Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection
Condensed Lab Overview

Janek Bevendorff[1], Ian Borrego-Obrador[2], Mara Chinea-Ríos[2], Marc Franco-Salvador[2], Maik Fröbe[3], Annina Heini[4], Krzysztof Kredens[4], Maximilian Mayerl[5], Piotr Pęzik[4], Martin Potthast[6], Francisco Rangel[2], Paolo Rosso[7], Efstathios Stamatatos[8], Benno Stein[1], Matti Wiegmann[1(✉)], Magdalena Wolska[1], and Eva Zangerle[5]

[1] Bauhaus-Universität Weimar, Weimar, Germany
pan@webis.de
[2] Symanto Research, Valencia, Spain
[3] Friedrich Schiller University Jena, Jena, Germany
[4] Aston University, Birmingham, UK
[5] University of Innsbruck, Innsbruck, Austria
[6] Leipzig University and ScaDS.AI, Leipzig, Germany
[7] Universitat Politècnica de València, Valencia, Spain
[8] University of the Aegean, Samos, Greece
https://pan.webis.de

Abstract. The paper gives a brief overview of three shared tasks which have been organized at the PAN 2023 lab on digital text forensics and stylometry hosted at the CLEF 2023 conference. The tasks include authorship verification across discourse types, multi-author writing style analysis, profiling cryptocurrency influencers with few-shot learning, and trigger detection. Authorship verification and multi-author analysis continue and advance from past editions of PAN and influencer profiling and trigger detection are new tasks with novel research questions and evaluation resources. All four tasks alilgn with the goals of all shared tasks at PAN: to advance the state of the art in text forensics and stylometry while ensuring objective evaluation on newly developed benchmark datasets.

1 Introduction

PAN is a workshop series and a networking initiative for stylometry and digital text forensics. The workshop's goal is to bring together scientists and practitioners studying technology to analyze texts regarding their originality, authorship, trust, and ethicality. Since its inception in 2009, PAN has been the venue for 69 shared tasks on computational challenges related to authorship analysis, computational ethics, and determining the originality of a piece of writing.

A. Arampatzis et al. (Eds.): CLEF 2023, LNCS 14163, pp. 459–481, 2023.
https://doi.org/10.1007/978-3-031-42448-9_29

Over the years, the respective organizing committees have assembled and studied 60 datasets evaluation resources,[1] nine of which are community contributions.

The 2023 edition of PAN at CLEF continues in the same spirit and presents four new shared tasks. First, cross-discourse type authorship verification asks if two given documents are written by the same or by different authors, where one document is in a written (essays, emails) and one in a spoken (interviews, speech transcriptions) register. The task iterates on the previous edition by defining a much more difficult setting based on the resources established last year. 10 participants submitted solutions. Second, multi-author writing style analysis asks at which position in the document the authorship changes. The task iterates on the previous edition by presenting a completely new dataset of Reddit comments while relying on the established problem definition. 6 participants submitted solutions. Third, profiling cryptocurrency influencers with few-shot learning requests participants to profile the influence, interest, and intent of Twitter users given at most 10 tweets from their timelines. The task proposes a completely new challenge, including a new evaluation resource for author profiling in a new, and difficult, few-shot setting, i.e., only little data is available to make a decision. 27 participants submitted solutions. Fourth, trigger detection asks to assign a warning label to a given fan fiction document if it contains potentially harmful content. The task presents a completely new problem, including a new evaluation resource for computational ethics. 6 participants submitted solutions.

PAN is committed to reproducible research in IR and NLP, hence all participants are asked to submit their software (instead of just their predictions) through the submission software TIRA. With the recent updates to the TIRA platform [11], all submissions to PAN were made as publicly available docker containers. In the following sections, we briefly outline the 2023 tasks and their results.

2 Cross-Discourse Type Authorship Verification

Authorship verification is the task of deciding whether a document has been written by a certain author. In general, a number of documents of known authorship by the author in question are available and the task aims at identifying stylistic similarities/differences between the known document and the disputed text. In its simplest form, only one document of known authorship is given and, in that case, authorship verification can be seen as determining whether two texts have been written by the same author [23]. Any authorship attribution case can be decomposed into a series of authorship verification tasks, therefore focusing on authorship verification is fundamental in testing the ability of computational approaches to recognize the writing style characteristics of authors.

One factor that may affect the difficulty of the authorship verification task is the length of the considered texts. In addition, it is critical to examine whether there are thematic similarities among the involved documents since the topic factor may be misleading (e.g., two documents may appear to be similar due to

[1] https://pan.webis.de/data.html.

a common theme rather than the writing style). It is even more challenging in cases the documents belong to different genres or discourse types (e.g., essay vs. email) that considerably affect the stylistic properties of documents.

Several previous editions of PAN included authorship verification tasks [1, 2, 41, 41, 62, 64]. There were also attempts to focus on *cross-domain authorship attribution* where the documents of known and unknown authorship belong to different domains (e.g., thematic areas or genres) [1, 2, 64]. Recent PAN editions focused on fan-fiction texts (i.e., non-professional fiction published online by fans of well-known works) where the documents of known and unknown authorship come from different fandoms (e.g., Harry Potter, Sherlock Holmes) allowing us to build large-scale datasets. The obtained results indicate that this task can be handled with relatively high accuracy [1, 2]. In the last edition of PAN, a more challenging scenario was considered, focusing on *cross-discourse type authorship verification* where the documents of known and unknown authorship belong to different discourse types (i.e., essays, emails, text messages, and business memos) [62]. The discourse type also affects the text length (e.g., essays are much longer than text messages). The obtained results indicate that it is extremely difficult to recognize the writing style characteristics related to the personal style of authors across discourse types.

In the current edition of PAN, we continue to focus on cross-discourse type authorship verification of document pairs. In contrast to previous versions of the task where only discourse types of written language were used, we also consider oral language. This provides the opportunity to study the ability of authorship verification methods to handle the different forms of expression in written and oral language.

Dataset

A new dataset has been created based on the recent Aston 100 Idiolects Corpus in English[2] including a rich set of discourse types written by around 100 individuals. All individuals have similar ages (18–22) and are native English speakers. The topic of text samples is not restricted. Part of this corpus was also used to build the datasets of the PAN-2022 edition of the task [62]. In more detail, we consider four discourse types: two from written language (i.e., emails and essays) and two from oral language (i.e., interviews and speech transcriptions). All possible six combinations of document pairs are examined.

