Linguistic Approaches to the Analysis of Online Terrorist Threats

Julien Longhi

Chapter 12

Linguistic approaches to the analysis of online terrorist threats

Julien Longhi

CY Cergy Paris Université, AGORA, IDHN

Institut Universitaire de France (IUF)

Abstract

This chapter focuses on online terrorist threats. For many, the best training for a forensic linguist is a course in descriptive and applied linguistics because each case will normally require a different selection of tools from the linguistic toolbox. Coulthard and Johnson (2007) provide examples of forensic analysis focusing on morphological meaning, syntactic complexity, lexico-grammatical ambiguity, lexical meaning, pragmatic meaning, speech-to-writing transformation, narrative analysis, and features of non-native language usage. Since 2015 I have been using a longitudinal linguistic approach to the analysis of online language crimes relating to security issues (Longhi, 2008, 2012, 2108), in which all levels of linguistic analysis have proved to be useful in detecting, identifying, and characterising a threatening phenomenon. This chapter will illustrate the application of different linguistic methodologies, namely, textometry and semantic analysis, together with various tools, by drawing on the detailed analysis of an exemplary case of online terrorist threat taken from my collaboration with the French Gendarmerie.

Key words: Terrorist threats; Forensic linguistics; Stylometry; Stylistics; Textometry; Deep learning

1 Introduction

According to Dean and Bell (2012, pp. 11-12), 'Web 2.0 social media technologies have allowed terrorism to become a massive "dot.com" presence on the internet.' The question of online terrorist threats is a topic of growing interest. While computer sciences have already invested a lot in this field, especially in terms of digital traces and the analysis of computer networks, linguistics has only recently taken an interest in this subject. Bérubé et al. (2020) confirm that forensic sciences have taken a growing interest in digital traces as the latter are 'invaluable sources of information, although using them effectively poses certain challenges' (p. 8).

1.1 Internet, technologies, and information

According to a report issued by the United Nations Office on Drugs and Crime (UNODC) in collaboration with the United Nations Counter-Terrorism Implementation Task Force:

> Technology is one of the strategic factors driving the increasing use of the Internet by terrorist organizations and their supporters for a wide range of purposes, including recruitment, financing, propaganda, training, incitement to commit acts of terrorism, and the gathering and dissemination of information for terrorist purposes. While the many benefits of the Internet are self-evident, it may also be used to facilitate communication within terrorist organizations and to transmit information on, as well as material support for, planned acts of terrorism, all of which require specific technical knowledge for the effective investigation of these offences. (UNODC, 2012, p. 1)

In the fourth chapter of the above report (entitled 'Investigations and intelligence-gathering', pp. 53-72), the authors admit that:

> Effective investigations relating to Internet activity rely on a combination of traditional investigative methods, knowledge of the tools available to conduct illicit activity via the Internet and the development of practices targeted to identify, apprehend and prosecute the perpetrators of such acts ... [thus] different

> types of investigative techniques, both traditional and specifically relating to digital evidence, are employed in unison. (UNODC, 2012, p. 53)

In this chapter we will work within the realm of *forensic computing* and thereby analyse complex data generated from the 'increased use of multimedia, combined with the rapid expansion of the Internet' (McKemmish, 1999, p. 5). Textual data—for example, from social media, blogs, and forums—require processing adapted to natural language. Indeed, if we consider the technological changes that have affected the way in which law enforcement agencies conduct their criminal investigations and gather intelligence (Bérubé et al., 2020, p. 8), natural language processing techniques can contribute to analysing online terrorist threats. Given their multifaceted nature, it is thus important for linguists to have a certain competence in computer science in order to be able to analyse these threats.

1.2 Language processing

Chaski and Chemylinski (summarised in Chaski, 2005) explain the usefulness of the 'method for decomposing the data into smaller chunks so that a larger set of variables can be used for the discriminant analysis' (Chaski, 2005, p. 11). They note that, on average, the method produces very good results with over 95% accuracy. The work of Ainsworth and Juola (2019) allows us to take authorship attribution analysis further. They argue that 'language can be analysed by its objectively identifiable features' (p. 1173):

- forensic authorship analysis is objective, repeatable, and reproducible but needs to be based on empirical methods that clearly show how they work;
- applying the method of using big amounts of document sets in which the identity of the author is known, researchers can perform a number of calculations, measurements, and comparisons, which make it possible to compare texts in a corpus from different perspectives (word length, sentence length, term frequencies, grammatical categories, etc).

Let us consider the concept of threats from a linguistic point of view. In a study by Ascone and Longhi (2017), 'starting from the rhetorical pattern characterising the jihadist propaganda, where threat represents only a facet of the jihadist discourse' (p. 96), we examined both the content and form of propositions conveying a threat. Ascone

and Longhi's analysis yielded the following results that were subsequently confirmed by Ascone (2018):

- two magazines we looked at, *Dabiq and Dar-al-Islam*, presented different discourses and different types of threat but 'the Islamic State's propaganda aims at reinforcing the reader's adhesion to the jihadist ideology, and at inviting him/her to act against its enemies in the name of the jihadist ideology' (p. 6);
- four kinds of threat were identified: direct threat against enemies, direct threat against Muslims, the description of threatening events, and incitement to commit violent acts against the enemy.

