Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, Style Change Detection, and Trigger Detection

Extended Abstract

Janek Bevendorff[1], Berta Chulvi[2], Elisabetta Fersini[3], Annina Heini[4], Mike Kestemont[5], Krzysztof Kredens[4], Maximilian Mayerl[6], Reyner Ortega-Bueno[2], 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, Weimar, Germany
pan@webis.de, magdalena.wolska@uni-weimar.de
[2] Universitat Politècnica de València, Valencia, Spain
[3] Universitty Milano-Bicocca, Milan, Italy
[4] Aston University, Birmingham, UK
[5] University of Antwerp, Antwerp, Belgium
[6] University of Innsbruck, Innsbruck, Austria
[7] Leipzig University, Leipzig, Germany
[8] Symanto Research, Nuremberg, Germany
[9] University of the Aegean, Mytilene, Greece
http://pan.webis.de

Abstract. The paper gives a brief overview of the four shared tasks to be organized at the PAN 2022 lab on digital text forensics and stylometry hosted at the CLEF 2022 conference. The tasks include authorship verification across discourse types, multi-author writing style analysis, author profiling, and content profiling. Some of the tasks continue and advance past editions (authorship verification and multi-author analysis) and some are new (profiling irony and stereotypes spreaders and trigger detection). The general goal of the PAN shared tasks is to advance the state of the art in text forensics and stylometry while ensuring objective evaluation on newly developed benchmark datasets.

1 Introduction

PAN is a workshop series and a networking initiative for stylometry and digital text forensics. The workshop's goal is to bring together scientists and practitioners studying technologies which analyze texts with regard to originality, authorship, trust, and ethicality. 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.

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 331–338, 2022.
https://doi.org/10.1007/978-3-030-99739-7_42

Over the years, the respective organizing committees of the 54 shared tasks have assembled evaluation resources for the aforementioned research disciplines that amount to 51 datasets plus nine datasets contributed by the community.[1] Each new dataset 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 2022 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 [7]. We briefly outline the upcoming tasks in the sections that follow.

2 Authorship Verification

Authorship verification is a fundamental task in author identification and all questioned authorship cases, be it closed-set or open-set scenarios, can be decomposed into a series of verification instances [6]. Previous editions of PAN included across-domain authorship verification tasks where texts of known and unknown authorship come from different domains [2,3,21]. In most of the examined cases, domains corresponded to topics (or thematic areas) and fandoms (non-professional fiction that is nowadays published online in significant quantities by fans of high-popularity authors or works, so-called fanfiction). The obtained results of the latest editions have demonstrated that it is feasible to handle such cases with relatively high performance [2,3]. In addition, at PAN'15, cross-genre authorship verification was partially studied using datasets in Dutch and Spanish covering essays and reviews [21]. However, these are relatively similar genres with respect to communication purpose, intended audience, or level of formality. On the other hand, 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). In such cases, it is very challenging to distinguish the authorial characteristics that remain intact along DTs.

Cross-DT Author Verification at PAN'22

For the 2022 edition, we will focus on the cross-DT authorship verification scenario. In more detail, we will use a new corpus in English comprising writing samples from around 100 individuals composing texts in the following DTs: essays, emails, text messages, and business memos. All individuals have similar age (18-22) and are native English speakers. The topic of text samples is not restricted while the level of formality can vary within a certain DT (e.g., text messages may be addressed to family members or non-familial acquaintances). The new edition of author verification task at PAN'22 will allow us to

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

study the ability of stylometric approaches to capture elements of authorial style that remain stable across DTs even when very different forms of expression are imposed by the DT norms. The task will also focus on the ability of the submitted approaches to compare long texts of known authorship with short texts of unknown authorship. As concerns the experimental setup, it will be similar to the last edition of PAN and the same evaluation measures (AUROC, c@1, F_1, $F_{0.5u}$, and Brier score) will be used [2].