Since the length of emails can be very short, we concatenate consecutive messages (ordered by date) so that at least text samples of at least 2,000 characters are obtained. In addition, since separate interview utterances are included in the corpus, we also concatenate consecutive utterances to obtain text samples of at least 2,000 characters. All text samples in the corpus have been pre-processed to replace named entities with general tags. This helps to reduce the topic factor.

In order to provide training and test datasets, we first split the available individuals into two non-overlapping sets of equal size. In more detail, the text

[2] https://fold.aston.ac.uk/handle/123456789/17.

Table 1. Statistics of the PAN'23 datasets used in the cross-discourse type authorship verification task.

	Training	Test
Text pairs		
Positive	4,418 (50.0%)	4,828 (50.0%)
Negative	4,418 (50.0%)	4,828 (50.0%)
Email - Speech transcription	1,036 (11.7%)	1,074 (11.1%)
Essay - Email	1,454 (16.5%)	1,618 (16.8%)
Essay - Interview	884 (10.0%)	938 (9.7%)
Essay - Speech transcription	256 (2.9%)	206 (2.1%)
Interview - Email	4,564 (51.7%)	5,214 (54.0%)
Speech transcription - Interview	642 (7.3%)	606 (6.3%)
Text length (avg. chars)		
Email	2,308	2,346
Essay	9,894	10,770
Interview	2,503	2,501
Speech transcription	2,395	2,537

samples of 56 individuals are used for the training dataset and the test dataset is obtained from another set of 56 individuals. Both sets of authors have similar gender distribution. Each dataset comprises a set of document pairs and in each pair, the documents belong to different discourse types. Given that the distribution of text samples over the discourse types is not balanced, the distribution of document pairs over the six possible combinations of discourse types is not homogeneous as can be seen in Table 1. However, it is similar between training and test datasets. In addition, both datasets are balanced regarding same-author and different-author pairs. This is also true when each specific combination of discourse types is considered separately.

Evaluation Setup and Results

The evaluation framework is similar to the one used in recent shared tasks at PAN [1,2,62]. Formally, one has to approximate the target function $\phi : (d_k, d_u) \rightarrow \{T, F\}$, d_k being a text of known authorship and d_u being a text of unknown or disputed authorship. If $\phi(d_k, d_u) = T$, then the author of d_k is also the author of d_u and if $\phi(d_k, d_u) = F$, then the author of d_k is not the same as the author of d_u. In the current edition of the task, d_k and d_u belong to different discourse types of written or oral language.

For each text pair of the test dataset, participants have to produce a scalar score a_i (in the $[0, 1]$ range) indicating the probability both texts are written by the same author. It is possible for participants to leave text pairs unanswered by submitting a score of precisely $a_i = 0.5$. As concerns the set of evaluation

Table 2. Final results for the cross-discourse type authorship verification task at PAN'23. Submitted systems are ranked by their mean performance across five evaluation metrics. The best result per column is shown in bold.

Systems	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
Ibrahim, et al. (reduced-graph) [19]	**0.616**	**0.572**	0.617	0.562	0.746	**0.623**
Ibrahim, et al. (resolving-globe) [19]	**0.616**	**0.572**	0.617	0.562	0.746	**0.623**
Guo, et al. (irregular-strategist) [14]	0.581	0.557	0.621	**0.571**	0.742	0.614
Ibrahim, et al. (golden-ottoman) [19]	0.598	0.546	0.622	0.550	0.744	0.612
BASELINE (cngdist)	0.516	0.499	**0.666**	0.555	0.741	0.595
Petropoulos (graceful-chianti) [40]	0.526	0.514	0.624	0.549	0.743	0.591
Petropoulos (clever-daemon) [40]	0.525	0.516	0.622	0.550	0.743	0.591
BASELINE (galicia22)	0.504	0.502	0.650	0.552	0.740	0.589
Valdez Valenzuela, et al. (GNN-SHORT) [70]	0.511	0.508	0.655	0.555	0.705	0.587
Sun, et al. (SDML epoch 8) [66]	0.504	0.502	0.632	0.546	0.747	0.586
Sun, et al. (SDML epoch 24) [66]	0.505	0.501	0.601	0.536	0.749	0.578
Guo, et al. (uniform-reward) [14]	0.595	0.555	0.460	0.527	0.723	0.572
Valdez Valenzuela, et al. (GNN-FULL) [70]	0.517	0.512	0.628	0.549	0.644	0.570
Sun, et al. (SDML epoch 35) [66]	0.511	0.508	0.558	0.526	0.749	0.570
Valdez Valenzuela, et al. (GNN-MED) [70]	0.503	0.502	0.602	0.534	0.709	0.570
BASELINE (najafi22)	0.601	0.569	0.466	0.543	0.595	0.555
Huang, et al. (isochoric-paint) [18]	0.563	0.563	0.511	0.550	0.563	0.550
Liu, et al. (coincident-sound) [30]	0.548	0.548	0.544	0.547	0.548	0.547
Lv (radioactive-copyright) [33]	0.553	0.553	0.504	0.540	0.553	0.541
Huang, et al. (steel-coriander) [18]	0.500	0.500	0.651	0.551	0.500	0.540
Li, et al. (wan-ocean) [28]	0.500	0.500	0.646	0.550	0.500	0.539
Lv, et al. (tender-bugle) [33]	0.551	0.551	0.501	0.537	0.551	0.538
Lv, et al. (cold-rotor) [33]	0.550	0.550	0.465	0.524	0.550	0.528
Qiu, et al. (corn-mall) [42]	0.540	0.540	0.421	0.499	0.540	0.508
Qiu, et al. (poky-deck) [42]	0.540	0.540	0.421	0.499	0.540	0.508
Liu, et al. (perpendicular-field) [30]	0.534	0.534	0.421	0.493	0.534	0.503
Liu, et al. (foggy-raster) [30]	0.533	0.533	0.424	0.493	0.533	0.503
BASELINE (compressor)	0.506	0.051	0.626	0.076	**0.750**	0.402
Sanjesh, et al. (calm-lyrics) [58]	0.525	0.500	0.030	0.068	0.729	0.370
Sanjesh, et al. (null-midpoint) [58]	0.523	0.499	0.031	0.066	0.730	0.370
Sanjesh, et al. (Multi-Feature Classifier) [58]	0.501	0.01	0.000	0.000	0.750	0.252

measures, the set of measures used in the last edition of PAN is also adopted. These include the area under ROC (AUROC), c@1 that rewards unanswered cases over wrong predictions, F_1, $F_{0.5u}$, and the complement of Brier score (so that higher scores correspond to better performance) [62]. The average of these diverse measures is used as the final score to rank participants.

