information

Article

Shedding Light on the Dark Web: Authorship Attribution in Radical Forums

Leonardo Ranaldi [1,2,*], Federico Ranaldi [2], Francesca Fallucchi [1] and Fabio Massimo Zanzotto [2]

[1] Department of Innovation and Information Engineering, Guglielmo Marconi University, 00193 Rome, Italy
[2] Department of Enterprise Engineering, University of Rome Tor Vergata, 00133 Rome, Italy
[*] Correspondence: l.ranaldi@unimarconi.it

Abstract: Online users tend to hide their real identities by adopting different names on the Internet. On Facebook or LinkedIn, for example, people usually appear with their real names. On other standard websites, such as forums, people often use nicknames to protect their real identities. Aliases are used when users are trying to protect their anonymity. This can be a challenge to law enforcement trying to identify users who often change nicknames. In unmonitored contexts, such as the dark web, users expect strong identity protection. Thus, without censorship, these users may create parallel social networks where they can engage in potentially malicious activities that could pose security threats. In this paper, we propose a solution to the need to recognize people who anonymize themselves behind nicknames—the authorship attribution (AA) task—in the challenging context of the dark web: specifically, an English-language Islamic forum dedicated to discussions of issues related to the Islamic world and Islam, in which members of radical Islamic groups are present. We provide extensive analysis by testing models based on transformers, styles, and syntactic features. Downstream of the experiments, we show how models that analyze syntax and style perform better than pre-trained universal language models.

Keywords: natural language processing; machine learning; deep learning; dark web; jihadist forum; radicalization; extremism

Citation: Ranaldi, L.; Ranaldi, F.; Fallucchi, F.; Zanzotto, F.M. Shedding Light on the Dark Web: Authorship Attribution in Radical Forums . *Information* **2022**, *13*, 435. https://doi.org/10.3390/info13090435

Academic Editor: Kostas Stefanidis

Received: 18 August 2022
Accepted: 13 September 2022
Published: 14 September 2022

1. Introduction

People typically use many social websites on the Internet, happy to reveal their true identities because they want friends and colleagues to be able to easily find them. People who share this content are not worried about privacy [1]. The same happens on forums and on social media, where users want to be recognized when expressing and sharing personal ideas and judgments about any topic. In these cases, they conceal their identity under weak nicknames [2].

On the contrary, many extremist and terrorist groups aim to hide their identity when using the Internet as a vital means to spread their ideology and to exchange and reinforce their beliefs [3]. This activity may pose a security threat because it increases the risk of individuals committing violent acts against society.

Originally, extremist groups used propaganda, and later, they began to spread printed magazines, followed by centralized websites to disseminate their views and information. These media, over time, have been replaced by interactive discussion forums (such as the Ansar Al-Mujahidin Jihadist Forum) and social media platforms such as Twitter and Facebook [4].

Nowadays, these traditional communication channels on the Internet are continuously monitored by national security agencies [7] and international security agencies [5,6]. Therefore there has been a migration of these kinds of activities to dark places on the web [8], where it is challenging to de-anonymize activist identities [9] and consequently possible security threats.

In this research, we aim to mitigate the de-anonymization problem by working on the authorship attribution (AA) task within radical Islamist forums on the dark web and using pre-trained language models (i.e., transformers [10] and KERMIT [11], a syntax-based neural network). AA can be a curse for users seeking to protect their anonymity by changing their name and using various nicknames and, simultaneously, a blessing for law enforcement agencies seeking to track users. AA is a long-studied task for identifying plagiarism in literary works, music, and many other contexts [12–14]. In this paper, we explore the performance of a set of pre-trained language models in authorship attribution tasks on an unseen corpus [15] from the dark net. We apply ML models based on neural networks and syntactic and semantic features built with POS-tagging, KERMIT [11], and transformer-based models [10], respectively, to tackle the challenging task.

Our contribution is twofold. First, we propose the AA task in non-traditional contexts, using as the only source of information the posts written by users of a dark web forum, a dark and radical place. Second, we propose the various transformer models, such as BERT [16], XLNet [17], ERNIE [18], and Electra [19], stylistic classifiers based on the bleaching text model [20], and neural networks based on part of speech tags, with syntactic neural networks based on KERMIT [11]. The code is available at the following GitHub repository (https://github.com/LeonardRanaldi/AuthorshipAttribution (accessed on 14 September 2022)). The analysis revealed that the transformer-based models perform worse than the lexical and syntactic features extracted using KERMIT, although this specific AA dataset is not completely stylistic. Indeed, it is also content-based, as different authors discuss different topics. This shows that AA task can be solved with fair results by lightweight syntactic models can solve the AA task with good results and without requiring special feature-engineering mechanisms, producing flexible models in various domains. The rest of the paper is organized as follows: Section 2 introduces the most well-known ML works used for authorship attribution; Section 3 describes the dataset used in the experiments, which are described in Section 4; and finally, Section 4.7 illustrates a new possible approach for the AA task.

2. Background and Related Work

For centuries, authors have been able to write anonymously with the notion that their true identity would never be uncovered. This has changed in recent years, however, as natural language processing (NLP) methods, heavily influenced by machine learning heuristics, have begun to be used extensively to unearth regularities in the symbols and dynamics that are created among social media users [21].

Task tracking in the virtual world is done in two ways: by tracking users and tasks and analyzing their written expressions [22]. Tasks influenced mainly by writing style are solved with excellent results by learning-based models.

The authorship attribution task (AA) is the task of identifying the author of a given document. This task can be described in two different configurations: (a) binary classification, where we try to guess whether an author produced a text or not; (b) multi-class classification, where each class corresponds to an author, and we try to attribute a text to one of the authors. Initially, the AA task was approached as a binary classification, and in the case of multiple authors, a voting classifier was implemented [23,24]. In this configuration, the basic models were limited to two authors, and the more fluent ones had to support a voting mechanism with multiple authors. This solution was workable but not simple to implement and control. For this reason, the evolution of ML techniques has revisited the problem as a multi-class classification problem [14].

Independent of task configuration, AA largely relies on the process of extracting features related to the content or style of an author [**?**]. The latest technologies propose extensive use of heuristics based on deep learning (DL) methods for AA tasks. DL methods make less and less use of previous feature extraction steps in favor of pre-trained representations of the text that can be used universally, regardless of the task. The following sections describe these methods being applied to the AA task on datasets from literary works, social media, forums, and more hidden places on the web.

2.1. Machine Learning Methods

In AA, the term frequency–inverse document frequency (TF-IDF) is often used to analyze style, measuring the importance of a term concerning a document or a collection of documents. TF-IDF is used at the word level, at the level of N-grams of characters, capturing the words, stems, or combinations of words or letters that an author uses.

It seems that style and content are the keys to solving this task. Stylometry reflects the characteristics and style of the author [14]. The assumption on which this statement is based is that each author has his or her writing style, such as the use of punctuation, average word length, sentence length, and the way a point is developed.

Features that allow a description of text style are often used as logistic regression (LR) inputs [12,13], because these models allow the use of prior knowledge, trying to emulate current transfer learning. Potha and Stamatatos in [26], to address a shared task proposed in an early edition of PAN (https://pan.webis.de/ (accessed on 14 September 2022)), implemented a system-based common N-grams (CNG) and reinforced it with dissimilarity functions. These works had as their cornerstone the study of style. Later, Allison and Guthrie, in [27], proposed the addition of an optional text pre-processing step by implementing stop-words removal and stemming. This additional step greatly influences previous methodologies but introduces syntactic patterns to solve AA tasks.

