Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification

Condensed Lab Overview

Abinew Ali Ayele,[1] Nikolay Babakov,[2] Janek Bevendorff,[3] Xavier Bonet Casals,[4] Berta Chulvi,[5] Daryna Dementieva,[6] Ashaf Elnagar,[7] Dayne Freitag,[8] Maik Fröbe,[9] Damir Korenčić,[10] Maximilian Mayerl,[11] Daniil Moskovskiy,[12] Animesh Mukherjee,[13] Alexander Panchenko,[12] Martin Potthast,[14] Francisco Rangel,[15] Naquee Rizwan,[13] Paolo Rosso,[5,16] Florian Schneider,[1] Alisa Smirnova,[17] Efstathios Stamatatos,[18] Elisei Stakovskii, Benno Stein,[19] Mariona Taulé,[4] Dmitry Ustalov,[20] Xintong Wang,[1] Matti Wiegmann,[19] Seid Muhie Yimam,[1] and Eva Zangerle[21]

[1] Universität Hamburg, Germany, [2] Universidade de Santiago de Compostela, Spain [3] Leipzig University, Germany, [4] Universitat de Barcelona, Spain, [5] Univ. Politècnica de València, Spain, [6] Technical University of Munich, Germany, [7] University of Sharjah, United Arab Emirates, [8] SRI International, USA, [9] Friedrich Schiller University Jena, Germany, [10] Ruđer Bošković Institute, Croatia, [11] University of Applied Sciences BFI Vienna, Austria, [12] Skoltech & AIRI, Russia, [13] Indian Institute of Technology Kharagpur, India, [14] University of Kassel, hessian.AI, and ScaDS.AI, Germany [15] Symanto Research, Spain, [16] ValgrAI - Valencian Graduate School and Research Network of AI, Spain, [17] Toloka, Switzerland, [18] University of the Aegean, Greece, [19] Bauhaus-Universität Weimar, Germany, [20] JetBrains, Serbia, [21] University of Innsbruck, Austria

pan@webis.de pan.webis.de

Abstract The goal of the PAN lab is to advance the state of the art in text forensics and stylometry through an objective evaluation of new and established methods on new benchmark datasets. In 2024, we organized four shared tasks: (1) multi-author writing style analysis, which we continue from 2023; (2) multilingual text detoxification, a new task that aims to re-formulate text in a non-toxic way for multiple languages; (3) oppositional thinking analysis, a new task that aims to discriminate critical thinking from conspiracy narratives and identify their core actors; and (4) generative AI authorship verification, which formulates the detection of AI-generated text as an authorship problem. PAN 2024 concluded as one of our most successful editions with 74 notebook papers by 147 participating teams.

1 Introduction

PAN is a workshop series and a networking initiative for stylometry and digital text forensics. PAN hosts computational shared tasks on authorship analysis, computational ethics, and the originality of writing. Since the workshop's inception in 2007, we organized 73 shared tasks[1] and assembled 57 evaluation datasets[2] plus nine datasets contributed by the community. In 2024, we organized four tasks that concluded in 74 notebook papers by 147 participating teams.

First, the *Multi-Author Writing Style Analysis* task asks to, given a document, determine at which positions the author changes. This task was revamped for 2023 with a new dataset and structured around topical heterogeneity as an indicator of difficulty. We continued the task in 2024 with minor modifications since it attracts consistent participation of high technical quality and the problem is still relevant and offers room for improvements. A total of 15 teams submitted notebook papers to *Multi-Author Writing Style Analysis*. The task details are described in Section 2.

Second, the new *Multilingual Text Detoxification* task asks to, given a toxic piece of text, re-write it in a non-toxic way while saving the main content as much as possible. The task was prepared for 9 languages—English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic—and had cross-lingual and multilingual challenges. A total of 31 teams submitted their solutions to *Multilingual Text Detoxification* resulting in 12 notebook papers. The task details are described in Section 3.

Third, the new *Oppositional Thinking Analysis* task asks, given an online message, to first distinguish between critical and conspiracy texts, and second, to detect the elements of the oppositional narratives. A total of 83 teams submitted their solutions to *Oppositional Thinking Analysis* resulting in 18 notebook papers. The task details are described in Section 4.

Fourth, the new *Generative AI Authorship Verification* task asks, given one text authored by a human and one by a machine, to pick out the human-written one. Detecting AI-generated text is a task of high urgency and, as an authorship task, it falls deeply within PAN's expertise. We formulate AI-detection as a verification task and collaborate with the ELOQUENT Lab to generate a total of 70 different verification datasets to benchmark the PAN submissions. A total of 34 teams submitted to *Generative AI Authorship Verification*, resulting in 29 notebook papers. The task details are described in Section 5.

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 [32], a majority of the submissions to PAN are publicly available as docker containers. In the following sections, we briefly outline the 2024 tasks and their results.

[1] Find PAN's past shared tasks at `pan.webis.de/shared-tasks.html`
[2] Find PAN's datasets at `pan.webis.de/data.html`

2 Multi-Author Writing Style Analysis

The analysis of writing styles is the foundation of authorship identification tasks. The multi-author writing style analysis task, as part of PAN@CLEF, continues to develop challenges in this crucial field of research. Over the years, the task has evolved significantly: from identifying and grouping individual authors [108] to detecting whether a document has been written by a single or multiple authors [127, 55, 146] and identifying the actual number of authors [145], and finally, to paragraph-level style change detection [141, 142, 143].

In the PAN'24 multi-author writing style analysis task, participants were asked to identify all positions of writing style changes within a given text. Specifically, for each pair of consecutive paragraphs, the task was to compute whether there is a change in writing style between the two paragraphs. The dataset used for this task is split into three subsets of increasing difficulty: *Easy*: Each document contains a variety of topics, therefore, topic information can be used for detecting changes in writing style. *Medium:* The topics contained in a document are more homogeneous, requiring the approaches to focus more on writing style to solve the detection task. *Hard:* The paragraphs in a document are of a single topic. We control for topical diversity to ensure that, particularly in the hard dataset, topical differences cannot be used as a proxy signal for authorship and that the focus remains on stylistic cues for detecting changes in writing style.

Data Set and Evaluation

The dataset used for the multi-author writing analysis task is based on user posts on Reddit[3]. We selected posts from the following subreddits to ensure that a variety of topics is used for the creation of the datasets: *r/worldnews*, *r/politics*, *r/askhistorians*, and *r/legaladvice*. After extracting posts from these subreddits, we applied cleaning steps, such as removing quotes, whitespace, emojis, or hyperlinks. The cleaned user posts were then split into paragraphs.

To generate documents for the dataset, we used paragraphs from a single Reddit post to ensure minimal topical coherence between paragraphs of the generated document. Each document was composed of paragraphs written by a randomly selected number of two to four authors. For each paragraph, we extracted and computed semantic and stylistic feature vectors to characterize the paragraph. The paragraphs were then concatenated based on the similarity of their feature vectors. This mixing approach allowed us to control for topical and stylistic similarity, enabling the creation of more coherent documents and allowing us to adjust the difficulty of the multi-author writing style task. For the three datasets, we configured the similarity threshold for consecutive paragraphs to be (1) relatively large for the *easy* dataset, (2) moderate for the *medium* dataset, and (3) small for the *hard* dataset. Each of the easy, medium, and hard datasets contains 6,000 documents. We provided participants with training, test, and validation splits for all three datasets. The training sets contain 70% of the

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

Table 1: Overall results for the multi-author analysis task, ranked by average F_1 performance across all three datasets. Best results are marked in bold.

Team	Easy F_1	Medium F_1	Hard F_1
fosu-stu [80]	0.987	**0.887**	0.834
nycu-nlp [68]	0.964	0.857	**0.863**
no-999 [139]	**0.991**	0.830	0.832
huangzhijian [50]	0.985	0.815	0.826
text-understanding-and-analysi [46]	**0.991**	0.815	0.818
bingezzzleep [135]	0.985	0.818	0.807
openfact [63]	0.981	0.821	0.805
chen [20]	0.968	0.822	0.807
baker [134]	0.976	0.816	0.770
gladiators [56]	0.956	0.809	0.783
khaldi-abderrahmane	0.905	0.806	0.641
karami-sh [117]	0.972	0.664	0.642
riyahsanjesh [113]	0.825	0.712	0.599
liuc0757 [72]	0.696	0.717	0.503
lxflcl66666 [66]	0.606	0.455	0.484
foshan-university-of-guangdong [73]	0.517	0.394	0.352
Baseline Predict 1	0.466	0.343	0.320
Baseline Predict 0	0.112	0.323	0.346
Baseline Random	0.414	0.506	0.495

documents in each dataset, while the test and validation sets contain 15% each. The test sets were withheld for the evaluation phase of the competition.

The performance of the submitted approaches is evaluated per dataset by macro-averaged F1-score value across all documents.

