A Appendix for 'VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web'

A.1 Datasheet

We release three large-scale datasets for authorship verification (DarkReddit+, SilkRoad1 and Agora) and one smaller dataset for authorship identification (DarkReddit+). We collectively refer to these datasets as the VeriDark benchmark. This datasheet follows the structure and guidelines provided by Gebru et al. [2018].

A.1.1 Motivation

Authorship analysis research traditionally uses corpora featuring literary texts and, more recently, social media texts from the Web. There is a lack of datasets for authorship research in the cybersecurity context, where texts from the Dark Web would be useful. Our motivation for releasing the VeriDark datasets is to bridge this gap and facilitate authorship analysis research into this domain, for both researchers and law enforcement analysts.

The VeriDark datasets[11] were created by a group of researchers from the Bitdefender Theoretical Machine Learning Research team and from the University of Bucharest. The project was partly funded by The Executive Unit for the Financing of Higher Education Research Development and Innovation (UEFISCDI) through the PN-III-P2-2.1-PTE-2019-0532 grant.

A.1.2 Composition

Our authorship verification datasets are comprised of document pairs. Specifically, each instance (example) consists in pairs of strings, where each string represents a user's comment. Each instance has two possible labels: 'same author' or 'different authors'. The author identification dataset is composed of documents. Specifically, each instance consists in a string, which represents a user's comment. There are 10 possible labels, each label representing the identity of a user.

The sizes of the authorship verification datasets are approximately 100K (DarkReddit+), 600K (SilkRoad1) and 4M (Agora), while the size the author identification dataset is approximately 7K. The complete statistics are listed in Table 1 from the main article. The train, validation and test splits are already provided and publicly released. The train, validation and test split percentages are 90%, 5% and 5%. Authors in the test and validation splits do not appear in the train split, making the setup an open set authorship verification problem.

Potential noise in the data comes from users who have multiple accounts or multiple users sharing the same account. For more details, see Section 4.4 from the main article. A user's comment may appear in several authorship verification pairs, but the comment pairs are unique. Moreover, some comment pairs may be very close in edit distance.

The datasets are self-contained and do not rely on other external resources. They have been made available on Zenodo, an open source platform for sharing datasets. Due to ethical concerns regarding the potential misuse of our benchmark, access to the datasets is restricted. Any person or organization who wants to be granted access must disclaim the intended usage for the datasets and must acknowledge to use it in an ethical manner. More details about accessing the data are given in Section A.1.7.

Due to the nature of the discussion platforms, the datasets include sensitive topics such as drugs, illicit substances or pornography. There may be instances of offensive and violent language. We do not recommend the use of these datasets for tasks other than authorship verification and identification, such as language modeling, as they may contain triggering content.

Our benchmark does not identify any subpopulation such as age, gender, race, political opinion, etc. However, it could contain sensitive information such as race or ethnic origins, sexual orientations, religious beliefs, political opinion, locations, financial or health data. We provide a contact form in the 'About' section of our leaderboard page, where users can raise issues about the dataset or request the removal of examples containing personal or sensitive content: https://veridark.github.io/about.

[11] https://veridark.github.io

The raw data on which the VeriDark datasets was build contains usernames, PGP keys, signatures and messages. To impede identifying individuals in real life based on these characteristics, we anonymized the usernames and removed the PGP-related information from the text. This information could be retrieved by third parties by inspecting the publicly available raw data archives we started from, but we do not publish this raw data and instead upload the preprocessed datasets. We therefore argue that it is difficult to directly identify an individual solely based on the VeriDark datasets and methods trained on these datasets.

However, at the heart of the application of authorship methods lies the possibility of attributing texts from the VeriDark datasets to authors whose identity is already known. This means that the VeriDark datasets could be indirectly used to identify a person, by using external datasets, models and knowledge. This may be a typical scenario for law enforcement agencies seeking to uncover criminal activities. This use case can be viewed as an ethical application of the technology, by weighing privacy concerns against security concerns. However, other use cases may cause harm to individuals, such as targeting vulnerable categories using the Dark Web for security reasons or detecting law enforcement agents. We ultimately believe that releasing these datasets together with the appropriate safeguards reduces the chance of negative social impacts, while making it a valuable resource for the cybersecurity stakeholders. We discuss the ethical implications in more detail in Section 5 from the main article.

A.1.3 Collection process

The data was directly observable in the form of text comments written by users in community forums (subreddit or DarkNet marketplaces).