Using the IMDb62 dataset [28], Sari et al. [14] achieved 95.9% accuracy by proposing an N-gram character classifier and including stylometric features. Although the stylometric features have given decent boosts to performance, the feature engineering is still strongly dependent. Soler-Company and Wanner [29] also showed that including syntactic and discursive features can help achieve state-of-the-art (SOTA) performance in author and genre identification. Bacciu et al. [30] combined sources of input features by solving AA using ensemble learners composed of, for example, several SVM classifiers. Each classifier is trained on distinct concepts related to style, content, and author profile.

2.2. Deep Learning Methods

Although feature engineering is constantly updated, it can be very cumbersome and not always decisive. Deep-learning-based (DL) methods have achieved SOTA results through convolutional neural networks (CNNs). Ruder et al. [31] explored word-level and character-level CNNs for AA and found that character-level CNNs tend to outperform other simple SVM-based approaches.

Zhang et al. [32], by proposing a model based on syntactic trees (rather than surface syntax, as in previous work), proposed a CNN model that outperforms other approaches on blog authorship and IMDb62 datasets.

Like CNNs, recurrent neural networks (RNNs) and derivatives have also been widely used in AA. Long short-term memory (LSTM) and gated recurrent unit (GRU) [33], at both the sentence and item level or using multi-headed recurrent neural networks (RNNs) [34], have also been used. These deep-network-based approaches are as high performing as they are computationally expensive. The high computational costs come from the long learning phases, so having pre-trained models available could be the key to increasing performance and accuracy in knowledge transfer and declining the costs.

In parallel with the wide use of transformers [10] in downstream text classification tasks, these approaches have also become widespread in AA. Barlas and Stamatatos [35] addressed AA by exploiting pre-trained language models (BERT [16], ELMo [36], ULMFiT [37]). The dataset on which these models were used is CMCC [38]. This work shows that transformer-based models, specifically BERT, perform better on large vocabularies and outperform multi-headed RNNs. Fabien et al. [39], in an extensive analysis on IMDb62, Enron corpus [40], and Blog corpus [41], propose a modification of BERT by adding stylistic features, which obtains better results than BERT alone. Manolache et al. [42] showed the potential of BERT knowledge transfer by fine-tuning the version of the dataset provided in the PAN-2020 shared task and adapting it to the DarkNet dataset extracted from the social Reddit. While closer to the dark web regarding topics, this work also proposes data sourced from surface web sources. There remains a need to find the truth even from submerged sources like those on the dark web.

2.3. Tor and the Dark Web

The dark web is the web of hidden services. The Tor browser [43] is the most popular tool for accessing the hidden part of the web and achieving anonymity on the Internet. Tor users can surf the web with many guarantees of anonymity. However, in addition to users, services can also be hidden, so websites are often hidden behind the Tor system. In this case, the site is a so-called hidden service. Neither the user nor the server knows each other's IP address in the dark web since the connection is made at a meeting point that connects the two entities using Tor. In addition, the dark web hosts several forums on the most diverse topics. Some are illegal, such as drug markets, hacking activities, or child pornography. In other cases, extremists meet in forums to share information that may be a security threat.

AA is a difficult task in a hostile environment, such as forums, and even more so if they are uncontrolled and some different authors and texts differ in size. Many researchers have already focused on AA in unexplored places like the dark web. Ho and Ng [22] analyzed the stylometric features of texts published in several dark web forums. They tried to link ten authors within dark web forums by extracting stylometric features and particular fingerprints, such as familiar words or typos of an author, using support vector machine (SVM) as a classifier. Spitters et al. [44], proposing a dataset extracted from the marketplace, investigated sellers as authors of item descriptions, and added character-level n-grams to stylometric features, using SVM as a classifier. Swain et al. [45] give an overview of recent ML-features-based approaches to AA techniques, describing several papers that focus on this area, specifically the category, language, domain, features, and techniques that have been used. This survey shows that naive Bayes and SVM are the most commonly used classifiers, English is the most analyzed language, and lexical and syntactic features are the most popular. The work mentioned is strongly influenced by applying pre-processing based on accurate feature engineering methodologies. Feature engineering poses static and unadaptable models in different domains.

Recently, Ranaldi et al. [46] attempted to unravel the adaptive capabilities of transformers in tasks from the dark web, and on tasks derived from the surface web, revealing that they fail to contextualize text. The authors argue that syntactic and lexical neural networks outperform pre-trained transformers. The task they address is not simple as it concerns the classification of legal and illegal activities. These activities may not be easily generalized by pre-trained language models.

In this paper, we propose a study of AA techniques on dark web forums. Following the analysis proposed in the literature, [8], a subset of the "Islamic Network forum" dataset is used as learning data in order to build a model capable of tracing the most radical users that may represent a security threat. After exploring the dataset and showing that the classes are divisible by topic, as each author seems to be talking about a topic, we show that pre-trained language models do not always perform well when dealing with semantic tasks. Instead, lightweight syntactic stylistic models can solve these tasks with promising results.

3. Dataset

This research was based on data from the Islamic Network's [15] forum. The Islamic Network forum aims to unite dedicated individuals to practice their Muslim faith. The discussion forum, which is in the English language, includes several topics of interest to Muslims, ranging from theology to current world events [47]. The original dataset was constructed by extracting forum posts from 2004 through the year 2010. The dataset consists of 91,874 posts with 13,995 discussions and 2082 active users.

The most active users and those with the highest radical scores [47], who are thus likely to represent potential threats, were extracted, taking into account the periods of highest interest [8], topics of interest [48], and sentiment analysis [49,50].

Therefore, following the directions in [8,47], the most radical and followed forum members were selected. After selecting the most radical and followed forum members, a second sorting was done. The second sorting was done on members dealing with topics related to terrorist activities, religious propaganda activities, and violent activities, as suggested in [48]. This second sorting produced the final dataset, which allowed us to focus our work on the ten members representing potential threats and with whom it was easier to work and investigate.

There are many topics covered in this dataset. In the examples in Table 1, some situations characteristic of the Islamic world have been shown. For example, user "Member A" is active on the topic of Jihad. In contrast, "Member B" seems to condemn the behavior of infidels in the Arabic lexicon *"kaafir"*. "Member C" is more idealistic and appreciates the efforts of good men in a community *"ummah"*, extolling marriage as Allah's *"ni'aam"* (favor, support, kindness). Although there do not appear to be explicit incitements to hatred and violence, these are still hot topics that can be channels for veiled and potentially threatening messages.

Table 1. Sample of posts by some active members of the Islamic Network forum anonymized with the name "Member x".

