Radical Text Detection based on Stylometry

Álvaro de Pablo[a], Óscar Araque[b] and Carlos A. Iglesias[c]

Intelligent Systems Group, Universidad Politécnica de Madrid, Madrid, Spain

Abstract: The Internet has become an effective tool for terrorist and radical groups to spread their propaganda. One of the current problems is to detect these radical messages in order to block them or promote counter-narratives. In this work, we propose the use of stylometric methods for characterizing radical messages. We have used a machine learning approach to classify radical texts based on a corpus of news from radical sources such as the so-called ISIS online magazines Dabiq and Rumiyah, as well as news from general newspapers. The results show that stylometric features are effective for radical text classification.

1 INTRODUCTION

Recent terrorist attacks have revealed a vulnerability in modern society. As a natural consequence, countering and preventing violent terrorism and radicalization has become a major international security priority. Terrorism is nowadays a major threat to security in the EU Member States (Europol, 2016).

In this context, cyberspace is of special interest, since the effective usage of the Internet has been one of the differential aspects of modern terrorist groups. Their use range from psychological warfare, publicity, and propaganda to more instrumental actions such as recruitment, mobilization, fundraising, information sharing, coordination, and networking (Weimann, 2004).

Terrorist organizations have used different media strategies. According to New York Times (Shane and Hubbard, 2014), Islamic State of Iraq and Syria (ISIS) media strategy can be named Jihad 3.0 since they follow a sophisticated multidimensional strategy (Al-Rawi, 2018; Feakin and Wilkinson, 2015), including not only social media (e.g., Twitter) but also video games (Al-Rawi, 2018), online magazines (e.g., Inspire, Dabiq and Rumiyah), high-quality videos, audio reports in SoundCloud or publication of battle summaries in JustPaste, to name a few.

Since understanding terrorist organizations is a requirement for detecting and countering them, a large

body of research has studied radical presence in social media (Correa and Sureka, 2013; Fernandez et al., 2018), including Twitter, YouTube and Facebook, online magazines (Ingram, 2017) and videogames (Al-Rawi, 2018).

In this work, we provide a first attempt to characterize the writing style of terrorist groups in online magazines. This has several important applications, such as the detection (Cohen et al., 2018) of online terrorist propaganda and misinformation (Wilner, 2018) as well as the authorship attribution for counterterrorism purposes (Kijewski et al., 2016).

To address this research objective, we have used a rich source of data, the official ISIS online magazines Dabiq and Rumiyah. The methodology followed has been to create a corpus of radical texts with the articles of these magazines. In addition, we have mined articles related to radicalism from the online magazine of the Arab satellite news network Al Jazeera, which is considered an "alternative" medium (Iskandar, 2006; Al-Sadi, 2012). Another corpus of nonradical nor alternative articles has been created with articles from international magazines (e.g., New York Times, CNN) related to radical events. Then, a stylometric analysis of the articles has been carried out, and a machine-learning classifier has been trained with stylometric features to evaluate its relevance.

The remainder of this paper is organized as follows. Sect. 2 reviews related work. Sect. 3 provides a background on the style metrics used in the analysis. Sect. 4 presents the methodology followed for collecting the dataset as well as a stylometric analysis of non-radical and radical news. Sect. 5 reports

[a] https://orcid.org/0000-0002-7561-9997

[b] https://orcid.org/0000-0003-3224-0001

[c] https://orcid.org/0000-0002-1755-2712

524

the experiments to classify texts based on stylometric features. Traditional approaches are compared to evaluate the increment in performance of using these features. Finally, Sect. 6 summarizes the results from our analysis and highlights the implications of our results.

2 RELATED WORK

Stylometry techniques[1] have been applied to authorship attribution, authorship verification, authorship profiling, stylochronometry, and adversarial stylometry (Neal et al., 2017).

In the context of radicalism detection, several works have already used stylometric features.

The approach most closely related to our approach is that of Ashcroft et al. (Ashcroft et al., 2015). They describe a machine learning system for classifying Jihadist messages on Twitter using stylometric, temporal, and sentiment features. They use the stylometric features described in (Narayanan et al., 2012), which include i) length of words/characters, vocabulary richness (based on Yule's K^3 and frequency of *hapax legomena*, *dis legomena*, etc.), word shape (frequency of words with different combinations of upper and lower case letters), word length (frequency of words that have 1–20 characters), frequency of different elements (words, letters, digits, punctuation, hashtags, special characters, function words, and syntactic category pairs). Nevertheless, this study does not detail the impact of stylometric features on the classification. In addition, most stylometric metrics are based on word and hashtag frequencies.