Two simple approaches are used as baselines: a compression-based approach based on Prediction by Partial Matching (PPM) [67] and a naive distance-based character n-gram model [21]. In addition, two submissions from the previous edition of the task at PAN-2022 are also used as baselines [62]. One of them is based on a pre-trained language model (T5) combined with a convolutional neural network [39] while the other uses a graph-based Siamese network [34]. We received submissions from 11 research teams and a total number of 27 runs (i.e., at most three runs per participant were allowed). The performance of each

run was evaluated using the TIRA experimentation framework. The evaluation results on the test dataset of all submitted software and the baselines can be seen in Table 2.

The difficulty of the task and the specific dataset including discourse types from both written and oral language is reflected in the obtained results. In general, the performance of most submitted systems is quite low, nearly surpassing a random guess baseline. The most successful approaches are based on pre-trained language models (e.g., BERT) enhanced by contrastive learning. However, a naive baseline based on character n-grams is quite competitive. A more detailed analysis of the evaluation results and the submissions is available in the task overview paper [63].

3 Multi-Author Writing Style Analysis

Authorship identification tasks are based on the intrinsic analysis of writing styles. Multi-author writing style analysis of multi-author documents aims to identify text positions at which the authorship changes based on an intrinsic style analysis. With advancing task definitions, data sets, and evaluation procedures, this PAN task has evolved steadily since 2016. The task in 2016 was to identify individual authors within a document and group these fragments [56]. In 2017, participants were asked to assess whether a given document is multi-authored. We asked participants to identify the positions of style changes if the document was indeed multi-authored [69]. For the challenges between 2018 and 2021, we asked participants to predict whether a given document is single- or multi-authored [22]. Additionally, we asked for the number of authors of multi-author documents [81]. In 2020 and 2021, we asked participants to detect paragraph-level style changes for multi-author documents [80]. In 2021, participants had to assign all paragraphs of the text uniquely to some author [77]. In 2022, participants were asked to identify all positions of writing style changes both on the paragraph- and the sentence-level [78].

Multi-Author Writing Style Analysis at PAN'23

Methods for multi-author writing style analysis are the key enabling technology for author identification tasks. The analysis of writing styles allows for performing intrinsic plagiarism detection (i.e., detecting plagiarism without the use of a reference corpus). As part of PAN@CLEF, we continue to develop benchmarks and challenges to advance research in this important field.

The multi-author writing style analysis task at PAN'23 asks participants to identify all positions of writing style change on the paragraph level for a given text. For each pair of consecutive paragraphs, the goal is to assess whether there was a style change between those paragraphs. In previous years, we employed different tasks of different complexity, that were carried out on the same data sets. However, the previously used data sets exhibited substantial topic diversity, which allowed the participants to leverage topic information as a style change

Table 3. Overall results for the style change detection task. The best result for each data set is given in bold.

Systems	Easy F_1	Medium F_1	Hard F_1
Ye et al. [76]	0.983	0.830	**0.821**
Hashemi et al. [15]	**0.984**	**0.843**	0.812
Kucukkaya et al. [24]	0.982	0.810	0.772
Huang et al. [17]	0.968	0.806	0.769
Chen et al. [6]	0.914	0.820	0.676
Jacobo et al. [20]	0.793	0.591	0.498

signal. Therefore, at PAN'23, we provide three data sets of increasing difficulty w.r.t. the multi-author writing style analysis task: *Easy*: The paragraphs of a document cover a variety of topics, allowing approaches to make use of topic information to detect authorship changes. *Medium:* The topical variety in a document is small (though still present), forcing the approaches to focus more on style to effectively solve the detection task. *Hard:* All paragraphs in a document are on the same topic.

Data Set and Evaluation

As a departure from the data sets of previous years, the data sets for this year's edition of the Multi-Author Writing Style Analysis task are based on user posts on Reddit[3]. In an effort to generate both realistic and diverse texts for the data sets, we chose parts of Reddit (so-called *subreddits*) that tend to generate longer and more meaningful discussions by users to extract our data from. The following subreddits were chosen: *r/worldnews*, *r/politics*, *r/askhistorians*, and *r/legaladvice*.

Like in previous years, we performed various cleaning steps to ensure the documents generated for the task consisted of well-formed texts. Quotes, all forms of markdown, multiple line breaks or whitespaces, frequently used emojis, hyperlinks as well as trailing and leading whitespaces were removed.

Following this, the collected user posts were split into paragraphs, and then documents for the data sets were generated from the paragraphs of a single given Reddit post. This was done to ensure at least a basic level of topical coherence for all the paragraphs in the final document. To generate style changes, a random set of authors for the given post was chosen, and paragraphs written by those authors were concatenated to form the final document. For the first time this year, this mixing of paragraphs into documents was not done fully randomly, but instead uses a newly developed procedure that allows us to (1) generate more topically and stylistically coherent documents and (2) tweak the difficulty of the produced data set. For this, both semantic as well as stylistic properties of

[3] https://www.reddit.com/.

the paragraphs were extracted into a feature vector, and paragraphs were then mixed based on the similarity of those vectors, where those similarities were configured to be (1) relatively large for the *easy* data set, (2) moderate for the *medium* data set, and (3) small for the *hard* data set.

All generated documents were written by between two and four authors, with an even distribution of the number of authors over the documents. Overall, each data set consists of 6,000 documents. Like in previous years, training, test, and validation splits are provided for all three data sets, with the test sets being withheld until the evaluation phase of the competition. The training sets contain 70% of the documents in each data set, while the test and validation sets contain 15% each.

The effectiveness of the models is evaluated independently on the three datasets using macro-averaged F1-score value across all documents.

Results

The Multi-Author Writing Style Analysis task received six software and notebook paper submissions. The individual results achieved by the participants are presented in Table 3. For both the *easy* and *medium* data set, the submission by Hashemi et al. achieved the highest performance, while the approach by Ye et al. performed best on the *hard* data set. Further details on the approaches taken can be found in the overview paper [79].