Results

The task received 16 valid software submissions. The results achieved by the participants are shown in Table 1. The best average F_1 across the three datasets was achieved by the fosu-stu team. For the easy dataset, teams no-999 [139] and text-understanding-and-analysi [46] achieved the highest F_1 score (0.991), for the medium dataset, fosu-stu [80] reached an F_1 score of 0.887, and for the hard dataset, team nycu-nlp [68] achieved a F_1 of 0.863. All submissions were able to outperform the three simple baselines: a random baseline, one that predicted a style change for each pair of paragraphs, and one that predicted no style change for each pair of paragraphs. Further details on the approaches taken can be found in the overview paper [144].

3 Multilingual Text Detoxification

Text detoxification is a subtask of text style transfer where the style of text should be changed from toxic to neutral while preserving the content. As lan-

Table 2: The statistics of all ParaDetox datasets used in the TextDetox shared task. The human detoxified references were collected either via crowdsourcing or locally hired native speaker. For English and Russian, the previously collected train data was available during all shared task's phases. For other languages, 1 000 samples per language were divided correspondingly into development and test parts.

Language	**Source of Toxic Samples**	**Annotation Process**	**Train**	**Dev**	**Test**
English	[53]	Crowdsourcing+Manual	11 939	400	600
Russian	[11, 115]	Crowdsourcing+Manual	8 500	400	600
Ukrainian	[16]	Crowdsourcing	—	400	600
Spanish	[96, 124, 97]	Crowdsourcing	—	400	600
German	[133, 106, 107]	Manual	—	400	600
Hindi	[82]	Manual	—	400	600
Amharic	[8, 7]	Manual	—	400	600
Arabic	[90, 40, 87, 89]	Manual	—	400	600
Chinese	[77]	Manual	—	400	600

guage modeling advances, there is growing concern about the potential unintended consequences of this technology. One such concern is the possibility of harmful or biased texts, which could perpetuate negative stereotypes or misinformation [64]. This has led to a growing interest in AI safety and the need for approaches to mitigating these risks [17]. This presents a major challenge for researchers and practitioners in language model safety, who need to develop effective detoxification techniques that can be applied to many languages. Previously, the first parallel corpus for such a task was released for English [75] and Russian [27] that built a foundation for the RUSSE-2022 Text Detoxification shared task.

In PAN 2024, we extend our data and challenges even to more languages. The participants were asked to develop text detoxification systems for 9 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. For each language, the prepared dataset was split into two parts: (i) development and (ii) test. For the train part, we did not provide any training data except for English and Russian that was publicly available from the previous work [75, 27, 26]. Thus, in the shared task, the participants were asked to do experiments in two setups:

- **Cross-lingual setup**: In the *development* phase, participants were provided 400 toxic sentences per each language. They have to experiment with various techniques for cross-lingual detoxification.
- **Multilingual setup**: Then, in the *test phase*, we released parallel dev data and asked participants to perform detoxification on 600 samples per language (3600 instances in total). At this phase, participants were able to utilize par-

allel training corpora to improve their approaches and perform multilingual detoxification for any subset of languages.

For both phases, an automatic leaderboard was open to provide the participants scores of the adequacy and the proximity to the human references of their outputs. However, the **final** leaderboard was based on a human evaluation with crowdsourcing of subsamples from the test dataset. The human judgment gave a fair assessment of responses and prevented participants from over-tuning on automated metrics.

Data Set and Evaluation

Multilingual ParaDetox for 9 languages The full picture of the collected ParaDetox data for all target languages is presented in Table 2. While the methods of collecting human annotations vary across languages—some data were gathered via crowdsourcing, others by hiring local native speakers—the quality of the texts was uniformly verified by experts to ensure three key attributes as introduced in [28, 75]: (i) the style of new paraphrases is genuinely non-toxic, (ii) the main content is preserved, and (iii) the new texts are fluent.

For each language for the shared task's phases:

- During the *development* phase: 400 *only* toxic parts were available for participants to perform cross-lingual experiments.
- During the *test* phase: (i) 400 ParaDetox instances were fully released; (ii) participants should provide their final solutions for 600 toxic parts of the test dataset.

For English and Russian during all phases, additional training parallel datasets were available from previous work [75, 27, 26]. All the data is available online for public usage.[4]

Automatic Evaluation For both phases, we provided the leaderboard based on an automatic evaluation setup. We evaluate the outputs based on three parameters—style of text, content preservation, and conformity to human references—combining them into the final **J**oint score:

- **Style Transfer Accuracy (STA)** ensures that the generated text is indeed more non-toxic. It was estimated with XLM-R [22] `large` instance fine-tuned for the binary toxicity classification task for our target languages. The model determined the degree of non-toxicity in the texts.
- **Content Similarity (SIM)** is the cosine similarity between LaBSE embeddings [31] of the source texts and the generated texts.
- **Fluency (ChrF1)** is used to estimate the proximity of the detoxified texts to human references and their fluency.

[4] https://huggingface.co/textdetox

Human Evaluation We selected 100 random original toxic samples per each language from the *test* part of our dataset and performed human evaluation via Toloka crowdsourcing platform.[5] The concept of the human evaluation mirrored the approach used in the automatic evaluation. Each project type focused on assessing one of the three key qualities of detoxification; style transfer accuracy, content similarity, or fluency:

- **Style Transfer Accuracy**: we employed a pairwise comparison between the original toxic text and the generated detoxified text. Participants were tasked with determining which text was more toxic: the left text, the right text, or neither.
- **Content Similarity**: participants were shown pairs of texts (toxic phrase followed by detoxified phrase) and asked to indicate if the sense was similar, responding with "yes" or "no".
- **Fluency**: individual sentences were evaluated for intelligibility and correctness. Annotators could respond with "yes", "partially", or "no", corresponding to scores of 1, 0.5, and 0, respectively. The fluency score for a text pair was determined by comparing the detoxified text's score to the original. If the detoxified text had a higher or equal fluency score, the pair received a 1; otherwise, it received a 0.

Final Joint Score (J) For both automatic and human evaluation setups, the J score was the aggregation of the three above metrics. The metrics **STA**, **SIM** and **FL** were subsequently combined into the final **J** score used for the final ranking of approaches. Given an input toxic text x_i and its output detoxified version y_i, for a test set of n samples:

$$\mathbf{J} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{FL}(x_i, y_i),$$

where $\mathbf{STA}(y_i)$, $\mathbf{SIM}(x_i, y_i)$, $\mathbf{FL}(x_i, y_i) \in [0, 1]$ for *automatic* and $\in \{0, 1\}$ for *human* evaluation for each text detoxification output y_i.

We calculated all the metrics separately per each language. In the end, we calculated the **Average** score of 9 **J**oint scores per all languages that were used to compile the leaderboard.

Results

We received 20 submissions for the development phase leaderboard and 31 submissions for the test phase leaderboard; the final manually evaluated leaderboard was based on 17 submissions who confirmed their participation in the competition [93, 130, 110, 149, 95, 78, 34, 123, 91, 63, 105, 102, 99]. The final leaderboard based on human assessments is presented in Table 3.

Almost all of the participants used the current SOTA LLMs, among which are GPT-3.5 [92] and Llama-3 [3] models; to enhance the model's performance

[5] https://toloka.ai

Table 3: Results of the *human* final evaluation of the TextDetox test phase. Scores are sorted by the average **J**oint score. Scores are sorted by the average **J**oint score across all 9 languages. Baselines are highlighted with gray, Human References are highlighted with green.

Team	**Avg**	**System**
Human References	0.851	Human paraphrases from our multilingual ParaDetox
SomethingAwful	0.774	Few-shot LLaMa-3 prompting+mT0-XL
adugeen	0.741	Fine-tuned mT0-XL with ORPO [43]
VitalyProtasov	0.723	Preprocessing+mT0-large
nikita.sushko	0.712	Fine-tuned mT0-XL+postprocessing
erehulka	0.708	Few-shot LLaMa-3 prompting
bmmikheev	0.685	Few-shot LLaMa-3 prompting+GPT-3.5 post-eval.
mkrisnai	0.681	Few-shot GPT-3.5 prompting
d1n910	0.654	Few-shot Kimi.AI prompting
Yekaterina29	0.639	Fine-tuned mT5-XL
estrella	0.576	Tree of Thought GPT3.-5 prompting
gleb.shnshn	0.564	Zero-shot LLaMa-3-70b prompting
Delete	0.560	Elimination of toxic keywords
mT5	0.541	Fine-tuned mT5-XL
shredder67	0.524	Fine-tuned mT5-XL
razvor	0.516	Few-shot LLaMa-3 prompting
ZhongyuLuo	0.513	Translation+BART-detox&ruT5-detox
gangopsa	0.500	Fine-tuned T5&BART+token-level editing
Backtranslation	0.411	Translation of data to English+BART-detox
maryam.najafi	0.177	Mistral-7b with PPO
dkenco	0.119	Few-shot Cotype-7b prompting

on the task of detoxification participants tested both zero-shot and few-shot prompting methods. Among smaller models, there were used mT5 [137] and mT0 [88]—these models were usually finetuned using ad hoc filtering and data augmentation techniques, for instance, as RAG and backtranslation. Additionally, region-specific LLMs were also employed: Cotype-7b [86] and Kimi.AI [2].