Message	Member
So she wants to be widowed after few years? Jihad is a part of the deen so when she asks a husband to fulfil part of the deen then there's nothing wrong with this.	Member A
It includes many things. Also going for Jihad doesn't mean getting killed. if that's what she means he has to be killed then this is forbidden since she's asking something unknown.	Member A
If the companions and tabi'een had a consensus as is claimed, anyone who would go against the consensus would be a kaafir	Member B
It is just to many things for me personally to justify sistsers attending mixed universites of the kuffar. Looking up to and learning from a kaafir, should be done by strong capable muslim men who won't begin to learn, love respect the kuffar. (you should not let the weaker women/children be in this setting especially YOUNG ones who have never been exposed to such things).	Member B
Salam, It is encouraging to see that there are women whom do appreciate the efforts of the good men of this ummah. Marriage is definitely from the tremendous ni'aam(favors, support, kindness) of Allah subhanahu wa ta'ala.	Member C
May Allah open the doors of al-jannah for the women whom strive to please their husbands and the men who do the things to make that an easy task, ameen.	Member C

3.1. Topic Analysis

The studies [8,47,48] that were conducted on the final dataset construction affirm intense radicalization and thus attachment to specific topics by some users. Therefore, in order to better analyze the specific topics, a topic model was produced.

Based on the analyses in Figure 1, some users seem to have strong attitudes toward specific topics. The optimal number of topics based on latent Dirichlet allocation (LDA) [51] is ten, just like the authors considered (see Figure 1). In addition to LDA, Jaccard similarity and consistency were calculated for each topic. Consistency, in this case, measures a single topic by the degree of semantic similarity between high-scoring words in the topic (these words co-occur in the text corpus). The ideal number of topics that will maximize coherence and minimize topic overlap based on Jaccard similarity is ten.

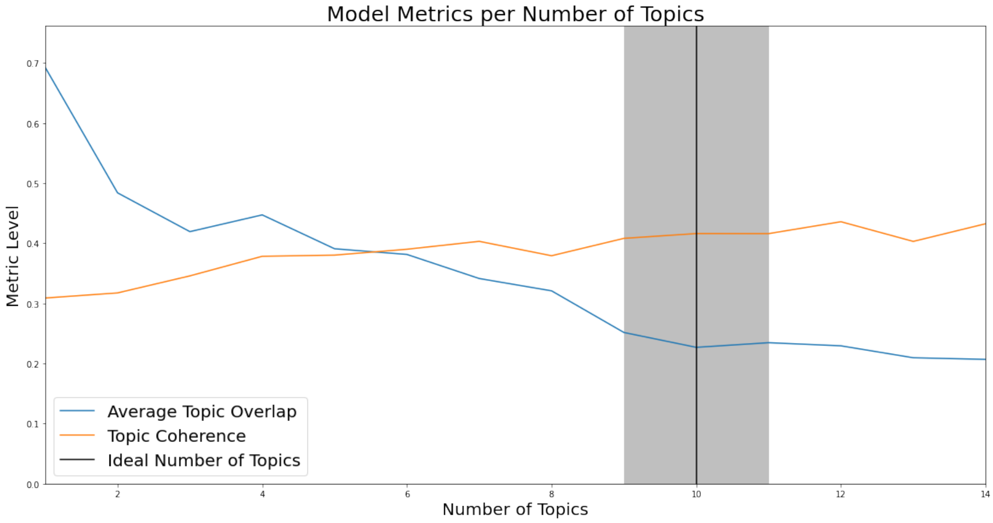


Figure 1. Latent Dirichlet allocation (LDA) [51] across different numbers of topics, Jaccard similarity, and coherence for each.

Continuing the LDA analysis, the topics were identified. We observed that each user has specific traits, as seen in Figure 2, where the distances between topic clusters and segregation are shown, respectively.

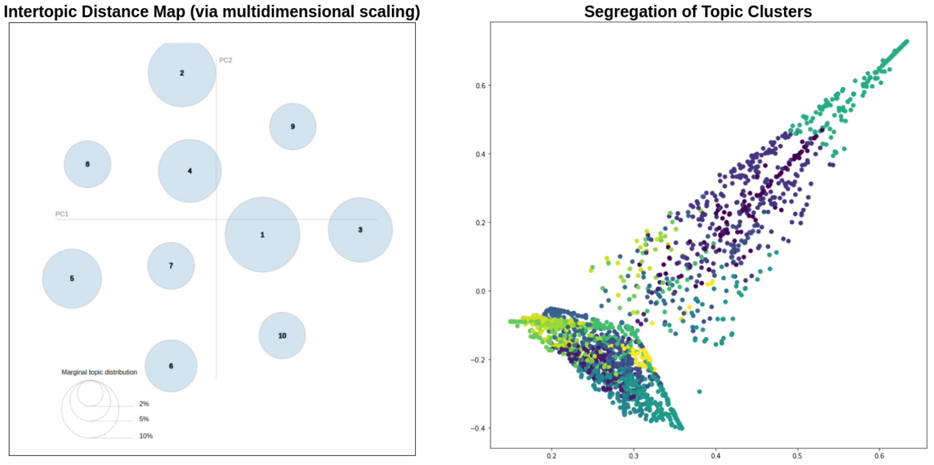


Figure 2. On the left, the intertopic distance map; each bubble represents a single topic, and the bubble size represents prevalence. On the right, the segregation of topics into clusters; each point color represents the topic number.

To show that each member talks about a particular topic, and on the other hand, that each topic is dealt with by a particular user, we continue the analysis. Hence, for each member we analyzed the posts. After numerous tests, k-means clustering on the document-topic probability matrix was used by fixing the number of clusters to ten. Then, singular value decomposition (SVD) [52] was used to visualize the clusters on two dimensions. Figure 2 shows the intertopic distance map obtained by multidimensional scaling on the left and right, and the segregation of topics into clusters. The intertopic distance map provides an overview through multidimensional scaling of the ten topics present and the relevance of each. The intertopic distance map does not provide a view of the position of the posts relative to the topics or clusters. However, through the post-segregation in the ten topics, we can observe their distribution and ranking.

3.2. Dataset Building

Downstream of the extraction work, three datasets were created from the original Islamic Network dataset. In the first dataset, the first three users with more posts and higher radical scores were sampled. The second dataset sampled the first five users with the same criteria, and the third dataset sampled the first ten users following the same criteria. To make the datasets balanced and not create situations with users with few posts, all examples for each user in all three datasets were balanced. Each dataset contains 1400 posts per user, and the average frequency of the number of tokens was studied (see Figure 3). The constructed datasets are accessible and shared on the GitHub repository (https://github.com/LeonardRanaldi/AuthorshipAttribution (accessed on 12 September 2022)).

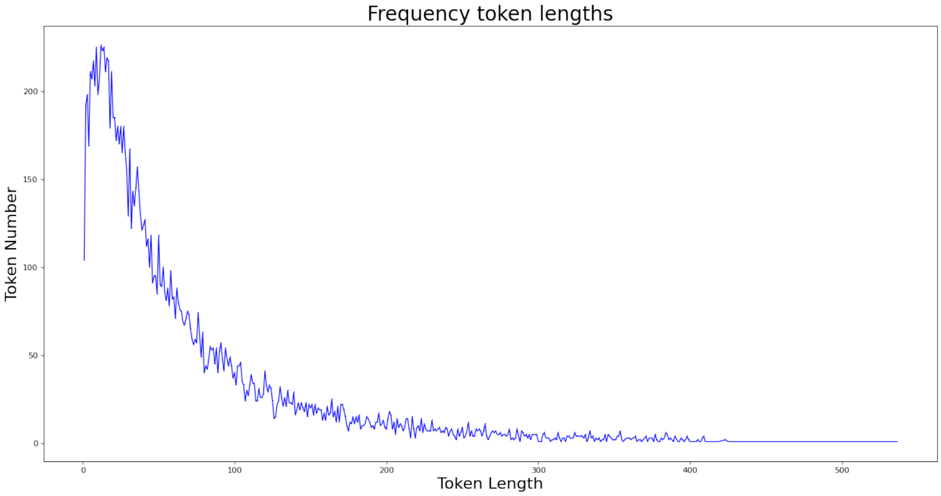


Figure 3. Frequency of the number of sample tokens with ten authors from the Islamic Network Forums dataset.

