the harvested posts to the extended word lists, and to flag hits as potential markers of leakage. Admittedly, many of such matches would probably be the result of false positives (e.g., due to ironic statements), but it is argued that the levels of false positives hopefully can be kept at an acceptable level if attention is restricted to websites or forums that are known to contain violent extremism related content.

When it comes to identification warning behaviour, this is quite complex since the definition given above covers quite a broad and complex range of phenomena. In Cohen et al. (2014), this is therefore simplified into three subcategories of identification: (i) identification with radical action (i.e., warrior

mentality), (ii) identification with role-model, and (iii) group identification. For the group identification, it is argued that identification with a group (or cause) can be expressed through a usage of positive adjectives in connection with mentioning the in-group. To detect positiveness, the use of sentiment analysis or opinion mining techniques is envisioned, while references to the in-group is hypothesised to be identifiable by counting the relative frequencies of first person plural pronouns such as 'we' and 'us', and checking whether such frequencies are higher than some predefined threshold. Identification of role-models or other radical thinkers is argued to be likely to be detectable through frequent quotations and mentions, but also possibly through similarities in language, since it is not uncommon to pick up terminology of one's role model, or even adapt to similar sentence structures. For this reason, authorship analysis algorithms or content analysis algorithms could potentially be one piece of the puzzle for a working system for the detection of identification linguistic markers. Identification with radical action (i.e., warrior mentality) can probably be spotted through a certain terminology in the same manner as the intent discussion above, while a sense of moral obligation could be captured through usage of words related to duty, honour, justice, etc.

Finally, for linguistic markers of fixation, we propose to simply count the relative frequency of key terms relating to named entities such as persons, organisations, etc., after making use of named entity recognisers to identify the named entities. When named entities of interest obtain a relative frequency which is higher than some pre-specified threshold, this can be used as an indication of a potential fixation with or to the named entity. Not setting the threshold extremely high would of course generate a large number of false positives on its own, but by combining it with other linguistic markers, this is hypothesised to be a useful marker.

The article by Cohen et al. (2014) has suggested a number of natural language processing techniques for detecting linguistic markers that can be useful for intelligence analysts when analysing the content of websites or social media in order to identify online warning behaviours. However the techniques are never implemented or evaluated, making it hard to judge the usefulness of such linguistic markers. In this chapter, we therefore take the proposed techniques one step closer to a real system by implementing some of the ideas into a prototype system. The inner working of this prototype system is described in more detail in following sections, but first we review some related work.

RELATED WORK

The idea to detect terrorism-related content on the Internet is not new, as it is believed that the detection of terrorist activities on the Internet may help prevent future terrorist attacks (Elovici, Kandel, Last, Shapira, & Zaafrany, 2004). In fact, there are currently calls for research projects by the EU research and innovation programme, Horizon 2020[1], which aim at detecting and analysing terrorism-related content on the Internet with the purpose of fighting terrorism. Previous attempts described in research literature include work carried out in the EU Seventh Framework Programme (FP7) project, INDECT[2] as well as the articles by Brynielsson et al. (2013), Elovici et al. (2004), and Last, Markov, and Kandel (2006), in which data mining and machine learning algorithms are proposed for learning to classify the textual content of websites as either terror-related or non-terror-related.

In the paper by Elovici et al. (2004), a methodology is proposed that allows for processing all Internet service providers (ISPs) traffic in real-time, allowing for large-scale monitoring. This can be contrasted with the methodology presented in Brynielsson et al. (2013), in which a much smaller subset of web-

sites, which are linked to known extremist websites, is targeted for analysis. Moreover, it is in the latter approach suggested that only user-generated content on the websites is analysed, not the web traffic as such. This is a quite important difference since the approach used in the prototype system described in this chapter, and the approach outlined in Brynielsson et al. (2013) analyse what people write; while the approach suggested in Elovici et al. (2004) builds on what information people access.

The approach described in Elovici et al. (2004) assumes that terror-related content usually viewed by terrorists and their supporters can be used by data mining tools to learn a 'Typical-Terrorist-Behaviour', but this is in our view problematic since many people who are not terrorists may be interested in reading content concerning the same topics as terrorists. For example, security researchers, intelligence analysts and journalists will be interested in reading information on terrorism websites, and are likely to be classified as 'suspected terrorists' using an approach such as the one proposed in Elovici et al. (2004). They argue that "missing one real terrorist in a haystack of legitimate users may be more costly than suspecting several legitimate users of being active terrorists" (p. 20). This may be true, but it is in general very hard for the general public to get information on how well surveillance systems work, making it hard to judge how much of our privacy it is worth losing just for the sake of a potentially more secure society.