The majority of the participants overcame the baselines and even a couple of solutions outperformed human references. Still, for not-so-rich-resource languages such as Ukrainian, Chinese, Amharic, and Hindi human detoxified paraphrases remained the gold standard. At the same time, various experiments from participants illustrate that vanilla usage of LLMs for the detoxification task does not achieve high results. At least more advanced prompting techniques and fine-tuning on the downstream task with our provided data boosted the performance significantly achieving such interesting SOTA results.

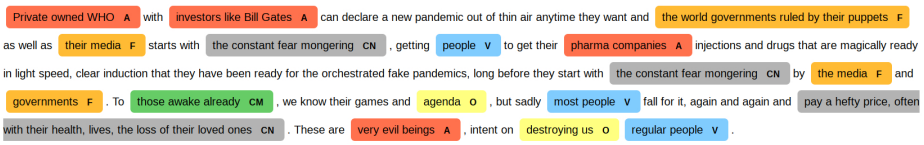

Figure 1: A Telegram text annotated with elements of oppositional narrative.

4 Oppositional Thinking Analysis: Conspiracy Theories vs Critical Thinking Narratives

Conspiracy theories are complex narratives that attempt to explain the ultimate causes of significant events as cover plots orchestrated by secret, powerful, and malicious groups [29]. A challenging aspect of identifying conspiracy with NLP models [109, 100, 101, 35, 62, 33] stems from the difficulty of distinguishing critical thinking from conspiratorial thinking in automatic content moderation. This distinction is vital because labeling a message as conspiratorial when it is only oppositional could drive those who were simply asking questions into the arms of the conspiracy communities.

At PAN 2024 we aim at analyzing oppositional thinking, and more concretely, at discriminating conspiracy from critical narratives from a *stylometry* perspective. The task will address two new challenges for the NLP research community: (1) to distinguish the conspiracy narrative from other oppositional narratives that do not express a conspiracy mentality (i.e., critical thinking); and (2) to identify in online messages the key elements of a narrative that fuels the inter-group conflict in oppositional thinking. Accordingly, we propose two sub-tasks:

- **Subtask 1** is a binary classification task differentiating between (1) critical messages that question major decisions in the public health domain, but do not promote a conspiracist mentality; and (2) messages that view the pandemic or public health decisions as a result of a malevolent conspiracy by secret, influential groups.

- **Subtask 2** is a token-level classification task aimed at recognizing text spans corresponding to the key elements of oppositional narratives. Since conspiracy narratives are a special kind of causal explanation, we developed a span-level annotation scheme that identifies the goals, effects, agents, and the groups-in-conflict in these narratives.

For the second task, a new fine-grained annotation scheme was developed with the goal of identifying, at the text span level, how oppositional and conspiracy narratives use inter-group conflict. The annotation was performed for the described 5,000 binary-labeled messages per language. We identify the following six categories of narrative elements at the span level (see Figure 1):

- **Agents**: the hidden power that pulls the strings of the conspiracy. In critical messages, agents are actors that design the mainstream public health policies: Government, WHO, . . . ;

Table 4: Overall results for subtask 1 on Conspiracy theories vs Critical thinking narratives in English (EN) in terms of Matthew's correlation coefficient (MMC).

Team	EN MCC	Team	EN MCC	Team	EN MCC
IUCL	0.838	nlpln	0.784	lnr-lladrogal	0.725
AI_Fusion	0.830	RalloRico	0.777	lnr-fanny-nuria	0.725
SINAI	0.829	LasGarcias	0.775	MarcosJavi	0.719
ezio	0.821	zhengqiaozeng	0.775	lnr-cla	0.716
hinlole	0.819	ALC-UPV-JD-2	0.772	lnr-jacobant.	0.716
Zleon	0.819	LorenaEloy	0.771	MUCS	0.716
virmel	0.819	lnr-alhu	0.770	lnr-aina-julia	0.715
inaki	0.814	NACKO	0.769	LaDolceVita	0.707
yeste	0.812	paranoia-pulverizers	0.768	alopfer	0.705
auxR	0.808	DiTana	0.765	lnr-luqrud	0.705
Elias&Sergio	0.803	FredYNed	0.764	LNR-JoanPau	0.705
theateam	0.803	dannuchihaxxx	0.764	lnr-carla	0.700
trustno1	0.798	lnr-detectives	0.763	lnr-Inetum	0.698
DSVS	0.797	TargaMarhuenda	0.761	lnr-antonio	0.685
sail	0.796	Trainers	0.759	LluisJorge	0.678
ojo-bes.	0.796	thetaylorswift	0.757	anselmo-team	0.672
RD-IA-FUN	0.796	locasporlnr	0.757	lnr-pavid	0.595
Baseline BERT	0.796	lnr-adri	0.755	LNRMADME	0.546
aish_team	0.791	TokoAI	0.754	lnr-mariagb.	0.506
rfenthusiasts	0.790	ede	0.753	LNR_08	0.442
Dap_upv	0.789	lnr-verdnav	0.752	Kaprov	0.370
oppositional_opposition	0.789	lnr-dahe	0.748	lnr_cebusqui	0.048
RD-IA-FUN	0.789	epistemologos	0.748	jtommor	0.040
miqarn	0.788	lucia&ainhoa	0.747	eledu	0.459
CHEEXIST	0.787	pistacchio	0.741	david-canet	0.631
tulbure	0.787	lnr-BraulioP.	0.739	lnr-guilty	0.659
XplaiNLP	0.787	Marc_Coral	0.739	lnrANRI	0.755
TheGymNerds	0.785	Ramon&Cajal	0.728	ROCurve	0.800

- **Objectives**: parts of the narrative that answer the question "What is intended by the agents of the conspiracy theory or by the promoters of the action being criticized from a critical thinking perspective?";
- **Consequences**: parts of the narrative that describe the effects of the agent's actions;
- **Facilitators**: the facilitators are those who collaborate with the conspirators; in critical messages, facilitators are those who implement the measures dictated by the authorities;
- **Campaigners**: in conspiracy messages, the campaigners are the ones who uncover the conspiracy theory; in critical messages, campaigners are those who resist the enforcement of laws and health instructions; and
- **Victims**: the people who are deceived into following the conspiratorial plan or the ones who suffer due to the decisions of the authorities.

Table 5: Overall results for subtask 1 on Conspiracy theories vs Critical thinking narratives in Spanish (ES) in terms of Matthew's correlation coefficient (MMC).

Team	ES MCC	Team	ES MCC	Team	ES MCC
SINAI	0.742	NACKO	0.646	Ramon&Cajal	0.58
auxR	0.720	ALC-UPV-JD-2	0.646	lnr-fanny-nur.	0.58
RD-IA-FUN	0.702	DSVS	0.646	lnr-antonio	0.57
Elias&Sergio	0.697	RD-IA-FUN	0.644	LluisJorge	0.56
AI_Fusion	0.687	locasporlnr	0.643	lnr-cla	0.56
zhengqiaozeng	0.687	DiTana	0.637	lnr-jacobant.	0.56
virmel	0.685	lnr-BraulioPaula	0.635	lnr-pavid	0.55
trustno1	0.684	Dap_upv	0.630	alopfer	0.55
Zleon	0.682	TheGymNerds	0.630	LNRMADME	0.54
ojo-bes	0.681	MUCS	0.629	lnr-carla	0.54
tulbure	0.672	LasGarcias	0.624	LorenaEloy	0.54
sail	0.671	lnr-dahe	0.619	CHEEXIST	0.53
nlpln	0.668	lnr-adri	0.619	lnr-guilty	0.52
Baseline BERT	0.668	hinlole	0.619	eledu	0.50
pistacchio	0.667	RalloRico	0.610	lnr-mariagb.	0.49
rfenthusiasts	0.665	lnr-aina-julia	0.61	dannuchihaxxx	0.47
XplaiNLP	0.662	lnr-verdnav	0.61	lnr-detectives	0.40
yeste	0.660	thetaylorswift	0.60	LNR_08	0.06
oppositional_opposition	0.660	lnr-alhu	0.60	jtommor	0.01
epistemologos	0.656	lnr-luqrud	0.60	lnr-Inetum	0.00
miqarn	0.656	lnr-lladrogal	0.59	Marc_Coral	0.00
theateam	0.655	ede	0.59	MarcosJavi	-0.03
ezio	0.653	Fred&Ned	0.59	lnr_cebusqui	-0.41
lucia&ainhoa	0.652	LaDolceVita	0.59	david-canet	-0.50
TargaMarhuenda	0.651	LNR-JoanPau	0.59	lnrANRI	-0.61
TokoAI	0.651	anselmo-team	0.58	ROCurve	-0.64
paranoia-pulver.	0.649				

Data Set and Evaluation

For the creation of the corpus, we first manually compiled a list of 2,273 public Telegram channels in **English** and **Spanish** that contain oppositional non-mainstream views on the COVID-19 pandemic. We retrieved and filtered messages from the channels based on a set of oppositional and conspiracy keywords related to COVID-19. Then the messages were cleaned by removing duplicates, short texts, and texts with a large proportion of non-regular words (such as URLs and mentions). Finally, the messages were ranked using an index of quality based on the properties of a message and its channel. The index is composed of several criteria capturing the prevalence of COVID-19 topics and the channel's activity.