Mencarini and Sensidoni (Mencarini and Sensidoni, 2017) analyzed some terrorist or cybercriminal statements belonging to groups like Anonymous, ISIS, Al Qaeda, or the Muslim Brotherhood with some style metrics which include readability, vocabulary, registers and slangs, grammatical tenses and more. The application of the stylometry to these kinds of texts showed differences between different terrorist groups in aspects like the difficulty of reading or lexical diversity.

Regarding the analysis of radical online magazines from ISIS, Dabiq, and Rumiyah, as well as al Qaeda's Inspire, they have been subject to a number of studies.

Bisgin et al. (Bisgin et al., 2019) analyze Dabiq's propagandist elements by studying the entities mentioned in the articles.

Vergani and Bliuc (Vergani and Bliuc, 2018) have analyzed the language differences between ISIS and al-Qaeda in their journals Dabiq and Inspire, respectively. Their analysis is based on the tools Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) and Recursive Inspection of Text (RIOT) (Boyd, 2015) to calculate the frequency of texts in language categories (e.g., article, pronoun), psychological dimensions (e.g., affect, emotions, cognitive mechanisms) or moral foundations (e.g., harm, fairness, authority). The main insight of this study is that ISIS propaganda is more effective in mobilizing individuals who are more authoritarian and more religious than that of al-Qaeda.

Sikos et al. (Sikos et al., 2014) analyze an Arabic translation of al-Qaeda's Inspire magazine since most of the magazines are published in English together with an Arabic translation. They use stylometry features using LIWC to analyze authorship of Inspire's issues, concluding that the issues 1-9 were produced by one group of editors while a different group of editors produced issues 10 and 11.

Wignell et al. (Wignell et al., 2017) analyze Dabiq and Rumiyah's style in terms of the topics addressed and the images included, using a mixed-methods approach. They conclude that the distribution of topics and images is quite consistent in both magazines. In addition, they always include ISIS's core values, intolerance and a hostile world.

Johnston and Weiss (Johnston and Weiss, 2017) develop a deep learning model for detecting Sunni propaganda. They use a dataset of Sunni propaganda with the journals Dabiq, Rumiyah and Inspire as curated from the Clarion project[2], and another benign dataset that includes Wikipedia articles and news. In this case, their approach is similar to ours for building the dataset, but they do not address stylometry.

3 ON STYLOMETRY

In this section, the main metrics used for the stylometric analysis of the texts are presented.

Readability Index. (Kincaid et al., 1975) is a style metric that measures how easy or difficult reading a text is. We have used two popular metrics for texts written in English, Fog Count, and Flesch (Kincaid et al., 1975), both based on the number of syllables of the words within the text. These metrics relate the value obtained with the US Grade Level required to read a text.

---

[1] The interested reader in stylometry techniques and applications should consult (Neal et al., 2017) for a recent survey of this topic.

[2] https://clarionproject.org/

Vocabulary Richness. Measures the lexical diversity of the texts. The most relevant metric is Type Token Ratio (TTR) (Šišková, 2012), which is calculated as shown in Eq. 1.

$$TTR = \frac{types}{tokens} \quad (1)$$

The main problem of TTR is that the longer a text is, the smaller TTR. To solve this, the metrics MTLD and HD-D have been proposed, as described below.

The **Measure of Textual Lexical Diversity (MTLD)** (McCarthy and Jarvis, 2010) is a sequential algorithm for measuring the Vocabulary Richness of a text. It is calculated as follows:

$$MTLD = \frac{tokens}{Segments + Partial\ segment} \quad (2)$$

where *Segments* is the number of segments of the text that have a TTR lower than a limit (usually 0.72), and *Partial Segment* is the last segment of the text.

Hypergeometric Distribution of the Diversity index (HD-D) (McCarthy and Jarvis, 2010). It uses the hypergeometric distribution to determine the probability of a particular word in a text occurring at least once in a random sample of a particular length. The probability is calculated as follows:

$$P(X = x) = \frac{\binom{d}{x}\binom{N-d}{n-x}}{\binom{N}{n}} \quad (3)$$

where N is the population size, n is the sample size, d is the number of the elements belonging to the request type and x is the number of elements in the sample that belongs to the request category.