3 Author Profiling

Author profiling is the problem of distinguishing between classes of authors by studying how language is shared by people. This helps in identifying authors' individual characteristics, such as age, gender, or language variety, among others. During the years 2013-2021 we addressed several of these aspects in the shared tasks organised at PAN.[2] In 2013 the aim was to identify gender and age in social media texts for English and Spanish [14]. 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 [12]. In 2015, apart from age and gender identification, we addressed also personality recognition on Twitter in English, Spanish, Dutch, and Italian [16]. In 2016, we addressed the problem of cross-genre gender and age identification (training on Twitter data and testing on blogs and social media data) in English, Spanish, and Dutch [17]. In 2017, we addressed gender and language variety identification in Twitter in English, Spanish, Portuguese, and Arabic [15]. In 2018, we investigated gender identification in Twitter from a multimodal perspective, considering also the images linked within tweets; the dataset was composed of English, Spanish, and Arabic tweets [13]. In 2019 the focus was on profiling bots and discriminating bots from humans on the basis of textual data only [11]. We used Twitter data both in English and Spanish. Bots play a key role in spreading inflammatory content and also fake news. Advanced bots that generated human-like language, also with metaphors, were the most difficult to profile. It is interesting to note that when bots were profiled as humans, they were mostly confused with males. In 2020 we focused on profiling fake news spreaders [9]. The easiness of publishing content in social media has led to an increase in the amount of disinformation that is published and shared. The goal was to profile those authors who have shared some fake news in the past. Early identification of possible fake news spreaders on Twitter should be the first step towards preventing fake news from further dissemination. In 2021 the focus was on profiling hate speech spreaders in social media [8]. The goal was to identify Twitter users who can be considered haters, depending on the number of tweets with hateful content that they had spread. The task was set in English and Spanish.

[2] To generate the datasets, we have followed a methodology that complies with the EU General Data Protection Regulation [10].

Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO)

With irony, language is employed in a figurative and subtle way to mean the opposite to what is literally stated [18]. In case of sarcasm, a more aggressive type of irony, the intent is to mock or scorn a victim without excluding the possibility to hurt [4]. Stereotypes are often used, especially in discussions about controversial issues such as immigration [22] or sexism [19] and misogyny [1]. At PAN'22 we will focus on profiling ironic authors in Twitter. Special emphasis will be given to those authors that employ irony to spread stereotypes, for instance, towards women or the LGTB community. The goal will be to classify authors as ironic or not depending on their number of tweets with ironic content. Among those authors we will consider a subset that employs irony to convey stereotypes in order to investigate if state-of-the-art models are able to distinguish also these cases. Therefore, given authors together with their tweets, the goal will be to profile those authors that can be considered as ironic, and among them those that employ irony to convey stereotypical messages. As an evaluation setup, we will create a collection that contains tweets posted by users in Twitter. One document will consist of a feed of tweets written by the same user.

4 Multi-author Writing Style Analysis

The goal of the style change detection task is to identify—based on an intrinsic style analysis—the text positions within a given multi-author document at which the author switches. Detecting these positions is a crucial part of the authorship identification process and multi-author document analysis; multi-author documents have been largely understudied in general. This task has been part of PAN since 2016, with varying task definitions, data sets, and evaluation procedures. In 2016, participants were asked to identify and group fragments of a given document that correspond to individual authors [20]. In 2017, we asked participants to detect whether a given document is multi-authored and, if this is indeed the case, to determine the positions at which authorship changes [23]. However, since this task was deemed as highly complex, in 2018 its complexity was reduced to asking participants to predict whether a given document is single- or multi-authored [5]. Following the promising results achieved, in 2019 participants were asked first to detect whether a document was single- or multi-authored and if it was indeed written by multiple authors, to then predict the number of authors [26]. Based on the advances made over the previous years, in 2020 we decided to go back towards the original definition of the task, i.e., finding the positions in a text where authorship changes. Participants first had to determine whether a document was written by one or by multiple authors and, if it was written by multiple authors, they had to detect between which paragraphs the authors change [25]. In the 2021 edition, we asked the participants first to detect whether a document was authored by one or multiple authors. For two-author documents, the task was to find the position of the authorship change and for multi-author documents, the task was to find all positions of authorship change [24].

Multi-author Writing Style Analysis at PAN'22

The analysis of author writing styles is the foundation for author identification. As previous research shows, it also allows distinguishing between authors in multi-authored documents. In this sense, methods for multi-author writing style analysis can pave the way for authorship attribution at the sub-document level and thus, intrinsic plagiarism detection (i.e., detecting plagiarism without the use of a reference corpus). Given the importance of these tasks, we foster research in this direction through our continued development of benchmarks: the ultimate goal is to identify the exact positions within a document at which authorship changes based on an intrinsic style analysis. Based on the progress made towards this goal in previous years and to entice novices and experts, we extend the set of challenges: (i) Style Change Basic: given a text written by two authors and that contains a single style change only, find the position of this change, i.e., cut the text into the two authors' texts on the paragraph-level, (ii) Style Change Advanced: given a text written by two or more authors, find all positions of writing style change, i.e., assign all paragraphs of the text uniquely to some author out of the number of authors assumed for the multi-author document, (iii) Style Change Real-World: given a text written by two or more authors, find all positions of writing style change, where style changes now not only occur between paragraphs but at the sentence level. For this year's edition, we will additionally introduce a new corpus that is based on publicly available social media data to show the performance of the approaches based on different data sources.