We developed an annotation schema to differentiate between the messages criticizing the mainstream views on COVID-19 and the messages evoking the existence of a conspiracy. A message was labeled "conspiracy" if any of these four criteria were met: (1) it framed COVID-19 or a related public health strategy as the result of the agency of a small and malevolent secret group; (2) it claimed that the pandemic is not real (e.g. a plandemic); (3) it accused critics of the

Table 6: Overall results for subtask 2 on the Text-span recognition of elements of oppositional narratives, in English (EN) and Spanish (ES), in terms of macro-averaged span-F1

Team	EN span-F1
tulbure	0.6279
Zleon	0.6089
hinlole	0.5886
oppositional_opposition	0.5866
AI_Fusion	0.5805
virmel	0.5742
miqarn	0.5739
TargaMarhuenda	0.5701
ezio	0.5694
zhengqiaozeng	0.5666
Elias&Sergio	0.5627
DSVS	0.5598
CHEEXIST	0.5524
rfenthusiasts	0.5479
ALC-UPV-JD-2	0.5377
Baseline BETO	0.5323
Dap_upv	0.5272
aish_team	0.5213
SINAI	0.4582
Trainers	0.3382
nlpln	0.3339
ROCurve	0.2996
TokoAI	0.2760
DiTana	0.2756
TheGymNerds	0.2070
epistemologos	0.1709
theateam	0.1503
LaDolceVita	0.0726
kaprov	0.0150

Team	ES span-F1
tulbure	0.6129
Zleon	0.5875
AI_Fusion	0.5777
CHEEXIST	0.5621
virmel	0.5616
miqarn	0.5603
DSVS	0.5529
TargaMarhuenda	0.5364
Elias&Sergio	0.5151
hinlole	0.4994
Baseline BETO	0.4934
Dap_upv	0.4914
zhengqiaozeng	0.4903
ALC-UPV-JD-2	0.4885
ezio	0.4869
nlpln	0.4672
rfenthusiasts	0.4666
SIANI	0.4151
TheGymNerds	0.3984
DiTana	0.3004
ROCurve	0.2649
TokoAI	0.1878
epistemologos	0.1657
LaDolceVita	0.1056
theateam	0.0994
oppositional_opposition	0.0037

conspiracy theory of being a part of the plot; (4) it divided society into two: those who know the truth (the conspiracy theorists) and those who remain ignorant. A message was labeled "critical" if it opposed publicly accepted understandings of events but had none of these four characteristics of the conspiratorial mindset.

Using this annotation scheme, 5,000 messages per language were annotated as "conspiracy" or "critical" thinking. For these messages, we performed anonymization by removing sensitive and identifiable information such as nicknames, user IDs, and e-mail addresses. The average text length is 128 tokens for Spanish texts and 265 tokens for English texts that tend to elaborate more on conspiracy theories.

Each message was annotated by three linguists and the inter-annotator agreement (IAA) was calculated. Disagreements were discussed with the social psychologist who created the annotation scheme. For English messages, the IAA in terms of Krippendorf's α is 0.79 for "conspiracy" messages and 0.60 for "criti-

cal" messages, while the average observed percentage of agreement between the three annotators is 91.4%, and 80.3%, respectively. For Spanish messages, Krippendorf's α is 0.80 for "conspiracy" messages and 0.70 for "critical" messages, corresponding to the percentage agreements of 90.9% and 84.9%.

For the second task, a new fine-grained annotation scheme was developed with the goal of identifying, at the text span level, how oppositional and conspiracy narratives use intergroup conflict. The annotation was performed for the described 5,000 binary-labeled messages per language.

In the process of span-level annotation, each of the 5,000 Spanish and English messages were annotated by two linguists. Currently, the annotation instructions are being discussed and improved and, to this end, we are using the Gamma (γ) measure of the IAA test [83], yielding a first average γ of 0.43. The following batch had an average gamma of 0.53, and the last one had a γ of 0.61. We deemed this a good agreement because it is close to or above the average agreement of other highly conceptual span-level schemes [24, 132]. A detailed description of the dataset can be found in [60].

The official evaluation metric for subtask 1 (critical vs. conspiracy classification) is Matthew's correlation coefficient (MCC) [21], while the official metric for subtask 2 (span-level detection of narrative elements) is macro-averaged span-F1 [23].

Results

A total of 83 teams submitted their runs for subtasks 1 and 2, resulting in 18 notebook submissions [51, 131, 44, 9, 25, 112, 4, 150, 30, 111, 36, 6, 81, 147, 47, 128, 71]. In the tables above we illustrate the ranking per language. Concretely, Table 4 and Table 5 show the overall results obtained for subtask 1 on Conspiracy theories vs critical thinking narratives, in terms of Matthew's correlation coefficient; while Table 6 shows the results of subtask 2 on Text-span recognition of elements of oppositional narratives, in terms of macro-averaged span-F1.

We will analyze in detail the results and describe the models of the participants in the task overview paper [61].

5 Voight-Kampff Generative AI Authorship Verification

Authorship verification is a fundamental task in author identification. All cases of questioned authorship can be decomposed into a series of verification instances, be it in a closed-set or open-set scenario [59]. Since PAN has been continuously organizing Authorship verification tasks [119, 13, 12, 118], we are well-equipped to tackle a timely and highly important issue: identification of machine authorship in contrast to human authorship.

Authorship identification of generative AI "in the wild" where a single document is disputed without reference is an open-set problem and the hardest

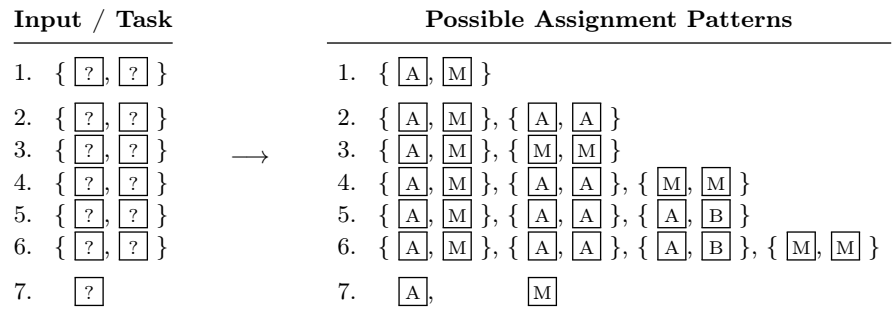

Figure 2: Hierarchy of authorship verification problems from "easiest" (1) to "hardest" (7), involving LLM-generated text. Ignoring mixed human and machine authorship, the difficulty arises from the pairing constraints imposed by the possible assignment patterns. [M] denotes LLM-generated text, while [A] and [B] denote human-authored text (same letter meaning same human author).

formulation of the task. Although the literature suggests limited success in solving this problem given the current generation of LLMs, it is questionable whether this will remain so with improving technology. Setting aside mixed human and machine authorship, we have broken down all possible formulations of the problem with increasing levels of difficulty to get a more fundamental understanding of the task at hand and the feasibility of potential solutions. Figure 2 visualizes the cascade of all problem variants from easiest (Task 1) to most difficult (Task 7). In the easiest case, two documents with unknown authorship are given, yet we guarantee that exactly one is generated by a human [A], and the other by a machine [M], respectively. This constraint is relaxed in the following variants where, for example, both texts may also stem from a machine, {[M], [M]}. In the hardest case, a single text is given, which could be either [A] or [M].

For the 2024 task on "Generative AI Authorship Verification," we follow the "easiest" formulation of the task in order to establish a feasibility baseline. The task description reads: "Given two texts, one authored by a human, one by a machine: pick out the human. "

The task is organized in collaboration with the ELOQUENT Lab [54] in a builder-breaker style, in which PAN participants build systems to identify machine authorship, while ELOQUENT participants supply datasets trying to break the systems.

Data Set

In addition to the ELOQUENT-provided data, we collected 1,359 articles of major 2021 U.S. news headlines from Google News. We chose this time period specifically as it predates the release of GPT-3.5 so that we could be reasonably certain the articles were actually human-authored. We used GPT-4-Turbo to generate a bullet-point summary of each article and the summaries were then

Table 7: Overview of the 65 dataset variants provided as baseline datasets. All variants contain the same 271 human texts and (roughly) one machine generated text per LLM used. Discarding erroenous generations, this results in 3,441 pairings each for main and cross-domain variants, 600 for both unicode variants and short texts, 543 for german texts, 542 for the Kaggle prompt, 272 for both contrastive decoding (* using Llama2-13B).

	ChatCat	**Bloomz**	**Gemini Pro** *with temp.*		**Text-Bison**	**GPT**			**Llama2**		**Mistral**		**Qwen-1.5**
Variation / Obfuscation	7B	7B	0.6	0.9	002	2-OI	3.5	4	7B	70B	7B	8x7B	72B
Main	x	x	x	x	x	x	x	x	x	x	x	x	x
Unicode sub. (machine)			x	x		x	x	x	x	x			
Unicode sub. (both)			x	x		x	x	x	x	x			
Cross-domain	x	x	x	x	x	x	x	x	x	x	x	x	x.
Short text	x	x	x	x	x	x	x	x	x	x	x	x	x
German text (machine)					x			x					
Contr. decoding ($\alpha = 0.1$)										x*			
Contr. decoding ($\alpha = 0.6$)										x*			
Kaggle prompt			x					x					

given to a selection of 13 downstream large language models to write new articles from them.