HD-D algorithm is shown below.

$$HDD = \sum_{i=0}^{n} \frac{1}{42} * P(X = type_i) \quad (4)$$

where $P(X = type_i)$ is the probability given by the hypergeometric distribution for any type in the text.

Formality Metrics. (Heylighen and Dewaele, 1999) measure the degree of formality of a text. The main metrics are Adjective Score and F Score.

Adjective Score (Fang and Cao, 2009) analyzes the adjective density of a text. Its value is usually under 10%. The higher the Adjective Score, the higher the degree of formality. Its calculation is as follows:

$$Adj\ Score = \frac{Number\ of\ Adjectives}{tokens} * 100 \quad (5)$$

F Score (Heylighen and Dewaele, 1999) is another formality metric. It takes into account the POS (Part Of Speech) tagging processes. The higher the F Score, the higher the degree of formality.

$$F = \frac{\frac{N + Adj + Prep + Dt - Pro - Vb - Adv - In}{Tokens} * 100 + 100}{2} \quad (6)$$

where N is *Nouns*, *Adj* is *Adjectives*, *Prep* is *Prepositions*, *Dt* is *Determiners*, *Pron* is *Pronouns*, *Vb* is *Verbs*, *Adv* is *Adverbs* and *In* is *Interjections*.

Coherence Measure. (Foltz et al., 1998) is an index that evaluates the coherence of a text. The Coherence index of a text is calculated by adding the semantic similarity of every sentence and the successive one, as shown in Eq. 7.

$$Coherence = \frac{\sum_{i=0}^{n-1} Coh(sent_i, sent_{i+1})}{sentences} * 100 \quad (7)$$

where *sentences* is the number of sentences within the text, $sent_i$ is the sentence number i within the text and $Coh(sent_i, sent_{i+1})$ is the semantic similarity of a sentence with the following in the text.

4 ANALYSIS OF THE DATA

4.1 Dataset Collection

The field of radicalization detection has an inherent need for representative and reliable data with which it is possible to develop computational models. Nevertheless, annotated data is scarce, which makes it difficult for some research advancements. Additionally, many of the existing datasets have been extracted from Twitter (Fernandez et al., 2018). While Twitter is indeed a representative source of data, it also has some generalization issues that proper radicalization detection studies should take into account. In this way, intending to explore alternative data sources, we have collected a dataset based on magazines.

Following previous work that tackles the use of Dabiq for characterizing radical language (Nouh et al., 2019; Bisgin et al., 2019), we have considered both the Dabiq and Rumiyah magazines, which are used by ISIS radical organization. Original publications have been collected from an online resource dedicated to radicalization online[3]. The obtained dataset consists of the 15 issues of Dabiq and the 13 issues of Rumiyah (276 texts), and 349 articles from Al Jazeera, which can be considered as an alternative source (Iskandar, 2006; Al-Sadi, 2012).

---

[3]http://www.jihadology.net

Table 1: Statistics of the collected dataset.

	CNN	NYT	Al Jazeera	Dabiq	Rumiyah
Avg. no. of words/article	692.565	197.501	680.169	969.823	612.286
Avg. no. of sentences/article	31.025	9.04	28.263	43.0	41.720
Avg. no. of word appearances/article					
Allah	0.005%	0.001%	0.004%	1.799%	2.628%
Jihad	0.012%	0.012%	0.013%	0.318%	0.257%
Khilafah/Caliphate	0.037%	0.038%	0.033%	0.314%	0.325%
Iran	0.228%	0.221%	0.228%	0.029%	0.013%

As a balance of the previous radical data, we have selected two online newspapers that address ISIS-related issues using a neutral tone: CNN and The New York Times. These newspapers are freely accessible and can be obtained through their APIs. In total, 383 articles have been collected from CNN, and 765 from The New York Times. For the dataset compensation, 300 random texts from the CNN dataset and 300 random texts from The New York Times have been analyzed. The same preprocessing has been applied to all the obtained data: normalization of capital letters, numbering, and contraction tokenization (e.g., *I've*, *we'll*). The processed data contains documents with lower case tokens. In Table 1, some statistics of the obtained dataset are shown.

4.2 Analysis

This section focuses on the analysis of the dataset described previously. We follow a systematic approach to analyze different stylometry metrics for every issue of the magazines Dabiq and Rumiyah and provide a comparison with the average metrics of the news dataset. From now on, *D.n* refers to *"Dabiq: Issue n"*, while *R.n* refers to *"Rumiyah: Issue n"*.

