Shared Tasks on Authorship Analysis at PAN 2020

Janek Bevendorff,[1] Bilal Ghanem[2], Anastasia Giachanou[2], Mike Kestemont,[3] Enrique Manjavacas,[3] Martin Potthast,[4] Francisco Rangel,[5] Paolo Rosso,[2] Günther Specht,[6] Efstathios Stamatatos,[7] Benno Stein,[1] Matti Wiegmann,[1] and Eva Zangerle[6]

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

pan@webis.de   http://pan.webis.de

Abstract The paper gives a brief overview of the four shared tasks that are to be organized at the PAN 2020 lab on digital text forensics and stylometry, hosted at CLEF conference. The tasks include author profiling, celebrity profiling, cross-domain author verification, and style change detection, seeking to advance the state of the art and to evaluate it on new benchmark datasets.

1 Introduction

PAN is a series of scientific events and shared tasks on digital text forensics and stylometry, bringing together scientists, industry professionals, and public institutions from information retrieval and NLP to work on challenges in authorship analysis, originality, and computational ethics. Since its inception in 2007, PAN has hosted 22 shared tasks at 21 different events with continually increasing reception within the community. The latest installment of PAN at CLEF 2019 had a strong focus on authorship analysis, featuring tasks on author profiling, celebrity profiling, authorship attribution, and style change detection. Continuing in 2020, PAN will again organize four shared tasks in these domains. The first task, profiling fake news spreaders on Twitter, addresses the critical societal problem of fake news from the perspective of author profiling, by studying stylistic deviations of users inclined to spread them. The second task, cross-domain authorship verification, studies the stylistic association between authors and their works in a setting without the interference of domain-specific vocabulary. The third task, celebrity profiling, analyzes the presumed influence that celebrities have on their followers to study whether celebrities can be profiled based on their followership. The fourth task, style change detection, continues the research on multi-author documents by attempting to separate segments of a document based on authorship.

A milestone in PAN's development has been the development of the TIRA platform, switching from the traditional submission of answers to *software* submissions. The guaranteed availability of all submitted software greatly enhances the reproducibility of methods and PAN is committed to continue this endeavor.

2 Author Profiling

Author profiling distinguishes between classes of authors by studying how language is shared by people. This helps in identifying profiling aspects such as age, gender, and language variety, among others. In the years 2013-2018, we addressed several aspects in the shared tasks we organized at PAN.[1] In 2013, the aim was to identify gender and age in social media texts for English and Spanish [22]. The corpus included chat lines of potential pedophiles with the purpose of investigating the robustness of the best-performing systems also from this perspective (i.e., identifying the age of the pedophiles). Age classes included a gap in between: 10s (13-17), 20s (23-27), 30s (33-48). Results in both languages and in both subtasks were below 70% accuracy.

In 2014, the aims of the shared task were twofold: to address age identification from a continuous perspective (without gaps between the age classes), and to include other genres such as blogs, Twitter and reviews (in Trip Advisor), both in English and Spanish. The best results were obtained on Twitter, where users showed a more spontaneous way to communicate [20]. In 2015, apart from age and gender identification, we addressed also personality recognition in Twitter in English, Spanish, Dutch and Italian. The best results (above 80% accuracy) were obtained on English data [24]. 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. The best results were obtained on blogs for English with an accuracy above 75% for gender and below 60% for age identification [25]. In 2017, we addressed gender and language variety identification in Twitter, in English, Spanish, Portuguese and Arabic. The lowest results were obtained for Arabic with an accuracy of 80% for gender and 83% for language variety identification [23]. In 2018, our aim was to investigate if approaching gender identification in Twitter from a multimodal perspective (e.g., considering also images of the links in tweets) could improve results. The corpus was composed of English, Spanish, and Arabic tweets. Only for Arabic it was possible to improve accuracy (albeit less than 2%) [21].

Last, in 2019, in the shared task on bots and gender profiling, we aimed at investigating how difficult it is to discriminate bots from humans on the basis only of textual data, and what were the most difficult types of bots. We used Twitter data both in English and Spanish and the best-performing systems showed that it is possible to profile bots with an accuracy above 90%. Advanced bots that generated human-like language, also with metaphors, were the most difficult to be profiled. It is interesting to mention that when bots were profiled as humans, they were mostly confused with males [19]. The number of the participants in the several editions of the author profiling task can be seen in Figure 1.

Profiling Fake News Spreaders on Twitter at PAN'20

Fake news can be very harmful since they are usually created with the aim to manipulate public opinions and beliefs. Recently fake news detection has gained a lot of

[1] To generate the corpora, we followed a methodology that complies with the EU General Data Protection Regulation [18].