4 Author Profiling

Author profiling is the problem of distinguishing between classes of authors by studying how language is shared by people. This helps in identifying authors' individual characteristics, such as age, gender, or language variety, among others. During the years 2013–2022, we addressed several of these aspects in the shared tasks organized at PAN.[4] In 2013 the aim was to identify gender and age in social media texts for English and Spanish [50]. In 2014 we addressed age identification from a continuous perspective (without gaps between age classes) in the context of several genres, such as blogs, Twitter, and reviews (in Trip Advisor), both in English and Spanish [47]. In 2015, apart from age and gender identification, we addressed also personality recognition on Twitter in English, Spanish, Dutch, and Italian [52]. In 2016, we addressed the problem of cross-genre gender and age identification (training on Twitter data and testing on blogs and social media data) in English, Spanish, and Dutch [53]. In 2017, we addressed gender and language variety identification in Twitter in English, Spanish, Portuguese, and Arabic [51]. In 2018, we investigated gender identification on Twitter from a multimodal perspective, considering also the images linked within tweets; the dataset was composed of English, Spanish, and Arabic tweets [49]. In 2019 our

[4] To generate the datasets, we have followed a methodology that complies with the EU General Data Protection Regulation [45].

focus was on profiling and discriminating bots from humans on the basis of textual data only [46] and targeting both English and Spanish tweets. In 2020, we focused on profiling fake news spreaders [44], in two languages, English and Spanish. The ease of publishing content on social media has also increased the amount of disinformation that is published and shared. The goal of this shared task was to profile those authors who have shared some fake news in the past. In 2021 the focus was on profiling hate speech spreaders in social media [43]. The goal was to identify Twitter users who can be considered haters, depending on the number of tweets with hateful content that they had spread. The task was set in English and Spanish. Finally, in 2022, we focused on profiling irony and stereotype spreaders on English tweets [55]. The shared task goal was to profile highly ironic authors and those that employ irony to convey stereotypical messages, e.g. towards women or the LGTB community.

Profiling Cryptocurrency Influencers with Few-shot Learning

Cryptocurrencies have massively increased their popularity in recent years [59]. The promise of independence from central authorities, the possibilities offered by the different projects, and the new influencer-driven gold rush make cryptocurrencies a trendy topic in social media. Additionally, we believe that due to the early stage and complexity of the crypto ecosystem, many users trust social media influencers to bridge the gap in their lack of knowledge to later take investment decisions. As a consequence, profiling those influential actors becomes relevant.

Producing a sufficient number of high-quality annotations for author profiling is challenging. Profiling influencers, in particular, has high requirements in the economic and temporal cost, the psychological and linguistic expertise needed by the annotator, and the congenital subjectivity involved in the annotation task [3,68]. Additionally, in a real environment, i.e. when traders want to leverage social media signals to forecast the market, profiling needs to be done in real-time in a few milliseconds. This difficult, expensive, and high-speed data collection process implies data scarcity: models need to work with as little data as possible and still perform.

In this shared task, we aim to profile cryptocurrency influencers in social media from a low-resource perspective, that is, using little data. Moreover, we proposed to profile types of influencers also using a low-resource setting. Specifically, we focus on English Twitter posts for three different sub-tasks: (1) *SubTask1-Low-resource influencer profiling*: profile authors according to their degree of influence (non-influencer, nano, micro, macro, mega); (2) *SubTask2-Low-resource influencer interest profiling*: profile authors according to their main interests or areas of influence (technical information, price update, trading matters, gaming, other); and (3) *SubTask3-Low-resource influencer intent profiling*: profile authors according to the intent of their messages (subjective opinion, financial information, advertising, announcement).

Table 4. Datasets statistics including the per-class numbers of users, where the tasks are the following. SubTask1: Low-resource influencer profiling; SubTask2: Low-resource influencer interest profiling; and SubTask3: Low-resource influencer intent profiling.

Task	Partition	Total number of users per class
1	train	macro:32, mega:32, micro:30, nano:32, non-influencer:32
	test	macro:42, mega:45, micro:46, nano:45, non-influencer:42
2	train	technical information:64; trading matters:64; price update:64; gaming:64; other:64
	test	technical information:42; trading matters:112; price update:108; gaming:40; other:100
3	train	announcement:64; subjective opinion:64; financial information:64; advertising:64
	test	announcement:37; subjective opinion:160; financial information:43; advertising:52

Dataset and annotation

As in previous years, a new dataset has been created from English tweets posted by users on Twitter. We built the datasets as follows: first, we identified those who are crypto influencers, and next, we classified their interest and intent.

We identify crypto influencer candidates with two conditions: (1) user with tweets that contain the *ticker* hashtag for different crypto projects e.g. *$ETH, $BTC, $UNI etc.* ; and (2) tweets with mentions in the name of the crypto projects e.g. *Ethereum, Bitcoin, Uniswap.* Next, we extract the number of followers for those users. Finally, we use a follower scale to determine their influence grade. This scale adjusted as much as possible to the most commonly accepted definition of influencer tiers:[5]

- Non-influencer: Individuals with a minimal social media following; typically ranging from 0 to 1,000 followers. Lacks the ability to sway opinions or impact decisions through their online presence.
- Nano-influencers: Individuals with a small, dedicated social media following; typically ranging from 1,000 to 10,000 followers.
- Micro-influencers: Individuals with a moderately sized social media following ranging from 10,000 to 100,000 followers. They often have a more focused and engaged audience.
- Macro-influencers: Individuals with a substantial social media following; ranging from 100,000 to 1 million followers. They have a wide reach and may cover a broader range of topics or industries.
- Mega-influencers: Individuals with an extensive social media following; more than 1 million followers. They often have a significant impact on popular culture and possess considerable influence across multiple platforms.

For the interest and intent datasets, we applied the following criteria after the influencer identification. For each influencer, three human annotators classified

[5] https://zerogravitymarketing.com/the-different-tiers-of-influencers-and-when-to-use-each/.
https://twitter.com/latermedia/status/1385337617340829701.
https://izea.com/resources/influencer-tiers/.

Table 5. Participant and baseline results of the profiling cryptocurrency influencers shared task. Results in terms of macro F_1 for all three sub-tasks (ST), ordered by weighted average. Bold indicates the leading approach for each task.