4.2.1 Formality: Adjective Score

The Adjective Score remains constant in Rumiyah and Dabiq issues, although it is slightly slower in Rumiyah. The reason for this could be the differences between Rumiyah and Dabiq. Some authors have reported that Rumiyah replaced Dabiq because of a shift in ISIS's priorities, and it is more focused on directly attacking western countries, rather than encouraging recruits to migrate to Iraq and Syria (Comerford, 2016). Other authors (McKernan, 2016) also report that Rumiyah can be considered an inferior product based on recycled material and significantly shorter.

It can be highlighted the difference between propaganda news and standard news, being the formality of standards news significantly higher (around 7%). In conclusion, propaganda texts are less formal than standard news. Another finding of this experiment is that Dabiq is more formal than Rumiyah.

These results are aligned with the conclusions of the authors of the adjective score, Fang and Cao (Fang and Cao, 2009). They concluded that the variations of adjectives seem to be a reliable indicator to categorize text categories. This is confirmed for our experiment for the two text categories, radical propaganda, and news.

4.2.2 Formality: F Score

In contrast with the previous analysis, the F Score shows a very stable value across all the issues. In addition, the F Score of news (72.49%) is slightly higher, and only one issue achieves this value (D.2, with 74.69%).

Based on the F Score, values above 70% imply high formality. Thus, the analyzed news are very formal, and the vast majority of the terrorist texts have a standard formality grade. According to (Heylighen and Dewaele, 1999), a formal style is characterized by objectivity and cognitive load, while an informal style is more direct, subjective, less accurate, and less informative. This confirms our results, where subjective propaganda texts have a lower formality index.

4.2.3 Coherence

For measuring coherence, a Word Embeddings model has been used (Sun et al., 2016) for English texts. The average of the coherence of the news is around 60.77% and only two issues (D.5 and D.7) are far from that value. The higher value is the coherence of R.12, with 74.41%.

The results show that Dabiq is less coherent than Rumiyah. Specifically, the average of the coherence

of the issues of Dabiq is 64.33% (1 percentage point over the coherence of the news), and the average of Rumiyah is 69.84% (almost seven percentage points higher). Moreover, every issue of Rumiyah exceeds the coherence value of the news, but some of the issues of Dabiq do not exceed it.

Thus, almost all issues (except D.5, with 51.27% and D.7, with 59.98%) are above 60% of coherence. In fact, most of the values are above 70%.

Therefore, it can be said that both magazines and news are coherent, being the coherence of Rumiyah higher than the coherence of Dabiq. Also, ISIS magazines are a bit more coherent than news. In fact, every issue of Rumiyah is more coherent than newspapers, while there are few issues of Dabiq that are less coherent than newspapers.

4.2.4 Vocabulary Richness

Two different metrics have been used to measure the lexical diversity of the texts.

Firstly, the average of the application of the HD-D metric to the news texts is 74.21%, while the average of Dabiq and Rumiyah texts is around 71.70%, close to news average but below. Moreover, while the HD-D of Dabiq (with an average of 72.93%) is closer to the HD-D of the news in each of its issues (in fact, D.3 surpasses it), the HD-D of Rumiyah (with an average of 70.28 %) is always under this limit, and it does not get that close. This finding is aligned with some observations about the lower quality of Rumiyah, as commented previously. Even so, all the HD-D values are around 70%, which implies a high lexical diversity.

Regarding the metrics MTLD, it provides better discrimination between news and propaganda, since its value for news is around 92% while its average value for propaganda is around 67%. In addition, its value is lower in Rumiyah than in Dabiq.

In conclusion, both news and magazines have good lexical diversity, but news values are higher in this aspect. Moreover, Dabiq has a better vocabulary richness than Rumiyah. These conclusions are supported by both metrics, although MTLD is more discriminative.

4.2.5 Readability Index

Two different metrics have been used too for measuring the Readability Index of the texts.

The results of the magazines vary greatly throughout the issues, and all the metrics applied to the magazines follow the same path, having approximately the maximums and the minimums in the same issues.

As for news, the Fog Count is 14.27, and Flesch-Kincaid is 13.05, which implies in both cases that news require a college level of education (Kincaid et al., 1975).