Most readers are likely to have heard of intelligence agencies such as the National Security Agency (NSA), and various surveillance projects such as Echelon and PRISM. Although much secret information has become public knowledge due to the leak by Edward Snowden, not very much is known about the inner workings of these systems. For this reason, it is hard to compare our implementation of the prototype system with existing real systems.

IMPLEMENTATION OF LINGUISTIC MARKERS FOR DETECTING WARNING BEHAVIOURS

As explained above, the article by Cohen et al. (2014) discusses possible ways to implement a system for detecting warning behaviours based on some of the markers described by Meloy et al. (2012). Cohen et al. (2014) suggest using natural language processing techniques like lemmatisation, part of speech tagging, named entity recognition as well as lexical resources like WordNet. This is a common approach in natural language processing applications, which generally improves the quality of the resulting analysis. However, this is only true if the individual natural language processing components are accurate enough; if on the contrary, the natural language processing components are inaccurate, the errors will propagate and severely affect the end result.

This is a potential problem, since language used in social media is highly dynamic, noisy and productive, and tends to be highly multilingual. Such environments pose severe challenges for traditional natural language processing components and available natural language processing frameworks, since they are either based on precompiled lexica and rules (this tends to be the case for lemmatisers), or based on supervised machine learning where a classifier is trained on some manually annotated training data (which is normally the case for part of speech taggers and named entity recognisers). The problem with such approaches is that the trained classifiers and precompiled resources cannot keep up-to-date automatically with the constantly evolving linguistic environment in social media. Consequently, they will have limited recall, which of course is a serious deficiency in a monitoring scenario, since this means that there will be signs of violent actions that such a system will fail to identify.

Furthermore, adequate training data and lexical resources are not in abundance for languages other than English (and even for English, they can be domain-specific and less suitable to train models for application to security analysis in social media), and are costly and time-consuming to produce and maintain. We also note that the use of natural language processing components adds computational complexity to the system, which might be a limiting factor in a Big Data scenario where efficiency and scalability are important considerations.

Due to these concerns, we are aiming for a light-weight and resource-lean approach that presumes a minimum of pre-processing in order to be as efficient, scalable, and portable as possible. The proposed approach is based on simple lists of keywords, where a keyword can consist not only of single words, but also of multi-word units (e.g., the bigram 'Al-Shabab'). This means that we opt for recall rather than precision, and efficiency rather than representational sophistication. We substantiate this choice with the fact that the current application is a monitoring scenario, in which we cannot afford to miss relevant content; we thus argue that false positives are more acceptable than false negatives. This trade-off is an important consideration when dealing with monitoring tasks in user-generated content as we on the one hand would like to find as many of the real threats as possible, while on the other hand cannot afford a too high false alarm rate since this will be a burden for the analyst who in the end may stop using the system. Moreover, too low precision (i.e., low false alarm rate) also means an increased risk for privacy violations, something which is discussed in more detail in the section regarding ethical considerations.

We follow Cohen et al. (2014)'s definition of leakage in terms of 'intent to commit violent acts', and implement this marker as the combination of expressions of violent acts and intent. That is, it is not enough to mention violent acts in order for this marker to trigger; the user also has to express some form of intent in relation to the violent act. Violent acts are defined as three separate lists of terms: one relating to general expressions of violence (i.e., featuring terms like 'kill', 'assassinate', and 'bomb'), one featuring more specific terms referring to bomb ingredients, and one containing various types of weapons. Intent is defined by terms like 'going to', 'someone should', and 'plans'. This means that expressions like 'I am going to use an M-16' or 'someone should bomb the government' will trigger the marker, whereas expressions like 'I own a sniper rifle' or 'that was a good kill' will not.

