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

Extended Abstract

Janek Bevendorff,[1] Mara Chinea-Ríos,[8] Marc Franco-Salvador,[8] Annina Heini,[4]
Erik Körner,[1] Krzysztof Kredens,[4] Maximilian Mayerl,[6] Piotr Pęzik,[4]
Martin Potthast,[7] Francisco Rangel,[8] Paolo Rosso,[2] Efstathios Stamatatos,[9]
Benno Stein,[1] Matti Wiegmann,[1] Magdalena Wolska,[1] and Eva Zangerle[6]

[1]Bauhaus-Universität Weimar, Germany
[2]Universitat Politècnica de València, Spain
[4]Aston University, UK
[5]University of Antwerp, Belgium
[6]University of Innsbruck, Austria
[7]Leipzig University, Germany
[8]Symanto Research, Spain
[9]University of the Aegean, Greece

pan@webis.de   pan.webis.de

Abstract The paper gives a brief overview of the four shared tasks organized at the PAN 2023 lab on digital text forensics and stylometry to be hosted at the CLEF 2023 conference. The general goal of the PAN lab is to advance the state-of-the-art in text forensics and stylometry while ensuring objective evaluation of new and established methods on newly developed benchmark datasets. PAN's tasks cover four areas of digital text forensics: author identification, multi-author analysis, author profiling, and content analysis. Some tasks follow up on past editions (cross-domain authorship verification, multi-author writing style analysis) and some explore novel ideas (profiling cryptocurrency influencers in social media and trigger detection). As with the previous editions, PAN invites software submissions rather than run submissions; more than 400 pieces of software have been submitted from PAN'12 through PAN'22 combined, with recent evaluations running on the TIRA experimentation platform. This proposal briefly outlines our goals for PAN as a lab and our contributions proposed for PAN'23.

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 technologies that analyze texts with regard to originality, authorship, trust, and ethicality, among others. Since its inception 15 years back PAN has included shared tasks on specific computational challenges related to authorship analysis, computational ethics, and determining the originality of a piece of writing. Over the years, the respective organizing committees of the 64 shared tasks[1] have assembled evaluation resources for the aforementioned research disciplines that amount to 55 datasets[2] plus nine datasets contributed by the community. Each new dataset was compiled by the task's authors specifically for the given task and introduced new variants of author verification, profiling, or author obfuscation tasks as well as multi-author analysis and determining the morality, quality, or originality of a text. The tasks build incrementally on the experience and results of prior PAN shared tasks and extend them in meaningful ways by increasing complexity. The 2023 edition of PAN continues in the same vein, introducing new resources as well as previously unconsidered problems to the community. As in earlier editions, PAN is committed to reproducible research in IR and NLP therefore all shared tasks will ask for software submissions on our TIRA platform [9]. We briefly outline the upcoming tasks in the sections that follow.

2   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 [6]. The past editions of PAN considered the task of *cross-domain authorship verification*, where the texts of known and unknown authorship come from different domains [1, 2, 24]. In most of the examined cases, the domains corresponded to topics, thematic areas, or fandoms (non-professional fiction published online in significant quantities by fans of high-popularity authors or works, so-called fanfiction). The relatively high performance of the past submissions [1, 2] demonstrates that authorship in most of these cases can be successfully verified. However, it is not clear yet how to handle more difficult authorship verification cases where texts of known and unknown authorship belong to different discourse types (DTs), especially when these DTs have few similarities (e.g., argumentative essays vs. text messages to family members). Hence, the most recent edition of PAN adopted a new and very challenging scenario: *cross-discourse type authorship verification*. Here, documents belong to different discourse types (i.e., essays, emails, text messages, business memos) whose style depends on the level of formality, intended audi-

---

[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`

ence, and communicative purpose [23]. The relatively low obtained evaluation results show that the task is still exceedingly difficult.

Cross-Discourse Type Author Verfication at PAN'23

In its simplest form, authorship verification deals with determining whether two documents are written by the same author. In cross-discourse type authorship verification, introduced in the last edition of PAN [23], the two documents are of distinct DTs. Yet despite their differences, all documents in this and previous PAN editions are only forms of written language. At PAN'23, we will focus for the first time on (cross-discourse type) authorship verification where both written (e.g., essays, emails) and oral language (e.g., interviews, speech transcriptions) are represented in the set of discourse types. This will provide the opportunity to study the robustness and effectiveness of stylometric approaches in challenging and intriguing conditions. In addition, the ability of authorship verification methods to handle the different forms of expression in written and oral language will be highlighted. New training and evaluation datasets will be provided that cover DTs in both written and oral language. The same evaluation framework and measures as in the latest PAN editions of authorship verification tasks will be adopted [1, 23]. The evaluation includes well-known measures like the area under the ROC curve, F_1 score, and Brier score, as well as more specialized measures that take into account non-answers, like $c@1$ (a variant of accuracy rewarding non-answers) and $F_{0.5u}$ (a variant of F-score rewarding correctly predicted same-author cases in addition to non-answers).