The Flesch results of the magazines have the lowest values. In Dabiq's issues, only D.11 exceeds the Flesch of the news. On the other hand, most of Rumiyah issues exceed this value. In fact, most of the Dabiq values are between 8 and 12, and only a High School Grade Level is required for reading them. Nevertheless, Rumiyah values are higher than 12, and a College Grade Level is needed.

Fog Count results show similar values but higher than Flesch. In fact, a College Grade Level is required in almost all issues.

Moreover, for both metrics, Dabiq is easier to read than Rumiyah. Also, most Flesch values are under the values of the news, but in the Fog Count measure, there are more values above the values of the news.

4.2.6 Use of Words

Also, it has been measured how many common and uncommon words are used in the different analyzed texts. Please note that common words are those who are common in the original language (in this case, English). If a word is common or uncommon is given by an algorithm based on the Zipf Law (Montemurro, 2001).

The results show that terrorism news have the highest proportion of common words (98.32%). The common words of Rumiyah and Dabiq are 92.49% and 92.83%, respectively. On average, terrorist texts have 92.66%, more than five percentage points under news texts values.

The most frequent uncommon words of the news texts (1.68%) are words like "*Poway*", "*Haftar*" or "*Yisroel*". In the case of the two magazines, the most frequent uncommon words (7.34%) are "*Mujahidin*", "*Khilafah*" and "*Kufr*".

5 EXPERIMENTATION

In order to assess the effectiveness of the style features in radicalization detection, we have designed an experimental study that makes use of a machine learning model. In this way, we postulate the problem as a classification task that learns from the provided features. We quantify the performance using style features and compare such results over a baseline, which we define with bigrams. Following, we experiment with the fusion of both types of features, with the aim of complementing a typical textual representation (bi-

grams) with style-specific features. For delimiting the problem, we have cross-validated the number of bigram tokens, considering both the final accuracy of the method, as well as the time needed to train the models. With these constraints, we set the bigram approach to use the 100 most-common tokens.

Regarding the importance analysis of style metrics for radicalization detection, we measured the correlation between each style metric with the annotation of the documents (neutral or radical). The obtained results can be seen in Figure 1.

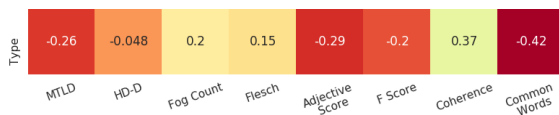| Type	MTLD	HD-D	Fog Count	Flesch	Adjective Score	F Score	Coherence	Common Words
	-0.26	-0.048	0.2	0.15	-0.29	-0.2	0.37	-0.42

Figure 1: Correlation between Style Metrics and Type of texts.

As can be seen, the style metrics with higher correlations to the text annotations are MTLD (Vocabulary Richness), Common Words, F Score, Adjective Score (Formality), and Coherence. It is worth to note that in the field of Readability Index, the computed correlations are not large. Even so, the Fog Count metric stands out over the Flesch metric, for which it has been included as a feature. Furthermore, Common Words can be stressed over the rest of the metrics in terms of the obtained correlation. This may indicate that such a metric represents a useful feature for radicalization detection.

In summary, the style metrics selected as features by the classifier were Fog Count, Adjective Score, Common Words, F Score, Coherence, and MTLD. It should be noted that these results fit well with what was stated in the previous section.

Table 2 shows the results obtained in the experiments. We evaluate the performance of bigrams and style metrics, as well as their combination. As a learning algorithm, we evaluate a popular model: Random Forest. We report the accuracy, recall, precision and f-score, averaged through cross-validation in a 5-fold fashion.

Table 2: Classifier results with Random Forest algorithm.

Random Forest Classifier			
Features	**Style**	**Bigrams**	**Bigrams + Style**
Accuracy	78.3	89.6	92.2
Precision	78.5	90.1	92.8
Recall	78.2	89.6	92.3
F-Score	78.3	89.6	92.3

As seen in the evaluation, Random Forest learner improves its f-score over the bigrams baseline when combines with style features. It is noteworthy that the performance obtained when using uniquely bigrams is higher than using style metrics uniquely. This is to be expected, as the bigrams features are more expressive than style features when the aim is to represent text. Also, we need to consider that the dimensionality of the bigrams representations (100) is much higher than the style features, which amount to a selection of 6.

Considering the improvement with the combination of bigrams and style features, it can be seen that the obtained improvement is 2.7% using the Random Forest, reaching 92.3%. This positive result indicates the effectiveness of the proposed style features when used with a standard text representation approach, obtaining a much more robust classification system.