Systems	Macro F_1		
	ST1 (Influence)	ST2 (Interest)	ST3 (Intent)
Cano-Caravaca (terra-classic)	61.14	63.15	**67.46**
Villa-Cueva et al. (stellar) [72]	58.44	**67.12**	64.46
(MRL-LLP)	57.44	62.00	65.74
Balanzá García (holo)	**62.32**	57.50	61.81
Giglou et al.(symbol) [12]	52.31	61.21	65.83
Cardona-Lorenzo (vechain)	55.51	60.16	60.28
Carbonell Granados (shiba-inu)	50.38	58.47	66.15
Ferri-Molla et al. (magic) [35]	57.14	55.68	61.62
Li et al.(neo) [29]	55.10	61.63	57.62
Iranzo Sánchez (iota)	54.43	64.55	50.62
t5-large (label tuning) - FS	49.34	56.48	59.91
Huallpa (hive)	52.94	51.48	59.08
Llanes Lacomba (api3)	49.18	46.07	63.12
Labadie et al.(dogecoin) [26]	50.80	51.72	52.59
Casamayor Segarra (tron)	50.13	49.77	53.43
user-char-lr	35.25	52.95	60.21
de Castro Isasi (terra)	48.74	44.60	54.83
Rodríguez Ferrero (harmony)	47.93	54.41	45.83
LDSE	50.20	44.92	51.96
Jaramillo-Hernández (waves)	55.06	42.35	49.21
Girish et al. [13]	37.92	46.66	50.42
Espinosa et al. (core) [9]	34.76	43.47	55.34
Coto et al. (ethereum)	46.68	–	55.94
García Bohigues (sushiswap)	46.64	19.23	22.58
t5-large (bi-encoders) - ZS	12.76	33.34	32.71
random	15.92	20.81	18.41
Kumar et al. [25]	50.21	–	–
Siino et al. (alchemy-pay) [60]	38.51	–	–
Siino et al. (nexo) [61]	38.34	–	–
Lomonaco et al. (wax) [31]	37.62	–	-
Valles Silva (solana)	15.92	–	–
Muslihuddeen et al. (icon) [38]	12.90	–	–

the interest and intent for a random tweet sample. We used majority voting to select the final class.

Table 4 presents the statistics of the datasets and the number of users for each class. Due to the low resources task nature, the number of tweets shared with our participants is small. For SubTask1 the maximum number of tweets is 10; for SubTask2 and SubTask3, the number of tweets per user is limited to 1.

On average more than 20 teams participated in each subtask. Most of the participants addressed our few-shot scenario using neural Transformers [71], including the best-performing system, which used DeBERTaV3 [16]. We compare the participants' results with different baselines covering diverse concepts such as transfer [73] and few-shot learning [8,36,37]:

- *random*: labels are randomly selected with equal probability.
- *t5-large (bi-encoders) - ZS* : Zero shot (ZS) text classification employing a t5-large model with bi-encoders [36].
- *t5-large (label tuning) - FS* : Few shot (FS) text classification employing a t5-large model with a label-tuning training strategy [36]
- *Character n-grams with logistic regression* (user-char-lr): We use [1..5] character n-grams with a TF-IDF weighting calculated using all texts.
- *Low-Dimensionality Statistical Embedding* (LDSE): This method [48] represents documents on the basis of the probability distribution of occurrence of their tokens in the different classes. The distribution of weights for a given document is expected to be closer to the weights of its corresponding class.

Results

Table 5 shows the participants' scores and baseline results. Our result analysis shows that around 46% of the final submissions outperformed our best baseline for each subtask. In addition, only one submission performed worse than the *random* baseline. Finally, the best systems could achieve an improvement of up to 10% absolute macro F_1 score compared to our best baselines.

Further details on the participants' approaches and results can be found in the task overview paper [7].

5 Trigger Detection

A trigger in psychology is a stimulus that elicits negative emotions or feelings of distress. In general, triggers include a broad range of stimuli—such as smells, tastes, sounds, textures, or sights—which may relate to possibly distressing acts or events of whatever type, for instance, violence, trauma, death, eating disorders, or obscenity. In order to proactively apprise the audience that a piece of media (writing, audio, video, etc.) contains potentially distressing material, the use of "trigger warnings"—labels indicating the type of potentially triggering content present— has become common not only in online communities but also in institutionalized education, making it possible for a sensitive audience to prepare themselves for the content and better manage their reactions. We cast this

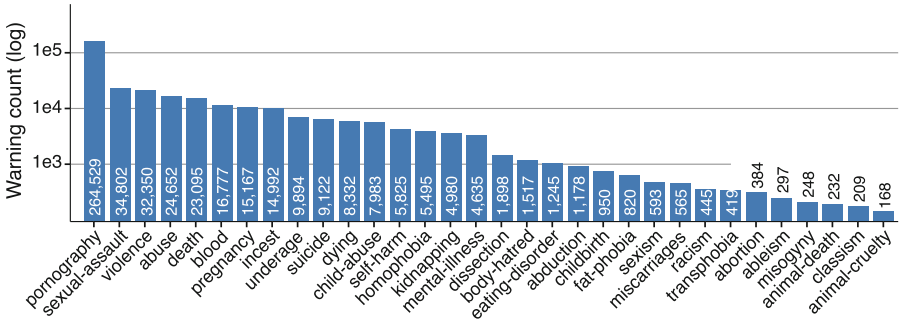

Fig. 1. Distribution of the 32 classes in the PAN23-trigger-detection dataset.

Table 6. Descriptive statistics of the training, validation, and test split of the dataset.

Training Dataset		**Validation Dataset**		**Test Dataset**	
Total Works	307,102	Total Works	17,104	Total Works	17,040
$<$ 512 words	15,233	$<$ 512 words	861	$<$ 512 words	813
$<$ 4,096 words	261,156	$<$ 4,096 words	14,571	$<$ 4,096 words	14,555
Mean no. words	2,400	Mean no. words	2,386	Mean no. words	2,388
Median no. words	2,126	Median no. words	2,115	Median no. words	2,101
90pct no. words	4,579	90pct no. words	4,550	90pct no. words	4,558

setting as a computational problem of identifying whether or not a given document contains triggering content, and if so, of what kind. In the present edition of the shared task, we asked participants to work with a corpus in which documents have been pre-tagged with content descriptors by the author (see below). Specifically, we modeled trigger detection as a multi-label document classification challenge of assigning each document all appropriate trigger warnings, but not more.

In this pilot edition of the Trigger Detection task at PAN 2023, our aim was to establish the computational problem of identifying whether or not a given document contains triggering content, and if so, of what type. As data, we created PAN23-trigger-detection, a new evaluation resource of fan fiction from Archive of our Own (Ao3) in which trigger warnings have been assigned by the authors, hence we rely on user-generated labels and follow the authors' understanding of triggers and which documents require a warning. The warnings are assigned via AO3's freeform content descriptor system ("tags"), which are custom, high-dimensional, and mostly contain non-warning descriptors, so we developed a distant-supervision strategy to detect if a freeform tag corresponds to one of 32 predefined warnings which we compiled from institutional content warning guidelines.

We formalize trigger detection as a multi-label document classification task as follows: Given a fan fiction document, assign all appropriate trigger warnings from the given set. The task is primarily evaluated with the standard measures

for multi-label classification, micro and macro F_1. In total, 6 participants submitted software to Trigger Detection 2023.

Dataset and Evaluation

For trigger detection 2023, we created a new evaluation resource, PAN23-trigger-detection, consisting of 341,246 fan fiction works downloaded by us from Archive of our Own (Ao3) and annotated in a multi-label setting with any of 32 warning labels. Figure 1 shows the distribution of the labels over the test dataset; Table 6 shows the statistics of the standard splits of our dataset.

