Instruments and Tools to Identify Radical Textual Content

Josiane Mothe, Md Zia Ullah, Guenter Okon, Thomas Schweer, Alfonsas Juršėnas, Justina Mandravickaitė

information

Article

Instruments and Tools to Identify Radical Textual Content

Josiane Mothe [1,*], Md Zia Ullah [1], Guenter Okon [2], Thomas Schweer [2], Alfonsas Juršėnas [3] and Justina Mandravickaitė [3]

[1] Institut de Recherche en Informatique de Toulouse, IRIT UMR5505 CNRS, Université de Toulouse, Componsante INSPE UT2J, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, F-31400 Toulouse, France; mdzia.ullah@irit.fr
[2] Institut für Musterbasierte Prognosetechnik (IfmPt), 46145 Oberhausen, Germany; analyse@bsmg-okon.de (G.O.); thomas.schweer@ifmpt.de (T.S.)
[3] Baltic Institute of Advanced Technology (BPTI), Pilies g. 16, 01124 Vilnius, Lithuania; alfonsas.jursenas@bpti.eu (A.J.); justina.mandravickaite@bpti.eu (J.M.)
* Correspondence: josiane.mothe@irit.fr; Tel.: +33-561556444

Abstract: The Internet and social networks are increasingly becoming a media of extremist propaganda. On homepages, in forums or chats, extremists spread their ideologies and world views, which are often contrary to the basic liberal democratic values of the European Union. It is not uncommon that violence is used against those of different faiths, those who think differently, and members of social minorities. This paper presents a set of instruments and tools developed to help investigators to better address hybrid security threats, i.e., threats that combine physical and cyber attacks. These tools have been designed and developed to support security authorities in identifying extremist propaganda on the Internet and classifying it in terms of its degree of danger. This concerns both extremist content on freely accessible Internet pages and content in closed chats. We illustrate the functionalities of the tools through an example related to radicalisation detection; the data used here are just a few tweets, emails propaganda, and darknet posts. This work was supported by the EU granted PREVISION (Prediction and Visual Intelligence for Security Intelligence) project.

Keywords: cybercrime; radical content detection; text analysis; text mining; information extraction; key-phrase extraction; graph-based representation

Citation: Mothe, J.; Ullah, M.Z.; Okon, G.; Schweer, T.; Juršėnas, A.; Mandravickaitė, J. Instruments and Tools to Identify Radical Textual Content. *Information* **2022**, *13*, 193. https://doi.org/10.3390/info13040193

Academic Editor: María N. Moreno García

Received: 24 November 2021
Accepted: 18 March 2022
Published: 12 April 2022

1. Introduction

The Internet and social networks are increasingly becoming a media of extremist propaganda. On homepages, in forums or chats, extremists spread their ideologies and world views, which are often contrary to the basic liberal democratic values of the European Union. It is not uncommon that violence is used against those of different faiths, those who think differently, and members of social minorities. Especially in times of crisis such as the Corona pandemic, conspiracy theorists and radicals are increasingly popular [1–3].

The text analysis carried out in the PREVISION project deals with problematic content in the field of Internet-based online communication. The focus is primarily on extremist or propagandistic content. The aim is to develop lexicons that can be used to (a) automatically determine the degree of radicalisation of propaganda material (keywords: ideologization and indoctrination) and (b) identify contents that give reasons to assume the preparation of politically motivated acts of violence (keywords: activism, recruitment). Discriminatory and offensive contents (e.g., hate speech) play only a secondary role in the analysis, although problematic subjects such as extremism, propaganda and political activism naturally overlap with problem areas such as conspiracy theories, disinformation, discrimination, glorification of violence and threats (See the point "category systems": Jünger, Jakob u. Chantal Gärtner, Datenanalyse von rechtsverstoßenden Inhalten in Gruppen und Kanälen von Messengerdiensten am Beispiel Telegram. Studie der Universität

Greifswald. Landeszentrale für Medien NRW (Hg.), Düsseldorf 2020). The project focuses on English-language Islamist content.

The following developments are currently noted. On the one hand, Islamist propaganda on the Internet is at present on the decline. This is probably primarily due to the military weakening of the "Islamic State," from which the online propaganda of the self-proclaimed caliphate has also suffered (See Islamistische Propaganda in sozialen Netzwerken geht zurück, https://www.zeit.de/gesellschaft/zeitgeschehen/2019-04/lageberichtislamismus-soziale-netzwerke-propaganda-rueckgang (accessed on 23 November 2021). The fact that Islamist network content is deleted more quickly or effectively is another important fact, for example, right-wing extremist content. It seems to indicate that right-wing extremist content is treated with greater tolerance. Looking at the deletion and blocking rates it is noticeable, that providers such as YouTube, Facebook and Instagram now have quite high deletion and blocking rates, while the rate for the messenger service Telegram is below average.

Nevertheless, political actors such as the Islamic State continue to rely on social networks, messenger services and video portals such as Facebook, Twitter, Instagram and Telegram for ideologisation, indoctrination and recruitment, with Islamist actors operating primarily on Telegram and right-wing extremist groups preferring Discord. Both have in common the desire to escape from state restrictions through bots, face accounts and closed groups (See [4], p. 32).

The target groups of extremist online communication are, among others, specifically young people, in particular people with experiences of marginalisation and discrimination. (Thus, in 5163 Facebook posts (including comments), 1877 keywords were found "which emphasized one's own victim role and the enemy image of the West." (See [4], p. 39)) Extremist groups of the most diverse nature show here overlapping in terms of content and language or use common interpretative patterns and argumentation. These include their anti-pluralistic understanding of society and politics and their claim for absoluteness. In their dichotomous worldview, they stylise their own group as victims or stigmatise the foreign group as the enemy ("us versus them" narrative) (Engelhorn, Jochen, Extremismus und Sprache: Ein Vergleich extremistischer Deutungsmuster am Beispiel der Finanzkrise. Grin-Verlag, Ravensburg 2008).

Identifying problematic content is, therefore, a challenging and prioritised task for any government agency. The goal could be an automated search and classification of extremist contents, also using machine learning and artificial intelligence [5]. (To this end, the project analyzed Islamist propaganda material (texts, videos), posts and chat communication, on the basis of which categories and algorithms were developed.) Akinboro et al. address the problem of derogatory and offensive posts on social media platforms and discuss methods that can be used to detect this content [6]. Approaches described in this study include natural language processing (NLP), the deep learning approach, multilevel classification, hybrid approaches and approaches for recognising multilingual contexts, which are becoming increasingly important. Problems identified in the study were related to data sparsity, word ambiguity and sarcastic meaning of words and sentences, as the context is usually more complex than the written word. Another specific problem regards photographs, which contain text. So far there is little knowledge on this specific topic to date, he said [6].

In addition, Akinboro et al. address the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), which is also known as the Five-Factor Model (FFM). The Big Five personality traits can be used to categorise a person's character—the combination makes the individual unique. The FFM serves as the basis for Facebook's Map-Reduce Back Propagation neural network (MRBPNN), with the accounts themselves and activity on Facebook being valuable indicators of a user's character traits [6]. Results from the 300-user sample and approximately 10,000 status updates collected for the "My Personality" project show that the MRBPNN is more efficient than the Back Propagation Neural Network (BPNN) and the Support Vector Machine (SVW) [6].

In this context, it is critical to note that extremists or terrorists have different personality traits or personality structures. The biographies analysed in this project confirmed the common belief in terrorism research i.e., that the life histories of violent Islamist perpetrators are extremely diverse. It is also a misconception to assume that "terrorists" are mentally ill i.e., "crazy". Almost all studies on radicalisation and psychopathology have shown, that terrorists—with the exception of the so-called 'lone wolves'—are no more or less mentally insane than the rest of the population [...] Almost always wrong are [...] assumptions that explain radicalisation with a single cause. "Terrorists are not all fatherless, uneducated or poor, nor do they always come from large families, have doctorates, or have rich parents" [7].

Furthermore, the risk assessment developed as part of Prevision is not about assessing whether or why a person becomes radical, but rather what potential degree of danger a person, already classified as extremist, has in terms of committing a politically motivated act of violence. "The majority of Salafists living in Europe [practice] their puritanical interpretation of religion in private [...] and categorically reject political activism and violence" [4]. Scruton distinguishes between "cognitive" and "violent" extremism in this context [8].

Ali et al. provide a framework for analysing traffic events and accidents. Data from sensors and additionally from social media accounts (such as Facebook and Twitter) are taken as a basis and these are analysed by using the "latent Dirichlet Allocation (OLDA)" and "Bidirectional Long Short-Term Memory (Bi-LSTM)" models. Because of the social media data, real-time data can more easily be captured. The results show that the models can structure and categorise unstructured social media data and the basis for this adequately represents the traffic flow or traffic situations with an accuracy of 97% [9].

OLDA or the sentence segmentation approach could also be applied to the analysis of extremist texts, especially with regard to the evaluation of chat messages or the labelling of relevant and non-relevant messages.

There are many forms of violent radicalisation [10] such as ultra-right-side (associated with fascist, racialist/racist, ultra-nationalist motives), politico-religious (associated with a political reading of religion and the defence of a religious identity, whatever the religion is), ultra-left side (articulated around anti-capitalism and the transformation of a political system perceived as a generator of social inequalities), unique cause ones (e.g., environmental, animal rights, anti-abortionists, homophobic, anti-feminist, etc.). There are also different media where radical people express themselves: social media, web 2.0, specific newspapers, chat boxes, forums in different languages or dialects (each media has also its specific form of expression e.g., tweets vs. newspapers). The Internet and social media are attractive to extremist groups for various reasons [11–13]:

- Internet and social media represent an ideal opportunity for self-expression and communication.
- Internet and social media can reach millions of addressees around the world in a very short time.
- Internet and social media offer an ideal opportunity for networking with like-minded people.
- On Internet and social media, social control, not only by the security authorities but also by the social environment is made considerably more difficult.

The linguistic approach can play a central role in the assessment of the degree of radicalisation and the potential danger of extremist groups.

Considering the question of how radicalisation processes can progress in the context of computer-mediated communication and how people in forums, chat rooms and social networks are susceptible to extremist and radical content, different explanatory models and theories can be used. These include the Social Identity Model of Deindividuation Effects (SIDE-Model) [14,15] and the Social Identity Approach [16]. In this research area, the focus is particularly on processes of group formation and group dynamics that can be triggered by certain characteristics of the Internet (such as the possibilities for anonymity) [17].

The SIDE-Model assumes that when an individual's identity is predominant, the perceived group homogeneity (of the group relevant to the individual) decreases and the individual orientates himself or herself primarily towards his or her norms and values. In contrast, when an individual's social identity (feeling of belonging to a group) is predominant, the perceived group homogeneity increases and the individual then orientate themselves primarily to the group norms and values. Circumstances of computer-mediated communication such as anonymity and identifiability play an important role in this context. Anonymity reinforces the described processes and a low level of identifiability leads to a person's orientation towards their norms and values. With increasing identifiability, however, the orientation of the individual towards group norms increases [18]. In short, it assumes that "processes of social identity and identification with groups can lead to group-conform behaviour" [17].

The theory of social identity assumes that individuals strive for positive social identity and acquire it through membership of one or more groups and the emotional significance of this membership [19]. Group members gain or lose prestige through a comparison with other (relevant) groups, which serves to strengthen their own social identity.