4. Methods

Authorship attribution (AA) aims to develop a classifier that predicts the authors of texts—in our case, the user who made a post. In this work, as introduced in the Section 2, we used the multiclass version of the task, where each text will be assigned to the most likely user. The code for the models presented in the following sections is open source and available on the GitHub repository.

Machine learning models that solve this task while achieving high accuracy values, as seen in Section 2, powerfully leverage syntactic information[32]. Stylistic [14] features are very successful, both in datasets where authors come from literary works and in contexts where authors are users of social media [1] or forums [22]. Downstream of this long series of works, it seems that the modern models based on transformers, fine-tuned on context, perform better than the old models [35].

Despite previous successes in solving the AA problem discussed in an earlier section, we argue that the syntactic information in an author's writing can partially characterize the author's "writing style". Specifically, even when writing the same content, two authors may prefer to use a different syntactic structure in constructing their sentences.

In order to study the role of syntax in the AA task, we propose models based on syntactic features, one based on lexical features and one based on purely stylistic features, and a set of pre-trained language models based on transformers. Figure 4 summarizes the basic steps, which are described in detail in the following sections.

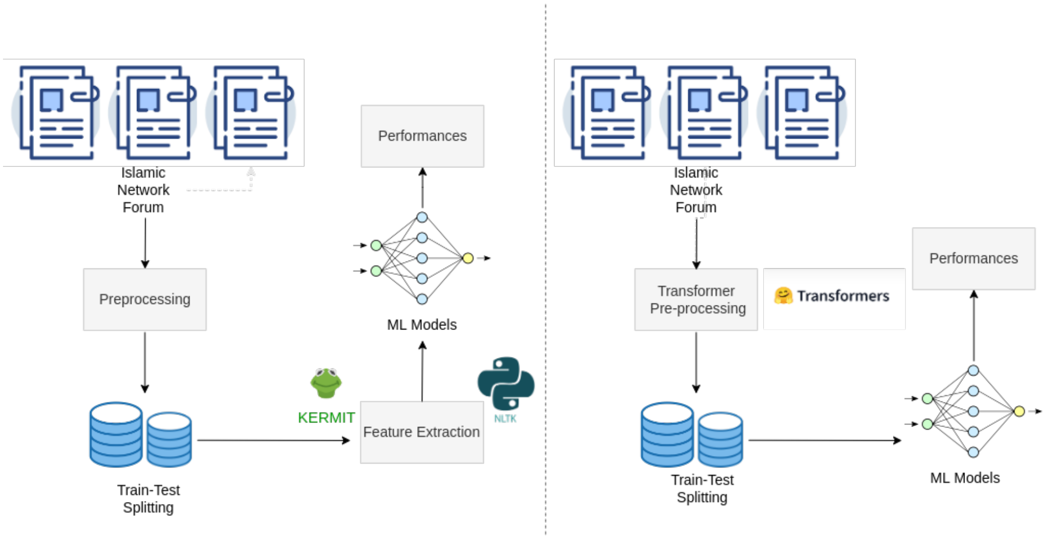


Figure 4. Diagram of the proposed model development pipeline. (**a**) Style- and lexical-based features and KERMIT. (**b**) Transformer models.

4.1. Style- and Lexical-Based Features

Term frequency, also known as bag of words (BoW) and term frequency-inverse document frequency (TF-IDF), is referred to as lexical-based features in this research. They are often used in the context of AA [12–14]. The main focus of these features is not on the topic, but rather on the frequency of words within a text. However, the TF-IDF approach attempts to overcome a problem typical of the BoW approach: rarely used words, which may be the most interesting, are obscured by more frequently used words.

BoW and TF-IDF were used downstream of two different pre-processing steps of the input text, focusing in both cases on unigrams (i.e., groups of words of length one).

BoW and TF-IDF took as input two different pre-processed versions of the text. The first version is tokenization, which is done using the process provided by the Natural Language Toolkit (NLTK) [53] libraries. The second version was done by extracting part-of-speech (POS) tags. Then, separately, BoW and TF-IDF were applied in the following way:

- Pre-processing was done using NLTK features. The text was tokenized, stopwords were removed, and surface cleaning was done. Then, BoW and TF-IDF were applied to the pre-processed input, respectively. Therefore, the models based on these two inputs were called BoW and TF-IDF plainness.
- Pre-processing was done using POS tags, which are unique labels assigned to each token (word) to indicate grammatical categories and other information, such as time and number (plural/singular) of words. POS tags were extracted using the method proposed by Toutanova et al. [54]. In order to differentiate between the input consisting of POS and the input not constructed by POS, BoW and TF-IDF that took pre-processing with POS were called bag of part-of-speech tags (BoPOS) and TF-IDF$_{POS}$.

BoW, BoPOS, TD-IDF, and TF-IDF$_{POS}$ representations were given as input to a feed-forward neural network (FFNN) consisting of: an input layer of the same size as the representations; two hidden layers of size 124 and 64, respectively; and finally, the output layer of a size equal to the number of classes.

Between each layer, the *ReLU* activation function and a dropout of 0.1 are used to avoid overfitting the training data. The full definition and development process can be noted on the left side of Figure 4.

Finally, bleaching text [20] is a model that captures the style at the word level. This model is designed to predict authors' gender and is multipurpose for dealing with the AA task. The style is captured in the following way: the model converts sequences of tokens into abstract sequences according to a list of pre-defined rules appropriately defined by the authors. Each sentence is represented by concatenating all the previous transformations.

We use a linear support vector machine (SVM) classifier [55] with binary baggage of the word representation for classification.

4.2. KERMIT

To investigate the role of syntax, the kernel-inspired encoder with recursive mechanism for interpretable trees (KERMIT) [11] was used. This model positively exploits parse trees in neural networks as it increases the performance of pre-trained transformers when used in combined models. The version used in the experiments encodes the parse trees into vectors of 4000 dimensions. KERMIT takes advantage of the parse trees produced by the CoreNLP parser [56]. The syntactic embeddings produced by KERMIT were given as input to a feed-forward neural network (FFNN) composed of two hidden layers of size 4000 and 2000, respectively, and finally, to an output layer of a size equal to the number of classes. Between each layer, the ReLU activation function and a dropout of 0.1 are used to avoid overfitting the training data.