3 Author Profiling

Author profiling is the problem of distinguishing between classes of authors by studying how language is shared by people. Profiling can help to identify 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.[3] In 2013 the aim was to identify gender and age in social media texts for English and Spanish [16]. 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 [14]. In 2015, apart from age and gender identification, we addressed also personality recognition on Twitter in English, Spanish, Dutch, and Italian [18]. In 2016, we addressed the problem of cross-genre gender and age identification in English, Spanish, and Dutch [19]. The training data was gathered from Twitter and the test data was gathered from blogs and social media data. In 2017, we addressed gender and language variety identification in Twitter in English, Spanish, Portuguese, and Arabic [17]. In 2018, we investigated gender identification on Twitter from a multi-modal perspective, considering also

---

[3]All our datasets comply with the EU General Data Protection Regulation [12].

the images linked within tweets; the dataset was composed of English, Spanish, and Arabic tweets [15]. From 2019 to 2022, we focused on a series of shared tasks related to profiling harmful information spreaders. In 2019 our focus was on profiling and discriminating bots from humans on the basis of textual data only [13] and targeting both English and Spanish tweets. In 2020, we focused on profiling fake news spreaders [11], both in English and Spanish. The ease of publishing content on social media has also increased the amount of disinformation that is published and shared and our goal was to profile those authors who have shared some fake news in the past. In 2021, we focused on profiling hate speech spreaders in social media [10], both in English and Spanish. 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. Finally, in 2022, we focused on profiling irony and stereotype spreaders on English tweets [20]. The 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 at PAN'23

Cryptocurrencies have massively increased their popularity in recent years [22]. 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. Profiling research is particularly interested in the cryptocurrency ecosystem to identify influential actors that motivate others into action.

Producing sufficiently many high-quality annotations for author profiling is challenging. Profiling influencers in particular has high requirements in the economic and temporal cost, psychological and linguistic expertise needed by the annotator, and the congenital subjectivity involved in the annotation task [3, 25]. 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 propose to profile types of influencers also using a low-resource setting. Specifically, we focus on English Twitter posts for three different sub-tasks: (1) *Low-resource influencer profiling*: profile authors according to their degree of influence (null, nano, micro, macro, mega); (2) *Low-resource influencer interest profiling*: profile authors according to their main interests or areas of influence (technical information, price update, trading matters, gaming, other); (3) *Low-resource influencer intent profiling*: profile authors according to the intent of their messages (subjective opinion, financial information, advertising, announcement). Participants need to choose carefully which models to apply to this under-resourced setting.

Concepts such as transfer learning [28] and few-shot learning [4, 8, 7, 27] are key to succeed.

4   Multi-Author Writing Style Analysis

Style change detection concerns itself with identifying positions within a given text document at which the writing style—and therefore, by extension, the author—changes. This task can be a constituent task of authorship identification and multi-author document analysis, and has applications in areas such as plagiarism detection. At PAN, the style change detection task has been studied since 2016, in various different forms. In 2016, participants had to identify the authors of fragments of a document, and group all fragments written by the same author together [21]. In 2017, the task was twofold [26]. First, participants had to determine whether a given document was written by one or by multiple authors. Second, for documents by multiple authors, they had to determine the exact positions within the documents where the author changes. In 2018, following feedback that the task posed in the previous year was too difficult, the problem was simplified to only identifying whether a document was written by one or more authors [5]. In the following years, we built on this and gradually made the task definitions more complex again. In 2019, participants had to first determine whether a document had a single or multiple authors, and, if it is multi-authored, determine the concrete number of authors involved in writing it [32]. In 2020, participants again had to determine whether a document is single- or multi-authored. For multi-authored documents, they also had to identify between which paragraphs in the document the author changes, and assign paragraphs to concrete authors [31]. The task posed in 2021 was very similar, but this time, we additionally provided participants with a simplified version of the task, where each document contained exactly one style change, and the participants had to determine between which paragraphs in the document this occurred [29]. Finally, in 2022, we added a more complex subtask where style changes could now occur not only between paragraphs but also between sentences [30].