The author (Thomas Schweer) proceeds from the following theses: Extremist propaganda on the Internet or in social media influences individual radicalisation processes. Hate speech or racist expressions and stigmatisation of ethnic, religious and social minorities by extremist actors on the Internet encourage concrete acts of violence. If fantasies of violence are expressed, the risk of concrete violent action increases. If communication in social media is stopped, the danger of concrete violent action increases. The degree of radicalisation of Internet-based propaganda can be measured by various indicators. In addition to the degree of radicalisation, the degree of dissemination and the target groups are also of analytical interest. Young people, in particular, are receptive to Internet-based propaganda [20]. Some of the exemplary indicators for the degree of radicalisation are stated as follows:

- The banning of a site by state authorities is an indicator of a high degree of radicalisation.
- If a site calls for violence against people and/or objects, this is an indicator of a high degree of radicalisation.
- If a propaganda site explicitly calls on people to join extremist/terrorist groups, this is an indicator of a high degree of radicalisation.
- If a propaganda site calls for sympathy with persons or groups who have been involved in politically motivated (violent) acts in the past, this can be considered as a high degree of radicalisation.

The more conspiratorial information is shared on the net (darknet), the higher the degree of radicalisation of the acting actors. Text analysis can also be extended to image analysis to identify extremist symbols. Symbols used can also provide clues to the degree of radicalisation of an individual or an actor.

Law Enforcement Agencies (LEAs) are lacking tools to help them in their daily analysis.

This paper presents a set of instruments and tools developed to help investigators to better address hybrid security threats, i.e., threats that combine physical and cyber attacks. The project organises five representative and complementary use cases, including the protection of public spaces and the fight of illicit trafficking of antiquities, in full compliance with privacy requirements applicable law. Here, we illustrate the tools to support security authorities in identifying extremist propaganda on the Internet and classifying it in terms of its degree of danger. This concerns both extremist content on freely accessible Internet pages and content in closed chats. We illustrate the functionalities of the tools through an example related to radicalisation detection; the data used here are just a few tweets, emails propaganda, and darknet posts. These tools have been designed and developed within the EU granted PREVISION project (http://www.prevision-h2020.eu/ (accessed on 23 November 2021) and https://cordis.europa.eu/project/id/833115/fr (accessed on 23 November 2021) Grant N° 833115) (Prediction and Visual Intelligence for Security Intelligence).

The main originality of the tools are as follows:

- The first series of tools we developed relies on a lexicon built thanks to key-phrase extraction. Recent related work has developed new algorithms for key-phrase extraction [21–23]. In the literature, these key phrases are used for information retrieval purposes. Here, we develop an original use of the key phrases, which is to rank documents, not according to a query, but rather according to the lexicon itself. Because the lexicon is representative of a sub-domain an LEA is interested in (e.g., religious extremism), it is then possible to order the texts according to their inner interest for the user.
- Document ordering is based on two means. One relies on expert knowledge of the importance of some criteria. In this, our models are more task-oriented than the usual models.
- Current search engines do not explain to the user the results that they retrieve. As opposed to that, here, the link between the lexicon and the text is highlighted so that the user can understand the reason for the document order (and can agree/disagree with the results).
- Visual tools complement the underlying representation of texts and help the user understand what the texts are about by an overview of the main important terms.
- One fundamental point of PREVISION is that it integrates all the elementary tools into a complete processing chain that is directly usable by LEAs; such a platform does not exist where the user keeps control of the system results.

The tools' functionalities and results are illustrated considering an example use case in link with radicalisation; this is described in the methodology Section 4. Section 3 presents an overview of the methods and tools we developed. Section 4 presents the data set used for the illustrative use case as well as the development framework. Section 5 details the results when using the methods and tools on the use case. Finally Section 6 discusses and concludes this paper.

2. Related Work

Many related works focus on detecting radical content by casting the problem into a classification one. The problem is then to predict whether content is radical or not [24–26]. The usual classification methods used are mainly supervised ones such as support vector machine, random forest, Naive Bayes, Adaboost or neuron networks like Bert [24,27,28]. While the former methods generally rely on manually defined features, the latter relies on features that are automatically extracted [24,27,29]. Lexicons are core for feature extraction in this domain [30–32].

Cohen et al. [29] presented tools and methods to detect traces as weak signals for violent radicalisation on extremist web forums; they focused on linguistic markers. Based on a study of behavioural markers for radical violence, the authors give an overview of text analysis techniques that could be used such as translation, sentiment analysis, and author recognition. The authors also detail some linguistic markers associated with the behavioural markers. Nouh et al. [26] also shown there are distinguishable textual, psychological, and behavioural properties that can be extracted from radical contents. Furthermore, the authors used vector embedding features to detect such contents. Ashcroft et al. [24] focused on detecting messages for radical propaganda on Twitter. The authors consider a classification method to learn whether a tweet is supportive of Jihadist groups. To represent the texts, the authors defined three types of features: stylometric features, time-based features and sentiment-based features. Araque and Iglesias [33] use an emotion lexicon for radicalisation detection. They extract emotion-driven features considering using this lexicon as well as word embedding. From their experiments, the authors conclude that both are useful for radicalised content detection.

Gaikwad et al. [27] also provide a comprehensive review of the literature on online extremism detection.

Some platforms integrate different text analysis tools, helping users to understand large document collections.

Tétralogie (https://atlas.irit.fr/PIE/Outils/Tetralogie.html accessed on 31 March 2022) is one of them. Tétralogie consists of several agents that communicate with each other on users' demands to build overviews under the form of histograms, networks, and geographical maps. Meta-data and document content are extracted before being mined using various data analysis methods [34,35].

Tableau (https://www.tableau.com accessed on 31 March 2022) is an intelligence-driven business. It offers functionalities for the user to interact with the analysed data and by building visualisations with drag and drop, employing AI-driven statistical modelling.

Curatr is an online platform for the exploration and curation of ancient literature. The platform provides a text mining workflow. It "combines neural word embeddings with expert domain knowledge to enable the generation of thematic lexicons, allowing researchers to curate relevant sub-corpora from a large corpus" [36].

InfraNodus (https://infranodus.com/ accessed on 31 March 2022) is a tool that reveals the relations and patterns in data [37]. It helps analyse texts by displaying term relationships under the form of networks. Different text formats can be analysed. The tool uses network analysis and graph visualisation for generating insights from any text. It includes topic modelling and data mining functionalities to perform sentiment analysis and term clustering for example.

Methods for automatic keyword extraction are mainly of two categories: supervised and unsupervised. The problem of keyword extraction is generally cast into a binary classification problem for supervised methods and as a ranking problem for unsupervised methods [38]. Supervised methods are considered better than non-supervised ones, specifically for domain-specific data [39]. The main advantage of unsupervised methods, however, is they do not need any training data and can produce results in any domain. State-of-the-art unsupervised approaches for key-phrase extraction are mainly based on TF-IDF [40], clustering, and graph-based ranking [38]. El-Beltagy and Rafea [40], for example, proposed KP-Miner, which achieved the best performance among unsupervised models in SemEval 2010 tasks [41]. This method modified TF-IDF to compute the score of key-phrase candidates. Another unsupervised approach is Rapid Automatic Keyword Extraction (Rake), developed by Rose et al. [23]. Rake considers the word degree, word frequency, and the ratio of the degree to frequency to weigh the candidate keywords. Campos et al. [22] proposed the key-phrase extraction algorithm YAKE, considering some local features from the single document such as the frequency of the word, their position in the sentence, and the context around. This simple key-phrase extraction algorithm generates high-quality candidates key-phrases.

In our Jargon detection tool (See Section 3.1), we utilize Yake due to its simple computation and ability to extract high-quality candidate key-phrases. We also include Rake in our Jargon tool to compare the extracted results with Yake.

3. Overview of Methods and Tools for Identification of Radical Content

To answer the LEAs need for text analysis in the task of radicalisation detection, we consider several aspects:

- Detection of the main key-phrases of the domain to build a domain-oriented lexicon;
- Scoring texts according to a lexicon considering the matching between individual texts and a chosen lexicon or set of key-phrases;
- Evaluating the risk of radicalisation of a suspect based on the texts written;
- Visualisation of the results in a way that the user can understand the results provided by the algorithms and tools

Each of these aspects are detailed in the next sub-sections.

3.1. Jargon Detection

The jargon detection tool aims to automatically extract key-phrases from texts. The key-phrases not only provide a compact representation of content, but they can also serve as seeds in various other text processing tasks. Some of the applications of key-phrases can be

collecting or filtering documents, clustering documents, identifying possibly related people because of common jargon used, or scoring texts according to target vocabulary.

A jargon term refers to a word or a key phrase used in an unusual context, specifically by a suspect. For instance, a suspect uses a specific term (a jargon term) that may be shared with other suspects in the same case or more or less often in the different documents s/he wrote; the jargon term can be a clue to detect the implicit interaction between the suspects. Thus, the use cases such as the detection of radicalisation can employ this tool.

key-phrases are extracted from sequences of one or more words (or n-grams). The key phrase extraction algorithms consist mainly of several steps [22]: (a) pre-treatment of the text (e.g., parsing, POS tagging, stopword removal, stemming or lemmatizing), (b) candidate key-phrase extraction where a candidate for key-phrases are extracted from the texts, and (c) candidate key-phrase scoring and ranking for selection. The key phrase extraction algorithms differ from how they implement this essential steps [21–23,42]. The candidate key-phrases are represented with features such as their frequency, positions in the text, and word vectors using language models trained on open data sets such as Wikipedia to get the common context of terms. Then, the final scores for the candidate key-phrases are computed using the extracted features. The key phrase extractor sorts the candidate key-phrases according to the final scores. Moreover, on top of the extraction of terms, we also use the word embedding [43] to find other potentially interesting key-phrases that are semantically related and used in the different contexts of these extracted terms.

We developed the jargon detection module by adapting the existing key-phrase extractors [22,23] and by also introducing our own algorithms [21]. The developed tool detects the candidate jargon key-phrases and presents them to the user for selection and inclusion in the jargon lexicon. It also highlights the extracted key-phrases in the input documents.

3.2. Scoring Texts and Language Contents

Texts can be scored either in an unweighted or in a weighted way.

3.2.1. Scoring with an Unweighted Lexicon

In the unweighted scoring module, each key phrase has the same influence on the final score.

Firstly, the unweighted scoring module computes the exact occurrence of each key phrase (i.e., the frequency of the key phrase) of the lexicon in the text. Secondly, each term of the matched key phrase is assigned to a weight defined as the ratio of the term frequency to the length of the text (i.e., the total number of terms). Thirdly, the score of each matched key phrase is calculated by summing up the weight of its constituted terms. Finally, the text score is estimated by summing up the score of the matched key phrases of the lexicon in the text. We also handle the cases when the same term appears in multiple key phrases.

This scoring module also estimates the percentage of the lexicon found in the text (i.e., the ratio of the matched key phrases to the total number of key phrases in the lexicon).

3.2.2. Scoring with a Weighted Lexicon

Official pages are usually characterised by error-free spelling, orthography, and grammar. Accordingly, the taxonomies used should be able to identify these quite easily. It becomes more difficult when evaluating texts on social media. There, scene- or youth-typical language, slang terms or codewords are often used. Spelling errors are also common. The question here is how to deal with these challenges.

One possibility would be to work with a "phonetic" or fuzzy string search. Furthermore, algorithms can be developed that take into account an error tolerance determined by the developer when searching for keywords (so-called "triggers"). Furthermore, scene-typical terms and code words can be included in the taxonomies and assigned to a score.

The idea is to determine what is specific to the phenomenon that is monitored. Each keyword or phrase is assessed to determine whether it is a trigger or neutral feature. Triggers correspond to the abnormal content. Only trigger features are included in the scoring.

Each keyword or phrase marked as a trigger feature is assigned to a scoring value that lies between 0 and 1.

The analysis of a text can be triggered in different ways.

- Manually: in this case, some abnormal text has been identified and automated classification is targeted.
- Crawler searches the web for content to select the relevant texts that include the so-called "Main Trigger". Only main triggers start an automated analysis.

In addition to the trigger characteristics, so-called "main triggers" are used. These are required if a free Internet search is carried out, as otherwise, the texts to be examined would take on an inflationary dimension. In a free Internet search, only those texts are filtered out in which the algorithm identifies one or more "main triggers".