4.3. Transformer Models

Transformer-based architectures achieve state-of-the-art (SOTA) results in many downstream text classification tasks. For example, even in the AA task, a transformer-based model [10], more precisely BERT with the addition of dense layers and a softmax activation, achieves SOTA results in well-known datasets in the AA domain [39]. In this work, six transformers-based encoders were tested. The proposed models differ in corpus and parameters set in the pre-training phase:

- $BERT_{base}$ [16], the bidirectional encoder representations from transformers architecture that is trained on the BooksCorpus [57] and English Wikipedia.
- Multi-lingual $BERT_{multi}$ [58], which is similar to BERT but is trained on a Wikipedia dump of 100 languages.
- XLNet [17], which unlike the previous is based on a generalized autoregressive pre-training technique that allows the learning of bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order. XLNet is trained on Wikipedia, BookCorpus [57], Giga5 [59], Clueweb [60], and Common Crawl [61].
- ERNIE [18] introduced a language model representation that addresses the inadequacy of BERT and utilizes an external knowledge graph for named entities. ERNIE is pre-trained on the Wikipedia corpus and Wikidata knowledge base.
- ELECTRA [19] proposes a mechanism for "corrupting" the input token by replacing it with a token that potentially fits the place. The training is done by trying to guess whether each token is a corrupted input or not. To make its performance comparable to BERT, they trained the model on the same dataset that BERT was trained on.
- DistilBERT [62] is a pre-training method for a smaller, more general-purpose language representation model than BERT. In DistilBERT, a distillation of knowledge is proposed during the pre-training phase by reducing the size of BERT by 40%, while retaining 97% of its language comprehension capabilities and being 60% faster.

Pre-processing and tokenization and related encoders of the proposed models were performed using codes provided by Huggingface's transformers library [63].

To use transformers as classifiers, we used a fully connected classification layer on top of the $[CLS]$ token. For the training phase, the number of epochs was set to 5. Moreover, we used the Adam optimizer [64] and the cross-entropy loss function, a learning rate of 2×10^{-5}, a linear schedule with no warmup step, a batch size of 32, and gradient norm clipping of 1.0. We also limited the maximum sequence length to 512 tokens, as proposed by Fabien et al. [39]. The number of tokens set covers all examples of the datasets used, as shown in Figure 3.

Finally, the last transformer-based model consists of two versions of BERT specifically fine-tuned for sequence classification task [65], called *BERTForSequenceClassification* (Bert4SC). Pre-trained versions of $BERT_{base}$ and $BERT_{multi}$ were used in $Bert4SC_{base}$, and compared with $Bert4SC_{multi}$. $Bert4SC_{base}$ and $Bert4SC_{multi}$ were fine-tuned on the training dataset.

The full definition and development process is shown in Figure 4b.

4.4. Experimental Setup

This section describes the general set-up of our experiments and the specific configurations adopted. The models described in Section 4 were tested on the dataset described in Section 3. Each experiment was repeated five times, initializing the models with NNs with five different seeds to ensure that the models based on NNs do not produce anomalous results. Three datasets were split into training and test sets (7:3 ratio) after stratified random shuffling with the same random seed for all experiments. For the evaluation of systems, the average accuracy was considered. The average number of correct predictions over the total number of predictions of the results was obtained in each of the five runs with five different seeds. When choosing the metrics, we chose accuracy because in all the configurations used, the data sets are balanced, so this metric is the most explanatory. In addition, to interpret the best models, we show the confusion metrics for each dataset (see Figure A1).

4.5. Results

We explored the performance of the models described in Section 4 on the dataset described in Section 3. The results reported in Table 2 show that natural language processing (NLP) algorithms based on lexical features combined with feed-forward neural network (FFNN) perform better in terms of accuracy.

Table 2. Comparison of the different performances of the proposed models on the authorship attribution task on the Islamic Network dataset.

Models	Methods	3 Authors	5 Authors	10 Authors
Style- and lexical-based	BoW	84.1 (± 1.5)	73.2 (± 1.2)	58.3 (± 0.8)
	BoPOS	**84.9 (± 0.85)**	**74.5 (± 1.4)**	57.6 (± 0.9)
	TF-IDF	83.7 (± 0.6)	74.1 (± 1.7)	57.8 (± 1.2)
	TF-IDF$_{POS}$	74.6 (± 0.55)	72.5 (± 0.86)	**58.9 (± 1.21)**
	Bleaching text	70.3 (± 0.5)	70.8 (± 0.5)	56.9 (± 0.5)
Syntax-based	KERMIT	83.95 (± 1.84)	70.94 (± 1.35)	56.34 (± 1.52)
Transformer-based	$BERT_{base}$	62.1 (± 1.45)	47.6 (± 1.62)	32.7 (± 0.94)
	$BERT_{multi}$	40.2 (± 1.14)	29.1(± 0.92)	23.3 (± 1.23)
	XLNET	45.7 (± 1.32)	31.7 (± 2.23)	19.9 (± 1.53)
	ELECTRA	61.4 (± 1.15)	45.7 (± 1.22)	32.1 (± 0.85)
	ERNIE	62.1 (± 1.37)	42.4 (± 1.43)	29.4 (± 0.93)
	DistilBERT	59.8 (± 1.55)	46.9 (± 1.62)	32.5 (± 1.23)
BERT$_{fine-tuning}$	$\boldsymbol{Bert4SC_{base}}$	*80.7* (± 1.33)	*69.8* (± 1.27)	*58.2* (± 1.24)
	$\boldsymbol{Bert4SC_{multi}}$	*76.2* (± 2.12)	*63.3* (± 1.23)	*50.8* (± 1.46)

The results show that in two of the three configurations, the BoPOS models obtained the best results in terms of accuracy. We can deduce from these results that style is a relevant feature in the authorship attribution (AA) task, as was already mentioned in Section 2. Although style seems very important, purely stylistic models, such as bleaching text, while they achieve good results, do not stand out among the excellent ones.

The deep syntax proposed by KERMIT in two out of three subtasks is in line with the best results. Hence, by comparing the deep syntax of KERMIT and the shallow syntax of a generic BoPOS, we can see that the results do not differ radically. However, a less heavyweight representation seems to be better in this case.

Regarding a weakness of the transformers, it seems that they fail to generalize into definitely unseen domains. They also do not seem adaptable to the task when a heavy fine-tuning step is not performed, as in the version proposed by Fabien et al. [39] and

Manolache et al. [42]. Although transformer-based models did not originate for AA-related tasks, as we saw in Figures 1 and 2, the topics are very distinct among authors with semantically different concepts.

Finally, we tested two fine-tuned versions of BERT: $Bert4SC_{base}$ and $Bert4SC_{multi}$. We preferred BERT because it is the most widely known of the transformers and because it has already been studied in the AA task. After an appropriate fine-tuning phase, these two models properly constructed to solve downstream classification tasks appear to achieve significantly better performance than the basic configurations of transformer-based models. Although fine-tuning seems to have given an excellent boost to the models, they still have not reached the performances achieved by BoPOS and TF-IDF on text POS.

4.6. Discussion

Transformer-based models are greatly influencing the NLP world. It seems that transformers are adept at solving semantic [66], syntactic [67], and even stylistic tasks [68]. Transformer-based models, in general, are adapted to the specific task with an appropriate fine-tuning phase [69]. These steps seem to work very well, and transformers solve downstream tasks optimally even when the tasks are purely stylistic [39,42]. The key to success would be pre-training on massive corpora with conspicuous resources, as defined in Section 4.3. The resources cited for each model are all texts taken from the surface web. This statement does not prove that pre-trained language models, such as BERT, XLNet, and other transformer-based models pre-trained on surface web data, cannot generalize and apply their knowledge in definitely unseen domains. What can be said is that there is a different language between the two worlds [70]. This language cannot always be captured by pre-trained language models, which seem to be affected by unknown domains [46,71].

Therefore, we can say that one of the possible explanations concerns the origin of the sources. The dataset of Islamic Network forums originates in a definitely unseen domain for transformers, the dark web. Ranaldi et al. [46] studied the behavior of a set of pre-trained natural language comprehension models on genuinely new and unexplored data provided by classification tasks on a dark web corpus, arguing that syntactic and lexical neural networks outperform pre-trained transformers.