Multi-Author Writing Style Analysis at PAN'23

Traditionally, writing style analysis has focused on single-author documents. However, more recent research, including that conducted at previous editions of PAN, has shown that writing style analysis can effectively be employed for detecting author changes within a document. This can be used to partition a document into parts that have been written by different authors, which can be applied to areas such as plagiarism detection. In previous editions of PAN, our participants developed a range of different techniques for detecting author changes in documents. However, the datasets used in those editions of the task exhibited a large variety of topics, also within single documents. This allowed approaches to indirectly exploit topic changes to make the task easier.

In the 2023 edition, we have therefore paid special attention to developing datasets that do not exhibit this problem. We will provide participants with datasets of three difficulty levels: (1) "Easy dataset": The paragraphs of a document cover a variety of topics, allowing approaches to make use of topic information to detect authorship changes. (2) "Medium dataset": The topical variety in a document is small (though still present) forcing the approaches to focus more on style to solve the detection task effectively; (3) "Hard dataset": All paragraphs in a document are on the same topic. Similar to most tasks in recent editions, style changes are once again limited to occur between paragraphs (i.e., each paragraph belongs to a single author).

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, such as 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" have become common. Trigger warnings are labels that indicate which type of triggering content is present. They are frequently used in online communities and in institutionalized education and allow a sensitive audience to prepare for the content to better manage their reactions. In the planned series of shared tasks on triggers, we propose a computational problem of identifying whether or not a given document contains triggering content, and if so, of what type.

Identifying Violent Content at PAN'23

In the pilot edition of the task at PAN'23, we will focus on a single trigger type: violence. As data we will use a corpus of fanfiction (millions of stories crawled from `fanfiction.net` and `archiveofourown.org` (Ao3)) in which trigger warnings have been assigned by the authors, that is, we do not define "violence" as a construct ourselves here, but rather rely on user-generated labels. We unify the set of label names where necessary and create a balanced corpus of positive and negative examples. The problem is formulated as binary classification at the document level as follows: Given a piece of fanfiction discourse, classify it as triggering or not triggering, that is, in the PAN'23 edition of the task, assign the trigger warning "violence" if appropriate. Standard measures of classifier quality will be used for evaluation.

Acknowledgments

Bibliography

[1] 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, vol. 12880, pp. 419–431 (2021)

[2] 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, vol. 12260, pp. 372–383 (2020)

[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] 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)

[5] 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: CLEF 2018 Labs and Workshops, Notebook Papers (2018)

[6] 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)

[7] Mueller, T., Pérez-Torró, G., Franco-Salvador, M.: Few-shot learning with siamese networks and label tuning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8532–8545 (2022)

[8] Müller, T., Pérez-Torró, G., Basile, A., Franco-Salvador, M.: Active few-shot learning with fasl. arXiv preprint arXiv:2204.09347 (2022)

[9] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World, Springer (2019), https://doi.org/10.1007/978-3-030-22948-1_5

[10] 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: CLEF 2021 Labs and Workshops, Notebook Papers (2021)

[11] 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)

[12] Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Language and Law / Linguagem e Direito **5**(2), 95–117 (2019)

[13] 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)

[14] Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at PAN 2014. In: CLEF 2014 Labs and Workshops, Notebook Papers (2014)

[15] 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)

[16] 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)

[17] 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)

[18] 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)

[19] 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 (Sep 2016), ISSN 1613-0073

[20] 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)

[21] 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)

[22] 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)

[23] Stamatatos, E., Kestemont, M., Kredens, K., Pezik, P., Heini, A., Bevendorff, J., Stein, B., Potthast, M.: Overview of the Authorship Verification Task at PAN 2022. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (Sep 2022)

[24] Stamatatos, E., Potthast, M., Pardo, F.M.R., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. 9283, pp. 518–538 (2015)

[25] Troiano, E., Padó, S., Klinger, R.: Emotion ratings: How intensity, annotation confidence and agreements are entangled. arXiv preprint arXiv:2103.01667 (2021)

[26] 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, Notebook Papers (2017)

[27] Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) **53**, 1–34 (2020)

[28] Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**, 1–40 (2016)

[29] 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)

[30] Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2022. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (Sep 2022)

[31] 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)

[32] 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)