The analysis of online threats requires both precise linguistic criteria and computer tools. For example, the question of enunciation devices, the links between explicit and implicit aspects of discourse, and sociolinguistic dimensions must be considered before conducting any data computer processing.

1.3 Looking for clues

While linguistics alone cannot solve crimes, a better understanding of language data brings additional clues that can be useful for criminal investigations. Digital tools, as we will see in the case study, are useful in extracting clues from the linguistic traces that make up texts. According to Margot (2014), 'it is only based on the knowledge of the possible versions of the facts, or propositions, that the value as a clue or the quality of the information provided by the trace can be measured or assessed' (p. 79). This chapter tries to show the link between trace and clue and how textometric analysis works. A linguistic trace fits Margot's (2014) definition:

> A trace is only an object with no meaning of its own. Its link to a case, and to reasonable hypotheses explaining its presence, in a way gives it its fundamental *raison d'être*. It is the observed result that makes the reasoning possible, an inference about a past fact. Thus, a trace becomes a sign when it is used for investigative purposes, or a clue when it is involved in reconstruction or demonstration. (p. 86).

The term *sign* is particularly interesting because it is the founding term of linguistics, as used by Ferdinand de Saussure when he developed linguistics as a part of

semiology. To link signs, clues, and evidence, we can follow Ainsworth and Juola (2019) who explain that when we look for clues, 'linguistic analysts examine systematic language variation on many levels' (pp. 1168-1169). Language usage patterns can be called 'style markers', and analyses based on these markers become 'forensic stylistics'. Ainsworth and Juola (2019) distinguish different levels:

- Patterns of punctuation, spacing, and spelling which can reflect an idiolect, regional dialects or slang;
- Grammatical choices;
- Narrative structures, levels of formality/informality, the use of irony, sarcasm, hyperbole, etc.

All these makers can be considered qualitatively and quantitatively, which works very well with the textometric method that will be presented and employed in the remainder of this article.

Thus, these reflections on the links between traces, clues, and evidence intersect with the debates and discussions that can take place in the fields of stylistics and stylometry. In his thesis, Wright (2014) investigates the advantages and disadvantages of the different approaches, highlighting three points: '(i) objectivity and reliability, (ii) theoretical and linguistic validity and explanation, and (iii) accessibility to lay judges and juries' (p.19).

If qualitative stylistic approaches can be seen as too subjective, 'much of this criticism comes from the United States, where the admissibility of expert evidence is determined in relation to the standards of the Daubert Criteria' (Wright, 2014, p. 19). Thus, stylometric approaches are 'considered to be more objective, empirical, replicable, and ultimately more reliable than their stylistic counterparts', but they can hardly give information about theoretical aspects of linguistic variation.

2 Related works

For Chen, Zhou, Reid and Larson (2011, p. 1, cited by Dean, & Bell, 2012, p. 15), terrorism informatics 'draws on a diversity of disciplines from Computer Science, Informatics, Statistics, Mathematics, Linguistics, Social Sciences, and Public Policy and their related sub-disciplines'. They point out that different approaches (for example, data mining, data integration, language translation technologies, image and video

processing) can be used in the prevention, detection, and remediation of terrorism. Three problems can appear with this information about digital aspects when applied to online terrorist threats: the amount of data, the specificities of computer processing, and the way in which linguistic treatments of digital corpora can be computerised.

2.1 The amount of data

According to UNODC (2012, p. 60), 'there is a vast range of data and services available via the Internet which may be employed in an investigation to counter terrorist use of the Internet'. Whether as part of internal or external communication, terrorist organisations produce a large amount of textual data. Communications between members of the organisation, communication intended for potential recruits, propaganda messages, or apologias for terrorism are all analysable data. However, the first difficulty concerns identifying relevant data among a myriad of messages. UNODC recommends 'a proactive approach to investigative strategies and supporting specialist tools, which capitalizes on evolving Internet resources, promotes the efficient identification of data and services likely to yield the maximum benefit to an investigation' (p. 60). Of all the existing resources, and because we are interested in analysing textual data, we will turn to natural language analysis methods and tools.

2.2 The specificities of computer processing of textual data

To discuss this point in more detail we will start with the work of McKemmish (1999) who has set out the rules of forensic computing, the observance of which 'is fundamental to ensuring admissibility of any product in a court of law' (p. 3).

For him, essentially, the rules of forensic computing are:

- Rule 1: An original should be handled as little as possible

During the examination of original data there should be minimum application of forensic computer processes. As McKemmish considers this rule as the single most important in forensic computing, I think it is therefore necessary to consider that the data processing which one will apply to textual corpora must be carried out with discernment and that the tool must be put at the service of the analyst.

- Rule 2: Every change should be accounted for

Every change that occurs during a forensic examination needs to be documented in terms of its nature, extent, and reason for it: I think that, in the case of textual data analysis, this rule essentially concerns the corpus. Thus, by describing how the corpus is set up, its possible enrichment, and its relation to the problem raised, the analyst can take a measured look at the case study.