For identification and fixation, we list a number of targets that could be relevant in the context of online violent extremism, ranging from individuals and organisations to ideologies and controversial topics: abortion, Anders Behring Breivik, communism, counter-jihad, cultural Marxism, EDL (European Defence League), infidels, Islamism, Jews, Kahanism, neo-nationalism, white supremacy, Charlie Hebdo, ISIS (Islamic State in Iraq and Syria), Anwar Al-Awlaki, Heil Hitler, and a list of famous mass murderers. Note that this is by no means an exhaustive list of targets that could be relevant for real-world monitoring scenarios; this list is simply meant as an example of topics that could be of relevance for the sake of demonstration. Considerably more effort would have been required from analysts and knowledge engineers to identify the relevant targets before a system like this could be put in real-world use.

For each mention of one of these targets, we count the occurrences of appreciative and depreciative expressions (i.e., we do a rudimentary form of sentiment analysis (Pang, Lee, & Vaithyanathan, 2002) of the targets), and define identification as expressions featuring a target in an appreciative context, and fixation as expressions featuring a target in a depreciative context. That is, an expression like 'I support the views of Breivik' would trigger the identification marker, whereas an expression like 'Charlie Hebdo is awful' would trigger the fixation marker. In reality, it would make sense to ensure that the thresholds of relative frequencies have to be passed before a marker (i.e., identification or fixation) is triggered –

since mentioning a named entity such as Charlie Hebdo once cannot be counted as a fixation. However, in this first prototype system this has been neglected in order to simplify the implementation.

Note that since we do not presume that the input data will be morphologically normalised (i.e., all words will be in lemma or basic form), we include morphological variants in the keyword lists (e.g., violence list contains both 'kill' and 'killed' and the list for Charlie Hebdo contains both 'Charlie Hebdo' and 'Charlie Hebdo's'). We also do not make any distinction between different parts of speech in the keyword lists, which means that the violence list contains both verbs like 'kill' and nouns like 'assassination'. The lack of part of speech tagging means that some words with polysemic meaning will trigger more often than they would have if part of speech tagging was used (e.g., the word 'conflict' can be both a noun and a verb); but as argued above, avoiding part of speech tagging allows for faster processing.

The following is a complete list of the 39 different targets used in the prototype implementation:

- violence – intent
- weapons – intent
- bomb ingredients – intent
- abortion – positive
- abortion – negative
- Anders Behring Breivik – positive
- Anders Behring Breivik – negative
- Anwar Al-Awlaki – positive
- Anwar Al-Awlaki – negative
- Charlie Hebdo – positive
- Charlie Hebdo – negative
- Columbine – positive
- Columbine – negative
- communism – positive
- communism – negative
- counter-jihad – positive
- counter-jihad – negative
- cultural Marxism – positive
- cultural Marxism – negative
- EDL – positive
- EDL – negative
- Heil Hitler – positive
- Heil Hitler – negative
- infidels – positive
- infidels – negative
- ISIS – positive
- ISIS – negative
- Islamism – positive
- Islamism – negative
- Jews – positive
- Jews – negative
- Kahanism – positive

- Kahanism – negative
- mass murderers – positive
- mass murderers – negative
- neo-nationalism – positive
- neo-nationalism – negative
- white supremacy – positive
- white supremacy – negative

Vocabulary Variation

The arguably most difficult problem when dealing with natural language in online data, and in particular when using keyword-based approaches is vocabulary variation, which is the situation when different people use different terms to refer to the same thing. As an example, consider the expressions of appreciation; there are literally hundreds, if not thousands of ways to express appreciation in English, and it would be impossible for an analyst to list them all a priori. Furthermore, language use is productive (especially in social media), which means that new expressions are invented and modified continuously. Consequently, even if we somehow could list all possible expressions of appreciation in English at this particular moment, the list would become outdated basically as soon as it was compiled (Karlgren, 2006). Keeping à jour with the productivity of language use is one of the most difficult challenges for social media monitoring in general.

One way to approach this challenge is to use unsupervised machine learning techniques that can learn to identify semantically similar terms in the data by simply reading lots of text. Such techniques are generally known as 'distributional semantic models' (Sahlgren, 2006; Turney & Pantel, 2010), and they work by collecting statistics on how terms in the data co-occur with each other. These statistics are used to identify terms that have similar co-occurrence profiles; if two terms have co-occurred significantly with the same *other* terms it means that they often can be substituted by each other in context, which is often used as criterion for semantic relationship in linguistics (Murphy, 2003).