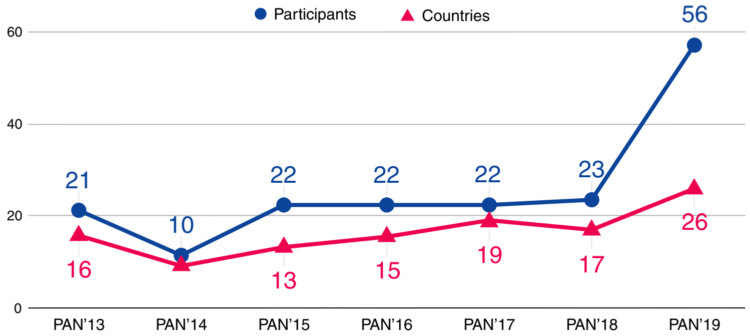

Figure 1. Evolution of the number of participants and countries in the author profiling task.

attention from the research community. Indeed, their early detection can prevent further dissemination of false claims and rumors, but it's a hard and time-consuming task, since they involve manual annotation. Recent approaches that have been proposed [7, 6, 16] are effective in detecting false claims that already have been disseminated, but not the newly emerging ones. In addition, these models do not take into account the role of users that unintentionally or intentionally share the false claims and who play a critical role in their propagation. To this end, in this task, we aim at identifying and profiling fake news spreaders on social media as a first step towards preventing fake news from being propagated among online users.

We propose a new task that focuses on fake and real news spreaders detection. The detection of accounts that are possible spreaders of fake news is very important for the field of misinformation detection. These accounts could be operated by laymen [2], "professional" trolls [4], and even bots [14]. The fake news spreaders might be identified from several possible perspectives: textual, semantic, sentiment, social variables, etc. A previous work [5] showed that word embeddings and style features are important to profile such accounts, whereas other information, such as hashtags are not useful.

Given a user with her corresponding tweet stream, the task is to identify the user as faker (fake news spreader), or legitimate user (real news spreader). For the evaluation setup, we create a collection of Twitter accounts, each with a sample of tweets from her timeline. The collection has been created in English and Spanish, and it is balanced. Thus, we are going to use accuracy to evaluate the performance of the systems.

3 Celebrity Profiling

Celebrity profiling is author profiling applied to celebrities. Celebrities can contribute much to author profiling research: they are prolific social media users, often supplying extensive writing samples as well as personal details. Celebrities build a consistent public persona either themselves or with the help of public relations agents. In addition, celebrities are in a unique position within their communities: they are highly influential on their followers, frequently considered trustworthy and reliable, and they act as hubs for like-minded people on social media. Celebrity Profiling [30] is the newest addition to PAN's shared tasks. In 2019 [31], the goal was to determine the demographics age, gender, occupation, and fame from the timelines of celebrities on Twitter. Eight participants submitted solutions, which, given sufficient training data, performed well on demographics with a coherent separability by topic or domain. Poor performance was achieved in cases where certain demographics are rare (e.g., non-binary genders), or where they are underrepresented (e.g., age groups for very low and high ages). Also domain-invariant demographics, like the scientific creative occupations, posed problems. The results of the first shared task on celebrity profiling are coherent with most of the related work in author profiling, authorship analysis, and computational stylometry in general: the domain-specific vocabulary is the primary discriminator and demographic differences are often reflected by topics.

Celebrity Profiling at PAN'20

The unique contributions of celebrities on social media towards author profiling research is their domain-variant claim-to-fame and the varying degree of influence they exert on their followers. The formation of closely connected communities around celebrities, who are also under their influence, allows us to investigate the role of author characteristics, domain, and demographic on language use. For the upcoming edition of celebrity profiling, we focus on separating a celebrity author's textual characteristics from domain-specific language use, using the demographics as an indicator. Instead of predicting the authors demographics from his text alone, we use the texts of highly influenced individuals, while the prediction targets remain largely the same as last year (age, gender, occupation). The results of this shared task will help us to determine for the first time, whether and to what extent an influencer's demographics and characteristics can be predicted from his or her followers. Tangible applications, besides academic interest, include methods to profile users with few own text samples, and to judge influence exerted between users in a community.