- Rule 3: Rules of evidence should be complied with

Rules of evidence should be complied with when applying or developing forensic tools and techniques. From my point of view, one of the fundamental precepts of forensic computing is the need to ensure that the application of tools and techniques does not lessen the final admissibility of the product. Consequently, the type of tools and techniques used, as well as the way in which they are applied, is an important factor in ensuring compliance with the relevant rules of evidence. This point is very important and we will recall it when looking at our case study. Even assisted by a reputable computer tool, the linguist must always contact investigators and, more generally, those who know the case to which the corpus relates. Linguistics thus interacts with other analysis components and can provide new information linked to specific skills.

- Rule 4: An expert should not exceed their knowledge

A forensic computer expert should be aware of the scope and possible limitations of their experience. I think that this partly supports the previous rule by spelling it out. Suppose the researcher must keep in mind the limitations of their disciplinary point of view in relation to the investigation (and therefore consider other parameters outside their field of expertise). In that case, they must also keep in mind the degree of confidence in these tools and methods and question the status of their results in terms of, among other things, what is certain proof or, on the contrary, only a possibility or a probability.

From these rules and their description for the purposes of our context, we can reflect more precisely on the computerisation of our task, namely, the analysis of online terrorist threats.

2.3 How far can we computerise the processing of complex concepts?

While ‘forensic science is constantly evolving and transforming in response to the numerous technological innovations in recent decades’ (Bérubé et al., 2020, p. 1), in forensic linguistics determining authorship on the basis of habits of style questions the validity of stylometry. According to Totty, Hardcastle and Pearson (1987), some difficulties appear with such methods and tools. First, in forensic applications, at times ‘stylometry has been used to test the validity of claims by convicted persons that records of interviews, containing full or partial confessions, which formed part of the prosecution evidence at their trial, had been fabricated, in whole or in part’ (p. 17). In fact, utterances are often made by the person concerned in answer to questions and are taken down at the time by a police officer, so their validity as evidence is based on statements made by police officers that the records are true and accurate accounts of what was said.

The second difficulty linked to corpus linguistics concerns the scope of corpora and admissibility criteria. For example, in politics, the appearance of the political tweet as a genre of discourse (Longhi 2013, 2018) has modified certain parameters for setting up political corpora. Nevertheless, if certain precautions are taken, digital textual data are an important source for the researcher, and also for the investigator. As a scientific discipline, stylometry must consider the variety of texts, and thus the variety of size, as a constraint. The work of Lam, Demange and Longhi (2021) opens up promising prospects in this respect.

The second difficulty relates to the authenticity of the corpora. The question of the author, and more broadly of the speaker or the enunciator, has long been dealt with by linguistics, and it seems that a good knowledge of the theories of enunciation would provide guarantees for the analysis conducted (Benveniste, 1966; Ducrot, 1981, among others). In the study case that will be described later, the analysis focuses on anonymous texts, which is a subject of interest to linguistics and stylometry in particular.

3 Description and explanation of the most significant methodologies

In this section I will present the type of mixed method used to explore corpora which combines qualitative with quantitative processing. This approach allows for a concrete

analysis of corpora as presented at point 4, while bearing in mind the need for thoroughness and replicability.

3.1 A methodological mix: combining qualitative and quantitative approaches

Ascone's methodology combines linguistic analysis with a computer-based approach. As illustrated by a study I conducted together with Ascone (Ascone, & Longhi, 2017, p. 87), an iterative approach to jihadist propaganda may consist of four stages:

1) Stage 1: A preliminary qualitative analysis of the jihadist ideology, the radicalisation process, and the linguistic characteristics of hate speech has been essential to understanding the jihadist discourse and putting forward initial hypotheses. It is very important that the researcher knows the corpus, has read it, or at least browsed it if it is very voluminous, and knows the different criteria for structuring and setting up the corpus (extraction choices, variable selection). Using certain concepts or a qualitative analysis grid, the researcher or analyst can highlight a certain number of phenomena, make assumptions about the corpus, or extract a certain number of characteristics according to the knowledge acquired during the literature review.
2) Stage 2: A quantitative analysis whose goal is to verify the hypotheses' validity. The analysis can be performed with the help of software—for example, using a pre-established lexicon to identify the discourse themes. This instrumentation is reasoned because it is based on specific observables and described in the literature. However, at the same time, this approach allows unexpected results to emerge while objectifying analyses.
3) Stage 3: A deeper qualitative analysis of the themes. Textual statistics or data mining tools make it possible to return to the corpus. This criterion is fundamental for linguistic analysis. The results yielded by the tools are only a way of looking at textual data differently, by reorganising them or exploring them according to a certain hypothesis.
4) Stage 4: A final quantitative analysis to test the hypotheses and results yielded by the qualitative analysis. The last instrumentation can be useful to interact

again with the corpus via a tool—Steps 3 and 4 can be performed repeatedly since the instrumentation should be seen as an aid to the researcher or analyst.