Of the original 1,359 human-authored articles, participants were given 1,087 together with their machine counterparts from 13 LLMs to calibrate their systems. The remaining 272 articles and generations from 15 LLMs were kept back for testing, resulting in 3,984 test cases, which together form the "main" portion of the test set.

To further test the robustness of the submitted systems, we generated multiple variants of the original pairs. In particular, we: (1) amended the prompt to generate German instead of English texts (this was already part of the "main" test set, but not communicated to the participants); (2) replaced 15 % of the characters in (a) the machine texts and (b) both the human and machine texts with Unicode lookalike characters; (3) shuffled the test case pairs to break the topic coherence; (4) used contrastive decoding [121] instead of top-k / top-p sampling; (5) cropped texts to 35 words; and (6) used the prompt from a previous Kaggle competition on LLM detection [57] to generate more faithful paraphrases of the original articles, instead of using the stripped-down bullet point summaries.

In total, we created 65 test set variations from 13 (15) different LLMs, which are summarized in Table 7, with ELOQUENT providing another five. A more detailed description is available in the joint task overview paper [15].

Evaluation

At test time, participants were given pairs of human and LLM texts and had to calculate a score between 0 and 1, indicating which text was more likely to be human-authored. Scores less than 0.5 mean the left text is human and scores greater than 0.5 mean the right text is human. A score of exactly 0.5 could be given to signal a non-decision. We borrowed this evaluation scheme from previous installments of the PAN Authorship Verification Task.

We rank systems by their macro-average effectiveness across all $n = 70$ dataset variants (including ELOQUENT submissions) discounted by half a standard deviation (estimated from the scores with $n - 1$ DoF), which penalizes unstable systems that are not robust against text obfuscations or other text variations. We use the macro average over datasets since all datasets have different numbers of examples, yet we consider them equally important as performance indicators.

Also in line with previous task installments, we compute the effectiveness for each dataset variant as the average of the established evaluation measures in authorship verification (all with comparable 0–1 scales). In particular:

- Roc-Auc: The area under the Receiver Operating Characteristic curve.
- Brier: The complement of the Brier score (mean squared loss)
- C@1: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.
- F_1: The harmonic mean of precision and recall.
- $F_{0.5u}$: A modified $F_{0.5}$ measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives.

Submitted Systems

In total, our task attracted 34 teams to submit systems in addition to the baseline systems we provided. Table 8 shows the best-performing system of each team that submitted notebook papers and a brief description of their approach.

Baselines We provided implementations of six baseline systems to compare submitted systems against four state-of-the-art zero-shot LLM detection baselines and two adapted authorship verification baselines.

The zero-shot LLM detection baselines are: (1) Binoculars [42], (2) DetectLLM (both NPR and LRR scoring mode), (3) DetectGPT [85], and (4) Fast-DetectGPT [10]. All three were provided in two variants using either Falcon-7B [5] or Mistral-7B [52] to estimate text perplexities. The required text perturbations for DetectGPT and DetectLLM-NPR were generated with T5-3B [104].

The two authorship verification baselines were adapted to the LLM detection task by splitting each text in half and comparing the two halves against each other under the assumption that LLM texts are stylistically more self-similar than human texts. The baselines provided are a compression model (PPMd CBC) [114, 41] and short-text authorship unmasking [58, 14].

Table 8: The score is the mean of all evaluation measures across all other metrics on the main dataset corrected by half a standard deviation to correct for spread.

Team	Score	System
Tavan [125]	0.924	Ensemble: LoRA-trained LLM + Binoculars
J. Huang [46]	0.921	BERT with multiscale PU loss [126]
Lorenz [76]	0.886	SVM with TF-IDF features
M. Guo [39]	0.884	LSTM embeddings + GPT-2 PPL
Z. Lin [69]	0.851	Finetuned BERT + R-Drop
Abburi [1]	0.843	Ensemble: RoBERTa + E5 + GPT-2 Perplexity
Miralles [84]	0.806	Entropy and text features + XGBoost
Yadagiri [138]	0.806	Finetuned BERT + linguistic features
Lv [79]	0.804	Finetuned DeBERTa with Reptile meta learning
Gritsai [37]	0.796	Ensemble: LoRA-trained LLMs
Cao [18]	0.778	*Finetuned BERT*
L. Guo [38]	0.763	BERT and text features + Bi-LSTM
Binoculars 1	0.741	*Baseline Binoculars (Falcon-7B) [42]*
B. Huang [45]	0.735*	Finetuned BERT + R-Drop [67]
Valdez-Valenzuela [129]	0.727*	Graph Neural Network + BERT
Ye [140]	0.722	T5 with LM head trained to predict class
Chen [19]	0.694	Ensemble: 2x BERT + GPT-2 (PPL)
W. Huang [49]	0.683	Perplexity of GPT-2 trained on LLMs + SVM
Qin [103]	0.680*	Ensemble: BERTs + R-Drop
Binoculars 2	*0.671*	*Baseline Binoculars (Mistral-7B) [42]*
DetectLLM 1	*0.654*	*Baseline DetectLLM LRR (Mistral-7B) [120]*
Petropoulos [98]	0.641	RoBERTa embeddings + Bi-LSTM
Fast-DetectGPT 1	*0.638*	*Baseline Fast-DetectGPT (Mistral-7B) [10]*
Wu [136]	0.608	BERT embeddings + extra Transformer block
Text Length	*0.604*	*Baseline Text length*
Z. Lin [70]	0.565	T5 with LM head trained to predict class
Zhu [148]	0.555	Finetuned DeBERTa
PPMd CBC	*0.544*	*Baseline PPMd Compression-based Cosine [114, 41]*
Sun [122]	0.531	BERT embeddings + CNN
DetectLLM 2	*0.512*	*Baseline DetectLLM NPR (Mistral-7B) [120]*
Lei [65]	0.504	LoRA-trained ChatGLM
Fast-DetectGPT 2	*0.500*	*Baseline Fast-DetectGPT (Falcon-7B) [10]*
Liu [74]	0.497	Preplexity of pre-trained GPT-2
DetectGPT 1	*0.488*	*Baseline DetectGPT (Mistral-7B) [85]*
K. Huang [48]	0.480	Siamese DeBERTa
DetectLLM 3	*0.468*	*Baseline DetectLLM NPR (Falcon-7B) [120]*
Unmasking	*0.467*	*Baseline Authorship Unmasking [58, 14]*
Sheykhlan [116]	0.460	Ensemble: BERT, RoBERTa, and Electra
DetectLLM 4	*0.460*	*Baseline DetectLLM LRR (Falcon-7B) [120]*
DetectGPT 2	*0.439*	*Baseline DetectGPT (Falcon-7B) [85]*
Ostrower [94]		[No software submitted]

* Scores estimated due to run failures on some dataset variants.

As an additional seventh baseline, we measured and compared the text lengths in characters. This baseline serves as both a quasi-random baseline and as a data sanity check.

Participant Systems While our baseline systems reproduce established methods in either authorship verification or intrinsic, zero-shot LLM detection, the participant systems cover a broad range of approaches. The most popular approach is to use a BERT-based classifier with some modification (like PU loss or R-Drop), bagging, and / or expansion of the given training data with other LLM detection datasets. Some systems use engineered features like perplexity, properties of token distributions, or stylometrics (exclusively or in addition to BERT-embeddings) as classifier (Linear, XGBoost, LSTM) inputs. Most of these classification methods apply a posterior comparison of scores similar to how we use Binoculars, although some participants also train models to directly discriminate between the pairings. In some cases, participants also developed zero-shot methods and adapted LLMs directly for the detection task, often using LoRa.

Results

Table 8 shows the ranking scores of the best system submitted by each participating team and the baselines. In total, 10 teams surpassed all baselines. The overall best submission (by Tavan and Najafi; mean score of 0.924) finetunes Mistral and Llama2 models, combining them into an ensemble with the Binoculars baseline [42]. This approach beats the original baseline by 0.183 points, though there appears to be no general best strategy for AI detection. The top 5 systems are a mixture of zero-shot perplexity estimators and supervised blackbox classifiers based on BERT or even linear classifiers.

On the individual datasets, we see that almost all submissions perform quite well on non-obfuscated text (ROC-AUC > 0.9). We must therefore conclude that even the most advanced LLMs still exhibit obvious stylistic idiosyncrasies which make their texts easy to distinguish from human ones. However, none of the systems is entirely robust against (unexpected) obfuscations and particularly short text samples are a big challenge for all systems. Some systems did not produce any output on the short texts due to a programming problem. For the final evaluation, the missing values were filled with the corresponding mean values from all other systems. Affected systems are marked with * in Table 8.

A more detailed description and analysis of the submissions and the results can be found in the joint PAN and ELOQUENT task overview paper [15].