4 Author Identification

Authentication is a major concern in today's global information society and in this sense it does not come as a surprise that author identification has been a long-running task at PAN. Author identification still poses a challenging empirical problem in fields related to information and computer science, but the underlying methods are nowadays also increasingly used as an auxiliary technology in more applied domains, such as literary studies or forensic linguistics. These communities crucially rely on trustworthy, transparent benchmark initiatives that reliably establish the state of the art in the field [17]. Author identification is concerned with the automated identification of the individual(s) who authored an anonymous document on the basis of text-internal properties related to language and writing style [27, 9, 12]. At different editions of PAN (since 2007), author identification has been studied in multiple incarnations: AUTHORSHIP ATTRIBUTION: given a document and a set of candidate authors, determine which of them wrote the document (2011-2012, 2016-2020); AUTHORSHIP VERIFICATION: given a pair of documents, determine whether they are written by the same author (2013-2015); AUTHORSHIP OBFUSCATION: given a document and a set of documents from the same author, paraphrase the former so that its author cannot be identified anymore (2016-2018); OBFUSCATION EVALUATION: devise and implement performance measures that quantify safeness, soundness, and/or sensibleness of an obfuscation software (2016-2018).

For the next edition, we shall continue working with 'fanfiction' [11, 10]. This term refers to the global phenomenon of non-professional authors taking up the production of fiction in the tradition of well-known cultural domains, called 'fandoms', such as J.K. Rowling's Harry Potter or Sherlock Holmes [8]. The abundance of data is a major advantage, as fanfiction is nowadays estimated to form the fastest growing form of online writing [3]. Fan writers actively aim to increase their readership and on most platforms (e.g., archiveofourown.org or fanfiction.net), the bulk of writings can be openly accessed, although the intellectual rights are not unproblematic [29]. The multilingualism of the phenomenon is another asset, extending far beyond the Indo-European languages that are the traditional focus of shared tasks. Finally, fanfiction is characterized by a relative wealth of author-provided metadata, relating to the textual domain (the fandom), period of production, and intended audience.

Cross-domain Authorship Verification at PAN'20

In 2020, we shall visit the task of authorship verification again: as opposed to authorship attribution, which requires a carefully balanced classification setup, authorship verification is a more fundamental task. Authorship verification can be formalized as the task of approximating the target function $\phi : (D_k, d_u) \rightarrow \{T, F\}$, where D_k is a set of documents of known authorship by the same author and d_u is a document of questioned authorship. If $\phi(D_k, d_u) = T$, then the author of D_k is also the author of d_u and if $\phi(D_k, d_u) = F$, then the author of D_k is not the same with the author of d_u. In cross-domain settings, D_k and d_u do not share topic, genre or even language (in our case the fandom is different). A simple form of the verification task is to only consider the case where D_k is singleton, thus only pairs of documents are examined. Given a training set of such text pairs, verification systems can be trained and calibrated to analyze the authorship of unseen pairs. Such verifiers produce a score in the form of a bounded scalar between 0 and 1, indicating the probability of the test item being a same-author pair (rather than a binary choice).

The nature of the relationship between the training set and test set and their exact composition is crucial to the difficulty of the task. For PAN'20, we shall vary these along a number of dimensions. (I) The ratio of same-author pairs (SA) over the number of different-author (DA) pairs: while this ratio is extremely low in real-world settings, computational systems benefit from under-sampling DAs to achieve a better balance. (II) Systems are known to be very sensitive to changes in domain and topic: whether or not train and test pairs are extracted from the same fandom(s) will strongly affect performance [1]. Including multiple fandoms into training and/or test pairs is another valuable aspect for experimentation. (III) Overfitting on specific authors is a real danger during training: allowing authors to contribute more than one text during the construction of training pairs might affect performance. Likewise, one explicitly can vary the number of test authors (if any) that have not been encountered in training. (IV) Text length is another challenge [13]: short documents are more difficult to analyze and text pairs that significantly differ in length also present an important obstacle.

We shall extract a number of datasets exploring these aspects from a recent large-scale crawl from an established fan platform fanfiction.net, that contains over 5.8M stories, in 44 languages, distributed over about 10,300 fandoms. We intend to apply various techniques to estimate the degree of topical divergence between individual fandoms. These estimates will be useful to construct datasets of varying complexity. The large size of these datasets will be a novel contribution to the state of the art: whereas a larger number of different authors typically degrades the performance of authorship attributors [15], the same is not necessarily true for verification systems, that are intrinsically better suited to learn from a variety of authorial styles [13]. Finally, our aim is to also release these datasets outside of the strict TIRA environment, in order to further lower the barrier for experimentation and stimulate the data's wider adoption in the community.

5 Style Change Detection

The goal of the style change detection task is to identify the text positions within a given multi-author document at which the author switches, based on an intrinsic style analysis. Detecting these positions is a crucial part of the authorship identification process, and for multi-author document analysis in general—documents which have not been studied a lot to date.