Since there was no authoritative (closed-set) label set, we complied the 32 labels for our dataset from two institutionally-prescribed trigger lists: the University of Reading list of "themes that require trigger warnings" [54] and the University of Michigan list of content warnings [32]. The two largely overlapping guidelines list 21 categories of triggering concepts, including health-related (*eating disorders, mental illness*), sexually-oriented (*sexual assault, pornography*) as well as verbal (*hate speech, racial slurs*), and physical abuse (*animal cruelty, blood, suicide*). The lists were preprocessed to unfold compound categories into individual elements (e.g. "Animal cruelty or animal death" → "animal cruelty", "animal death") and lower-cased. The final set of trigger warnings comprises 35 categories, although we removed the rarest three labels since there were too few annotated documents with those labels in the final dataset.

We initially downloaded all ca. 10 million works released between August 13, 2008 (the platform launch) and August 09, 2021, from `archiveofourown.org` and extracted the document text and metadata (i.e. the freeform tags) from the scraped HTML. To download the HTML page of each work, we scraped the output of the search function to get the work ID and then constructed a direct URL to that works page. Since the search function was limited to 10,000 works per page, we constructed queries to search for all works released on one particular day, for each day in the release window.

We annotated all works in our corpus with appropriate trigger warnings by replacing each freeform tag assigned to the work with all corresponding warnings or removing the freeform tag if there is no corresponding warning. The underlying replacement table, which maps from freeform tag to trigger warning, was created by (1) manually annotating 2,000 most common tags, (2) efficiently identifying sub-structures of the tag graph that indicate a trigger warning, annotating each node in the structure with that warning, and (3) merging both results, giving priority to the manual annotations. This method is presented in more detail by Wiegmann et al. [75].

From the resulting corpus of annotated fan fiction works, we sampled pan23-trigger-detection by discarding all works that had no warning assigned, were originally published pre-2009 (as opposed to posted since AO3 also archives works from older fan fiction sites), had freeform tags that could not decidedly be mapped, was not in English (ca 8% of the works), had less than 50 or more than 6,000 words (outliers; ease of computation), less than 2 or more than 66 freeform tags (confidence threshold), less than 1,000 hits (views), less than 10

Table 7. Participant scores of the Trigger Detection task at PAN 2023. Shown are only the core metrics. The table is sorted by macro F_1, the primary metric. Bold indicates the leading approach for each metric.

Systems	Macro			Micro			Acc
	Prec	Rec	F_1	Prec	Rec	F_1	
Sahin et al. [57]	0.37	0.42	**0.352**	0.73	0.74	0.74	0.59
Su et al. [65]	**0.54**	0.30	**0.350**	0.80	0.71	**0.75**	**0.62**
XGBoost baseline	0.52	0.25	0.301	**0.88**	0.57	0.69	0.53
Cao, H. et al. [5]	0.24	0.29	0.228	0.43	0.79	0.56	0.18
Cao, G. et al. [4]	0.28	0.22	0.225	0.58	0.66	0.62	0.32
Felser et al. [10]	0.11	**0.63**	0.161	0.27	**0.82**	0.40	0.27
Shashirekha et al. [27]	0.10	0.04	0.048	0.82	0.50	0.63	0.52

kudos (likes; popularity threshold). We also removed all (near) duplicates. The resulting dataset (cf. Table 6) had 341,246 fan fiction works remaining, from which we stratified sampled 90:5:5 intro training, validation, and test datasets, i.e. we kept the label distribution equal across the standard splits.

We evaluate the submitted approaches through precision, recall, and F_1 at both micro- and macro average, as well as with subset accuracy, which measures accuracy on a per-example level (i.e. if all labels of one example are set correctly). We slightly favor the macro over the micro F_1 scores due to the label imbalance. We also favor recall over precision, since trigger warning assignment is a high-recall task where false negatives cause more harm than false positives. As a baseline approach, we supplied an XGBoost approach trained on TF·IDF document vectors of a max. 1,000 examples per-class random down-sampling of the training data.

Submissions and Results

The 6 submissions to trigger detection 2023 utilized a broad set of techniques, from hierarchical transformer structures to strategic feature engineering via semi-supervised topic modeling. Table 7 shows the final results, ordered by macro F_1. Most submissions focussed on improving the long document aspect of the task (most documents are longer than the input size of the SotA classification models) by using chunking and hierarchical neural networks and coping with the label imbalance (the most common label (*pornography*) is an order of magnitude more common than the other labels) by using adapted class balancing or custom loss functions. The best-performing approaches used hierarchical transformers to use the strong contextualization of the architecture while overcoming its input size limitation.

Sahin et al. submitted a hierarchical transformer architecture that achieved the top macro F_1 score (by a slim margin of 0.002) and second in micro F_1 and accuracy while having a relatively high recall within the top approaches. The

approach first segments the document into chunks (200 words with 50 words overlap) and then pre-trains a RoBERTa transformer on the chunks to learn the genre. The architecture then embeds all chunks of a document using the pre-trained transformer, followed by an LSTM for each label (in a one-vs-all setting) which predicts the class from a sequence of chunk-embeddings (RoBERTa's `[CLS]` token). To cope with the label imbalance, the approach up-scales the class weight in the loss function for the rare half of the classes.

Su et al. also submitted a siamese transformer that achieved the second-best macro F_1 score (by a slim margin of 0.002) and the top scores in micro F_1 and accuracy while notably favoring precision over recall. The approach first segments the documents into 505-word chunks encodes the first and last chunks using a pre-trained RoBERTa, mean-pools the contextual embeddings (ignoring the `[CLS]` token), and classifying based on the pooled embeddings using a 1D convolutional neural network.

Cao, H. et al. submitted a voting-based transformer that favors recall over precision. The approach segments the training documents into chunks, assigns each chunk the labels from its source document, and trains a single RoBERTa-based classifier on each chunk. To make predictions, the documents are again chunked, the labels for each chunk are predicted, and a label is assigned to the document if it is assigned to more than half of the chunks. The training data was dynamically over- and under-sampled: pornography was under-sampled to 5,000 and other labels to 2,000. Examples with rare labels were replicated 8–10 times.

Cao, G. et al. also submitted a voting-based transformer that achieved very balanced results, neither favoring macro over micro or precision over recall. The approach chunks and votes similarly to Cao, H. et al. but builds two different models to overcome the data imbalance, one for pornography and one for the other 31 classes. The pornography model was trained on a random selection of 40,000 works with and 40,000 works without the pornography warning. The model for the other labels removes works with only the pornography warning, under-samples frequent classes to 3,000 examples, and over-samples the rare labels by replicating works 4–6 times.