Acknowledgments

The work of Paolo Rosso, Damir Korenčić, and Berta Chulvi was in the framework of XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), funded by

MCIN/AEI / 10.13039/501100011033 and by the European Union NextGenerationEU / PRTR.

The work from Symanto has been partially funded by XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), Pro^{2}Haters – Proactive Profiling of Hate Speech Spreaders (CDTi IDI-20210776), 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 Janek Bevendorff, Matti Wiegmann, Maik Fröbe, Martin Potthast, and Benno Stein has been funded as part of the OpenWebSearch project by the European Commission (OpenWebSearch.eu, GA 101070014).

Bibliography

[1] Abburi, H., Pudota, N., Veeramani, B., Bowen, E., Bhattacharya, S.: Team Deloitte at PAN: Generative AI Text Detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[2] AI, M.: Kimi chatbot (2024), URL https://kimi.moonshot.cn, accessed: 2024-05-31
[3] AI@Meta: Llama 3 model card (2024), URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
[4] Albladi, A., Seals, C.: Detection of Conspiracy vs. Critical Narratives and Their Elements using NLP. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[5] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., Penedo, G.: The Falcon series of open language models. arXiv [cs.CL] (28 Nov 2023)
[6] Ansari, T., Ghazi, T., Alvi, F., Samad, A.: Decoding COVID-19 Narratives: Conspiracy or Critique? Working Notes of CLEF 2024, CEUR-WS.org (2024)
[7] Ayele, A.A., Dinter, S., Belay, T.D., Asfaw, T.T., Yimam, S.M., Biemann, C.: The 5Js in Ethiopia: Amharic hate speech data annotation using Toloka Crowdsourcing Platform. In: Proceedings of the 4th International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp. 114–120, Bahir Dar, Ethiopia (2022), URL https://ieeexplore.ieee.org/document/9971189
[8] Ayele, A.A., Yimam, S.M., Belay, T.D., Asfaw, T., Biemann, C.: Exploring Amharic hate speech data collection and classification approaches. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, (Sep 2023), URL https://aclanthology.org/2023.ranlp-1.6
[9] Balasundaram, P., Swaminathan, K., Sampath, O., Km, P.: Oppositional Thinking Analysis: Conspiracy Theories vs Critical Thinking Narratives. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[10] Bao, G., Zhao, Y., Teng, Z., Yang, L., Zhang, Y.: Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. arXiv [cs.CL] (8 Oct 2023)
[11] Belchikov, A.: Russian language toxic comments. https://www.kaggle.com/blackmoon/russian-language-toxic-comments (2019), accessed: 2023-12-14
[12] Bevendorff, J., Chulvi, B., la Peña Sarracén, G.L.D., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., Zangerle, E.: Overview of PAN 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Springer (2021)
[13] Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Pardo, F.M.R., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Wiegmann, M., Zangerle, E.: Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Springer (2020)
[14] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Generalizing unmasking for short texts. In: Proceedings of the 2019 Conference of the North, pp. 654–659, Association for Computational Linguistics, Stroudsburg, PA, USA (2019), https://doi.org/10.18653/v1/n19-1068

[15] Bevendorff, J., Wiegmann, M., Karlgren, J., Dürlich, L., Gogoulou, E., Talman, A., Stamatatos, E., Potthast, M., Stein, B.: Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. Working Notes of CLEF 2024, CEUR Workshop Proceedings (2024)
[16] Bobrovnyk, K.: Automated building and analysis of ukrainian twitter corpus for toxic text detection. In: COLINS 2019. Volume II: Workshop (2019), URL https://ena.lpnu.ua:8443/server/api/core/bitstreams/c4c645c1-f465-4895-98dd-765f862cf186/content
[17] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H.S., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D.: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. CoRR **abs/1802.07228** (2018)
[18] Cao, H., Han, Z., Ye, J., Liu, B., Han, Y.: Enhancing Human-Machine Authorship Discrimination in Generative AI Verification Task with BERT and Augmented Data. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[19] Chen, J., Kong, L.: Integrating Dual BERT Models and Causal Language Models for Enhanced Detection of Machine-Generated Texts. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[20] Chen, Z., Han, Y., Yi, Y.: Team chen at PAN: Integrating R-Drop and Pre-trained Language Model for Multi-author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[21] Chicco, D., Tötsch, N., Jurman, G.: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining **14**(1), 13 (Feb 2021), ISSN 1756-0381, https://doi.org/10.1186/s13040-021-00244-z
[22] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th ACL, ACL (2020), https://doi.org/10.18653/V1/2020.ACL-MAIN.747
[23] Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P.: SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1377–1414, International Committee for Computational Linguistics, Barcelona (online) (2020), https://doi.org/10.18653/v1/2020.semeval-1.186, URL https://aclanthology.org/2020.semeval-1.186
[24] Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P.: Fine-Grained Analysis of Propaganda in News Articles. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5636–5646, Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://doi.org/10.18653/v1/D19-1565, URL https://aclanthology.org/D19-1565
[25] Damian, S., Herrera-Gonzalez, B., Vazquez-Santana, D., Calvo, H., Felipe-Riverón, E., Yáñez-Márquez, C.: DSVS at PAN 2024: Ensemble Approach of Large Language Models for Analyzing Conspiracy Theories Against Critical Thinking Narratives. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[26] Dementieva, D., Babakov, N., Panchenko, A.: Multiparadetox: Extending text detoxification with parallel data to new languages. arXiv preprint arXiv:2404.02037 (2024)
[27] Dementieva, D., Logacheva, V., Nikishina, I., Fenogenova, A., Dale, D., Krotova, I., Semenov, N., Shavrina, T., Panchenko, A.: RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022), URL https://api.semanticscholar.org/CorpusID:253169495
[28] Dementieva, D., Ustyantsev, S., Dale, D., Kozlova, O., Semenov, N., Panchenko, A., Logacheva, V.: Crowdsourcing of parallel corpora: the case of style transfer for detoxification. Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021), CEUR Workshop Proceedings (2021), URL https://ceur-ws.org/Vol-2932/paper2.pdf
[29] Douglas, K.M., Sutton, R.M.: What are conspiracy theories? a definitional approach to their correlates, consequences, and communication. Annual Review of Psychology **74**(1), 271–298 (2023), URL https://doi.org/10.1146/annurev-psych-032420-031329
[30] Espinosa, D., Sidorov, G., Ricárdez-Vázquez, E.: Using BERT to Identify Conspiracy Theories. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[31] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. Proceedings of the 60th ACL, ACL (2022), https://doi.org/10.18653/V1/2022.ACL-LONG.62
[32] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: Continuous Integration for Reproducible Shared Tasks with TIRA.io. Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Springer (2023)

[33] Gambini, M., Tardelli, S., Tesconi, M.: The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset. Computer Communications **217**, 25–40 (2024), https://doi.org/10.1016/j.comcom.2024.01.027
[34] Gangopadhyay, S., Khan, M., Jabeen, H.: HybridDetox: Combining Supervised and Unsupervised Methods for Effective Multilingual Text Detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[35] Giachanou, A., Ghanem, B., Rosso, P.: Detection of conspiracy propagators using psycho-linguistic characteristics. Journal of Information Science **49**(1), 3–17 (2023), https://doi.org/10.1177/0165551520985486
[36] Gómez-Romero, J., González-Silot, S., Montoro-Montarroso, A., Molina-Solana, M., Martínez Cámara, E.: Detection of conspiracy-related messages in Telegram with anonymized named entities. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[37] Gritsai, G., Boyeva, G., Grabovoy, A.: Team ap-team at PAN: LLM Adapters for Various Datasets. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[38] Guo, L., Yang, W., Ma, L., Ruan, J.: BLGAV: Generative AI Author Verification Model Based on BERT and BiLSTM. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[39] Guo, M., Han, Z., Chen, H., Peng, J.: A Machine-Generated Text Detection Model Based on Text Multi-Feature Fusion. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[40] Haddad, H., Mulki, H., Oueslati, A.: T-hsab: A tunisian hate speech and abusive dataset. In: International conference on Arabic language processing, pp. 251–263, Springer (2019)
[41] Halvani, O., Winter, C., Graner, L.: On the usefulness of compression models for authorship verification. In: Proceedings of the 12th International Conference on Availability, Reliability and Security, vol. Part F1305, ACM, New York, NY, USA (29 Aug 2017), ISBN 9781450352574, https://doi.org/10.1145/3098954.3104050
[42] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., Goldstein, T.: Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. arXiv [cs.CL] (22 Jan 2024)
[43] Hong, J., Lee, N., Thorne, J.: ORPO: monolithic preference optimization without reference model. CoRR **abs/2403.07691** (2024), https://doi.org/10.48550/ARXIV.2403.07691, URL https://doi.org/10.48550/arXiv.2403.07691
[44] Hu, Q., Han, Z., Peng, J., Guo, M., Liu, C.: An Oppositional Thinking Analysis Method Using BERT-based Model with BiGRU. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[45] Huang, B., Zhong, C., Yan, K., Han, Y.: Author authentication of generative AI based on BERT by regularization method. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[46] Huang, J., Chen, Y., Luo, M., Li, Y.: Generative AI Authorship Verification Of Tri-Sentence Analysis Base On The Bert Model. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[47] Huang, J., Han, Z., Zhu, R., Guo, M., Sun, K.: Conspiracy Theory Text Classification Based on CT-BERT and BETO Models. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[48] Huang, K., Qi, H., Yan, K.: Voight-Kampff Generative AI Authorship Verification based on Contrastive Learning and Domain Adaptation. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[49] Huang, W., Grieve, J.: Authorial Language Models For AI Authorship Verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[50] Huang, Z., Kong, L.: Team huangzhijian at PAN: DeBERTa-v3 with R-Drop Regularization for Multi-Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[51] Huertas-García, Á., Martí-González, C., Muñoz, J., Ambite, E.: Small Language Models and Large Language Models in Oppositional thinking analysis: Capabilities and Biases and Challenges. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[52] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B. arXiv [cs.CL] (10 Oct 2023)
[53] Jigsaw: Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (2017), accessed: 2024-03-18
[54] Karlgren, J., Dürlich, L., Gogoulou, E., Guillou, L., Nivre, J., Sahlgren, M., Talman, A.: ELOQUENT CLEF Shared Tasks for Evaluation of Generative Language Model Quality: 46th European Conference on Information Retrieval (ECIR), Springer Nature Switzerland (2024), https://doi.org/10.1007/978-3-031-56069-9_63
[55] Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN 2018: Cross-domain Authorship Attribution and Style Change Detection. In: Working Notes of CLEF 2018, CEUR-WS.org (2018)
[56] Khan, A., Rai, M., Khan, K., Shah, S., Alvi, F., Samad, A.: Team Gladiators at PAN: Improving Author Identification: A Comparative Analysis of Pre-Trained Transformers for Multi-Author Classification. Working Notes of CLEF 2024, CEUR-WS.org (2024)