As an example, consider a term like 'shit', which in its derogatory sense, is often used in sentiment analysis as a negative term signalling depreciation – and is thus of use in the prototype implementation as a keyword for the fixation marker. A standard lexicon will list a number of terms that are often used as synonyms to 'shit', like 'bad', 'terrible' and perhaps 'poor'. However, no single lexical resource (or, for that matter, no single human analyst) will list all possible terms used in online/social media to signal appreciation. Looking up the term 'shit' in an online lexicon continuously trained on millions of web documents each day produces the following result:

*[shit: sh1t, sh*t, $hit, shyt, sht, sh-t, shite, shyte, crap, shiz, dogshit, horseshit, s--t, dog shit, crapola, crud, dipshit ...]*

'Shit' in the derogatory sense may not even be listed in traditional lexica, and even human analysts would be hard pressed to come up with all these alternative terms. This means that without the use of data mining tools like distributional semantic models, we would miss all expressions that use such productive terminology. Lexical productivity might not be such a severe problem when it comes to named entities like Anders Behring Breivik and ISIS, but it will be a significant problem when looking for expressions of, for example, violence, intent and appreciation/depreciation. This demonstrates the importance of

not solely relying on manually compiled knowledge bases such as WordNet or Wikipedia; we simply cannot afford to assume that vocabulary usage will remain static and predictable if we want to monitor language use in social media.

In order to handle problems with vocabulary variation, we use a distributional semantic lexicon to ensure our lists of terms are exhaustive and up-to-date with current language use. This is done by continuously consulting the online lexicon for new relevant terms; if any such terms are found, they are included in the keyword list in question. It is important to stress that each addition to the keyword lists needs to be confirmed by a human operator, since not all distributional neighbours will be relevant for the intended meaning of the keywords.

As an example, a distributional lexicon might suggest the terms 'ISIL' and 'Iowa' as related to 'ISIS'. The term 'ISIL' would be relevant in the present scenario, since it refers to the Islamic State in Iraq and the Levant, which is a synonym to ISIS. The term 'Iowa', on the other hand, would not be relevant, since it refers to another meaning of ISIS (i.e., the Iowa Student Information System).

TRIANGULATION WARNING BEHAVIOURS

The approach described above is both efficient and scalable, and manages to handle vocabulary variation by using a distributional semantic lexicon to ensure that the term lists are constantly up-to-date. As such, the proposed prototype system is geared towards high recall, and is designed to be able to capture as much relevant material as possible. However, it will also capture a considerable amount of irrelevant material, which will affect the precision of the system.

Recall is the proportion of relevant material that is detected by the system, whereas precision is the proportion of material detected by the system that is actually relevant. It is a well-known fact in information retrieval that there is a trade-off between recall and precision; a system that is geared towards recall will typically have lower precision, while a system that is geared towards precision will typically have lower recall. In the current application scenario, it is arguably more important to aim for high recall, since it may have severe consequences if we miss relevant information. However, the precision has to be reasonable as well since a too low precision will make the system more or less useless for the analysts due to the high false alarm rate. Moreover, it can cause unnecessary privacy implications as discussed further in the section on ethical considerations.

Now, looking at each individual marker by itself will likely lead to unacceptable levels of false positives (i.e., hits that are not relevant). As an example, expressions of violent intent is, unfortunately, prevalent in social media, both in expressions where the intent to commit violence is literal but perhaps not very realistic (e.g., as used by people who are upset about something), and in metaphorical expressions like 'we are going to kill tonight' (i.e., we are going to have a good time tonight). Furthermore, some keywords, like 'ISIS' may occur frequently due to their generally topical nature.

One way to reduce the amount of false positives (i.e., improve the precision) while still keeping the markers as general as possible to ensure high recall is to 'triangulate' warning behaviours based on the suggested markers. This means that we would look for content that trigger several different markers, and perhaps also over a period of time. As an example, a website that simultaneously triggers the leakage and fixation detectors would be more interesting than a website that only triggers the leakage detector. Furthermore, if the same website (i.e., trigger both the leakage and fixation markers) also triggers the

identification marker but at a different time, it should further increase our interest in that website. Since we are operating with three main types of linguistic markers, we refer to this process as 'triangulation'.