4.7. Future Works

In this paper, language models based on transformers and stylistic and syntactic models were proposed for the authorship attribution (AA) task on the Islamic Network forum dataset. Downstream of the results in Table 2, in the discussions presented in Section 4.5, we hypothesized why transformers used in a configuration with a slight fine-tuning step do not achieve as supportable performance as lexicon-based models. Therefore, we continued investigating to mitigate the possible problems of a slight fine-tuning phase. In this regard, we proposed two models based on two specially fine-tuned versions of BERT. At the same time, the results, although good, did not exceed the best results obtained.

In order to give a real answer to the performance of transformer-based models in future developments, we would like to understand what kind of language is present in text input from the dark web and how much of this language can be linked to learning sources as proposed by Carlini et al. [72].

Another major challenge involves stylistic investigation of the language used by more radical users, who may pose potential threats. The stylistic investigation could be done by looking into the sequences in which the symbols appear; more precisely, it could be done on the syntactic structures. Downstream of a syntactic analysis that could be developed with the KERMIT [11] framework, one could project the syntactic visualizations produced by the framework with the specialized visualizer KERMITviz [73].

The last (but not least) challenge lies in continuing the work by proposing an approach of linking dark web forum users to standard web forum users [9], enriching these works with explainable visualizations of the links. AA is a well-known field of study, but it is still an open problem, with a few hundreds to thousands of candidate authors in two parallel worlds. The ability to link the two worlds could have positive implications and aid threat monitoring by intelligence teams. In addition, the use of appropriately constructed visualizers, such as KERMITviz, could facilitate the analysis of hidden patterns and recurring structures, which would enable the solution of the AA task across heterogeneous contexts.

4.8. Limitations

In this paper, we presented an approach to the authorship attribution task in a definitely unseen domain. Ours is one of the first studies on pre-trained language models in the dark web domain. Like all work, our study suffers from some limitations. The main limitation is that we tested only a few tasks based on the absolute necessity of determining the intent of forum members since many quoted and paraphrased portions of sacred texts such as the Quran. Other members, on the other hand, preach violence and discuss many topics that may represent threats. In order to delineate all these nuances, we investigated thoroughly whether this could turn out to be a merely stylistic task, and the result is that it is not. After demonstrating that the authors we considered deal with different topics, we tried applying transformers, which perform very well on downstream tasks. However, these did not produce the desired results either. However, our results might be better if there was a model pre-trained on the dark web. By pre-trained, we mean a model trained with the classical masked language model and next sentence prediction tasks, as in [16]. Unfortunately, at this time, no giant corpus is yet available to try to pre-train a domain-specific version of BERT, ELECTRA, or ERNIE.

5. Conclusions

The increasing use of websites and social media to disseminate terrorist propaganda and communicate with like-minded individuals is both a challenge and an asset for intelligence analysts and others involved in counter-terrorism. There is a risk that this allows people with no apparent connections to terrorist groups to be radicalized. However, digital communication also leaves traces that analysts can potentially use to identify people who are about to commit actions related to violent extremism. Possible threats often hide behind anonymous masks and constantly changing nicknames. Identifying a disseminator of potential threats could be the key to many national-security-related activities.

In this paper, we propose an authorship attribution task in a sensitive reference context, namely the Islamic Network dataset. Our contribution is twofold. First, we analyze the Islamic Network dataset, from which we extract the most radical users who might pose a security threat. We define an authorship attribution task on an unseen domain (the dark web) and sensitive topics on the extracted dataset. Second, we study the performance of pre-trained language models on entirely different domains, including deep and shallow syntactic and stylistic models, without going through a long-winded feature-engineering phase. The fundamental reason for this avenue of our research is to build suitable models in all domains, and therefore not with domain-dependent features, which may incur overfitting. Behind a training phase with a feed-forward neural network, we can say that models based on style, lexicon, and syntax achieve sustainable performance. However, the transformer-based models, despite the fine-tuning, do not seem to perform well in the author attribution task in definitely unseen domains.

Author Contributions: Conceptualization, L.R. and F.M.Z.; methodology, L.R.; software, F.R.; validation, L.R., F.F. and F.M.Z.; formal analysis, L.R.; investigation & resources, L.R. and F.R.; data curation, L.R. and F.R.; writing—original draft preparation, F.R.; writing—review and editing, L.R.; supervision, F.M.Z.; project administration, L.R.; funding acquisition, F.M.Z. All authors have read and agreed to the published version of the manuscript.

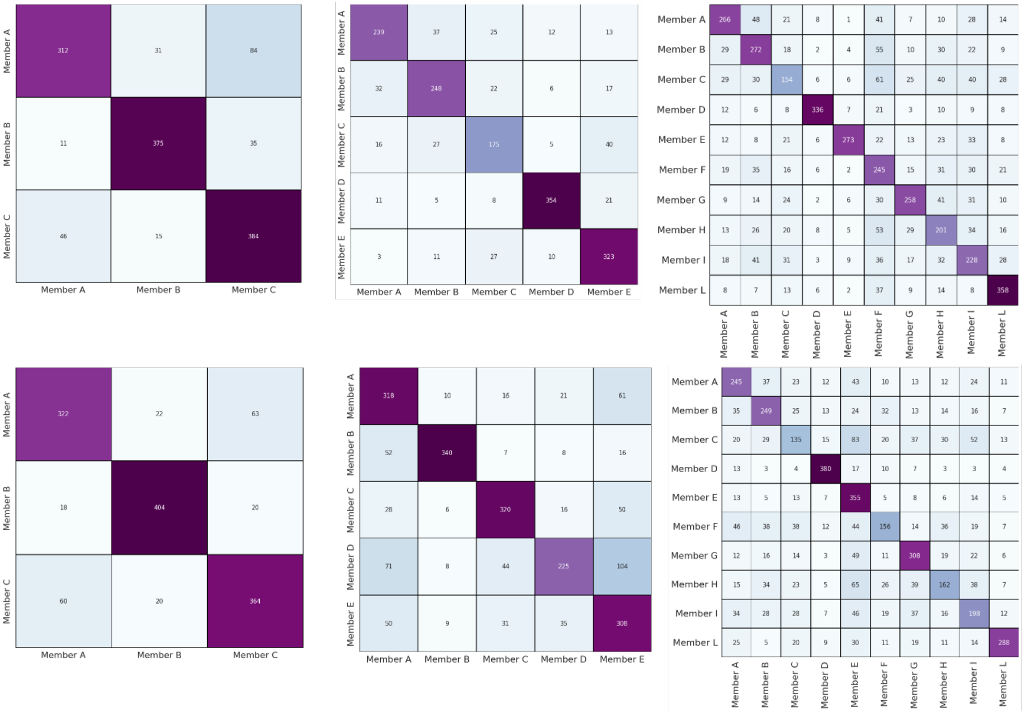
Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository.