The above-mentioned methodology may seem tedious because it requires the mobilisation of human expertise and IT tools in several stages. However, it makes it possible to overcome the difficulties raised by Bérubé et al. (2020): 'The amount of data is often too large for a traditional qualitative analysis, computational methods of network and content analysis have been used, depending on the research objectives' (p. 2). Therefore, we were able to explain that the proposed methodological mix makes it possible to benefit from the advantages of qualitative and quantitative methods.

This progressive methodology makes it possible to work on both raw and processed corpora at different levels, according to needs and tools. From a methodological point of view, we highlighted the role of corpora, the importance of their structuring, and the tools adapted to analyse them. We followed Ainsworth and Juola (2019) who explain how the community of forensic linguists can be structured thanks to this:

> the key to validating the science behind authorship attribution has been the development of accuracy benchmarks through the use of shared evaluation corpora on which practitioners can test their methodologies. These corpora consist of document sets with known 'ground truths' about their authorship (p. 1176).

We thus found practices close to what could be done in textual data analysis or data science, with empirical comparisons of methods and results based on common corpora. This type of practice seems to have the virtue of improving the reliability and transparency of studies in this field. Likewise, we tried to draw inspiration from the work and discussions in the textometry community on the one hand, and deep learning on the other, in order to improve the replicability and thoroughness of analyses.

3.2 Ensuring the replicability of the analysis in order to ensure its thoroughness

In the context of analysing corpora that include terrorist threats, it is crucial that the method used is replicable, the results are explicit and transparent, and the conclusions

can be clearly justified. This takes us to ‘Forensic reproducibility’ (Garfinkel, Farrell, Roussev, & Dinolt, 2009, p. 3) where two areas of interest stand out: reproducibility makes it possible ‘for one researcher or research group to validate that they have mastered a technique and then to go off in a different direction’. This criterion therefore allows the community to discuss, compare results, and make progress. The criterion of validation, and more generally of consensus, also makes it possible, in my view, to give a stronger social scope to academic disciplines, not least, in our case, corpus linguistics.

For the remainder of the article I will use the textometric approach which is the subject of regular work and discussions in the scientific community, including at the International Conference on Statistical Analysis of Textual Data or Journées d’analyse des données textuelles (JADT) which takes place in Europe every two years. Textometry offers an instrumented approach to corpus analysis, combining quantitative with qualitative analysis (Lebart, & Salem, 1994). Functionally, textometry implements differential principles, using statistics and probabilities. The approach highlights similarities and differences observed in a corpus according to different criteria (words, grammar, n-gram, among others.). Textometry employs contextual and contrastive modelling, which makes it possible to perform objective ‘measurements’ on texts, but also to measure distances, proximities, similarities, and differences between texts or parts of texts. Such a method makes it possible to ensure both the replicability of the results and their explanation (there is no ‘black box’ here).

4 Case study

My own case study (Longhi 2021) proposes a summary and an extension of this research with new analysis based on specific examples and the presentation of innovative methods (deep learning).

4.1 Characterising the themes running through criminal acts and the types of threat

The French Gendarmerie provided material to help investigators with cases involving criminal acts that concerned especially terrorist groups with links to the far left. I collected 23 anonymous texts from various websites in which authors claimed

responsibility for malicious acts. The analysis conducted aimed to help investigators by formulating hypotheses on the possible number of authors or the probability that, for example, texts x, y, z had one, two, or three authors. These characteristics should help investigators with other dimensions of forensic study.

The 23 texts contained 12,109 occurrences, 2,534 forms with an average of 526 occurrences per text. This corpus size, although modest, was well suited to textometric methods. Such an approach could already be useful in proposing a thematic classification of articles. It could be represented using the descending hierarchical classification which resulted from the Alceste method and was provided by the Iramuteq software (Reinert, 1990)—see Figure 12.1.