To avoid "flooding" the texts to be analysed, a kind of "pre-filter" would also be conceivable. If the text is attached to a link that can be assigned to state institutions, non-suspicious publishers or academic institutions, this text is automatically filtered out. This pre-filter can be considered as ANTI-TRIGGER.

In the next step, each trigger feature occurring in the analysed text is multiplied by the frequency of its occurrence. The product values are added and then the sum is divided by the number of "matches". This gives the scoring value. The closer the scoring value is to the value "1", the greater the probability that the content is related to the observed phenomenon (here radicalisation) in nature. In addition to the scoring value, a second value is calculated: the "share of observed phenomenon passages in the total text". This is the proportion of text content classified as problematic in relation to the total text.

3.3. Text Visualisation Tool

Collecting information, text analytics and Artificial Intelligence tools allow identification of trending topics in different media sources, while exploratory visual analytics tools provide means to identify the prevalence of topics in different sources, and their dynamics. Here, we present the Text Visualisation Tool (TVT) where some of the visual analytics techniques are realised and applied in the visual analysis of content.

Exploratory visual analytics techniques provide means to identify the prevalence of topics in different sources and their dynamics as well. Such techniques, among others, include co-occurrence network analysis and topic modelling. Text Visualisation Tool (TVT) combines the latter two text visualisation techniques for exploratory visual analysis. These techniques allow the LEAs to visually identify the change in the dynamics of narrative and activity in relation to different groups of interest. Word Co-occurrence Network and Topic Modelling are language agnostic. These two techniques can be used without additional linguistic tools, such as POS taggers or syntactic parsers and, therefore, can be used for textual data from different media sources. TVT input is text files, while output consists of picture files (*.png, *.svg). The results TVT provide are text visualisations based on statistical models (co-occurrence and topic models) and the decisions in terms of the meaning of these visualisations, as they are of an exploratory nature, are made by the user. The flow of TVT is presented in Figure 1.

Though network analysis is most often used for the relationships between people, it can also be applied to represent relationships between words, e.g., [44–46]. Word co-occurrence network analysis is a text length insensitive method, thus it is suitable even when the dataset is made of short texts or just texts of uneven length [47]. The co-occurrence word networks link keywords and term phrases that co-occur together. These networks reveal the semantic correlation among different terms. Firstly, a weighted adjacency matrix is generated with the rows and columns corresponding to words. The cells of the adjacency matrix are the reverse cross-product of the term-frequency inverse-document frequency (TFIDF) for overlapping terms between two documents [44]. See the example of co-occurrence network in Figure 2.

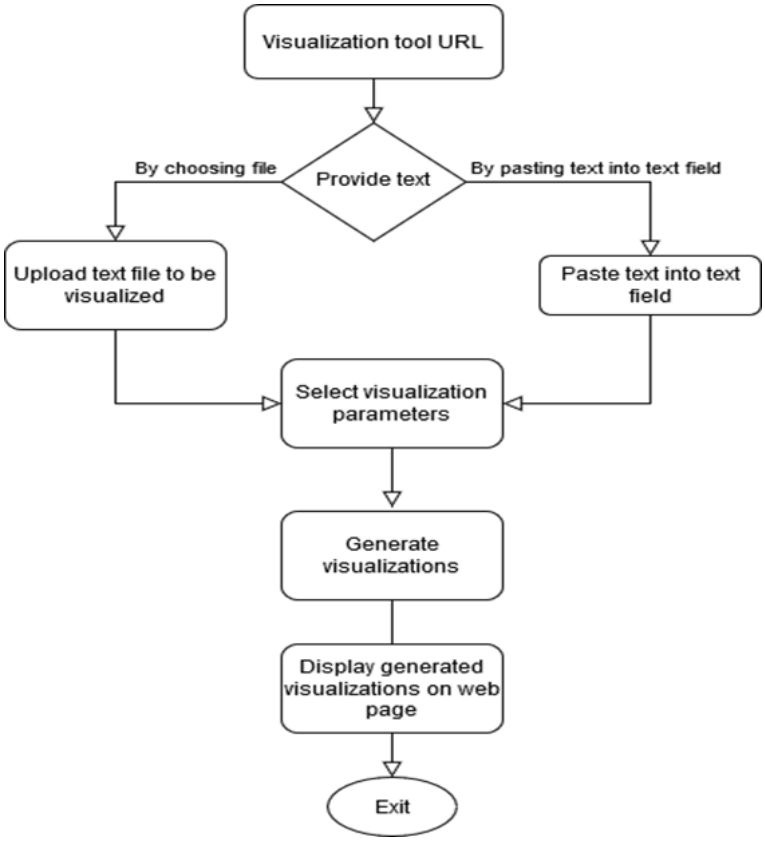

Figure 1. Text Visualisation Tool (TVT) workflow. The user opens TVT tool in a web user interface where he can provide the text to be analysed. One can also modify the text visualisation parameters (e.g., number of expected topics) or keep the default values. Finally (after the user clicks "generate") the visualisations are generated and displayed in the web user interface.

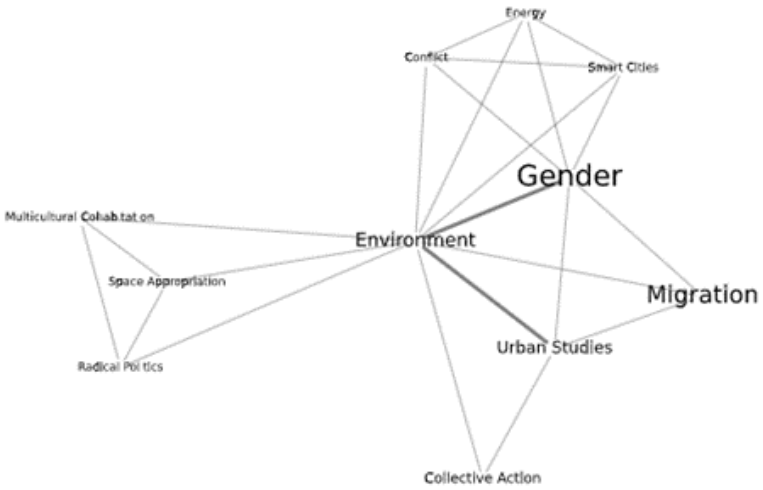

Figure 2. Example of co-occurrence network. Words that are considered as the most important are displayed in a bold and large font. Words are linked according to their co-occurrence in the texts.

Visualising text networks creates challenges because dense networks are very cluttered. Meaningful visualisation normally requires simplifications of the network. For example, networks may be drawn in such a way that the number of neighbours connecting to each term is limited. The criteria for limiting neighbours might be based on the absolute number of co-occurrences or more subtle criteria. Such filtering is realised in TVT as well.

Topic modelling is a frequently used statistical tool for detecting hidden semantic structures in a text. In topic modelling, a "topic" is viewed as a probability distribution over a fixed vocabulary [48]. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. Topic models can help to understand large collections of unstructured texts. See examples of topic models, visualised as word clouds [49] as a powerful visualisation tool.

We used the Latent Dirichlet Allocation (LDA) algorithm in TVT. LDA is a generative statistical model that views documents as bags of words (that is, the order does not matter) [50]. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions. LDA works in the following steps:
For each document *m*:

1. It is assumed that there are *k* topics across all of the documents in the corpus.
2. These *k* topics are distributed across document *m* by assigning each word a topic.
3. Word *w* in the document *m* is probabilistically assigned a topic based on two things:
    - what topics are in document *m*
    - how many times word *w* has been assigned a particular topic across all of the document.

This procedure is repeated a number of times for each document in the corpus. The number of topics depends on the size of the dataset as well as how focused it is on the subjects it presents. The number must be at least 2. As a rule of thumb, for very small datasets focused on a very specific subject (e.g., survey experiments) 3–10 topics should suffice; small datasets (size from a few hundred to a few thousand documents) 5–50 topics should be enough; for medium-sized datasets (10,000 to 100,000 documents) 60–100 topics should work well; for larger datasets 100 topics is a common default size [51].

4. Methodology

To illustrate the usefulness of the developed tools, we consider radicalisation detection as a use case. Moreover, we have integrated all these tools within a platform that we also present.

4.1. Dataset and Example Use Case

The data used in an analysis can come from various different sources. These can be texts from propaganda sites on the Internet (WWW or Darknet) but also posts in social media or emails (see also Section 5.5.2). All of these text files can contain a number of keywords or phrases that allow conclusions to be drawn about certain radicalisation processes.

In the example use case that we will use to illustrate the functionalities of the software components we have developed, we consider a use-case in which "0049855" is the ID of an anonymous suspect. For this suspect, the data collected from different sources include 12 tweets from Twitter, two emails, and two articles from propaganda sites. These text data will be analysed by different tools that we developed to estimate the level of radicalisation of the suspect.

The following text is an example of a tweet related to Islamist extremism that "0049855" wrote :

> *"From: Ibraim Jafari*
> *Time: Fri 29 Jan 09:22:23 +0000 2021*
> *@Guenter Seifert: My friend, there is no other way than violence to change the situation! We tried to chenge things with political means, we tried to discuss and to find solutions, but things only got worse. No, it can't get even worse! So let us carry the violent struggle to where it belongs to: To Minister B! Do not let our people down, my brother! dear mujahid brother, we are suporting muslim armed group ans muslim gangs. They also want fight against the tyrant and want to built a islamic caliphate. i want to come in paradise and I think you too. i am a holy warrior of the islamic state. our state is victorious and the khilafa is here."*

4.2. Methods

All the tools and methods we presented were implemented in the PREVISION platform. The tools are arranged in a comprehensive processing chain to help LEAs in their analyses. Indeed, "crime investigators work on the field—identifying, documenting, collecting, and interpreting evidence to solve complex cases. Today, more than ever before, they need to make sense of massive streams of heterogeneous data. The EU-funded PREVISION project aims to provide law enforcement agencies with advanced, almost-real-time, analytical support for multiple Big Data streams (coming from various data sources). The project allows for building dynamic and self-learning knowledge graphs that will help investigators become more aware of these fields and better address hybrid security threats, i.e., threats that combine physical and cyber attacks. The project organises five representative and complementary use cases, including the protection of public spaces and the fight of illicit trafficking of antiquities, in full compliance with privacy requirements, human rights and applicable law" (https://cordis.europa.eu/project/id/833115 accessed on 23 November 2021).

5. Results on the Radicalisation Use Case

In this section, we reconsider the tools and methods we presented in Section 4 and show how they apply to answer the use case.

5.1. Jargon Detection and Text Highlighting

The jargon detection tool serves other modules in the text analysis workflow. It helps LEAs to detect specific jargon used in a text or a set of texts. In PREVISION, this tool is integrated into two other modules: one that aims at building/updating the specific and topic-oriented lexicons and another one that scores the document considering the occurrences of key phrases from a lexicon for ranking according to their potential interest for LEAs. In the second case, jargon detected in each text is also highlighted (see Figure 3).