Our prototype implementation is build using existing commercial tools for text analysis and social media monitoring provided by Gavagai. As already discussed above, we use a distributional semantic model to ensure the lists of keywords are exhaustive and up-to-date. We also use a monitoring tool provided by Gavagai that reads large streams (i.e., several millions of documents per day) of social media content, primarily weblogs and discussion forums. The monitoring tool performs a keyword search in the incoming data for occurrences of the keywords included in the markers that already have been described. For each marker, the monitoring tool outputs a list of Uniform Resource Identifiers[3] (URIs) and the frequency of occurrence of the marker in the document. These lists of URIs are then input to an analysis script where it is possible to define how many or which markers that have to be triggered in order for an URI to be shown as potentially interesting to the user of the system.

SOLUTIONS AND RECOMMENDATIONS

In our prototype implementation, we have collected data from a number of data sources for a few days. The data has been compared against a number of basic linguistic markers consisting of lists of keywords as described above. Each time a keyword is triggered for a particular URI, a counter is increased. It is important to highlight that in a real-world system, the keywords that are used needs to be developed further; the current keywords are just to illustrate the concept that is used in the system. For each linguistic marker we obtain a list of URIs that has triggered the particular marker, and a count of how many times the URI has triggered the marker during the time period.

Now, it must be realised that most of the URIs that hit a single linguistic marker will be completely unrelated to both violent extremism in general and specific attacks that are about to take place. Since we currently use a list of keywords consisting of terms that may occur in many contexts, the false alarm rate for a single keyword will be high. However, the idea is that by combining several such linguistic markers, the precision of the obtained results will increase.

To just give an example, for a time period of three days, we got 130 hits for URIs that contained keywords matching both violence (intent), and at least one of bomb (intent) or weapons (intent). Adding an extra filter to these results so that they also should contain a negative sentiment against the Jews, we ended up with a list consisting of only four URIs. This demonstrates how triangulation can be used to reduce the pool of 'risky' individuals or URIs. Taking a closer look at the hits, these originated from the website of a Neoconservative Right magazine, a Christian blog, and two anti-Jewish blogs. Probably none of these sites would have caused an analyst to take any further action due to their content (unless they were added to a list of websites of interest beforehand). As a first selection, this approach is obviously much faster. We only need to check a list of four websites or URIs manually compared to reading through all data from the collected sources, corresponding to an enormous amount of work.

Ethical Considerations

There are many ethical and privacy concerns that come with this kind of technologies. Some important questions that should be raised before implementing any kind of surveillance techniques are: What are the benefits of this kind of surveillance, and what is the potential harm? Who should be entitled to use

this kind of techniques, and during what circumstances? Is the potential threat of violent extremism enough to warrant the use of these kinds of techniques?

First of all, it is important to think about what effect this kind of surveillance would have on citizens' privacy. As stated in Article 8 of the European Convention on Human Rights (ECHR), or more formally, the Convention for the Protection of Human Rights and Fundamental Freedoms, everyone has a right to a private life and there shall be no interference by a public authority with the exercise of this right (except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety, etc.).

For this reason, law enforcement agencies, security researchers, and any other actors involved in any kind of web surveillance has to balance the need for a secure society with a respect for online privacy, and make sure that this complies with laws and directives such as the EU data protection directive (or more officially Directive 95/46/EC on the protection of individuals with regard to the processing of personal data, and on the free movement of such data). Such directives obviously should be respected, but there are also additional ethical considerations that have to be made. Ultimately, if people start to censor themselves online due to the fear of being monitored by the authority or even private companies, important values of the society that we try to protect will be lost. For this reason, ethical and privacy concerns have to be considered, keeping the potential intrusion on people's privacy to an absolute minimum.

Basically, what our implemented prototype system does is that it processes publically available user generated content, and counts the number of occurrences of various terms. This is on a technical level not very different from what ordinary web spiders do when they index web pages for search engines. Few people would argue that web spiders and search engines pose a threat to online privacy. However, what mainly differs is the purpose. Search engines are intended to allow for general web searches while the purpose of our implemented system is to detect signs of radical violence, which may contain potentially sensitive information. Moreover, few would object against law enforcement officers using search engines manually to find suspicious websites; while probably more people would object against automating this endeavour in order to allow for continuous and large-scale applications of the same method.

While we are so used to search engines and manual search queries that we do not consider web spiders to be a threat against our online privacy (rather, we are more concerned about the search engines' use of tracking cookies for profiling purposes), automatic 'crawling' of crime- or terrorism-related content is something else that we are more uncomfortable with. Arguably, it is more of an intrusion on people's privacy to scan through their e-mails in order to sell targeted ads than to process publically available user generated content for any purpose. However, an important difference is that their customers in general have provided some kind of 'informed' consent.