5 Trigger Detection

A trigger in psychology is a stimulus that elicits negative emotions or feelings of distress. In general, triggers include a broad range of stimuli—such as smells, tastes, sounds, textures, or sights—which may relate to possibly distressing acts or events of whatever type, for instance, violence, trauma, death, eating disorders, or obscenity. In order to proactively apprise audience that a piece of media (writing, audio, video, etc.) contains potentially distressing material, the use of "trigger warnings"—labels indicating the type of triggering content present—have become common not only in online communities, but also in institutionalized education, making it possible for sensitive audience to prepare for the content and 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'22

In the first pilot edition of the task, 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 document level as follows: Given a piece of fanfiction discourse, classify it as triggering or not triggering. Standard measures of classifier quality will be used for evaluation.

Acknowledgments. The contributions from Bauhaus-Universität Weimar and Leipzig University have been partially funded by the German Ministry for Science and Education (BMBF) project "Shared Tasks as an innovative approach to implement AI and Big Data-based applications within universities (SharKI)" (grant FKZ 16DHB4021). The Cross-DT corpus was developed at the Aston Institute for Forensic Linguistics with funding from Research England's Expanding Excellence in England (E3) Fund. The work of the researchers from the Universitat Politècnica de València was partially funded by the Spanish MICINN under the project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31), and by the Generalitat Valenciana under the project DeepPattern (PROMETEO/2019/121). The work of Francisco Rangel has been partially funded by the Centre for the Development of Industrial Technology (CDTI) of the Spanish Ministry of Science and Innovation under the research project IDI-20210776 on Proactive Profiling of Hate Speech Spreaders - PROHATER (Perfilador Proactivo de Difusores de Mensajes de Odio).

References

1. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In: Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F. (eds.) NLDB 2018. LNCS, vol. 10859, pp. 57–64. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91947-8_6
2. Bevendorff, J., et al.: Overview of PAN 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 419–431. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_26
3. Bevendorff, J., et al.: Overview of PAN 2020: authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 372–383. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_25
4. Frenda, S., Cignarella, A., Basile, V., Bosco, C., Patti, V., Rosso, P.: The unbearable hurtfulness of sarcasm. Expert Syst. Appl. (2022). https://doi.org/10.1016/j.eswa.2021.116398
5. Kestemont, M., et al.: Overview of the author identification task at PAN 2018: cross-domain authorship attribution and style change detection. In: CLEF 2018 Labs and Workshops, Notebook Papers (2018)
6. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. J. Assoc. Inf. Sci. Technol. **65**(1), 178–187 (2014)
7. 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. TIRS, vol. 41, pp. 123–160. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22948-1_5
8. Rangel, F., De-La-Peña-Sarracén, G.L., Chulvi, B., Fersini, E., Rosso, P.: Profiling hate speech spreaders on twitter task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)
9. 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)
10. Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Lang. Law/Linguagem e Direito **5**(2), 95–117 (2019)
11. 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)
12. Rangel, F., et al.: Overview of the 2nd author profiling task at PAN 2014. In: CLEF 2014 Labs and Workshops, Notebook Papers (2014)
13. 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)
14. 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)
15. 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. In: Working Notes Papers of the CLEF (2017)
16. 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)
17. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: CLEF 2016 Labs and Workshops, Notebook Papers (2016). ISSN 1613–0073
18. Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowl. Inf. Syst. **40**(3), 595–614 (2014)
19. Rodríguez-Sánchez, F., et al.: Overview of exist 2021: sexism identification in social networks. In: Procesamiento del Lenguaje Natural (SEPLN), no. 67, pp. 195–207 (2021)
20. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN 2016–new challenges for authorship analysis: cross-genre profiling, clustering, diarization, and obfuscation. In: 7th International Conference of the CLEF Initiative on Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2016) (2016)
21. Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 518–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_49
22. Sánchez-Junquera, J., Chulvi, B., Rosso, P., Ponzetto, S.: How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. Appl. Sci. **11**(8), 3610 (2021)
23. Tschuggnall, M., et al.: Overview of the author identification task at PAN 2017: style breach detection and author clustering. In: CLEF 2017 Labs and Workshops, Notebook Papers (2017)

24. 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)
25. 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)
26. 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)