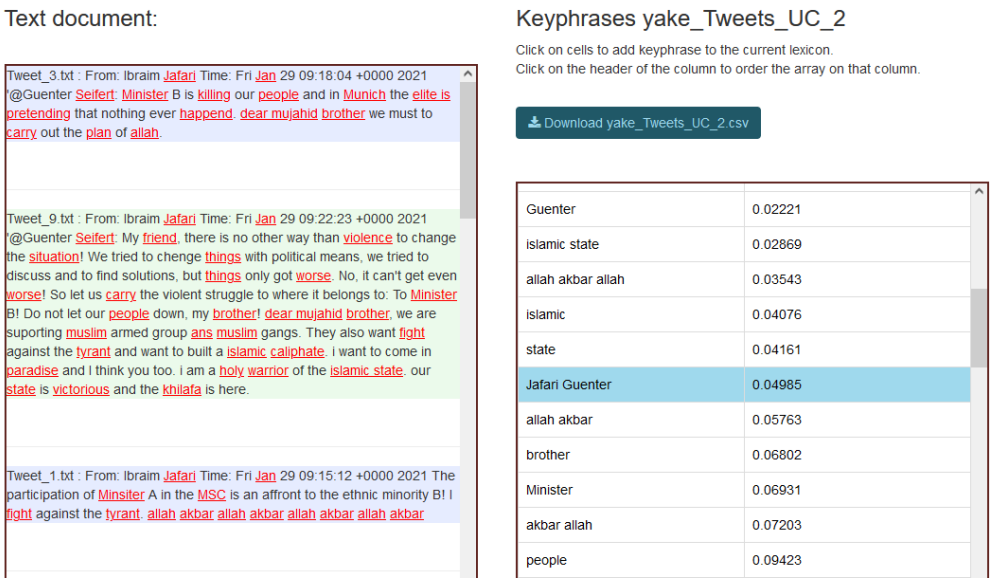

Figure 3. The terms from the selected jargon lexicon (**right-side part**) are automatically highlighted in the analysed text (**left-side part**).

Thanks to the jargon detection tool, an LEA can create a new lexicon or update an existing one using a set of text files on the domain of interest, here radicalisation.

In Figure 3, we can see the detected jargon key phrases on the right-side and the highlighted text of the input document on the left side. The LEA can choose the relevant jargon key-phrases by looking at the highlighted text and its apparent meaning.

5.2. Scoring Texts with Unweighted Lexicon on Radicalisation

The text analysis can use various sources of information:

- publicly available websites,
- social media (WhatsApp, Twitter, Facebook, etc.).

The quality of the sources is likely to vary widely. There are "Problematic" texts and images on sites whose owners do not have or support any extremist motives. In this context, reference should be made to information materials from security authorities, publications from research institutes or press material. To be able to sort out such content in advance, a list of links to such sources should be stored in the background of the software. There are websites for the groups, parties or organisations, banned in the EU or were classified as extremist sites. In these cases, a list should also be stored to identify those sites directly. The user could be advised by a "flag" that the corresponding page is already on the index.

For example, when computing the score of two matching key-phrases "brothers rise up" and "brothers come on" from the lexicon in the text, the term "brothers" appears twice. In the scoring process, we skip the weight calculation of repeated terms of the key-phrases if it has already been estimated for another key-phrase.

Table 1 illustrates the results when using the Islamist lexicon and three short texts. The scores (third column) range from 0 to 1. For making the scoring tool interpretable and transparent to LEAs, we visualise each matching key-phrases in the final score in terms of its frequency and score to the whole text (i.e., the ratio of the term frequency to the length of the text).

The module computes the score of any textual content based on the unweighted lexicon. The scoring module will estimate the level of radical content found in the text when using a lexicon on radicalisation. A lexicon consists of a list of keywords or key-phrases representing a particular domain (See Section 3.1). An unweighted lexicon refers to a list of key phrases without any weights.

Figure 4 shows an extract of an unweighted Islamist lexicon where we can observe that key-phrases are not assigned any weights.

Unweighted lexicon

keyphrase
#CountrysideCleanup
#MyJihad
#SpringJihad
airstrike
alal-Jihad
aleppo
aleppo
army
Ash-Sham
assad
attack
battlefield
break
claim

Figure 4. A part of an Islamist lexicon consisting of a list of unweighted keywords or key-phrases.

Table 1. The overall score of texts is explained to the user by showing the individual jargon keyphrases that occur in the text, an extract of which is presented here (unweighted lexicon case).

Document	Criteria	Score	Keyphrase	Frequency	Text Score
Tweet_3.txt	Text score	0.05	killing	1	0.02564
...	Share of lexicon	0.02	brother	1	0.02564
Tweet_9.txt	Text score	0.05	fight	1	0.00781
...	Share of lexicon	0.04	paradise	1	0.00781
			brother	2	0.01562
			armed	1	0.00781
			islamic state	1	0.01562
Islam_test.txt	Text score	0.03	army	18	0.00098
	Share of lexicon	0.27	attack	11	0.0006
			battlefield	1	0.00005
			claim	10	0.00054
			extreme	1	0.00005
			fight	9	0.00049
			kafir	3	0.00016
			libya	1	0.00005
			mosul	2	0.00011
			soldier	4	0.00022
			support	12	0.00065
			ummah	12	0.00065
			victory	6	0.00033
			war	20	0.00108

The heart of this scoring module is the lexicon. Indeed, this module can compute the scores for any considered lexicon. For example, in PREVISION, we have different lexicons such as Right-wing, Left-wing, Islamist, and Hate-speech lexicons.

5.3. Scoring with a Weighted Lexicon on Radicalisation

Over the past months, a variety of extremist content (Right-wing, Left-wing, and Islamist spectrum) from official websites, blogs, and chats were analysed. The extremist content was compared with "unproblematic" content. In this way, the first version of a taxonomy was developed that includes relevant keywords or phrases. Let us consider the following set of key phrases:

⇒ "i am a holy warrior of the islamic state"
⇒ "i will be a mujahid"
⇒ "mock the messenger"
⇒ "our islamic caliphate"
⇒ "allah"

Each keyword or phrase is assessed to determine whether it is a trigger or neutral feature. Triggers indicate radical content. Only trigger features are included in the scoring. For example:

⇒ "with our death"=Trigger
⇒ "allah"

Then, regarding the weighting, the higher the scoring, the higher the analyst evaluates the potential radical level of the respective text passage. For example:

⇒ "until we fall in battle (0.8)
⇒ "forth (0.3)"

With regard to the manual analysis, that means that an LEA-Analyst manually triggers the analysis. In this case, he already suspects that the present text contains radical material and wants the automated classification. The analysis can also be based on a crawler that searches the web for content to select the relevant texts that include the so-called Main Trigger. For example:

⇒ "black flag" = MAIN TRIGGER

An example of ANTI TRIGGER is a text on political education.
The following (see Table 2) is an example from a text analysis of 100 words.

Table 2. Extract of the scoring list of text analysis.

Key Words and Phrases	Category	Score	Main Trigger	Number of Matches	Total Score	Number of Words
In the name of allah	Neutral					
In the name of allah, the merciful, the gracious	Neutral					
there is only one god	Neutral					
achieve martyrdom	Trigger	0.8				
against kuffar and murtaddin	Trigger	0.8				
allah	Neutral					
allah akbar	Neutral					
allah willing	Neutral					
apostates of Islam	Trigger	0.7				
banner of the khilafah	Trigger	0.7				
become a martyr	Trigger	0.8				
black flag	Trigger	0.8	Main Trigger	1	0.8	2
brothers it's time to rise	Trigger	0.7				
brothers rise up	Trigger	0.7		2	1.4	6
caliphate	Trigger	0.7				
call of allah	Neutral					
call of allah and his messenger	Neutral					
claim your victory	Trigger	0.7		2	1.4	6
contradict the sharia	Trigger					
dear mujahid brother	Trigger	0.6				
death for paradise	Trigger	0.8				
demolish	Trigger	0.6				
fight against the tyrant	Trigger	0.7				
forth	Trigger	0.3		2	0.6	2

The weighted score is illustrated in Figure 5.

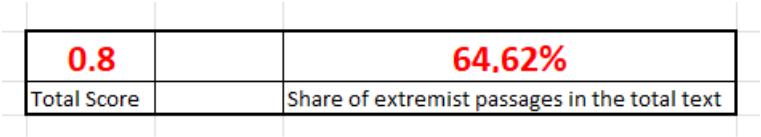

Figure 5. Example of a result provided by the tool when scoring texts.

The text can be classified as highly extremist. The lexicon can be constantly expanded, supplemented or shortened. It is also possible to re-evaluate items at any time to determine whether they should function as a trigger feature or as a neutral feature.

A lexicon is selected and then one or more text files (see Table 3) are selected. These can be stored in the database or a file system.

The tweet_9.txt in the Table 3 is the sample text from the point 2 Datasets. It contains the corresponding weighted keywords from the created Lexicon. This tweet has a score of 0.7 with a share of extremist content of 28.91% (See Figure 6). It is a tweet with strongly radical content.

Figure 6. This figure shows the API of the application that calculates the radicalisation score of text files for a weighted lexicon. First, the user selects the lexicon to be used and the file or directory to analyse (top part of the figure). As a result, the overall score of the texts is displayed.

Table 3. Text score is explained by the individual terms that occur in the text—weighted lexicon case. The same texts as in Table 1 are presented.

Document	Criteria	Score	Keyphrase	Category	Weight	Frequency	Total Words	Text Score	%-of-Share
Tweet_3.txt	Text score	0.6	dear mujahid brother	Trigger	0.6	1	3	0.6	7.69
	Share of extremist content (%)	7.69							
Tweet_9.txt	Text score	0.7	dear mujahid brother	Trigger	0.6	1	3	0.6	2.34
	Share of extremist content (%)	28.91	muslim armed group	Trigger	0.6	1	3	0.6	2.34
			muslim gangs	Trigger	0.6	1	2	0.6	1.56
			fight against the tyrant	Trigger	0.7	1	4	0.7	3.12
			i want to come in paradise	Trigger	0.7	1	6	0.7	4.69
			islamic caliphate	Trigger	0.7	1	2	0.7	1.56
			i am a holy warrior of the islamic state	Trigger	0.8	1	9	0.8	7.03
			our state is victorious	Trigger	0.8	1	4	0.8	3.12
			the khilafa is here	Trigger	0.8	1	4	0.8	3.12
islam_test.txt	Text score	0.64	forth	Trigger	0.3	7	7	2.1	0.04
	Share of extremist content (%)	1.29	jihad	Trigger	0.6	29	29	17.4	0.16
			jihad for allah	Trigger	0.6	1	3	0.6	0.02
			mujahidin	Trigger	0.6	38	38	22.8	0.21
			murtaddin	Trigger	0.6	16	16	9.6	0.09
			islamic state	Trigger	0.7	52	104	36.4	0.56
			kafir	Trigger	0.7	3	3	2.10	0.02
			kuffar	Trigger	0.7	8	8	5.6	0.04
			kufr	Trigger	0.7	14	14	9.80	0.08
			taghut	Trigger	0.7	16	16	11.2	0.09
Post_Example2.txt	Text score	0.61	strong explosive	Trigger	0.2	1	2	0.2	0.3
	Share of extremist content (%)	7.14	chlorix	Trigger	0.3	1	1	0.3	0.15
			chlorine	Trigger	0.6	7	7	4.2	1.04
			nitroglycerin	Trigger	0.6	7	7	4.2	1.04
			nitroglycerin	Trigger	0.6	7	7	4.2	1.04

5.4. Text Visualisation Tool—Application of Visual Analytics for Analysis of Radicalised Content

Figure 7 shows the topic model extracted from the dataset, which was introduced in Section 4. It combines artificially made tweets that represent a use case in radicalisation analysis.