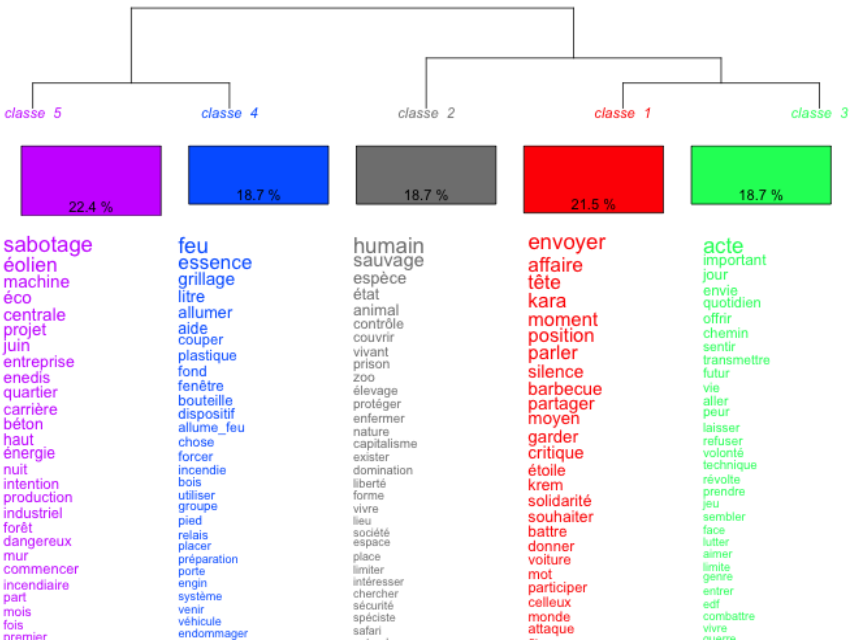

Figure 12.1 Descending hierarchical classification (themes) of the corpus

By means of the themes shown in figure 12.1 we were able to establish a link between terrorist threats and the subjects they applied to—e.g. energy sources, places related to animal husbandry, etc. Probst, Shkapenko, Tkachenko, & Cheruyakov (2018) explained that the communicative and semantic factors that determine the force of the motivating impact of a speech act of threat include i) the significance of punitive measures, ii) the possibility of punishment, and iii) the high probability of negative actions stated or implied by the producer.

When looking for examples in the corpus which reflected these criteria, the following quotes stood out:

1) We don't live in the past, we don't expect anything from the future, our revolts have no future, so they can't be put off until tomorrow.
2) We are answering the call for a dangerous June because it expresses these nuances well.
3) On Thursday night we broke into the ENEDIS building in Crest, which supplies the energy that allows this shitty world to turn. We poured 10 litres of petrol inside and lit it with handheld flares (have a plan B in case the handheld flares fail). 10 litres of petrol give one hell of a blast. By the time we got back out, the building was in flames. We found out later it was destroyed to a large extent.
4) On the night of 25 to 26 October 2016 we set fire to the car of a chief of the Tuilières gendarmerie which was parked in the barracks compound. We committed this act of sabotage in solidarity with the migrants in the Calais Jungle.
5) An incendiary device was placed and lit under each of the three Enedis vans parked in the company car park. We were in a hurry to get to the karaoke night so we didn't have much time to watch them burn, but we do hope we started a nice bonfire. ... They say June is going to be dangerous. Let's hope it's just the beginning.

Words such as 'revolts' that can't be 'put off' (1), 'a dangerous June' (2), 'lit it', 'flames', and 'destroyed' (3), 'set fire' and 'sabotage' (4), 'incendiary device', 'burn', and 'dangerous' (4) illustrate the way in which these acts echoed terrorist threats, reprisals, and violent acts which, from the point of view of their perpetrators, were the consequence of actions contrary to what they stood for.

4.2 Textometric and deep learning analysis[i]

However, to help investigators link authors of texts with perpetrators of malicious acts, we can resort to calculating specificities, making it possible to group texts according to their linguistic dimensions. The full analysis is developed in my study (Longhi, 2021).

For example, one can observe these distances between texts by focusing on grammatical variables as depicted in Figure 12.2.