The raw data for the SilkRoad1 and Agora datasets was provided in the form of scraped forums from the DarkNetMarkets dataset [Branwen, 2021]. We parsed the .html files that contained the word 'topic' in their names with the BeautifulSoup library. We thus gathered all the comments in a dictionary where each key is a user name and its value is the list of comments written by that user. Based on this dictionary, it is straightforward to generate pairs of comments written by the same author or by different authors. The two marketplace forums are a subset of the larger DarknetMarkets dataset hosting 40 such forums. They were chosen due to their large size and popularity.

The raw data for the DarkReddit+ dataset was retrieved from an archive containing all the comments written on Reddit up until 2015 [Baumgartner et al., 2020]. Similarly to the previous two datasets, we retrieved usernames and comments from the defunct subreddit /r/darknetmarkets, which was a gateway to the DarkNet marketplace commerce. While there are many other subreddits, we selected this subreddit specifically due to its sole focus on Dark Web activity.

The datasets were collected between May 14th 2022 and June 2nd 2022 by the same team of researchers involved in the whole research project. The data associated with the datasets was however created in the past decade. Specifically, the SilkRoad1 marketplace forum data was written between January 31st 2011 and October 2nd 2013. The Agora marketplace forum data was written between December 3rd 2013 and September 6th 2015. The DarkReddit+ comments were written between January 2014 and May 2015.

While no formal internal ethical review process was conducted, there were extensive discussions between the members of the research project and an internal cybersecurity team regarding the ethical implications of the project and the broader impact of our work.

The datasets were created using publicly available data crawled from comments written by users on the Web and the Dark Web. The individuals writing the comments were not informed that their data would be used for the purpose of this research work. We provide a contact form in the 'About' section of our leaderboard page, where users can raise issues about the dataset or request the removal of examples containing personal or sensitive content: https://veridark.github.io/about.

No data protection impact analysis has been performed.

A.1.4 Preprocessing

We performed several preprocessing steps on all the datasets. We anonymized the usernames and replaced them using the placeholder 'userX', where X is a number ranging from 1 to the number of authors. We removed the PGP signatures, keys and messages to further impede identifying individuals.

We removed comments that were not written in English using the langdetect[12] library. We then removed comments with less than 200 characters, due to little information available for determining authorship. This removal resulted in approximately two times less comments from the Dark Web (`Agora` 12.6M => 5.6M comments, `SilkRoad1` 1.7M => 800K comments) and almost four times less comments from Reddit (~560K => ~150K comments). Since we used these comments to create the same author and different author pairs, this preprocessing step reduced the `Agora` and `SilkRoad1` dataset sizes by a factor of 2, and `DarkReddit+` by a factor of 4. We also removed duplicate comment pairs from the authorship verification datasets.

The raw data used to obtain the datasets was saved, but not published, as it may contain sensitive information (original usernames, PGP-related information). We also do not provide the software used to preprocess the raw data, as it would make it easy to recreate our datasets with the additional sensitive information available.

A.1.5 Uses

The raw DarkNet data that two of our datasets were based on was used by other works which can be found at `https://www.gwern.net/DNM-archives#works-using-this-dataset`. Our datasets have never been used before in any authorship tasks.

Our datasets were created specifically for the authorship verification and identification tasks, but can also be preprocessed so that they can be used for authorship attribution. We do not recommend to use these datasets for training generative models for text (such as language models) or multimodal data involving text (i.e. visual language models). We also strongly condemn using these datasets for non-ethical uses, such as evading or unmasking law enforcement agencies, exposing vulnerable individuals (whistleblowers, journalist), etc.

A.1.6 Responsibility

We tried to limit sensitive information leakage from our datasets. The raw data from which these datasets were created is publicly available. We take responsibility for the published datasets.

A.1.7 Distribution

The datasets are available on the Zenodo platform at the links provided in Table 6. Due to ethical concerns regarding the potential misuse of our datasets, we require the users to complete the following information in the Zenodo access request page in order to be granted permission to use our datasets:

1. The name of the person requesting access, together with their affiliations, job title and an e-mail address. If the person holds an institutional e-mail address, we strongly recommend using it instead of a personal e-mail address.
2. The intended usage for the dataset.
3. An acknowledgement that the dataset will be strictly used in an ethical manner. Non-ethical uses of the dataset include, but are not limited to:
   - using the datasets for the task of Language Modeling or similar generative algorithms.
   - building algorithms that could aid criminals to evade law enforcement organizations.
   - building algorithms that have the aim of unmasking undercover law enforcement agents.
   - building algorithms that could interfere with the activity of law enforcement agencies.
   - building algorithms that could lead to violating any article of the United Nations Universal Declaration of Human Rights.
   - building algorithms with the purpose of exposing the identity of reporters, individuals in the political realms, leakers, whistleblowers, dissidents, or other persons who are seeking to express an opinion about what they perceive is a particular injustice in the world, without regard to what that injustice may be.
   - building algorithms that can help entities discriminate, or exacerbate bias against other persons on the basis of race, color, religion, gender, gender expression, age, national origin, familiar status, ancestry, culture, disability, political views, sexual orientation, marital status, military status, social status, or who have other protected characteristics.