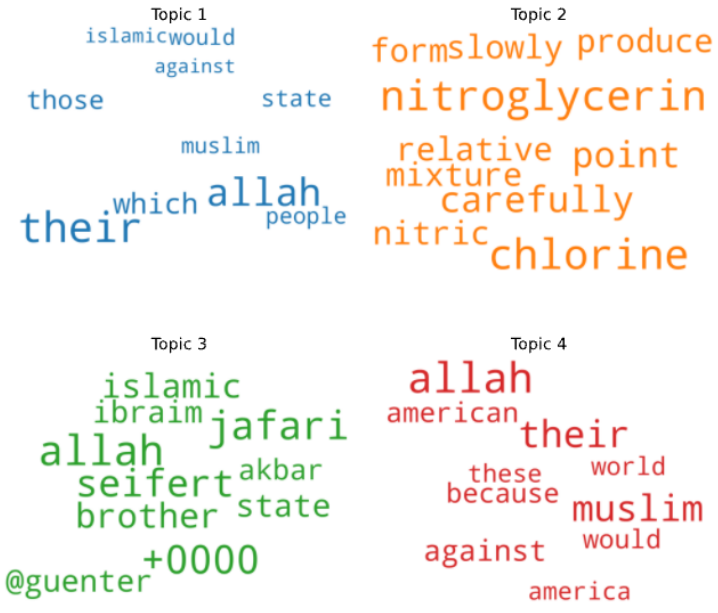

Figure 7. Topic models that were estimated from the dataset, which was introduced in Section 4.1. Each topic is displayed in a specific colour and the size of the word expresses it importance.

We also demonstrate here the application of visual analytics techniques, realised in TVT, for radical content exploration.

The representation of a word co-occurrence network is presented in Figure 8. Semantically related word clusters are introduced there, which could be used for complementing specialised lexicons.

In this way, analysing word cluster by cluster, useful words and word sequences can be added to the specialised lexicon for further analysis.

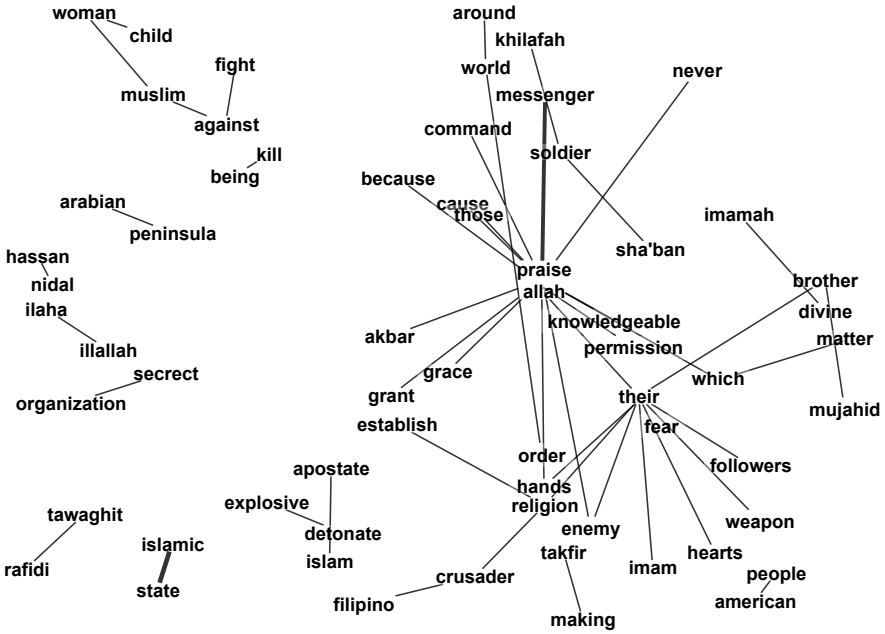

Figure 8. Word co-occurrence network discovered in the dataset which was defined in Section 4.1.

Topic modelling results are presented in Table 4, Figures 7 and 9. Figure 7 reports 4 topics extracted from data sample, Table 4 reports dominant topic for each document on the data sample while Figure 9 shows how topics are distributed in the data sample. As Figure 7 reveals, Topic 1 focuses on *Allah*, *muslim* and *people*, Topic 2—on *nitroglycerin*, *chlorine*, *produce*, Topic 3—on *Jafari*, *Allah* and *Seifert*, Topic 4—on *Muslim*, *Allah* and *America*.

Table 4. Dominant topic per document in the text sample introduced in Section 4.1.

Doc. Number	Dominant Topic	Topic Perc Contrib	Keywords	Text
0	0	99.99%	their, allah, which, state, those, people, isl…	[allah, expel, disbeliever, among, people, scr…
1	0	97.75%	their, allah, which, state, those, people, isl…	[ibraim, jafari, 09:20:46, + 0000, @guenter, se…
2	0	96.34%	their, allah, which, state, those, people, isl…	[guenter, seifert, 09:22:15, +0000, @ibraim, j …
3	0	93.15%	their, allah, which, state, those, people, isl…	[ibraim, jafari, 09:18:35, +0000, @guenter, se…
4	0	96.36%	their, allah, which, state, those, people, isl…	[ibraim, jafari , 09:15:12, +0000, participatio…
5	0	96.12%	their, allah, which, state, those, people, isl…	[guenter, seifert, 09:19:45, +0000, @ibraim, j …
6	1	99.72%	allah, their, which, muslim, those, would, peo…	[production, chlorine, chlorine, already, worl…
7	0	98.55%	their, allah, which, state, those, people, isl…	[ibraim, jafari, 09:22:23, +0000, @guenter, se…
8	0	96.73%	their, allah, which, state, those, people, isl…	[ibraim, jafari, 09:22:12, + 0000, @guenter, se…
9	0	92.36%	their, allah, which, state, those, people, isl…	[guenter, seifert, 09:16:12, +0000, @ibraim, j …
10	0	95.67%	their, allah, which, state, those, people, isl…	[09:22:41, +0000, @ibraim, jafari, @guenter, s…
11	3	99.99%	allah, their, muslim, against, american, becau…	[jihad, fight, against, tyrant, supporting, as…
12	0	98.82%	their, allah, which, state, those, people, isl…	[ibraim, jafari , 09:22:52, +0000, @guenter, se…
13	1	99.98%	allah, their, which, muslim, those, would, peo…	[guenter, against, kuffar, murtaddin, fighting…
14	0	96.00%	their, allah, which, state, those, people, isl…	[ibraim, jafari, 09:18:04, +0000, @guenter, se…
15	0	97.37%	their, allah, which, state, those, people, isl…	[guenter, seifert, 09:22:30, +0000, @ibraim, j …

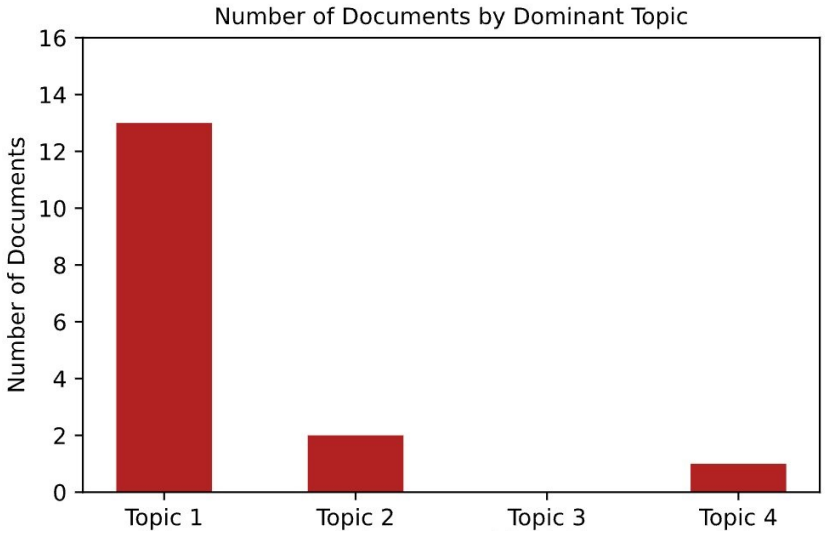

Figure 9. Distribution of discovered topics in the documents from the dataset (introduced in Section 4.1).

Figure 9 reports that Topic 1 is the most prevalent topic in the data sample as it is the dominant topic in 13 documents (out of 16). Topic 2 is dominant in 2 documents, while Topics 3 and 4 are the least prevalent topics in the selected data sample as topic 4 is dominant in only 1 document and topic 3 was not found in the selected document subset.

To elaborate, Topic 1 (Positive indexing starts from 0 in Python, the Topic 1 in Figure 9 is originally Topic 0 in Table 4. Therefore, Topic 0 ≡ Topic 1, Topic 1 ≡ Topic 2, etc.; also document 0 ≡ document 1, document 1 ≡ document 2, etc. To avoid the confusion, we will refer to indexing convention starting from 1.) is the dominant topic for documents no. 1–6, 8–11, 13, 15, 16. Meanwhile, Topic 2 is the dominant topic for document no. 7 and 14. Topic 4—for document 12 (see Table 4).

Topic modelling allows the discovery of common topics in a collection of documents, i.e., it gives a "bird's view" of the data and that is useful, especially in terms of large datasets. It shows a quick "summary" of the contents of the data, helps to identify documents associated with a specific "topic" for further analysis. Moreover, topic words can be added to specialised lexicons to use in the scoring model.

5.5. Machine Learning Model to Predict the Risk of Radicalisation

The objective of LEAs is to assess the risk of radicalisation of a suspect (or suspects) given the basic risk factors/protectors and online information, which are adequate, relevant, and limited.

To help LEAs assess the risk of radicalisation, CNRS-IRIT has proposed a tool with a two-step cascaded master scoring model and a knowledge base, where the master scoring model is a combination of a linear scoring model from basic risk factors/protectors and a set of linear/non-linear scoring models from online information (See Figure 10).

The knowledge base consists of the persons (suspects) list, basic data, online information collected using PREVISION crawler tool, jargon created for different categories by LEAs, assessment score of radicalisation performed earlier by LEAs. Once implemented in the final platform, the CNRS-IRIT tool uses an API to communicate to the knowledge base for retrieving the data and storing the assessment scores once verified by the LEAs.