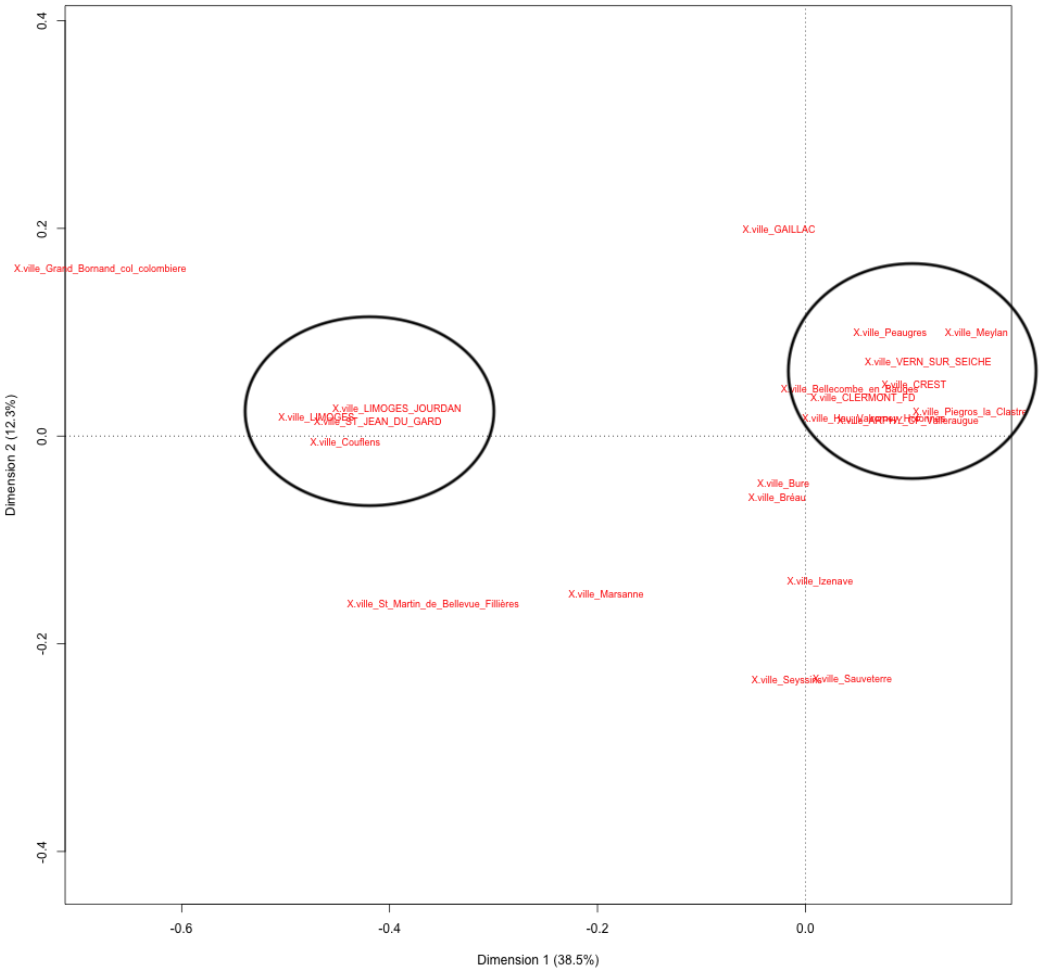

Figure 12.2 Graphic grouping of texts based on their grammatical characteristics

The analyst can thus formulate hypotheses on the proximities between texts, which they can compare with the survey data after returning to the corpus to examine these results in context. In order to increase the quantity of data to be analysed by relying on the most recent technologies, current projects focus on using deep neural networks. We will discuss this avenue in the next section.

Starting from the gendarmerie's initial data, I found some particularly interesting sites on which to test my approaches to linguistic analysis in security-related contexts (which were not necessarily akin to terrorism). The data retrieval work was carried out by Jeremy Demange, the engineer from CY's (Cergy Paris Université) digital humanities institute. The data came from the site https://nantes.indymedia.org. We

retrieved this site via a copy from the Common Crawl website (http://commoncrawl.org), which allowed us to avoid overloading the publisher site's server and also to obtain data quickly and easily. To retrieve the content, we used a suite of AWS tools such as Athena, which allowed us to retrieve all pages available on the site as of September-October 2020. This strategy allowed us to extract almost all of the articles published from 2003 (when https://nantes.indymedia.org was created) to September 2020. We were able to retrieve a total of 8,126 unique articles from the site and detected a total of 4,806 unique authors. The 'anonymous' author (without preprocessing) appears most often with a total of 287 frequencies. However, it should be noted that anyone can write any author's name on this site and be identified as *anonyme* ('anonymous'). Thus, I performed a preprocessing step to eliminate any results that appeared to me to be false—that is, uppercase characters and spaces were removed. Table 12.1 depicts the top 10 names of authors who wrote the most on the site (after preprocessing).

Author's name	Frequency
anonyme	407
zadist	232
nantesrévoltée	184
.	166
anonymous	152
radiocayenne	82
unsympathisantducci	67
...	63
x	48
*	44