Felser et al. submitted a multi-layer perceptron classifier based on fasttext-based document embeddings and coarse-grained label priors determined through a combination of semi-supervised topic modeling and supervised learning. This approach achieved the top micro and macro recall, although at the cost of precision on the test dataset. The document embeddings were created by training a fasttext model from the training data, generating the embeddings for each unique word in a document, scaling those by term frequency, and adding and norming those scaled word vectors over the document. The topic modeling features were created by, first, grouping the 32 labels semantically into 6 groups, second, bootstrapping a seeded LDA with the 50 most relevant bi-grams of each group (determined through a TF·IDF-like approach for n-gram weighting which also down-grades pornographic terms), and third, training a classifier to predict

the label from the topic model, where the classifiers label confidence serves as the final feature for the MLP.

Lastly, Shashirekha et al. present an LSTM-based approach using GloVE-embeddings and which is third in micro F_1 with very high precision but rather weak in macro averages.

A more extensive evaluation and comparison of the approaches and the insights they give us into trigger detection can be found in the extended overview paper [74].

Acknowledgments. The work from Symanto has been partially funded by the Pro^{2}Haters - Proactive Profiling of Hate Speech Spreaders (CDTi IDI-20210776), the XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), the OBULEX - *OBservatorio del Uso de Lenguage sEXista en la red* (IVACE IMINOD/2022/106), and the ANDHI - ANomalous Diffusion of Harmful Information (CPP2021-008994) R&D grants.

The work of Paolo Rosso was in the framework of the FairTransNLP research project (PID2021-124361OB-C31), funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

This work has been partially supported by the OpenWebSearch.eu project (funded by the EU; GA 101070014).

References

1. Bevendorff, J., et al.: Overview of PAN 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 419–431. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_26
2. Bevendorff, J., et al.: Overview of PAN 2020: authorship verification, celebrity profiling, profiling fake news spreaders on Twitter, and style change detection. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 372–383. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_25
3. Bobicev, V., Sokolova, M.: Inter-annotator agreement in sentiment analysis: machine learning perspective. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (2017)
4. Cao, G., et al.: A dual-model classification method based on RoBERTa for trigger detection. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
5. Cao, H., et al.: Trigger warning labeling with RoBERTa and resampling for distressing content detection. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
6. Chen, H., Han, Z., Li, Z., Han, Y.: A writing style embedding based on contrastive learning for multi-author writing style analysis. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
7. Chinea-Rios, M., Borrego-Obrador, I., Franco-Salvador, M., Rangel, F., Rosso, P.: Profiling cryptocurrency influencers with few-shot learning at PAN 2023. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2023)

8. Chinea-Rios, M., Müller, T., Sarracén, G.L.D.l.P., Rangel, F., Franco-Salvador, M.: Zero and few-shot learning for author profiling. arXiv preprint arXiv:2204.10543 (2022)
9. Espinosa, D., Sidorov, G.: Using BERT to profiling cryptocurrency influencers. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
10. Felser, J., Demus, C., Labudde, D., Spranger, M.: FoSIL at PAN?23: trigger detection with a two stage topic classifier. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
11. Fröbe, M., et al.: Continuous integration for reproducible shared tasks with TIRA.io. In: Kamps, J., et al. (eds.) ECIR 2023. Lecture Notes in Computer Science, vol. 13982, pp. 236–241. Springer, Berlin (2023). https://doi.org/10.1007/978-3-031-28241-6_20
12. Giglou, H.B., Rahgouy, M., Oskuee, J.D.M.M., Tekanlou, H.B.A., Seals, C.D.: Leveraging large language models with multiple loss learners for few-shot author profiling. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
13. Girish, K., Hegdev, A., Balouchzahi, F., Lakshmaiah, S.H.: Profiling cryptocurrency influencers with sentence transformers. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
14. Guo, M., Han, Z., Chen, H., Qi, H.: A contrastive learning of sample pairs for authorship verification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
15. Hashemi, A., Shi, W.: Enhancing writing style change detection using transformer-based models and data augmentation. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
16. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=XPZIaotutsD
17. Huang, M., Huang, Z., Kong, L.: Encoded classifier using knowledge distillation for multi-author writing style analysis. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
18. Huang, Z., Kong, L., Huang, M.: Authorship verification based on CoSENT. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
19. Ibrahim, M., et al.: Enhancing authorship verification using sentence-transformers. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
20. Jacobo, G., Dehesa, V., Rojas, D., Gómez-Adorno, H.: Authorship verification machine learning methods for style change detection in texts. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
21. Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., Daelemans, W.: Authenticating the writings of Julius Caesar. Expert Syst. Appl. **63**, 86–96 (2016)

22. Kestemont, M., et al.: Overview of the author identification task at PAN 2018: cross-domain authorship attribution and style change detection. In: CLEF 2018 Labs and Workshops, Notebook Papers (2018)
23. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. J. Assoc. Inf. Sci. Technol. **65**(1), 178–187 (2014)
24. Kucukkaya, I.E., Sahin, U., Toraman, C.: ARC-NLP at PAN 23: transition-focused natural language inference for writing style detection. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
25. Kumar, A., Saeed, A.A., Trinh, L.H.M.: Profiling cryptocurrency influencers with few shot learning. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
26. Labadie, R., Sarvazyan, A.M.: Reshape or update? metric learning and fine-tuning for low-resource influencer profiling. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
27. Lakshmaiah, S.H., Hegde, A., Balouchzahi, F.: Trigger detection in social media text. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
28. Li, J., Zhang, Q., Huang, M.: Author verification of text fragments based on the Bert model. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
29. Li, Z., Han, Z., Cai, J., Huang, Z., Huang, S., Kong, L.: CLEM4PCI: profiling cryptocurrency influencers with few-shot learning via contrastive learning and ensemble model. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
30. Liu, X., Kong, L., Huang, M.: Text-segment interaction for authorship verification using BERT-based classification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
31. Lomonaco, F., Siino, M., Tesconi, M.: Text enrichment with Japanese language to profile cryptocurrency influencers. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
32. LSA, U.: An Introduction to Content Warnings and Trigger Warnings (2023). https://sites.lsa.umich.edu/inclusive-teaching-sandbox/wp-content/uploads/sites/853/2021/02/An-Introduction-to-Content-Warnings-and-Trigger-Warnings-Draft.pdf. Accessed 10 May 2023
33. Lv, J., Han, Y., Dong, Q.: Application of R-drop in author authorship verification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
34. Martinez-Galicia, J.A., Embarcadero-Ruiz, D., Rios-Orduna, A., Gómez-Adorno, H.: Graph-based siamese network for authorship verification. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)
35. Mollá, I.F., Jordà, J.S.: Profiling cryptocurrency influencers with few-shot learning. Overview for PAN at CLEF 2023. In: Aliannejadi, M., Faggioli, G., Ferro, N.,

Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
36. Mueller, T., Pérez-Torró, G., Franco-Salvador, M.: Few-shot learning with Siamese networks and label tuning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 8532–8545 (2022)
37. Müller, T., Pérez-Torró, G., Basile, A., Franco-Salvador, M.: Active few-shot learning with fasl. arXiv preprint arXiv:2204.09347 (2022)
38. Muslihuddeen, H., Sathvika, P., Sankar, S., Ostwal, S., Kumar, D.A.: Profiling cryptocurrency influencers using few-shot learning. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
39. Najafi, M., Tavan, E.: Text-to-text transformer in authorship verification via stylistic and semantical analysis. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)
40. Petropoulos, P.: Contrastive learning for authorship verification using BERT and bi-LSTM in a Siamese architecture. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
41. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's shared tasks: In: Kanoulas, E., et al. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 268–299. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_22
42. Qiu, Y., Qi, H., Han, Y., Huang, K.: Authorship verification based on SimCSE. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
43. Rangel, F., De-La-Peña-Sarracén, G.L., Chulvi, B., Fersini, E., Rosso, P.: Profiling hate speech spreaders on Twitter task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)
44. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th author profiling task at PAN 2019: profiling fake news spreaders on Twitter. In: CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (2020)
45. Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Lang. Law/Linguagem Direito **5**(2), 95–117 (2019)
46. Rangel, F., Rosso, P.: Overview of the 7th author profiling task at pan 2019: bots and gender profiling. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)
47. Rangel, F., et al.: Overview of the 2nd author profiling task at PAN 2014. In: CLEF 2014 Labs and Workshops, Notebook Papers (2014)
48. Rangel, F., Rosso, P., Franco-Salvador, M.: A low dimensionality representation for language variety identification. In: Proceedings of the CICLING, pp. 156–169 (2016)
49. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in Twitter. In: CLEF 2019 Labs and Workshops, Notebook Papers (2018)
50. Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF 2013 Labs and Workshops, Notebook Papers (2013)
51. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter. Working Notes Papers of the CLEF (2017)

52. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Labs and Workshops, Notebook Papers (2015)
53. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: CLEF 2016 Labs and Workshops, Notebook Papers (2016). ISSN 1613–0073
54. Read., U.: Guide to policy and procedures for teaching and learning; Guidance on content warnings on course content ('trigger' warnings) (2023). https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/qap/trigger-warnings.pdf. Accessed 10 May 2023
55. Reynier, O.B., Berta, C., Francisco, R., Paolo, R., Elisabetta, F.: Profiling irony and stereotype spreaders on Twitter (irostereo) at PAN 2022. In: CLEF 2021 Labs and Workshops, Notebook Papers (2022)
56. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16–new challenges for authorship analysis: cross-genre profiling, clustering, diarization, and obfuscation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16) (2016)
57. Sahin, U., Kucukkaya, I.E., Toraman, C.: ARC-NLP at PAN 2023: hierarchical long text classification for trigger detection. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
58. Sanjesh, R., Mangai, A.: A Multi-feature custom classification approach to authorship verification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
59. Sawhney, R., Agarwal, S., Mittal, V., Rosso, P., Nanda, V., Chava, S.: Cryptocurrency bubble detection: a new stock market dataset, financial task & hyperbolic models. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5531–5545 (2022)
60. Siino, M., Tesconi, M., Tinnirello, I.: Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
61. Siino, M., Tinnirello, I.: XLNet with data augmentation to profile cryptocurrency influencers. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
62. Stamatatos, E., et al.: Overview of the authorship verification task at PAN 2022. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)
63. Stamatatos, E., et al.: Overview of the authorship verification task at PAN 2023. In: CLEF 2023 Labs and Workshops, Notebook Papers, CEUR-WS.org (2023)
64. Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 518–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_49
65. Su, Y., Han, Y., Qi, H.: Siamese networks in trigger detection task. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)

66. Sun, Y., Afanaseva, S., Patil, K.: Stylometric and neural features combined deep Bayesian classifier for authorship verification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
67. Teahan, W.J., Harper, D.J.: Using compression-based language models for text categorization. In: Croft, W.B., Lafferty, J. (eds.) Language Modeling for Information Retrieval. The Springer International Series on Information Retrieval, vol. 13, pp. 141–165. Springer, Netherlands (2003). https://doi.org/10.1007/978-94-017-0171-6_7
68. Troiano, E., Padó, S., Klinger, R.: Emotion ratings: How intensity, annotation confidence and agreements are entangled. arXiv preprint arXiv:2103.01667 (2021)
69. Tschuggnall, M., et al.: Overview of the author identification task at PAN 2017: style breach detection and author clustering. In: CLEF 2017 Labs and Workshops, Notebook Papers (2017)
70. Valenzuela, A.V., Adorno, H.G., Galicia, J.A.M.: Heterogeneous-graph convolutional network for authorship verification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
71. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
72. Villa-Cueva, E., Valles-Silva, J.M., Lopez-Monroy, A.P., Sanchez-Vega, F., Lopez-Santillan, J.R.: Integrating fine-tuned language models and entailment-based approaches for low-resource tweet classification. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
73. Weiss, K., Khoshgoftaar, T.M., Wang, D.D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016). https://doi.org/10.1186/s40537-016-0043-6
74. Wiegmann, M., Wolska, M., Potthast, M., Stein, B.: Overview of the trigger detection task at PAN 2023. In: CLEF 2023 Labs and Workshops, Notebook Papers, CEUR-WS.org (2023)
75. Wiegmann, M., Wolska, M., Schröder, C., Borchardt, O., Stein, B., Potthast, M.: Trigger warning assignment as a multi-label document classification problem. In: Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada (2023)
76. Ye, Z., Zhong, C., Qi, H., Han, Y.: Supervised contrastive learning for multi-author writing style analysis. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2023)
77. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)
78. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2022. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)
79. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2023. In: CLEF 2023 Labs and Workshops, Notebook Papers, CEUR-WS.org (2023)

80. Zangerle, E., Mayerl, M., Specht, G., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2020. In: CLEF 2020 Labs and Workshops, Notebook Papers (2020)
81. Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the style change detection task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)