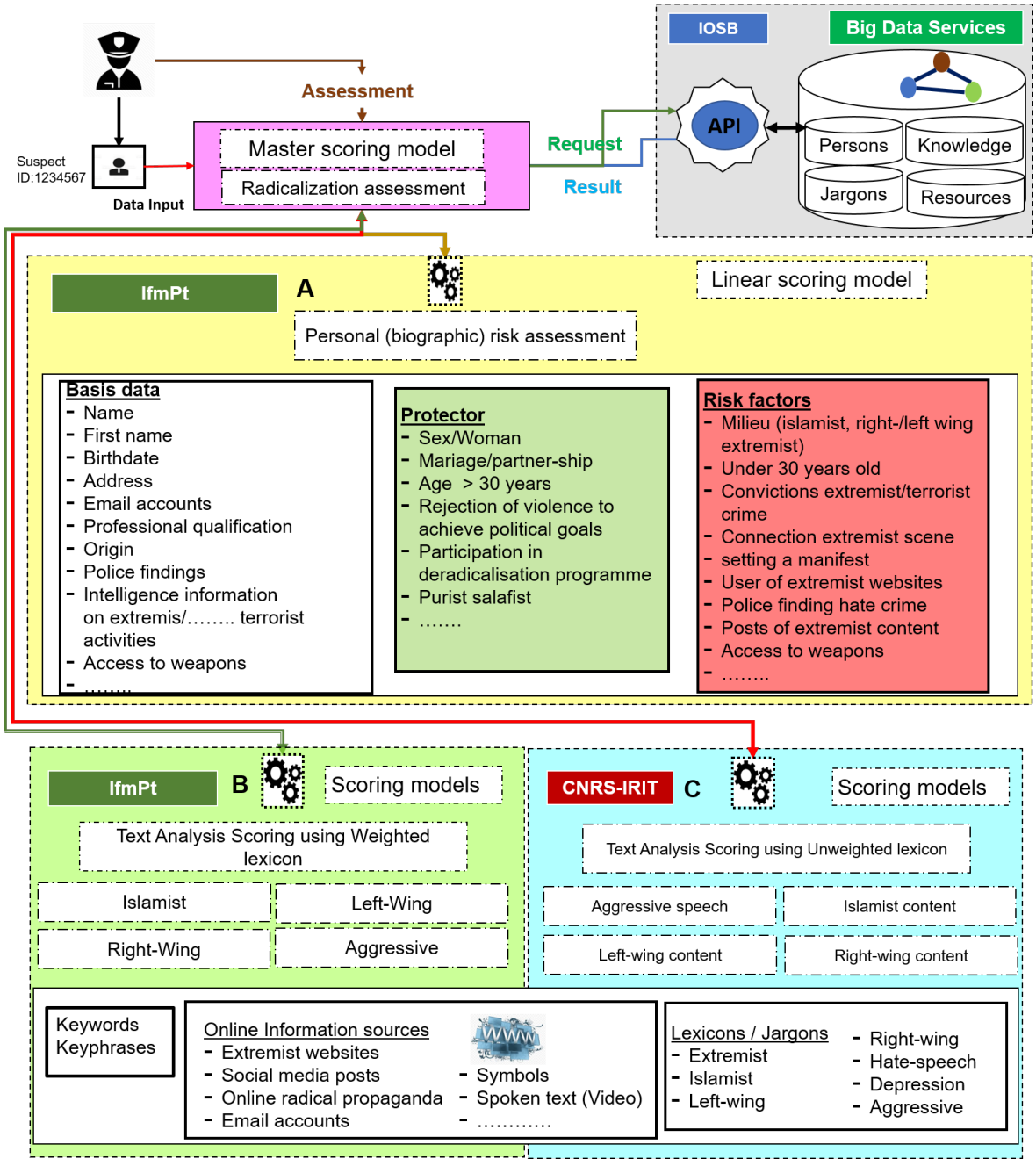

Figure 10. An architecture for helping LEAs is to assess radicalisation risk. In this schematic diagram, we can see the LEA interface, the Master scoring model, (**A**) Linear scoring model, (**B**) Text analysis with weighted lexicon model, (**C**) Text analysis with unweighted lexicon model, and the Knowledge-Base. Basically, (**A**) Linear scoring model computes the radicalisation score based on Person's biographical risk/protector factors, (**B**) Text analysis with weighted lexicon model computes the radicalisation score from online information (e.g., social media posts) using a weighted lexicon (See Section 5.3), and (**C**) Text analysis with unweighted lexicon model computes the radicalisation score from online information using an unweighted lexicon (See Section 5.2). Using a user interface, an LEA chooses one or more people (or persons) and instructs the Master scoring model to compute the radicalisation score. The master model works as a coordinator among all the models. Firstly, the master model receives the command from LEA, retrieves the background and online data from the Knowledge-base, and interacts with the (**A**–**C**) modules to compute the respective radicalisation scores. Secondly, three models update the computed scores back to the Knowledge-base. Finally, the master model visualises the radicalisation score over time to help LEAs in assessment (see sub-section 5.5.4).

5.5.1. Biographical Scoring from Basic Data

One way of looking at the future is through biographical research, i.e., by examining the biographies of people who have appeared before the police in the past because of (politically motivated) crimes. In this way, "typical" biographical patterns can be identified, on the basis of which statements can be made about future offender behaviour. Biographical research is now a common empirical tool in criminology [52,53], which is already being used in both extremism and mafia research.

RIVA (Risk Assessment of potential Islamist Violent Actors) lists individuals who have already been classified by the security authorities as extremists from the Islamist spectrum. The aim of the risk analysis is to assess the likelihood that the persons in question will commit a politically motivated act of violence in the near future. The aim is not the early detection of Islamist-motivated radicalisation.

As already mentioned, the forecasting tool described below focuses on Islamist extremism/terrorism. Its use for people from the right-wing or left-wing extremist spectrum requires adjustments to its content.

When collecting data, a distinction is made between basic data, protectors and risk factors. The basic data includes socio-demographic variables such as age, gender or marital status and information on political and criminal activities. The basic data is used to reconstruct the biographical history of a person as comprehensively as possible.

Not all basic data is used for risk analysis. Those variables that are selected for this purpose are divided into protectors and risk factors. Risk factors increase the danger of a politically motivated act of violence, protectors reduce respectively minimise the danger.

RIVA comprises seven categories with a total of 37 variables. Each variable was assessed to determine whether it is a protector or a risk factor. The classification as a risk factor/protector is based on the findings of national and international studies, the evaluation of biographies of violent Islamist offenders carried out as part of the project and discussions with experts in the field of "state protection". For example, previous research findings [54–57] indicate that violent acts motivated by Islamism are committed almost exclusively by men. Furthermore, most perpetrators are under thirty years of age [55] and are usually not married. Other factors include the attitude towards violence and the social context in which a person lives or communicates his or her attitude. Purists, for example, have an extremist attitude, but they do not proselytise or call for violence. They act primarily in their private environment. From a security perspective, therefore, they are rather unproblematic.

On the other hand, gender and age, as well as attitude, can be risk factors. If the person in question is a 25-year-old male Islamist, the risk of committing a politically motivated act of violence is increased. Particularly with regard to their attitude towards the (personal) perpetration of violence, clear differences can be seen within the scene. Political missionary actors carry their extremist attitudes into the public sphere and can thus contribute to the indoctrination and radicalisation of third parties. However, political-missionary actors, like purist actors, reject violence.

However, these are always people with extremist attitudes. However, their individual propensity to violence varies greatly. The characteristic "extremist attitude" is therefore not very meaningful in itself for assessing risk potential. At this point, the difference described above are necessary, or what has been said implies that the individual characteristics must be weighted. Thus, each protection or risk factor is multiplied by a factor x. The smaller the influence of the variable on the individual risk, the smaller the factor (1–10) with which the variable is multiplied.

In addition to attitude, the willingness and ability to use violence also plays a significant role. People who have already attracted attention through acts of violence in the past show that they have already exceeded a certain inhibition threshold. Spending time in combat zones and, in particular, participating in combat operations may also have contributed to a certain degree of blunting. Following the theory of differential opportunity, access to illegitimate means should be noted. This includes not only access to weapons and

explosives but also access to knowledge. Individuals who have undergone paramilitary training, i.e., who have been trained in the use of weapons and explosives, are significantly more likely to be able to carry out (more complex) attacks and thus have a significantly higher risk potential. Furthermore, concrete knowledge about, for example, a person's whereabouts or current activities plays a significant role in assessing risk. Actors about whom information is available that they have recently attempted to buy weapons and ammunition on the black market represent a very high risk (see Vienna assassin). The same applies to indications that a person is searching the net for instructions and materials to build bombs. For this reason, evidence that points to a concrete attack plan is not only highly valued, but these characteristics are so-called trigger characteristics. Triggers refer to the risk of concrete attack planning. The social network of the analysed person is also included in the risk analysis. In particular, contacts with jihadists or active fighters have a negative impact on the risk score. The same applies to contacts with high-ranking leaders or members of terrorist cells and organisations.

In the run-up to attacks, relevant behavioural anomalies can often be observed. These include social withdrawal. If a person's whereabouts are no longer known or the person ceases all (online) communication, extreme caution is called for. The person not only withdraws from the control of his or her social environment, but experience has also shown that going underground in violent cells rapidly accelerates the process of radicalisation.

Scaling:

As mentioned above, risk factors and protectors are determined from the base data. Risk factors are assigned the value -1, protectors the value +1. Each risk factor or protector is then multiplied by the factor x. The factor is based on whether the respective characteristic has been assigned a low, medium or high rating. This is to ensure that characteristics considered more relevant have a greater influence on the result. We decided to use a ten-point scale to determine a scoring value for the individual. In this way, a meaningful assignment of the individual characteristics can be made. Low e.g., married or not, Legal access to firearms(factor 1–3) Middle e.g., gender, military training (4–7) High e.g., Confession Video, Manifest, Former foreign fighter (8–10).

Basic Score.

The filled columns of each class (high, medium, low) are summed and divided by the number of protectors/risk factors to generate the basic score.

Each characteristic is assigned a corresponding factor. The value of a risk factor is multiplied by -1. The value of a protective factor is multiplied by +1. The resulting numbers are then added for the categories high, medium and low. The results from this calculation are then divided by the number of factors in the respective class. This makes the basic score.

Conclusive Score. In the first step, the score value is calculated separately for the "low", "middle" and "high" groups by summing up all values within a class and dividing this figure by the number of risk factors for that particular class. The calculated average values per class are then weighted with a factor (low: $\times 0.1$, middle: $\times 0.3$, high: $\times 0.6$). The results per class are then added together to give the Conclusive Score. This rating results from scientific and practical research and will be further tested during operational use. However, should within a scoring evaluation the suspect has at least one risk factor ranking at 9 or 10 (the two highest possible classed factors) such individual will be put on observation regardless of other factors decreasing the overall conclusive score. In such case, a special notice will be provided by the system to the operator.

The Conclusive Score avoids that too many risk factors in the lower and middle ranges are artificially reducing the overall risk or risk factors from the higher range from being given too little consideration.

If new information about a person is available, it is implemented into the system. The information can come from security authorities as well as from propaganda videos, chats, blogs, etc. The scoring system developed by IfmPt continuously and automatically performs a risk assessment. The operator (a police officer) decides on necessary operational

measures. The Risk Assessment is intended exclusively as a supporting tool; in no way does it replace the professional assessment by experienced police officers.

Currently, test runs are carried out with the created biographies in order to check the initial configuration for validity. RIVA is a dynamic system. New protectors and risk factors can be added, and existing ones can be excluded from the assessment. The weighting of individual characteristics or the three groups can also be adjusted. Simulation processes will be carried out continuously In order to assess accuracy, the system allows simulations based on modified configurations as outlined. It will also automatically adjust any risk assessment should any basic information proofs as being incorrect or invalid in the meantime. The degree of accuracy refers mainly to the weighing of the individual factors which is constantly reviewed (and also exchanged among LEAs as well as among LEAs and the system developer). The main importance though is the accuracy of the factors being entered into the system for which LEAs are responsible.

The system will also provide all details of the scoring evaluation from its first assessment i.e., any operator can trace the factors, which were individually entered into the system, how it was ranked and how the basic and conclusive score was detected. Such information is available at any point in time from the system and it also allows the operator to make a personal (human) review in detail if so required. The same applies to any adjustment to the risk assessment at a later point in time. Therefore for any suspected person, a fully detailed legend is available as to when and how the risk assessment was made and how it developed over time.

5.5.2. Linear/Non-Linear Scoring Models from Online Information

LEAs can categorise online Information into subsets such as extremist, Islamist, Left-wing, Right-wing, Hate-speech, etc. From these subsets of information, LEAs can use the PREVISION jargon detection tool to create the respective jargon/lexicons. Using specific jargon/lexicons (e.g., Islamist), online information (e.g., social media posts, emails, webpages, etc.), and occurrence of the keywords/key-phrases in online information, CNRS-IRIT has developed a scoring model stated in Section 4.1. The estimated score is explained in terms of the Jargon lexicon used in the analysed textual information (a document/post, a set of documents). Following the above approach, CNRS-IRIT developed and implemented different scoring models. The developed models could be used to score extremist content, recognise aggressive content, recognise left-wind content, recognise right-wing content, recognise the depression, etc.

5.5.3. Knowledge Base

The PREVISION knowledge base contains data that can be expressed in terms of the PREVISION ontology. It is a central sharing point of the platform for meta-information about and connections between individuals, artefacts, resources, and results. At present, it provides an HMI to manually insert pseudonymised data about persons, emails, social network accounts and other categories. In addition, read access through a REST API is available. Accordingly, write access is under development. Additionally, storage of the source of any piece of data will be assured, like the time of insertion, the name and version of the algorithm that produced it as well as technical details like lexicons or sets of training data that have been used. In this way, the transparency and explainability of any result will be guaranteed as far as possible.

In further development, the above-mentioned tools will be designed to make use of relevant information contained in the knowledge base and also store their result in it.

The user can assign weights to a selection of person properties (see Figure 11). There is also a default setup. Positive weights mitigate the radicalisation level of a person. Negative weights are assigned to properties that indicate radicalisation.