[12] `https://github.com/fedelopez77/langdetect`

Table 6: Digital object identifiers (DOI) and hosting links for all the VeriDark datasets. AV: authorship verification, AI: authorship identification

dataset name	task	DOI	Zenodo link
SilkRoad1	AV	10.5281/zenodo.6998371	https://zenodo.org/record/6998371
Agora	AV	10.5281/zenodo.7018853	https://zenodo.org/record/7018853
DarkReddit+	AV	10.5281/zenodo.6998375	https://zenodo.org/record/6998375
DarkReddit+	AI	10.5281/zenodo.6998363	https://zenodo.org/record/6998363

Access will be granted only if all the three pieces of information (requester information, intended usage and acknowledgement) are provided. Any personal information provided when requesting access to the datasets will be used only for deciding whether access is granted or not. We will not disclose your personal data. Access can be granted by any member of the VeriDark research project.

We strongly encourage the inclusion of an ethical statement and discussion in any work based on this dataset. We do not encourage the distribution of the dataset in its current form to any other parties without our consent.

The VeriDark datasets are hosted on the Zenodo platform since July 2022. The datasets are stored in .jsonl files. Each example is stored in the widely used JSON format.

Datasets identifiers and hosting links can be found in Table 6. To obtain the datasets, users must send a request with the required information filled in.

The VeriDark datasets are distributed under the CC-BY-4.0[13] license. We kindly request any paper using our datasets to cite our work.

A.1.8 Maintenance

The following people will be supporting/maintaining the dataset:

- Florin Brad (fbrad@bitdefender.com)
- Andrei Manolache (amanolache@bitdefender.com, andrei_mano@outlook.com)
- Elena Burceanu (eburceanu@bitdefender.com)
- Radu Ionescu (raducu.ionescu@gmail.com)
- Antonio Barbalau (abarbalau@fmi.unibuc.ro)
- Marius Popescu (popescunmarius@gmail.com)

As of October 2022, there is no erratum and all datasets are at version 0.1.0 on Zenodo. Future updates will be detailed in the Zenodo page of each dataset. They will also be listed in the GitHub repository https://github.com/bit-ml/VeriDark/tree/master/datasets.

The datasets will be updated to account for potential issues (wrong labels, duplicates, etc.) and examples containing sensitive information which receive a request for deletion. Please raise potential issues or request for deletions at the contact form on the 'About' page: https://veridark.github.io/about.

Older version of the datasets will be available on the Zenodo page of each dataset. People who want to extend the VeriDark datasets can contact any of the maintainers on their email addresses.

[13] https://creativecommons.org/licenses/by/4.0/

Table 7: Results on the `DarkReddit+` author identification dataset in terms of accuracy, when varying the amount of training/test data for the '*Other*' class. Each column name refers to the source of the '*Other*' training samples. The model from the **None** column has been trained on texts from the ten authors only. The first two rows show the results of the models evaluated on the `DarkReddit+` test set, while the last two rows show results on the `ClearReddit` test set. The first row of each section presents results on classifying texts belonging only to the ten original authors. For the second row of each section, texts from the '*Other*' class were introduced at test time from each source.

			Source for *Other* (2x data)					**Source for *Other* (3x data)**				
			None	**PAN**	**SR1**	**AG**	**DR+**	**None**	**PAN**	**SR1**	**AG**	**DR+**
Test Setup	**Dark**	W/o *Other*	84.6	83.17	82.33	82.11	81.37	84.6	82.29	84.05	81.50	78.33
		With *Other*	-	85.97	82.20	81.58	78.25	-	86.30	82.15	83.23	77.34
	Clear	W/o *Other*	81.2	80.35	77.31	79.28	71.08	81.2	74.21	77.02	77.52	68.65
		With *Other*	-	83.63	76.87	79.59	69.12	-	80.16	77.24	80.10	69.15

Table 8: Performance on the `VeriDark` datasets when training on the PAN dataset and fine-tuning on each of the `VeriDark` datasets.

	BERT-based			compression-based		
Metric	**DR+**	**SR1**	**AG**	**DR+**	**SR1**	**AG**
F1	72.9	81.9	85.2	61.1	53.6	52.7
F0.5	77.2	82.5	85.8	58.1	53.9	56.5
c@1	75.2	82.6	85.5	59.1	57.9	60.4
ROC	83.7	91.0	93.7	63.9	59.8	63.9
avg.	77.2	84.5	87.6	60.6	56.3	58.4

A.2 Introducing more samples from secondary sources for Authorship Attribution

Table 7 showcases the performance of the authorship identification baseline when trained and tested using a larger amount of samples for the '*Other*' class (two times and three times more data than in Table 4 in the main article.