This task has been part of PAN since 2016, with varying task definitions, datasets and evaluation procedures. In 2016, participants were asked to identify and group fragments of a given document that correspond to individual authors [26]. 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 [28]. However, this task was deemed as highly complex and hence, was relaxed in 2018, asking participants to predict whether a given document is single- or multi-authored [11]. Given the promising results achieved, in 2019, participants were asked to firstly detect whether a document was single- or multi-authored and, if it was indeed written by multiple authors, to predict the number of authors [32].

Style Change Detection at PAN'20

Given the key role of this task and the progress made in previous years, at PAN'20, we will continue to advance research in this direction. We aim to steer the task back to it's original goal: detecting the exact position of authorship changes. Therefore, the task for PAN'20 is to find the positions of style changes at the paragraph-level. For each pair of consecutive paragraphs of a document, we ask participants to estimate whether there is indeed a style change between those two paragraphs. This binary classification task will be performed on a dataset curated based on a publicly available dump of a Q&A platform to cover different types of documents at different lengths and topics. We will distill two different datasets: one featuring a rather narrow set of topics being discussed, and a second dataset containing a broad variety of topics. This setup allows for analyzing the performance of the developed approaches in dimensions of text length, topics, and the number of contributing authors.

Acknowledgments

The work of Anastasia Giachanou is supported by the SNSF Early Postdoc Mobility grant P2TIP2_181441 under the project Early Fake News Detection on Social Media, Switzerland. The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

Bibliography

[1] Bevendorff, J., Hagen, M., Stein, B., Potthast, M.: Bias analysis and mitigation in the evaluation of authorship verification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6301–6306. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1634, https://www.aclweb.org/anthology/P19-1634

[2] Darwish, K., Alexandrov, D., Nakov, P., Mejova, Y.: Seminar users in the arabic twitter sphere. In: International Conference on Social Informatics. pp. 91–108. Springer (2017)

[3] Fathallah, J.: Fanfiction and the Author. How FanFic Changes Popular Cultural Texts. Amsterdam University Press (2017)

[4] Ghanem, B., Buscaldi, D., Rosso, P.: Textrolls: Identifying russian trolls on twitter from a textual perspective. arXiv preprint arXiv:1910.01340 (2019)

[5] Ghanem, B., Paolo Ponzetto, S., Rosso, P.: Factweet: Profiling fake news twitter accounts. arXiv preprint arXiv: 1910.06592 (2019)

[6] Ghanem, B., Rosso, P., Rangel, F.: An emotional analysis of false information in social media and news articles. arXiv preprint arXiv:1908.09951 (2019)

[7] Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 877–880. ACM (2019)

[8] Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press (2014)

[9] Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval **1**(3), 233–334 (2006)

[10] Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019), http://ceur-ws.org/Vol-2380/

[11] 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 Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al. pp. 1–25 (2018)

[12] Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology **60**(1), 9–26 (2009)

[13] 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)

[14] Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)

[15] Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Digital Scholarship in the Humanities **26**(1), 35–55 (08 2010). https://doi.org/10.1093/llc/fqq013, https://doi.org/10.1093/llc/fqq013

[16] Popat, K., Mukherjee, S., Yates, A., Weikum, G.: Declare: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416 (2018)

[17] Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J.M., Köhler, J., Lötzsch, W., Müller, F., Müller, M.E., Paßmann, R., Reinke, B., Rettenmeier, L., Rometsch, T., Sommer, T., Träger, M., Wilhelm, S., Stein, B., Stamatatos, E., Hagen, M.: Who wrote the web? revisiting influential author identification research applicable to information retrieval. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. pp. 393–407. Springer International Publishing, Cham (2016)

[18] 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)

[19] Rangel, F., Rosso, P.: Overview of the 7th author profiling task at pan 2019: Bots and gender profiling. In: Cappellato L., Ferro N., Müller H, Losada D. (Eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2019)

[20] 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: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180 (2014)

[21] 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: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)

[22] Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179 (2013)

[23] 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)

[24] Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391 (2015)

[25] 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: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)

[26] 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: Fuhr, N., Quaresma, P., Larsen, B., Gonçalves, T., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16). Springer, Berlin Heidelberg New York (Sep 2016). https://doi.org/10.1007/978-3-319-44564-9_28

[27] Stamatatos, E.: A survey of modern authorship attribution methods. JASIST **60**(3), 538–556 (2009). https://doi.org/10.1002/asi.21001, https://doi.org/10.1002/asi.21001

[28] 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: Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al. pp. 1–22 (2017)

[29] Tushnet, R.: Legal fictions: Copyright, fan fiction, and a new common law. Loyola of Los Angeles Entertainment Law Review **17**(3) (1997)

[30] Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. In: 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics (Jul 2019)

[31] Wiegmann, M., Stein, B., Potthast, M.: Overview of the Celebrity Profiling Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)

[32] Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the style change detection task at PAN 2019. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380