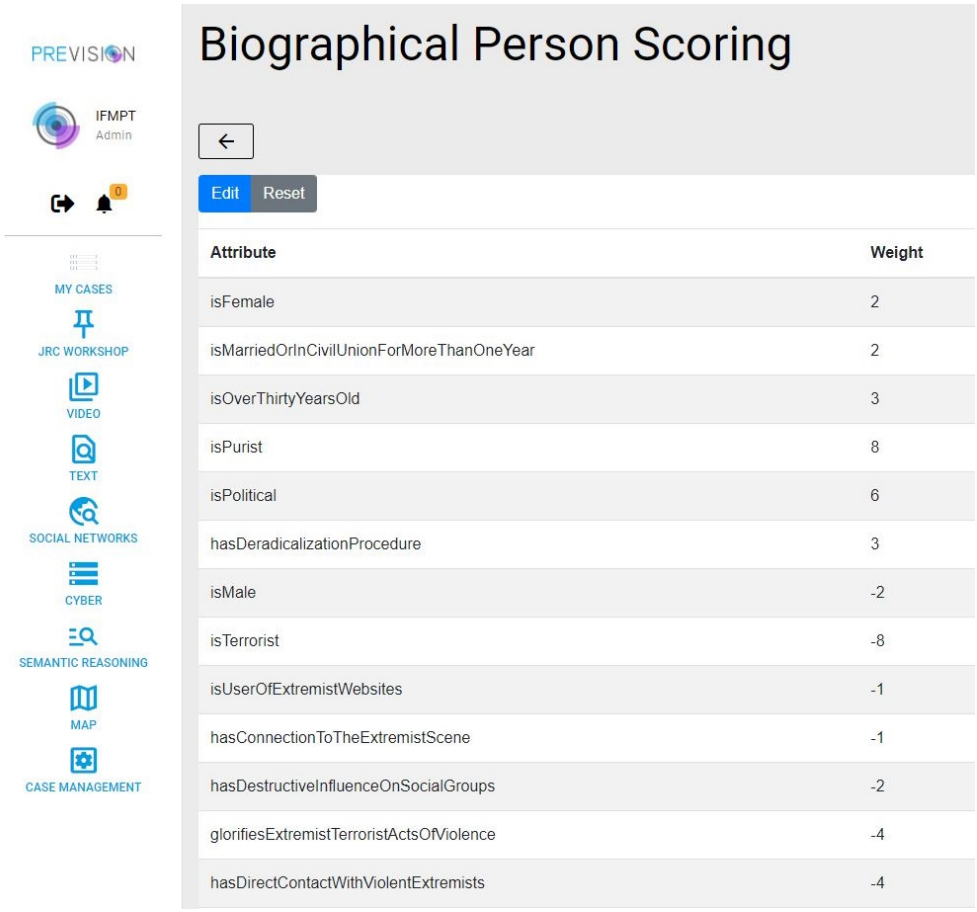

Figure 11. Example of the configuration of the Biographical Person Scoring in PREVISION. Each feature is weighted by LEAs which affects the calculation of the overall radicalisation probability of a suspect.

5.5.4. Master Scoring Model (LEAs)

Given the ID of a suspect provided by an LEA, this master model helps the LEA to retrieve her/his basic and online information from the knowledge base. Then, these data are forwarded the necessary to scoring models A, B, and C. Once the scores from the scoring model to the master model are returned, the master model combines the estimated scores of radicalisation risk. Once approved by the LEAs, the master model stores in the knowledge base the final score as well as the individual estimated scores from the models in A, B, and C with a timestamp. To approve the final score, LEAs can focus backwards and investigate deeper the intermediate predictions by the linear scoring model in A using basic risk factors and a set of linear/non-linear scoring models in B and C using online information and jargon/lexicons using the explanations (see Table 1: Explanation).

The proposed tool can also be used by LEAs to investigate the prognosis of radicalisation risk by visualising the assessment scores within a time frame (see Figure 12).

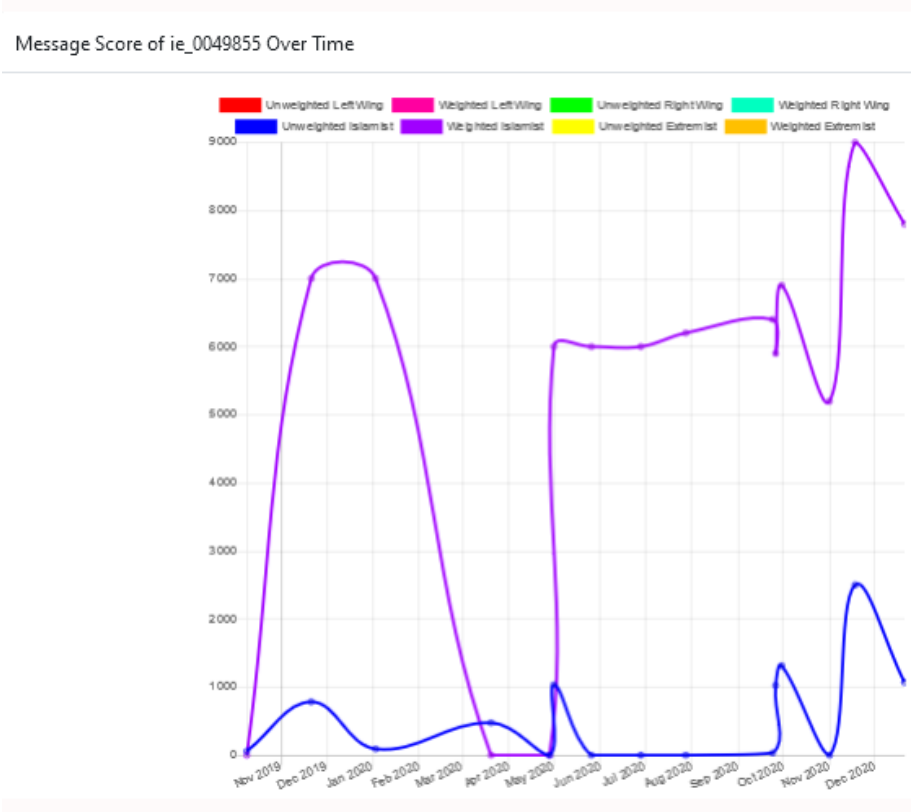

Figure 12. Visualisation of radicalisation scores over time. This plot shows the weighted and unweighted radicalisation scores change over time for the person ie_0049855, who was introduced in Section 4.1. The scoring schemes were introduced in Section 3.2.

6. Discussion, Implication, and Conclusions

PREVISION has created a new advanced innovative approach to the detection of radicalisation processes. It consists of several methods and tools that can be combined within a comprehensive process chain. First, a jargon detection tool is used to extract key phrases from texts. Once reviewed by an expert, these key phrases are stored in a lexicon that can be used in several ways. Jargon lexicons are used to score texts either in a weighted or unweighted way; depending on if some key phrases should be emphasised or not. Visualisation tools complement the analysis. They aim to help the user to understand the results and scores. Other visualisation tools help to consider the links between terms. We have illustrated the process considering an example use case on radicalisation detection. It is a "generic process", which will be further developed within the framework of the operational work. The first results have already been evaluated by the LEAs.

The tools that automatically analyse texts may "reproduce or amplify unwanted societal biases reflected in training data" [58]. Hovy and Prabhumoy outlined five sources where bias can occur in NLP systems: (1) the data, (2) the annotation process, (3) the input representations, (4) the models, and (5) the research design (or how the research is conceptualised) [59]. For example, it has been shown that part-of-speech tagging models have a lower accuracy for young people and ethnic minorities, vis-á-vis the dominant demographics in the training data, while the word embedding runs the risk of amplifying gender bias present in data [60].

To address the problem of the training data, the research community proposed the implementation of data sheets [58] or data statement [61] as a design solution and professional practice for NLP. As far as correcting bias is concerned there have also been attempts to reduce the bias by applying algorithms designed for this specific task [60]. Additional tools include using a Model Card for Model Reporting [62].

As in the case with other models, it is crucial to highlight that "[a] model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted biases" [58]. Mismatches would have especially severe consequences when machine learning is used in high stake domains such as law enforcement. Therefore, the tools must be tested and evaluated in the actual deployment context before it is used in operational conditions.

When developing the tools that have been presented in this paper, we considered the following measures:

- We used two types of data sources: the sources used to train word embeddings which are unknown (since we used pre-trained systems) and the sources used to extract the jargon, which are provided by the user or collected by the user and thus known. To address the lack of transparency on the first type of data sources, the tool shows both the Jargon that is extracted, the position of the jargon in the users' texts and provides the explanation of the score calculation for the scoring tools to the user. The data description is thus available.
- To address the risk of bias (describing the data types, features types, dictionaries, external resources used during training the tool): text data type was used across all tools, no pre-defined dictionary was used in the tools developed although the users can start from their own dictionaries. To model the pattern matching in the scoring tools, the Spacy language model was used, which code is available.
- To address the lack of transparency as to how the tool works (explaining the data representation, mathematical logic and rules underlying the models, exploring the features and the data that has an impact on the classification score): the tools tips have been added in the interface so the user knows how the tool works. The detailed explanation and mathematical calculations and rules underlying the models are also available.
- To further address the lack of transparency -explaining the confidence score of the model and the features that contribute to this confidence score. The explanation of the calculation of scoring tools is presented as a pop-up to the user.

During the course of the project, ethical and data protection issues and concerns, in particular, were discussed and examined extensively. Extensive research and explanations were necessary, especially in the area of Personal Risk Assessment (RIVA). Once again, it is important to point out that the tools and methods described are only used on police suspects.

With regard to key-phrase extraction and lexicons, compared to Curatr [36] which base the lexicons on word semantic similarities, in PREVISION, we opted for Yake [22] algorithm which is state of the art.

With regard to text analysis, both Tétralogie [35] and PREVISION can handle many different types of texts from tweets, emails, forums or journal articles. Both tools also integrate many different tools and can be used in many various applications and domains including using web documents [63]. PREVISION however focuses on security applications, which makes it quite unique. PREVISION also integrates some tools for image and video analysis for example that we did not detail in this paper.

Future work will focus on enhancing visual tools. For example, we would like to provide ontological-like representation so that terms would be replaced by concepts or visualisations could be made either at the key-phrase level or at the concept level. Interactive visualisations could also be helpful to LEAs where they would remove non-important words or maybe add new terms to consider in the analysis.

Author Contributions: All authors have read and agreed to the published version of the manuscript. All the authors have contributed to the paper. Among the main contributions are conceptualization, J.M. (Josiane Mothe) and M.Z.U.; methodology, G.O. and T.S.; software, M.Z.U.; validation, G.O., T.S., and M.Z.U.; formal analysis, G.O. and T.S.; investigation, G.O.; resources, G.O.; data curation, G.O.; writing—original draft preparation, J.M. (Josiane Mothe), M.Z.U., A.J., and J.M. (Justina Mandravickaitė); writing—review and editing, J.M. (Josiane Mothe), M.Z.U., and A.J.; visualization,

G.O., M.Z.U., and A.J.; supervision, J.M. (Josiane Mothe); project administration, J.M. (Josiane Mothe); funding acquisition, J.M. (Josiane Mothe).