A.3 Comparison to non-neural baseline

We compare our BERT-based approach to a competitive non-neural compression-based baseline [Teahan and Harper, 2003] used at the PAN competition. The results reported in Table 8 show that our approach significantly outperforms the baseline.

A.4 Using more chunk pairs for evaluation

For evaluating long text pairs (X, Y), we splitted each of them into N equal length chunks: $X = [X_1, X_2, ..., X_N]$ and $Y = [Y_1, Y_2, ..., Y_N]$. We then iteratively picked chunks (X_i, Y_i) for evaluation, then aggregated the individual scores. In theory, we could take all the possible combinations of chunks (X_i, Y_j), but this would result in an N^2 evaluations. To still measure the potential benefits of more data, we evaluated using twice more data by using the reverse (Y_i, X_i) chunk pair for each (X_i, Y_i) pair.

The results reported in Table 9 show that using twice as many pairs for evaluation results in slightly higher metrics across all the datasets. The results indicate that considering even more data from all the N^2 chunk pairs may further increase the results, but the trade-off between the evaluation time and performance must be considered.

Table 9: Performance on the VeriDark datasets when evaluating with the original chunk pairs as well as the reverse chunk pairs.

	DarkReddit+		SilkRoad1		Agora	
Metric	(X, Y)	$(X, Y) + (Y, X)$	(X, Y)	$(X, Y) + (Y, X)$	(X, Y)	$(X, Y) + (Y, X)$
F1	75.4	75.9	82.1	82.6	85.2	85.6
F0.5	73.9	74.3	83.4	83.9	85.8	86.3
c@1	74.5	75.0	83.1	83.5	85.5	86.0
ROC	82.7	83.5	91.4	91.8	93.7	94.0
avg.	76.6	77.2	85.0	85.5	87.6	88.0

A.5 Alternative view of Table 2

We provide an alternative view of Table 2 in Table 10, where the supercolumns represent test data, while the columns represent training data. This view makes it easier to compare models trained on different datasets and tested on the same dataset.

Table 10: Results on the VeriDark authorship datasets: DarkReddit+, Agora, SilkRoad1 and All (the previous three datasets aggregated). *The mean over 5 runs is reported, with the *std* line being the standard deviation for the *avg.* score.

Test		DarkReddit+				SilkRoad 1				Agora				All			
Train		**DR+**	**SR1**	**AG**	**All**	**DR+**	**SR1**	**AG**	**All**	**DR+**	**SR1**	**AG**	**All**	**DR+**	**SR1**	**AG**	**All**
Metric	*F1*	75.0	75.3	73.4	75.3	70.2	81.6	79.0	81.6	72.7	80.2	84.9	85.8	72.4	80.3	83.8	84.9
	F0.5	75.1	69.7	66.1	70.1	76.8	83.0	76.3	80.6	78.9	79.7	85.4	86.2	78.5	79.8	83.6	84.9
	c@1	75.1	71.5	67.3	71.7	74.8	82.6	78.3	81.7	76.2	80.2	85.3	86.1	76.0	80.3	83.9	85.2
	ROC	83.6	81.8	80.3	82.0	85.1	91.0	87.1	90.2	86.1	88.3	93.5	94.2	85.9	88.4	92.3	93.5
	avg.	77.2	74.6	71.8	74.7	76.7	84.6	80.2	83.5	78.4	82.1	87.3	88.1	78.2	82.2	85.9	87.1
	*std**	0.27	0.67	0.72	1.08	1.14	0.23	0.55	0.54	1.75	0.33	0.39	0.16	1.64	0.28	0.34	0.19

A.6 Pre-training on the PAN 2020 Authorship Verification dataset

Table 11: Performance on the VeriDark datasets when training on the PAN dataset and fine-tuning on each of the VeriDark datasets.

Metric	**DR+**	**SR1**	**AG**
F1	75.1	81.9	**82.4**
F0.5	75.4	81.8	**83.2**
c@1	75.2	82.2	**84.0**
ROC	83.5	90.8	**92.5**
avg.	77.3	84.2	**85.5**

We perform a pre-training step on the PAN 2020 authorship verification dataset, before training our network on the VeriDark datasets.

The initial experiments show that pretraining on the PAN dataset and fine-tuning on each of the VeriDark datasets gives similar results for DarkReddit+ and SilkRoad1 and slightly degrades the performance on Agora, as can be seen in Table 11.