Table 12.1 Names of authors listed in the articles

I trained a prototype Python model to test the author comparison on this corpus. I was able to achieve 93.48% accuracy (the model being improved). Figure 12.3 shows the details of the selected model.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None,)]                 0
_________________________________________________________________
expand_last_dim (ExpandLastD (None, 1)                 0
_________________________________________________________________
text_vectorization (TextVect (None, 512)               0
_________________________________________________________________
embedding (Embedding)        (None, 512, 64)           320064
_________________________________________________________________
dropout (Dropout)            (None, 512, 64)           0
_________________________________________________________________
conv1d (Conv1D)              (None, 508, 256)          82176
_________________________________________________________________
global_max_pooling1d (Global (None, 256)               0
_________________________________________________________________
dense (Dense)                (None, 256)               65792
_________________________________________________________________
re_lu (ReLU)                 (None, 256)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 671)               172447
_________________________________________________________________
classification_head_1 (Softm (None, 671)               0
=================================================================
Total params: 640,479
Trainable params: 640,479
Non-trainable params: 0
```

Figure 12.3. Prototype of authorship attribution model

For the analysis of similarities, an algorithm version was developed by Jérémy Demange in Python. The analysis focused on the first nine authors:

• author_anonymous
• author_zadist
• author_nantesrévoltée
• author_.
• author_anonymous
• author_radiocayenne
• author_…
• author_x

Figure 12.4 depicts the results of the analysis of the texts by the above-mentioned authors.

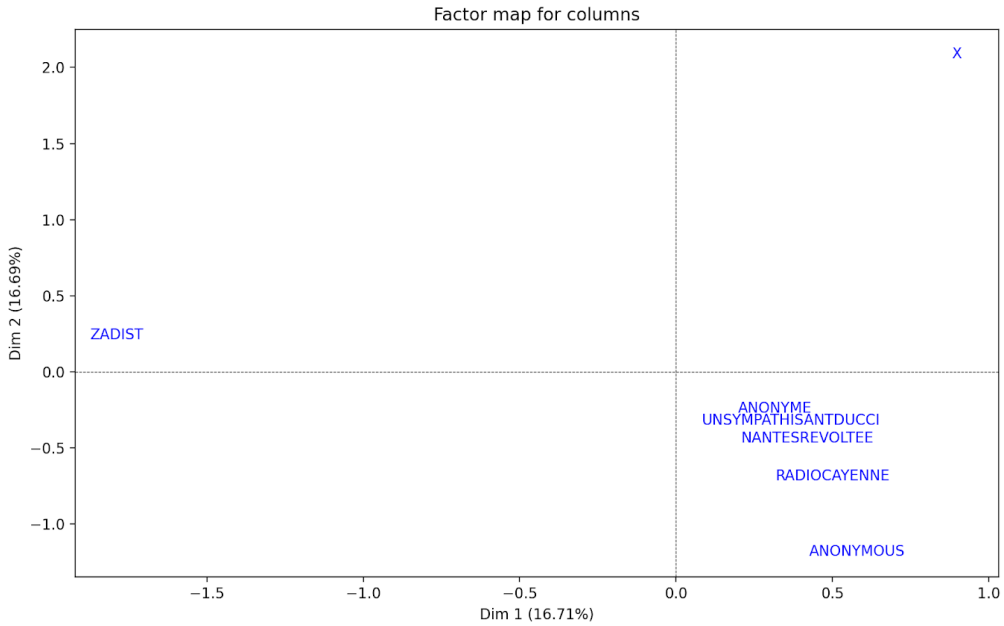

Figure 12.4 Authors connected by the analysis model

Therefore, we could observe the possible connection or distance between certain names of authors based on a larger body of text than at point 4.1. Of course, this work is in progress and needs further study, but there is promise in using this technology to help deal with online threats. Knowledge of the methods at 4.1 can help refine the results, for example, if we look at the themes of the texts or add information to the stylistic analysis of the authors, as shown in Figure 12.5.

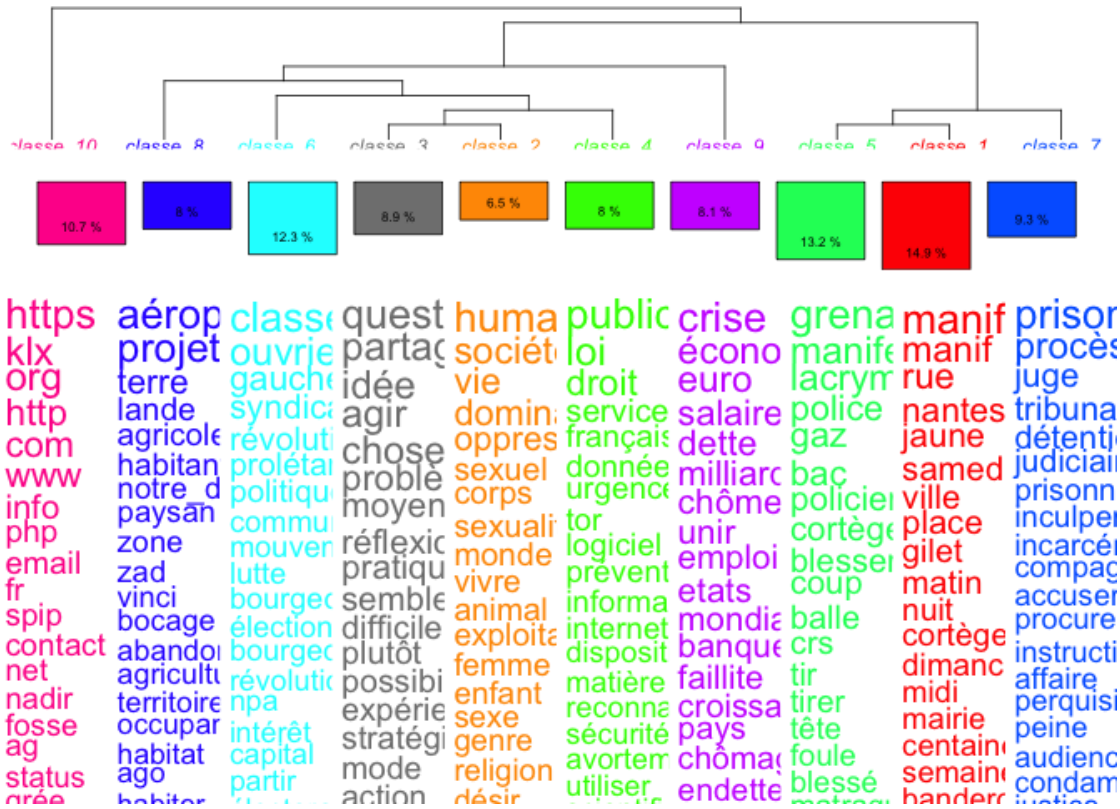

Figure 12.5 Descending hierarchical classification (themes) of the second corpus

This analysis allows, for example, to distinguish specific subjects (political, economic) and help investigators concentrate on the classes that would perhaps be more relevant to them—for example, class 5 which includes terms such as ‘grenade’, ‘bullet’, and ‘injure’, or class 2.

What I wish to highlight at the end of the analysis of this example is the possible complementarity between deep learning and textometry for the purposes of corpus analysis: while textometry allows for an efficient instrumentation of the analysis and the exploration of texts, it may present some limitations in producing clear results on certain issues (text authorship, content similarity, among others). Thus, textometry can be used to make connections, filter certain parts of the corpus, and orientate the analysis. Deep learning can then provide analysis procedures that are more efficient but also more explicit and thus easier to understand (the ‘black box’ aspect of some algorithms). Textometry can then be used at the end of the study to once again give meaning to and verify the results obtained.

5 Conclusions

This chapter has highlighted several dimensions of online threat analysis, particularly in the context of terrorism. The evolution of technologies and their efficient use for criminal purposes makes it necessary to take into account the linguistic aspects of these threats in a thorough and systematic way. To this end, I have presented the challenges of a method that combines qualitative and quantitative approaches and seeks to emphasise the replicability and thoroughness of such analyses. This serves a dual purpose: ensure the quality of analyses, but also provide institutions, professionals, and society at large with the assurance that these analyses are reliable and can be verified and redone.