Funding: This research was funded by European Union's Horizon 2020 research and innovation programme, H2020-EU.3.7. Secure societies Protecting freedom and security of Europe and its citizens, under GA No 833115. The paper reflects the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of PREVISION.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data presented in this paper are available from the authors. Please email to the authors for further information.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kalisch, M.; Stotz, P. Wer Liest Das Eigentlich? Die GELäufigsten Corona-Verschwörungstheorien und Die Akteure Dahinter. Available online: https://www.spiegel.de/netzwelt/web/corona-verschwoerungstheorien-und-die-akteure-dahinter-bill-gates-impfzwang-und-co-a-2e9a0e78-4375-4dbd-815f-54571750d32d (accessed on 7 November 2021).
2. Reinecke, S. Konjunktur der Verschwörungstheorien: Die Nervöse Republik. Available online: https://taz.de/Konjunktur-der-Verschwoerungstheorien/!5681544/ (accessed on 7 November 2021).
3. Berlin, B. Antisemitische Verschwörungstheorien Haben Während Corona Konjunktur. Available online: https://www.bz-berlin.de/berlin/antisemitische-verschwoerungstheorien-haben-waehrend-corona-konjunktur (accessed on 7 November 2021).
4. Fielitz, M.; Ebner, J.; Guhl, J.; Quent, M. *Hassliebe: Muslimfeindlichkeit, Islamismus und Die Spirale Gesellschaftlicher Polarisierung*; Amadeu Antonio Stiftung: Berlin, Germany, 2018; Volume 1.
5. Chen, H. *Dark Web: Exploring and Data Mining the Dark Side of the Web*; Springer Science & Business Media: Berlin, Germany, 2011; Volume 30.
6. Akinboro, S.; Adebusoye, O.; Onamade, A. A Review on the Detection of Offensive Content in Social Media Platforms. *FUOYE J. Eng. Technol.* **2021**, *6*. http://dx.doi.org/10.46792/fuoyejet.v6i1.591.
7. Neumann, P.R. *Der Terror ist unter uns: Dschihadismus, Radikalisierung und Terrorismus in Europa*; Ullstein eBooks: Berlin, Germany, 2016.
8. Scruton, R. *The Palgrave Macmillan Dictionary of Political Thought*; Springer: Berlin/Heidelberg, Germany, 2007.
9. Ali, F.; Ali, A.; Imran, M.; Naqvi, R.A.; Siddiqi, M.H.; Kwak, K.S. Traffic accident detection and condition analysis based on social networking data. *Accid. Anal. Prev.* **2021**, *151*, 105973.
10. Types de Radicalisation. Available online: https://info-radical.org/fr/types-de-radicalisation/ (accessed on 7 November 2021).
11. MDR.DE. Wie das Internet zur Radikalisierung Beiträgt | MDR.DE. Available online: https://www.mdr.de/wissen/bildung/extremismus-internet-online-radikalisierung-100.html (accessed on 7 November 2021).
12. Holbrook, D. A critical analysis of the role of the internet in the preparation and planning of acts of terrorism. *Dyn. Asymmetric Confl.* **2015**, *8*, 121–133.
13. Kahl, M. Was wir über Radikalisierung im Internet wissen. Forschungsansätze und Kontroversen. *Demokr. Gegen Menschenfeindlichkeit* **2018**, *3*, 11–25.
14. Reicher, S.D.; Spears, R.; Postmes, T. A social identity model of deindividuation phenomena. *Eur. Rev. Soc. Psychol.* **1995**, *6*, 161–198.
15. Spears, R.; Lea, M. Panacea or panopticon? The hidden power in computer-mediated communication. *Commun. Res.* **1994**, *21*, 427–459.
16. Tajfel, H.; Turner, J.C. The Social Identity Theory of Intergroup Behavior. In *Political Psychology: Key Readings*; Psychology Press/Taylor & Francis: London, UK, 2004; pp. 276–293.
17. Boehnke, K.; Odağ, Ö.; Leiser, A. Neue Medien und politischer Extremismus im Jugendalter: Die Bedeutung von Internet und Social Media für jugendliche Hinwendungs-und Radikalisierungsprozesse. In *Stand der Forschung und Zentrale Erkenntnisse Themenrelevanter Forschungsdisziplinen aus Ausgewählten Ländern. Expertise im Auftrag des Deutschen Jugendinstituts (DJI)*; DJI München Deutsches Jugendinstitut e.V.: Munich, Germany, 2015.
18. Kimmerle, J. SIDE-Modell im Dorsch Lexikon der Psychologie. Available online: https://dorsch.hogrefe.com/stichwort/side-modell (accessed on 7 November 2021).
19. Skrobanek, J. *Regionale Identifikation, Negative Stereotypisierung und Eigengruppenbevorzugung*; Das Beispiel Sachsen; VS Verlag für Sozialwissenschaften: Wiesbaden, Germany, 2004.
20. Knipping-Sorokin, R. Radikalisierung Jugendlicher über das Internet?: Ein Literaturüberblick, DIVSI Report. Available online: https://www.divsi.de/wp-content/uploads/2016/11/Radikalisierung-Jugendlicher-ueber-das-Internet.pdf (accessed on 4 April 2022).

21. Mothe, J.; Ramiandrisoa, F.; Rasolomanana, M. Automatic keyphrase extraction using graph-based methods. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, Pau, France, 9–13 April 2018; pp. 728–730.
22. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* **2020**, *509*, 257–289.
23. Rose, S.; Engel, D.; Cramer, N.; Cowley, W. Automatic keyword extraction from individual documents. *Text Mining Appl. Theory* **2010**, *1*, 1–20.
24. Ashcroft, M.; Fisher, A.; Kaati, L.; Omer, E.; Prucha, N. Detecting jihadist messages on twitter. In Proceedings of the 2015 European Intelligence and Security Informatics Conference, Manchester, UK, 7–9 September 2015; pp. 161–164.
25. Rowe, M.; Saif, H. Mining pro-ISIS radicalisation signals from social media users. In Proceedings of the Tenth International AAAI Conference on Web and Social Media, Cologne, Germany, 17–20 May 2016.
26. Nouh, M.; Nurse, J.R.; Goldsmith, M. Understanding the radical mind: Identifying signals to detect extremist content on twitter. In Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), Shenzhen, China, 1–3 July 2019; pp. 98–103.
27. Gaikwad, M.; Ahirrao, S.; Phansalkar, S.; Kotecha, K. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access* **2021**, *9*, 48364–48404.
28. Alatawi, H.S.; Alhothali, A.M.; Moria, K.M. Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access* **2021**, *9*, 106363–106374.
29. Cohen, K.; Johansson, F.; Kaati, L.; Mork, J.C. Detecting linguistic markers for radical violence in social media. *Terror. Political Violence* **2014**, *26*, 246–256.
30. Chalothorn, T.; Ellman, J. Affect analysis of radical contents on web forums using SentiWordNet. *Int. J. Innov. Manag. Technol.* **2013**, *4*, 122.
31. Jurek, A.; Mulvenna, M.D.; Bi, Y. Improved lexicon-based sentiment analysis for social media analytics. *Secur. Inform.* **2015**, *4*, 1–13.
32. Fernandez, M.; Asif, M.; Alani, H. Understanding the roots of radicalisation on twitter. In Proceedings of the 10th ACM Conference on Web Science, Amsterdam, The Netherlands, 27–30 May 2018; pp. 1–10.
33. Araque, O.; Iglesias, C.A. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access* **2020**, *8*, 17877–17891.
34. Mothe, J.; Chrisment, C.; Dkaki, T.; Dousset, B.; Karouach, S. Combining mining and visualization tools to discover the geographic structure of a domain. *Comput. Environ. Urban Syst.* **2006**, *30*, 460–484.
35. Dousset, B.; Mothe, J. Getting Insights from a Large Corpus of Scientific Papers on Specialisted Comprehensive Topics-the Case of COVID-19. *Procedia Comput. Sci.* **2020**, *176*, 2287–2296.
36. Leavy, S.; Meaney, G.; Wade, K.; Greene, D. Curatr: A platform for semantic analysis and curation of historical literary texts. In Proceedings of the Research Conference on Metadata and Semantics Research, Rome, Italy, 28–31 October 2019; pp. 354–366.
37. Paranyushkin, D. InfraNodus: Generating Insight Using Text Network Analysis. In Proceedings of the World Wide Web Conference, WWW'19, San Francisco, CA, USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 3584–3589. https://doi.org/10.1145/3308558.3314123.
38. Hasan, K.S.; Ng, V. Automatic Keyphrase Extraction: A Survey of the State of the Art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA, 22–27 June 2014; pp. 1262–1273.
39. Mahata, D.; Shah, R.R.; Kuriakose, J.; Zimmermann, R.; Talburt, J.R. Theme-Weighted Ranking of Keywords from Text Documents Using Phrase Embeddings. In Proceedings of the IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018, Miami, FL, USA, 10–12 April 2018; pp. 184–189. https://doi.org/10.1109/MIPR.2018.00041.
40. El-Beltagy, S.R.; Rafea, A.A. KP-Miner: Participation in SemEval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala, Sweden, 15–16 July 2010; pp. 190–193.
41. Kim, S.N.; Medelyan, O.; Kan, M.; Baldwin, T. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala, Sweden, 15–16 July 2010; pp. 21–26.
42. Litvak, M.; Last, M. Graph-based keyword extraction for single-document summarization. In Proceedings of the Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization, COLING, Manchester, UK, 23 August 2008; pp. 17–24.
43. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
44. Bail, C.A. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11823–11828.
45. Rule, A.; Cointet, J.P.; Bearman, P.S. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10837–10844.
46. Fabo, P.R.; Plancq, C.; Poibeau, T. More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing. In Proceedings of the LREC: The 10th Language Resources and Evaluation Conference, Portorož, Slovenia, 23–28 May 2016.

47. Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A. *A Practical Guide to Sentiment Analysis*; Springer: Berlin/Heidelberg, Germany, 2017.
48. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1608.
49. Murtagh, F.; Taskaya, T.; Contreras, P.; Mothe, J.; Englmeier, K. Interactive visual user interfaces: A survey. *Artif. Intell. Rev.* **2003**, *19*, 263–283.
50. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
51. Roberts, M.E.; Stewart, B.M.; Tingley, D. Stm: An R package for structural topic models. *J. Stat. Softw.* **2019**, *91*, 1–40.
52. Arlacchi, P. *Mafia von Innen. Das Leben des Don Antonio Corleone*; FISCHER: Taschenbuch, Germany, 1995.
53. Galliani, C. *Mein Leben für Die Mafia: Der Lebensbericht Eines Ehrbaren Anonymen Sizilianers*; Rowohlt: Hamburg, Germany, 1989.
54. Lara-Cabrera, R.; Gonzalez-Pardo, A.; Camacho, D. Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter. *Future Gener. Comput. Syst.* **2019**, *93*, 971–978.
55. Gilpérez-López, I.; Torregrosa, J.; Barhamgi, M.; Camacho, D. An initial study on radicalization risk factors: Towards an assessment software tool. In Proceedings of the 2017 28th International Workshop on Database and Expert Systems Applications (DEXA), Lyon, France, 28–31 August 2017; pp. 11–16.
56. Van Brunt, B.; Murphy, A.; Zedginidze, A. An exploration of the risk, protective, and mobilization factors related to violent extremism in college populations. *Violence Gend.* **2017**, *4*, 81–101.
57. Knight, S.; Woodward, K.; Lancaster, G.L. Violent versus nonviolent actors: An empirical study of different types of extremism. *J. Threat Assess. Manag.* **2017**, *4*, 230.
58. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H., III; Crawford, K. Datasheets for datasets. *arXiv* **2018**, arXiv:1803.09010.
59. Hovy, D.; Prabhumoye, S. Five sources of bias in natural language processing. *Lang. Linguist. Compass* **2021**, *15*, e12432.
60. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4349–4357.
61. Bender, E.; Friedman, B. Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science. 2019. Available online: https://aclanthology.org/Q18-1041/ (accessed on 23 November 2021)).
62. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 220–229.
63. Crimmins, F.; Smeaton, A.F.; Dkaki, T.; Mothe, J. TetraFusion: Information discovery on the Internet. *IEEE Intell. Syst. Their Appl.* **1999**, *14*, 55–62.