In order to attempt to make minimal harm to online privacy, we have run our prototype system for just a short period of time. We have minimised the number of people who have access to the results (i.e., matching URIs), and limited ourselves to just look at the top-N matching results in our evaluation after triangulating several linguistic markers. Additionally, we have for ethical reasons chosen not to show any of the collected URIs to the readers and we are only presenting the results in a way to avoid re-identification of any URIs or the creators behind the user generated content. All the collected URIs where deleted after the preparation of this chapter.

Whether or not this kind of systems should be used operationally is more of a political and legislative issue than a scientific question, but in case such systems are used operationally, it is important to use various safeguards in order to protect against potential misuse. Examples of such safeguards would be to only provide authorised users access to the system, encryption of all collected data, and protected logs

that would keep track of all operations and searches that the users of the system make. It could then be controlled so that the users of the system meet regulations for how such monitoring should be undertaken.

CONCLUSION

In this chapter, we have argued that the increasing use of websites and social media to spread terrorism propaganda and communicate with like-minded individuals, is both a challenge and an asset for intelligence analysts and other involved in counter-terrorism. There is a clear risk that this allows people who otherwise have no obvious connections to terrorist groups to be radicalised, but the digital communication also leave traces that potentially can be used by analysts to detect people who are about to commit violent extremism-related actions.

Psychologists have previously discovered a number of warning behaviours which often signal violent attacks on public figures, mass murders, and acts of lone wolf terrorism before they occur. The problem is that huge amounts of online textual data are generated each day, making it unfeasible (or even impossible) to manually read all available material. For this reason, we have made an attempt to implement a number of linguistic markers that might indicate that the writer shows signs of some of the warning behaviours described by Meloy et al. (2012). Our prototype system detects signs of linguistic markers automatically.

The potential advantages of such a system is that it would help intelligence analysts to focus on a smaller amount of user generated content of topical relevance for their work and thereby increase their efficiency, which in turn hopefully can lead to a more secure society. Our prototype system is a first step towards a tool that actually could be used by analysts in order to detect warning behaviours, but it also brings about ethical and privacy issues that have to be considered before putting such a system into use.

It is important to realise that the proposed system is not a way to replace the human analyst; this could never be the case for this kind of issues. Rather, we see it as a complementary way to manual approaches such as the excellent initiative by the U.K. Government (i.e., https://www.gov.uk/report-terrorism) where people can report illegal terrorism-related material they find online. Obviously, it is far from possible that all terrorist attacks can be discovered in advance using any natural language processing tool for social media analysis, since not all attackers will show warning behaviours that can be discovered on the Internet. However, there have been many cases where such warning behaviours have been identified after an attack have taken place – including recent cases such as the 2015 Charleston church shootings. Before the deed, the suspect posted photos and a manifesto on a white supremacist website. In such cases, this kind of tools could be very useful, given that a high enough classification performance can be achieved to detect such warning behaviours.

FUTURE RESEARCH DIRECTIONS

There are several directions for future work. Reducing the number of false positives and developing guidelines for how to handle false positives is one obvious direction. To reduce the number of false positives, the techniques used to detect the three types of linguistic markers described in this work needs to be refined. Detecting warning behaviours in written text is not a trivial problem and the methods and

techniques that we have used in this work are promising, but there is a need for further development if this approach should be used in practice. To detect warning behaviour such as fixation, it is important to also consider the subject's changes in behaviour over time – both when it comes to perseveration and negative characterisation. Using an approach where time is a parameter is therefore a desirable way forward when detecting fixation.

Analysing written text to obtain knowledge about the psychological meaning is not something new. Tausczik and Pennebaker (2010) describe how the text analysis tool, Linguistic Inquiry and Word Count (LIWC) counts words in psychologically meaningful categories. LIWC uses a number of different categories and calculates to what degree people are using the different categories in written text. This approach has been evaluated and tested in a number of different studies such as the one described in Cohn, Mehl, and Pennebaker (2004), and Davison, Pennebaker, and Dickerson (2000). Adding a deeper psychological meaning to the analysis could be another way forward when it comes to detecting leakage and identification warning behaviours.