[57] King, J., Baffour, P., Crossley, S., Holbrook, R., Demkin, M.: Llm – detect ai generated text (2023), URL https://kaggle.com/competitions/llm-detect-ai-generated-text
[58] Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Twenty-first international conference on Machine learning - ICML '04, pp. 489–495, ACM Press, New York, New York, USA (2004), ISBN 9781581138283, https://doi.org/10.1145/1015330.1015448
[59] Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology **65**(1), 178–187 (2014)
[60] Korenčić, D., Chulvi, B., Bonet, X., Mariona, T., Toselli, A., Rosso, P.: What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. expert systems. Expert System (2024)
[61] Korenčić, D., Chulvi, B., Bonet Casals, X., Taulé, M., Rosso, P., Rangel, F.: Overview of the oppositional thinking analysis pan task at clef 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[62] Korenčić, D., Grubišić, I., Toselli, A.H., Chulvi, B., Rosso, P.: Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs. In: Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online (2023), URL https://2022.multimediaeval.com/paper8969.pdf
[63] Księżniak, E., Węcel, K., Sawiński, M.: Team OpenFact at PAN 2024: Fine-Tuning BERT Models with Stylometric Enhancements. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[64] Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., Tsvetkov, Y.: Language generation models can cause harm: So what can we do about it? an actionable survey. CoRR **abs/2210.07700** (2022)
[65] Lei, H., Liu, X., Niu, G., Zhou, Y., Zhou, Y.: Generative AI Authorship Verification based on ChatGLM. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[66] Liang, X., Lei, H.: Team lxflcl66666 at PAN: Fine-Tuned Reasoning for Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[67] Liang, X., Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., Liu, T.: R-drop: Regularized dropout for neural networks. In: 34th Annual Conference on Neural Information Processing Systems 2021, NeurIPS (2021)
[68] Lin, T., Wu, Y., Lee, L.: Team NYCU-NLP at PAN 2024: Integrating Transformers with Similarity Adjustments for Multi-Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[69] Lin, Z., Han, Z., Kong, L., Chen, M., Zhang, S., Peng, J., Sun, K.: A Verifying Generative Text Authorship Model With Regularized Dropout. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[70] Lin, Z., Li, Y., Huang, J.: Voight-Kampff Generative AI Authorship Verification Based on T5. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[71] Liu, B., Han, Z., Cao, H.: An Approach to Classifying Conspiratorial and Critical Public Health Narratives. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[72] Liu, C., Han, Z., Chen, H., Hu, Q.: Team liuc0757 at PAN: A Writing Style Embedding Method Based on Contrastive Learning for Multi-Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[73] Liu, X., Chen, H., Lv, J.: Team foshan-university-of-guangdong at PAN: Adaptive Entropy-Based Stability-Plasticity for Multi-Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[74] Liu, X., Kong, L.: AI Text Detection Method Based on Perplexity Features with Strided Sliding Window. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[75] Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., Semenov, N., Panchenko, A.: ParaDetox: Detoxification with parallel data. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6804–6818, Association for Computational Linguistics, Dublin, Ireland (May 2022), https://doi.org/10.18653/v1/2022.acl-long.469, URL https://aclanthology.org/2022.acl-long.469
[76] Lorenz, L., Aygüler, F.Z., Schlatt, F., Mirzakhmedova, N.: BaselineAvengers at PAN 2024: Often-Forgotten Baselines for LLM-Generated Text Detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[77] Lu, J., Xu, B., Zhang, X., Min, C., Yang, L., Lin, H.: Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 16235–16250 (Jul 2023), URL https://aclanthology.org/2023.acl-long.898
[78] Luo, Z., Luo, M., Wang, A.: Multilingual Text Detoxification Using Google Cloud Translation and Post-Processing. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[79] Lv, J., Han, Y., Kong, L.: Meta-Contrastive Learning for Generative AI Authorship Verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)

[80] Lv, J., Yi, Y., Qi, H.: Team Fosu-stu at PAN: Supervised fine-tuning of large language models for Multi Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[81] Mahesh, S., Divakaran, S., Girish, K., Lakshmaiah, S.: Binary Battle: Leveraging ML and TL Models to Distinguish between Conspiracy Theories and Critical Thinking. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[82] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., Patel, A.: Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, p. 14–17, FIRE '19, ACM (2019), ISBN 9781450377508, https://doi.org/10.1145/3368567.3368584
[83] Mathet, Y., Widlöcher, A., Métivier, J.P.: The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment. Computational Linguistics **41**(3), 437–479 (Sep 2015), ISSN 0891-2017, https://doi.org/10.1162/COLI_a_00227, URL https://doi.org/10.1162/COLI_a_00227
[84] Miralles, P., Martín, A., Camacho, D.: Ensembling Normalized Log Probabilities. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[85] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: Zero-shot machine-generated text detection using probability curvature. International Conference on Machine Learning **202**, 24950–24962 (26 Jan 2023), https://doi.org/10.48550/arXiv.2301.11305
[86] MTS.AI: Cotype: Generative ai solutions (2022), URL https://mts.ai, accessed: 2024-05-31
[87] Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., Al-Khalifa, H.: Overview of osact4 arabic offensive language detection shared task. In: Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection, pp. 48–52 (2020)
[88] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., Tang, X., Radev, D., Aji, A.F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., Raffel, C.: Crosslingual generalization through multitask finetuning. In: Proceedings of the 61st ACL, ACL (2023), https://doi.org/10.18653/V1/2023.ACL-LONG.891
[89] Mulki, H., Ghanem, B.: Let-mi: An Arabic Levantine Twitter dataset for misogynistic language. In: Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., Touileb, S. (eds.) Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 154–163, Association for Computational Linguistics, Kyiv, Ukraine (Virtual) (Apr 2021), URL https://aclanthology.org/2021.wanlp-1.16
[90] Mulki, H., Haddad, H., Ali, C.B., Alshabani, H.: L-hsab: A levantine twitter dataset for hate speech and abusive language. In: Proceedings of the third workshop on abusive language online, pp. 111–118 (2019)
[91] Najafi, M., Tavan, E., Colreavy, S.: Marsan at PAN 2024 TextDetox: ToxiCleanse RL and Paving the Way for Toxicity-Free Online Discourse. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[92] OpenAI: Chatgpt: Optimizing language models for dialogue (2022), URL https://openai.com/blog/chatgpt, accessed: 2024-05-31
[93] Osipenko, M., Korchagin, M., Toleugazinov, A., Egorov, S., Udobang, J.: Fancy Transformers at PAN 2024 TextDetox: Surpassing the Baselines. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[94] Ostrower, B., Wessell, J., Bindal, A.: AI Authorship Verification: An Ensembled Approach. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[95] Peng, J., Han, Z., Zhang, H., Ye, J., Liu, C., Liu, B., Guo, M., Chen, H., Lin, Z., Tang, Y.: A Multilingual Text Detoxification Method Based on Few-shot Learning and CO-STAR Framework. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[96] Pereira-Kohatsu, J.C., Sánchez, L.Q., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in twitter. Sensors **19**(21), 4654 (2019), https://doi.org/10.3390/S19214654, URL https://doi.org/10.3390/s19214654
[97] Pérez, J.M., Furman, D.A., Alonso Alemany, L., Luque, F.M.: RoBERTuito: a pre-trained language model for social media text in Spanish. Proceedings of the 13th LREC, ELRA (2022), URL https://aclanthology.org/2022.lrec-1.785
[98] Petropoulos, P., Petropoulos, V.: RoBERTa and Bi-LSTM for Human vs AI generated Text Detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[99] Pletenev, S.: Memu_pro_kotow at PAN 2024 TextDetox: Uncensored Llama3 Helps to Censor Better. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[100] Pogorelov, K., Schroeder, D.T., Brenner, S., Langguth, J.: FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021. In: Working Notes Proceedings of the MediaEval 2021 Workshop Bergen, Norway and Online (2021)
[101] Pogorelov, K., Schroeder, D.T., Brenner, S., Maulana, A., Langguth, J.: Combining tweets and connections graph for fakenews detection at mediaeval 2022. In: Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023. (2023)