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

Appendix A



	Member A	Member B	Member C
Member A	312	31	84
Member B	11	375	35
Member C	46	15	384

	Member A	Member B	Member C	Member D	Member E
Member A	239	37	25	12	13
Member B	32	248	22	6	17
Member C	16	27	175	5	40
Member D	11	5	8	354	21
Member E	3	11	27	10	323

	Member A	Member B	Member C	Member D	Member E	Member F	Member G	Member H	Member I	Member L
Member A	266	48	21	8	1	41	7	10	28	14
Member B	29	272	18	2	4	55	10	30	22	9
Member C	29	30	154	6	6	61	25	40	40	28
Member D	12	6	8	336	7	21	3	10	9	8
Member E	12	8	21	6	273	22	13	23	33	8
Member F	19	35	16	6	2	245	15	31	30	21
Member G	9	14	24	2	6	30	258	41	31	10
Member H	13	26	20	8	5	53	29	201	34	16
Member I	18	41	31	3	9	36	17	32	228	28
Member L	8	7	13	6	2	37	9	14	8	358

	Member A	Member B	Member C
Member A	322	22	63
Member B	18	404	20
Member C	60	20	364

	Member A	Member B	Member C	Member D	Member E
Member A	318	10	16	21	61
Member B	52	340	7	8	16
Member C	28	6	320	16	50
Member D	71	8	44	225	104
Member E	50	9	31	35	308

	Member A	Member B	Member C	Member D	Member E	Member F	Member G	Member H	Member I	Member L
Member A	245	37	23	12	43	10	13	12	24	11
Member B	35	249	25	13	24	32	13	14	16	7
Member C	20	29	135	15	83	20	37	30	52	13
Member D	13	3	4	380	17	10	7	3	3	4
Member E	13	5	13	7	355	5	8	6	14	5
Member F	46	38	38	12	44	156	14	36	19	7
Member G	12	16	14	3	49	11	308	19	22	6
Member H	15	34	23	5	65	26	39	162	38	7
Member I	34	28	28	7	46	19	37	16	198	12
Member L	25	5	20	9	30	11	19	11	14	288

Figure A1. Confusion matrices of the top two models for each subtask. At the top are the confusion matrices for TFIDF+POS-tag-based models, at the bottom are the confusion matrices for BOW+POS-tag-based models.The authors were anonymized with the name "Member x".

References

1. Pillay, S.R.; Solorio, T. Authorship attribution of web forum posts. In Proceedings of the 2010 eCrime Researchers Summit, Dallas, TX, USA, 18–20 October 2010; pp. 1–7. [CrossRef]
2. Ranaldi, L.; Zanzotto, F.M. Hiding Your Face Is Not Enough: User identity linkage with image recognition. *Soc. Netw. Anal. Min.* **2020**, *10*, 56. [CrossRef]
3. Wagemakers, J. The Concept of Bay'a in the Islamic State's Ideology. *Perspect. Terror.* **2015**, *9*, 98–106.
4. Johansson, F.; Kaati, L.; Shrestha, A. Detecting Multiple Aliases in Social Media. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, ON, Canada, 25–29 August 2013; Association for Computing Machinery: New York, NY, USA, 2013; ASONAM '13, pp. 1004–1011. [CrossRef]
5. Reilly, B.C. Doing More with More: The Efficacy of Big Data in the Intelligence Community. *Am. Intell. J.* **2015**, *32*, 18–24.
6. Operation Onymous | Europol—Europol.europa.eu. Available online: https://www.europol.europa.eu/operations-services-and-innovation/operations/operation-onymous (accessed on 14 May 2022).
7. Relazione al Parlamento 2021-Sistema di Informazione per la Sicurezza della Repubblica—Sicurezzanazionale.gov.it. Available online: https://www.sicurezzanazionale.gov.it/sisr.nsf/relazione-annuale/relazione-al-parlamento-2021.html (accessed on 14 May 2022).
8. Park, A.J.; Beck, B.; Fletche, D.; Lam, P.; Tsang, H.H. Temporal analysis of radical dark web forum users. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 880–883. [CrossRef]
9. Arabnezhad, E.; La Morgia, M.; Mei, A.; Nemmi, E.N.; Stefa, J. A Light in the Dark Web: Linking Dark Web Aliases to Real Internet Identities. In Proceedings of the IEEE 40th International Conference on Distributed Computing Systems (ICDCS), Singapore, 29 November–1 December 2020; pp. 311–321. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
11. Zanzotto, F.M.; Santilli, A.; Ranaldi, L.; Onorati, D.; Tommasino, P.; Fallucchi, F. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 256–267. [CrossRef]
12. Madigan, D.; Genkin, A.; Lewis, D.; Fradkin, D. Bayesian Multinomial Logistic Regression for Author Identification. *Aip Conf. Proc.* **2005**, *803*, 509 .

13. Aborisade, O.; Anwar, M. Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers. In Proceedings 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 7–9 July 2018; pp. 269–276. [CrossRef]
14. Sari, Y.; Stevenson, M.; Vlachos, A. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 343–353.
15. AZSecure-data.org—azsecure-data.org. Available online: https://www.azsecure-data.org (accessed on 14 May 2022).
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1 pp. 4171–4186. [CrossRef]
17. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.G.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.
18. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *arXiv* **2021**, arxiv:2107.02137.
19. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the ICLR, Addis Ababa, Ethiopia, 30 April 2020.
20. van der Goot, R.; Ljubešić, N.; Matroos, I.; Nissim, M.; Plank, B. Bleaching Text: Abstract Features for Cross-lingual Gender Prediction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; Volume 2: Short Papers, pp. 383–389. [CrossRef]
21. Anwar, T.; Abulaish, M. A Social Graph Based Text Mining Framework for Chat Log Investigation. *Digit. Investig.* **2014**, *11*, 349–362. [CrossRef]
22. Ho, T.N.; Ng, W.K. Application of Stylometry to DarkWeb Forum User Identification. In Proceedings of the ICICS, Singapore, 29 November–2 December 2016.
23. Grieve, J. Quantitative Authorship Attribution: An Evaluation of Techniques. *Lit. Linguist. Comput.* **2007**, *22*, 251–270. [CrossRef]
24. Petrovic, S.; Petrovic, I.; Palesi, I.; Calise, A. Weighted Voting and Meta-Learning for Combining Authorship Attribution Methods. In Proceedings of the 19th International Conference, Madrid, Spain, 21–23 November 2018; pp. 328–335. [CrossRef]
25. tamatatos, E. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 538–556. [CrossRef]
26. Potha, N.; Stamatatos, E. A Profile-Based Method for Authorship Verification. In *Artificial Intelligence: Methods and Applications*; Likas, A., Blekas, K., Kalles, D., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 313–326.
27. Allison, B.; Guthrie, L. Authorship Attribution of E-Mail: Comparing Classifiers over a New Corpus for Evaluation. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008), Marrakesh, Morocco, 28–30 May 2008; European Language Resources Association (ELRA): Marrakech, Morocco, 2008.
28. Seroussi, Y.; Zukerman, I.; Bohnert, F. Authorship Attribution with Topic Models. *Comput. Linguist.* **2014**, *40*, 269–310._a_00173. [CrossRef]
29. Soler-Company, J.; Wanner, L. On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Association for Computational Linguistics: Valencia, Spain, 2017; Volume 2, Short Papers; pp. 681–687.
30. Bacciu, A.; Morgia, M.L.; Mei, A.; Nemmi, E.N.; Neri, V.; Stefa, J. Cross-Domain Authorship Attribution Combining Instance Based and Profile-Based Features. In Proceedings of the CLEF, Lugano, Switzerland, 9–12 September 2019.
31. Ruder, S.; Ghaffari, P.; Breslin, J.G. Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution. *arXiv* **2016**, arxiv:1609.06686v1.
32. Zhang, R.; Hu, Z.; Guo, H.; Mao, Y. Syntax Encoding with Application in Authorship Attribution. In Proceedings of the EMNLP, Brussels, Belgium, 31 October–4 November 2018; pp. 2742–2753.
33. Qian, C.; He, T.; Zhang, R. *Deep Learning Based Authorship Identification*; Stanford University: Stanford, CA, USA, 2017.
34. Bagnall, D. Author Identification using Multi-Headed Recurrent Neural Networks. *arXiv* **2015**, arxiv:1506.04891v2. https://doi.org/10.48550/ARXIV.1506.04891.
35. Barlas, G.; Stamatatos, E. Cross-Domain Authorship Attribution Using Pre-trained Language Models. *Artificial Intelligence Applications and Innovations*; Maglogiannis, I., Iliadis, L., Pimenidis, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 255–266.
36. Liu, Y.; Che, W.; Wang, Y.; Zheng, B.; Qin, B.; Liu, T. Deep Contextualized Word Embeddings for Universal Dependency Parsing. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2019**, *19*, 1–17. [CrossRef]
37. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 328–339. [CrossRef]
38. Goldstein-Stewart, J.; Winder, R.; Sabin, R. Person Identification from Text and Speech Genre Samples. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 336–344.