To this end, I have presented a model that combines textometry with deep learning: while textometry provides the means to measure, compare, and explore corpora, deep learning can then efficiently produce results on certain research questions that can initially be addressed from a statistical point of view. Textometry can also help to better understand the results provided by artificial intelligence algorithms, contextualising and exemplifying the results. I have thus highlighted examples of online threats involving violence, malicious acts, or reprisals. Being able to characterise the authors of such threats is therefore a major goal, particularly when it comes to dealing with terrorist acts and their consequences.

Endnotes

[i]The paragraph on deep learning was written in collaboration with Jeremy Demange, engineer at CY IDHN.

6 References

Ainsworth, J., & Juola, P. (2018). Who wrote this: Modern forensic authorship analysis as a model for valid forensic science. *Wash. UL Rev.*, *96*, 1161-1189.

Ascone, L. (2018). Textual analysis of extremist propaganda and counter-narrative: A quanti-quali investigation. *JADT*, June 2018, Rome, Italy. Retrieved from https://hal.archives-ouvertes.fr/hal-02317752

Ascone, L., & Longhi, J. (2017). The expression of threat in jihadist propaganda. *Fragmentum, 50,* 85-98.

Benveniste, E. (1966). *Problèmes de linguistique générale*. Paris: Gallimard.

Bérubé M., Tang T. U., Fortin F., Ozalp S., Williams M. L., & Burnap P. (2020). Social media forensics applied to assessment of post-critical incident social reaction: The case of the 2017 Manchester Arena terrorist attack. *Forensic Science. International*, *313*, 110364. doi: 10.1016/j.forsciint.2020.110364

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, *4*(1), 1-13.

Chen, H., Zhou, Y., Reid, E. F., & Larson, C. A. (2011). Introduction to special issue on terrorism informatics. *Information Systems Frontiers*, *13*(1), 1-3.

Dean, G., & Bell, P. (2012). The dark side of social media: review of online terrorism. *Pakistan Journal of Criminology*, *3*(4), 191-210.

Ducrot, O. (1981). Langage, métalangage, et performatifs. *Cahiers de linguistique*, *3*, 5-34.

Garfinkel, S., Farrell, P., Roussev, V., & Dinolt, G. (2009). Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation*, *6*, S2-S11.

Lam, T., Demange J., & Longhi, J. (2021). Attribution d'auteur par utilisation des méthodes d'apprentissage profond. *Proceedings of the Deep Learning for NLP workshop*, EGC 2021.

Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Longhi, J. (2013). Essai de caractérisation du tweet politique. *L'Information Grammaticale*, *136*, 25-32.

Longhi, J. (2018). *Du discours comme champ au corpus comme terrain. Contribution méthodologique à l'analyse sémantique du discours*. Paris: L'Harmattan.

Longhi, J. (2021). Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International*, *318*, 110564.

Margot, P. (2014). Traçologie: la trace, vecteur fondamental de la police scientifique. *Revue Internationale de Criminologie et de Police Technique et Scientifique*, *67*(1), 72-97.

McKemmish, R. (1999). *What is forensic computing?* Canberra: Australian Institute of Criminology. Retrieved from https://www.aic.gov.au/sites/default/files/2020-05/tandi118.pdf

Probst, N., Shkapenko, T., Tkachenko, A., & Chernyakov, A. (2018). Speech act of threat in everyday conflict discourse: Production and perception. In *Lege Artis*, *3*(2), 204-250. doi: 10.2478/lart-2018-0019

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia de Gerard de Nerval. *Bulletin of Sociological Methodology/Bulletin de méthodologie sociologique*, *26*(1), 24-54. doi: 10.1177/075910639002600103

Totty, R. N., Hardcastle, R. A., & Pearson, J. (1987). Forensic linguistics: The determination of authorship from habits of style. *Journal of the Forensic Science Society*, *27*(1), 13-28.

United Nations Office on Drugs and Crime (UNODC). (2012). *The Use of the Internet for Terrorist Purposes*. Retrieved from https://www.unodc.org/documents/terrorism/Publications/Use_of_Internet_for_Terrorist_Purposes/ebook_use_of_the_internet_for_terrorist_purposes.pdf

Wright, D. (2014). *Stylistics versus statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails* (Unpublished doctoral thesis), University of Leeds, England.