With all this, we can conclude that the style metrics presented in this paper can result in a benefit for machine learning models applied to the task of radicalization detection.

6 DISCUSSION

This paper presents a feature extraction method for detecting radical content in texts. As part of this, a comprehensive analysis of the presented texts' stylometry is done. Both radical magazines and international news are compared, drawing some interesting differences among them.

In the first place, it has been seen that the captured radical texts magazines are formal, characteristic that is shared with the international news. Even so, considering the two different formality metrics applied, the neutral news achieve higher formality scores than their radical equivalents. Interestingly enough, Dabiq issues tend to have more formal texts than the Rumiyah ones, regardless of the metric used for the study.

Attending to the coherence, both radical and neutral texts have obtained a score that lies in the 60%-70% range, indicating their coherence. Still, some differences can be observed when comparing radical and neutral texts. This is clearer when looking at the correlation with the annotation obtained by this style metric, which amounts to 0.35.

Continuing, both of the used metrics of Vocabulary Richness show that news have more lexical diversity than the terrorist texts analyzed. Again, such differences can explain the obtained correlations in Figure 1. As for the Readability Index, used metrics show that in many times terrorist texts are, in general, more comfortable to read than news texts. Consequently, we can conclude that the readability of the

texts can also offer information over the distinction between radical and non-radical content.

Lastly, the analysis of the use of common and uncommon words has interesting results. Non-radical news have a more significant proportion of common words in their texts in comparison to radical texts. In addition, the analyzed magazines have an equivalent proportion of these words, widely surpassed by news texts.

When performing a comparison between Dabiq and Rumiyah, we have found that the first one is more formal, has a larger lexical diversity, and is easier to read. Alternatively, Rumiyah has a higher coherence score.

7 CONCLUSIONS

The field of radicalization detection is nowadays facing multiple challenges. Regarding the available data, the scarcity of resources is relevant since limited datasets do not allow to develop computational models. In this sense, it is essential to gather more resources that represent reliable and robust sources of data. In light of this, this work includes the extraction of a dataset that is posteriorly used in the evaluation. This data combines radical content, coming from sources like Dabiq, Rumiyah, and Al Jazeera, as well as neutral content that has been extracted from international news outlets, CNN and The New York Times.

Considering the existing radicalization detection models, the relevant information for detecting radical content has not been exhaustively studied. While there are proposals that aim at exploiting different sources of information to detect radicalization (Saif et al., 2017), the application of the stylometry field in radicalization detection has not been thoroughly studied. This paper proposes the use of several stylometry metrics with the aim of characterizing radical and non-radical texts using this kind of information. To the extent of our knowledge, such an approach is novel in the field of radicalization detection.

Attending to the results obtained in the evaluation, it is safe to assume that by themselves, the style features can obtain fairly reasonable performances in this task. When compared to a more detailed representation of a text (e.g., bigrams), as expected, the mentioned style features perform poorly. This is due to their low dimensionality and, consequently, their lower representative power in comparison to an established technique such as bigrams. Nevertheless, as mentioned above, the trend of the field lies in combining several sources of knowledge in order to obtain

better, more robust representations on top of which learning algorithms can obtain better results. Thus, when evaluating the combination of these two types of features, we find that the overall performance is improved. These results indicate that the proposed features can enhance the performance of a classification system, complementing basic textual representations.

We believe that exploiting stylometric cues from the text is indeed an interesting research path. Previous work has also studied the viability of this kind of information source (Sikos et al., 2014).

As future work, we would like to focus on evaluating the style features utility on other domains, like radical content on social networks. To do so, several methods for extracting such metrics need to improve, since many texts in social networks are not as long as in magazines. This may difficult a successful analysis of style. Additionally, we would like to expand the textual representations by incorporating recent Natural Language Processing techniques such as word embeddings (Mikolov et al., 2013; Araque et al., 2017) and language models (Devlin et al., 2018), adapting them to the domain of radicalization detection.

ACKNOWLEDGEMENTS

REFERENCES

Al-Rawi, A. (2018). Video games, terrorism, and isis's jihad 3.0. *Terrorism and Political Violence*, 30(4):740–760.

Al-Sadi, M. R. (2012). Al jazeera television: Rhetoric of deflection. *Arab Media & Society*, 15.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236 – 246.

Ashcroft, M., Fisher, A., Kaati, L., Omer, E., and Prucha, N. (2015). Detecting jihadist messages on twitter. In *2015 European Intelligence and Security Informatics Conference*, pages 161–164. IEEE.