39. Fabien, M.; Villatoro-Tello, E.; Motlicek, P.; Parida, S. BertAA : BERT fine-tuning for Authorship Attribution. In Proceedings of the 17th International Conference on Natural Language Processing (ICON), Patna, India, 18–21 December 2020; NLP Association of India (NLPAI), Indian Institute of Technology Patna: Patna, India, 2020; pp. 127–137.
40. Klimt, B.; Yang, Y. The Enron Corpus: A New Dataset for Email Classification Research. In Proceedings of the 15th European Conference on Machine Learning, ECML'04, Pisa, Italy, 20–24 September 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 217–226. [CrossRef]
41. Deutsch, C.; Paraboni, I. Authorship attribution using author profiling classifiers. In *Natural Language Engineering*; Cambridge University Press: Cambridge, UK, 2022; pp. 1–28. [CrossRef]
42. Manolache, A.; Brad, F.; Burceanu, E.; Bărbălău, A.; Ionescu, R.C.; Popescu, M.C. Transferring BERT-like Transformers' Knowledge for Authorship Verification. *arXiv* **2021**, arxiv:2112.05125.
43. Dingledine, R.; Mathewson, N.; Syverson, P. Tor: The Second-Generation Onion Router. In Proceedings of the 13th Conference on USENIX Security Symposium—Volume 13, USENIX Association, SSYM'04, Anaheim, CA, USA, 9–11 August 2004; p. 21.
44. Spitters, M.; Klaver, F.; Koot, G.; van Staalduinen, M. Authorship Analysis on Dark Marketplace Forums. In Proceedings of the 2015 European Intelligence and Security Informatics Conference, Manchester, UK, 7–9 September 2015; pp. 1–8. [CrossRef]
45. Swain, S.; Mishra, G.; Sindhu, C. Recent approaches on authorship attribution techniques—An overview. In Proceedings of the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; Volume 1, pp. 557–566. [CrossRef]
46. Ranaldi, L.; Nourbakhsh, A.; Patrizi, A.; Ruzzetti, E.S.; Onorati, D.; Fallucchi, F.; Zanzotto, F.M. The Dark Side of the Language: Pre-trained Transformers in the DarkNet. *arXiv* **2022**, arxiv:2201.05613.
47. Scrivens, R.; Davies, G.; Frank, R.; Mei, J. Sentiment-Based Identification of Radical Authors (SIRA). In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; pp. 979–986. [CrossRef]
48. Zhang, Y.; Zeng, S.; Huang, C.N.; Fan, L.; Yu, X.; Dang, Y.; Larson, C.A.; Denning, D.; Roberts, N.; Chen, H. Developing a Dark Web collection and infrastructure for computational and social sciences. In Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics, Vancouver, BC, Canada, 23–26 May 2010; pp. 59–64. [CrossRef]
49. Chen, H.; Chung, W.; Qin, J.; Reid, E.; Sageman, M.; Weimann, G. Uncovering the Dark Web: A case study of Jjihad on the Web. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 1347–1359. [CrossRef]
50. Abbasi, A.; Chen, H.; Salem, A. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Trans. Inf. Syst.* **2008**, *26*, 1–34. [CrossRef]
51. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
52. Klema, V.; Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Trans. Autom. Control* **1980**, *25*, 164–176. [CrossRef]
53. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly: Beijing, China, 2009. http://my.safaribooksonline.com/9780596516499.
54. Toutanova, K.; Klein, D.; Manning, C.D.; Singer, Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003; pp. 252–259.
55. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
56. Zhu, M.; Zhang, Y.; Chen, W.; Zhang, M.; Zhu, J. Fast and Accurate Shift-Reduce Constituent Parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; Volume 1; Long Papers, pp. 434–443.
57. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv* **2015**, arxiv:1506.06724.
58. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is Multilingual BERT? *arXiv* **2019**, arxiv:1906.01502.
59. Parker, R.; Graff, D.; Kong, J.; Chen, K.; Maeda, K. *English Gigaword Fifth Edition ldc2011t07*; Technical Report; Linguistic Data Consortium: Philadelphia, PA, USA, 2011.
60. Callan, J.; Hoy, M.; Yoo, C.; Zhao, L. Clueweb09 Data Set. 2009. Available online: https://ir-datasets.com/clueweb09.html (accessed on 12 September 2022).
61. Crawl, C. Common Crawl. 2019. Available online: http://commoncrawl.org (accessed on 12 September 2022).
62. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arxiv:1910.01108.
63. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arxiv:1910.0.
64. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* **2015**, arxiv:1412.6980.
65. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*; Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 194–206.
66. Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, X.; Zhou, X. Semantics-aware BERT for language understanding. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020), New York, NY, USA, 7–12 February 2020.
67. Goldberg, Y. Assessing BERT's Syntactic Abilities. *arXiv* **2019**, arxiv:1901.05287.
68. Iyer, A.; Vosoughi, S. Style Change Detection Using BERT. In Proceedings of the CLEF, Thessaloniki, Greece, 22–25 September 2020.

69. Podkorytov, M.; Biś, D.; Liu, X. How Can the [MASK] Know? The Sources and Limitations of Knowledge in BERT. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8. [CrossRef]
70. Choshen, L.; Eldad, D.; Hershcovich, D.; Sulem, E.; Abend, O. The Language of Legal and Illegal Activity on the Darknet. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4271–4279. [CrossRef]
71. Ma, X.; Xu, P.; Wang, Z.; Nallapati, R.; Xiang, B. Domain Adaptation with BERT-based Domain Classification and Data Selection. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), Hong Kong, China, 3 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 76–83. [CrossRef]
72. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; HerbertVoss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. Extracting Training Data from Large Language Models. In Proceedings of the USENIX Security Symposium, Online, 11–13 August 2021.
73. Ranaldi, L.; Fallucchi, F.; Santilli, A.; Zanzotto, F.M. KERMITviz: Visualizing Neural Network Activations on Syntactic Trees. In *Metadata and Semantic Research*; Garoufallou, E., Ovalle-Perandones, M.A., Vlachidis, A., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 139–147.