[102] Protasov, V.: PAN 2024 Multilingual TextDetox: Exploring Cross-lingual Transfer in Case of Large Language Models. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[103] Qin, R., Qi, H., Yi, Y.: A model fusion approach for generative AI authorship verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[104] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv [cs.LG] (23 Oct 2019)
[105] Řehulka, E., Šuppa, M.: RAG Meets Detox: Enhancing Text Detoxification Using Open-Source Large Language Models with Retrieval Augmented Generation. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[106] Risch, J., Stoll, A., Wilms, L., Wiegand, M.: Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, pp. 1–12, Duesseldorf, Germany (2021)
[107] Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochumer Linguistische Arbeitsberichte, vol. 17, pp. 6–.9, Bochum, Germany (2016)
[108] 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)
[109] Ruffo, G., Semeraro, A., Giachanou, A., Rosso, P.: Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. Computer Science Review **47**, 100531 (2023), ISSN 1574-0137, https://doi.org/https://doi.org/10.1016/j.cosrev.2022.100531, URL https://www.sciencedirect.com/science/article/pii/S157401372200065X
[110] Rykov, E., Zaytsev, K., Anisimov, I., Voronin, A.: SmurfCat at PAN TexDetox 2024: Alignment of Multilingual Transformers for Text Detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[111] Sahitaj, A., Sahitaj, P., Mohtaj, S., Möller, S., Schmitt, V.: Towards a Computational Framework for Distinguishing Critical and Conspiratorial Texts by Elaborating on the Context and Argumentation with LLMs. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[112] Sánchez-Hermosilla, I., Panizo Lledot, A., Camacho, D.: A Study on NLP Model Ensembles and Data Augmentation Techniques for Separating Critical Thinking from Conspiracy Theories in English Texts. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[113] Sanjesh, R., Mangai, A.: Team riyasanjesh at PAN: Multi-feature with CNN and Bi-LSTM Neural Network approach to Style Change Detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[114] Sculley, D., Brodley, C.E.: Compression and machine learning: A new perspective on feature space vectors. In: Data Compression Conference (DCC'06), pp. 332–341, IEEE (2006), ISBN 9780769525457, ISSN 1068-0314,2375-0359, https://doi.org/10.1109/dcc.2006.13
[115] Semiletov, A.: Toxic Russian Comments: Labelled comments from the popular Russian social network. https://www.kaggle.com/alexandersemiletov/toxic-russian-comments (2020), accessed: 2023-12-14
[116] Sheykhlan, M., Abdoljabbar, S., Mahmoudabad, M.: Team karami-kheiri at PAN: Enhancing Machine-Generated Text Detection with Ensemble Learning Based on Transformer Models. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[117] Sheykhlan, M., Abdoljabbar, S., Mahmoudabad, M.: Team karami-sh at PAN: Transformer-based Ensemble Learning for Multi-Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[118] Stamatatos, E., Kestemont, M., Kredens, K., Pezik, P., Heini, A., Bevendorff, J., Potthast, M., Stein, B.: Overview of the Authorship Verification Task at PAN 2022. In: CLEF 2022 Labs and Workshops, CEUR-WS.org (2022)
[119] Stamatatos, E., Potthast, M., Pardo, F.M.R., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Springer (2015)
[120] Su, J., Zhuo, T.Y., Wang, D., Nakov, P.: DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. arXiv [cs.CL] (23 May 2023)
[121] Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., Collier, N.: A contrastive framework for neural text generation. arXiv [cs.CL] (13 Feb 2022)
[122] Sun, G., Yang, W., Ma, L.: BCAV: A Generative AI Author Verification Model Based on the Integration of Bert and CNN. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[123] Sushko, N.: PAN 2024 Multilingual TextDetox: Exploring Different Regimes For Synthetic Data Training For Multilingual Text Detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)

[124] Taulé, M., Nofre, M., Bargiela, V., Bonet, X.: Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish. LREC (2024)
[125] Tavan, E., Najafi, M.: Marsan at PAN: BinocularLLM and Fusing Binoculars' Insight with the Proficiency of Large Language Models for Cutting-Edge Machine-Generated Text Detection. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[126] Tian, Y., Chen, H., Wang, X., Bai, Z., Zhang, Q., Li, R., Xu, C., Wang, Y.: Multiscale positive-unlabeled detection of ai-generated texts. CoRR **abs/2305.18149** (2023), https://doi.org/10.48550/ARXIV.2305.18149
[127] Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the author identification task at PAN 2017: style breach detection and author clustering. In: CLEF 2017 Labs and Workshops (2017)
[128] Tulbure, A., Coll Ardanuy, M.: Conspiracy vs critical thinking using an ensemble of transformers with data augmentation techniques. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[129] Valdez-Valenzuela, A., Gómez-Adorno, H.: Team iimasnlp at PAN: Leveraging Graph Neural Networks and Large Language Models for Generative AI Authorship Verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[130] Vallecillo-Rodríguez, M., Martín-Valdivia, A.M.: SINAI at PAN 2024 TextDetox: Application of Tree of Thought Strategy in Large Language Models for Multilingual Text Detoxification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[131] Vallecillo-Rodríguez, M., Martín-Valdivia, M., Montejo-Ráez, A.: SINAI at PAN 2024 Oppositional Thinking Analysis: Exploring the fine-tuning performance of LLMs. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[132] Weimer, A.M., Barth, F., Dönicke, T., Gödeke, L., Varachkina, H., Holler, A., Sporleder, C., Gittel, B.: The (In-)Consistency of Literary Concepts. Operationalising, Annotating and Detecting Literary Comment. Journal of Computational Literary Studies **1**(1) (Dec 2022), ISSN 2940-1348, https://doi.org/10.48694/jcls.90, URL https://jcls.io/article/id/90/, number: 1 Publisher: Universitäts- und Landesbibliothek Darmstadt
[133] Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language (2018)
[134] Wu, B., Han, Y., Yan, K., Qi, H.: Team baker at PAN: Enhancing Writing Style Change Detection with Virtual Softmax. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[135] Wu, Q., Kong, L., Ye, Z.: Team bingezzzleep at PAN: A Writing Style Change Analysis Model Based on RoBERTa Encoding and Contrastive Learning for Multi-Author Writing Style Analysis. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[136] Wu, Z., Yang, W., Ma, L., Zhao, Z.: BertT: A Hybrid Neural Network Model for Generative AI Authorship Verification. Working Notes of CLEF 2024, CEUR-WS.org (Sep 2024)
[137] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. Proceedings of the NAACL-HLT 2021, ACL, https://doi.org/10.18653/V1/2021.NAACL-MAIN.41, URL https://doi.org/10.18653/v1/2021.naacl-main.41
[138] Yadagiri, A., Kalita, D., Ranjan, A., Bostan, A., Toppo, P., Pakray, P.: Team cnlp-nits-pp at PAN: Leveraging BERT for Accurate Authorship Verification: A Novel Approach to Textual Attribution. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[139] Ye, Z., Zhong, Y., Huang, C., Kong, L.: Team no-999 at PAN: Continual Transfer Learning with Progress Prompt for Multi-Author Writing Style Analysis". Working Notes of CLEF 2024, CEUR-WS.org (2024)
[140] Ye, Z., Zhong, Y., Huang, Z., Kong, L.: Token Prediction as Implicit Classification for Generative AI Authorship Verification. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[141] 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, CEUR-WS.org (2021)
[142] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2022. In: CLEF 2022 Labs and Workshops, CEUR-WS.org (2022)
[143] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2023. In: CLEF 2023 Labs and Workshops, CEUR-WS.org (2023)
[144] Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the Multi-Author Writing Style Analysis Task at PAN 2024. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[145] 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 (2020)
[146] 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 (2019)
[147] Zeng, Z., Han, Z., Ye, J., Tan, Y., Cao, H., Li, Z., Huang, R.: A Conspiracy Theory Text Detection Method based on RoBERTa and XLM-RoBERTa Models. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[148] Zhu, Y., Kong, L.: AI Authorship Verification Based On Deberta Model. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[149] Zinkovich, V., Karpukhin, S., Kurdiukov, N., Tikhomirov, P.: nlp_enjoyers at Multilingual Textual Detoxification (CLEF-2024. Working Notes of CLEF 2024, CEUR-WS.org (2024)
[150] Zrnić, L.: Conspiracy theory detection using transformers with multi-task and multilingual approaches. Working Notes of CLEF 2024, CEUR-WS.org (2024)