Bisgin, H., Arslan, H., and Korkmaz, Y. (2019). Analyzing the dabiq magazine: The language and the propaganda structure of isis. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 1–11. Springer.

Boyd, R. (2015). Recursive inspection of text scanner [internet].

Cohen, S. J., Kruglanski, A., Gelfand, M. J., Webber, D., and Gunaratna, R. (2018). Al-qaeda's propaganda decoded: A psycholinguistic system for detecting variations in terrorism ideology. *Terrorism and political violence*, 30(1):142–171.

Comerford, M. (2016). What isis lost in dabiq. *New Statesman*.

Correa, D. and Sureka, A. (2013). Solutions to detect and analyze online radicalization: a survey. *arXiv preprint arXiv:1301.4916*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Europol, T. (2016). European union terrorism situation and trend report 2016. *Europol*.

Fang, A. C. and Cao, J. (2009). Adjective density as a text formality characteristic for automatic text classification: A study based on the british national corpus. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, volume 1.

Feakin, T. and Wilkinson, B. (2015). *The Future of Jihad: What Next for ISIL and Al-Qaeda?* Australian Strategic Policy Institute.

Fernandez, M., Asif, M., and Alani, H. (2018). Understanding the roots of radicalisation on twitter. In *Proceedings of the 10th ACM Conference on Web Science*, pages 1–10. ACM.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Heylighen, F. and Dewaele, J.-M. (1999). Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*.

Ingram, H. J. (2017). An analysis of inspire and dabiq: Lessons from aqap and islamic state's propaganda war. *Studies in Conflict & Terrorism*, 40(5):357–375.

Iskandar, A. (2006). Is al jazeera alternative? *The Real (Arab) World: Is Reality TV Democratizing the Middle East?*, 1(2):249.

Johnston, A. H. and Weiss, G. M. (2017). Identifying sunni extremist propaganda with deep learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE.

Kijewski, P., Jaroszewski, P., Urbanowicz, J. A., and Armin, J. (2016). The never-ending game of cyberattack attribution. In *Combatting Cybercrime and Cyberterrorism*, pages 175–192. Springer.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

McCarthy, P. M. and Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

McKernan, B. (2016). Isis' new magazine rumiyah shows the terror group is 'struggling to adjust to losses'. *The Independent*.

Mencarini, M. and Sensidoni, G. (2017). Multilanguage semantic behavioural algorithms to discover terrorist related online. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, volume 1816. CEUR.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Montemurro, M. A. (2001). Beyond the zipf–mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3-4):567–578.

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., and Song, D. (2012). On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314. IEEE.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6):86:1–86:36.

Nouh, M., Nurse, J. R., and Goldsmith, M. (2019). Understanding the radical mind: Identifying signals to detect extremist content on twitter. *arXiv preprint arXiv:1905.08067*.

Saif, H., Dickinson, T., Kastler, L., Fernandez, M., and Alani, H. (2017). A semantic graph-based approach for radicalisation detection on social media. In *European semantic web conference*, pages 571–587. Springer.

Shane, S. and Hubbard, B. (2014). Isis displaying a deft command of varied media. *New York Times*, 30.

Sikos, J., David, P., Habash, N., and Faraj, R. (2014). Authorship analysis of inspire magazine through stylometric and psychological features. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 33–40. IEEE.

Šišková, Z. (2012). Lexical richness in efl students' narratives. *Language Studies Working Papers*, 4:26–36.

Sun, F., Guo, J., Lan, Y., Xu, J., and Cheng, X. (2016). Sparse word embeddings using l1 regularized online learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2915–2921. AAAI Press.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Vergani, M. and Bliuc, A.-M. (2018). The language of new terrorism: differences in psychological dimensions of communication in dabiq and inspire. *Journal of Language and Social Psychology*, 37(5):523–540.

Weimann, G. (2004). *www.terror.net: How modern terrorism uses the Internet*, volume 116. DIANE Publishing.

Wignell, P., Tan, S., O'Halloran, K., and Lange, R. (2017). A mixed methods empirical examination of changes in emphasis and style in the extremist magazines dabiq and rumiyah. *Perspectives on Terrorism*, 11(2):2–20.

Wilner, A. S. (2018). Cybersecurity and its discontents: Artificial intelligence, the internet of things, and digital misinformation. *International Journal*, 73(2):308–316.