REFERENCES

Bjelopera, J. P. (2013). *American jihadist terrorism: Combating a complex threat. CRS Report for Congress*. Washington, DC: Congressional Research Service.

Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C., & Svenson, P. (2013). Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, *2*(11). doi:10.1186/2190-8532-2-11

Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246–256. doi:10.1080/09546553.2014.849948

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693. doi:10.1111/j.0956-7976.2004.00741.x PMID:15447640

Davison, K. P., Pennebaker, J. W., & Dickerson, S. S. (2000). Who talks? The social psychology of illness support groups. *The American Psychologist*, *55*(2), 205–217. doi:10.1037/0003-066X.55.2.205 PMID:10717968

Elovici, Y., Kandel, A., Last, M., Shapira, B., & Zaafrany, O. (2004). Using data mining techniques for detecting terror-related activities on the web. *Journal of Information Warfare*, *3*(1), 17–29.

Europol. (2012). *TE-SAT 2012: European Union terrorism situation and trend report*. European Law Enforcement Agency.

Greenwald, G. (2013, July 31). XKeyscore: NSA tool collects 'nearly everything a user does on the internet'. *The Guardian*. Retrieved from http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data

Karlgren, J. (2006). New text - New conversations in the media landscape. *ERCIM News*. Retrieved from http://www.ercim.eu/publication/Ercim_News/enw66/karlgren_2.html

Last, M., Markov, A., & Kandel, A. (2006). Multi-lingual detection of terrorist content on the web. *Lecture Notes in Computer Science*, *3917*, 16–30. doi:10.1007/11734628_3

Meloy, J. R. (2011). Approaching and attacking public figures: A contemporary analysis of communications and behaviour. In C. Chauvin (Ed.), *Threatening communications and behaviour: Perspectives on the pursuit of public figures* (pp. 75–101). Washington, DC: The National Academies Press.

Meloy, J. R., Hoffmann, J., Guldimann, A., & James, D. (2012). The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences & the Law*, *30*(3), 256–279. doi:10.1002/bsl.999 PMID:22556034

Meloy, J. R., Hoffmann, J., Roshdi, K., & Guldimann, A. (2014). Some warning behaviors discriminate between school shooters and other students of concern. *Journal of Threat Assessment and Management*, *1*(3), 203–211. doi:10.1037/tam0000020

Meloy, J. R., Mohandie, K., Knoll, J. L., & Hoffmann, J. (2015). The concept of identification in threat assessment. *Behavioral Sciences & the Law*, *33*(2-3), 213–237. doi:10.1002/bsl.2166 PMID:25728417

Meloy, J. R., & O'Toole, M. E. (2011). The concept of leakage in threat assessment. *Behavioral Sciences & the Law*, *29*(4), 513–527. doi:10.1002/bsl.986 PMID:21710573

Murphy, M. L. (2003). *Semantic relations and the lexicon - antonym, synonymy and other paradigms*. Cambridge, U.K.: Cambridge University Press. doi:10.1017/CBO9780511486494

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA: Association for Computational Linguistics.

Ravndal, J. A. (2013). Anders Behring Breivik's use of the Internet and social media. *Journal EXIT-Deutschland*. Retrieved from http://journals.sfu.ca/jed/index.php/jex/article/view/28

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* (Doctoral dissertation). Department of Linguistics, Stockholm University, Sweden.

Schmid, G. (2001). *Report on the existence of a global system for the interception of private and commercial communications* (ECHELON Interception System, 2001/2098[INI]). European Parliament.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi:10.1177/0261927X09351676

Torok, R. (2013). Developing an explanatory model for the process of online radicalisation and terrorism. *Security Informatics*, *2*(6), 1–10.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188.

Weimann, G. (2012). Lone wolves in cyberspace. *Journal of Terrorism Research*, *3*(2). doi:10.15664/jtr.405

ENDNOTES

1. Horizon 2020 is the biggest European Union research and innovation programme ever with nearly €80 billion of funding available over seven years (2014 to 2020).
2. The INDECT project (Intelligent information system supporting observation, searching and detection for security of citizens in urban environment) is a research project, allowing new, advanced and innovative algorithms and methods aiming at detecting and counteracting threats and criminal activities, affecting citizens' safety. Please refer to http://www.indect-project.eu/ for more information.
3. The URI is a string of characters used to identify a name of a resource, such as a blog post.