Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection

Extended Abstract

Janek Bevendorff,[1] BERTa Chulvi,[2] Gretel Liz De La Peña Sarracén,[2] Mike Kestemont,[3] Enrique Manjavacas,[3] Ilia Markov,[3] Maximilian Mayerl,[4] Martin Potthast,[5] Francisco Rangel,[6] Paolo Rosso,[2] Efstathios Stamatatos,[7] Benno Stein,[1] Matti Wiegmann,[1] Magdalena Wolska,[1] and Eva Zangerle[4]

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

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

Abstract The paper gives a brief overview of the three shared tasks to be organized at the PAN 2021 lab on digital text forensics and stylometry hosted at the CLEF conference. The tasks include authorship verification across domains, author profiling for hate speech spreaders, and style change detection for multi-author documents. In part the tasks are new and in part they continue and advance past shared tasks, with the overall goal of advancing the state of the art, providing for an objective evaluation on newly developed benchmark datasets.

1 Introduction

The PAN workshop series has been organized since 2007 and 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 51 shared tasks have assembled evaluation resources for the aforementioned research disciplines that amount to 48 datasets plus nine datasets contributed by the community.[1] Each new dataset introduced new variants of author identification, profiling, and author obfuscation tasks as well as multi-author analysis and determining the morality, quality, or originality of a text. The 2021 edition of PAN continues in the same vein, introducing new resources and previously unconsidered problems to the community. As in earlier editions, PAN is committed to reproducible research in IR and NLP and all shared tasks will ask for software submissions on our TIRA platform [10].

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

2 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, and language variety, among others. During the years 2013-2020 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 [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 (training on Twitter data and testing on blogs and social media data) in English, Spanish, and Dutch [19]. In 2017, we addressed gender and language variety identification in Twitter in English, Spanish, Portuguese, and Arabic [17]. 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 [15].

In 2019 the focus was on profiling bots and discriminating bots from humans on the basis of textual data only [13]. 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 [11]. 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.

Haters: Profiling Hate Speech Spreaders on Twitter at PAN'21

Hate speech (HS) is commonly defined as any communication that disparages a person or a group on the basis of some characteristic, such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or others [8]. Given the huge amount of user-generated content on the Web and, in particular, on social media, the problem of detecting and, if possible, contrasting the HS diffusion, is becoming fundamental, for instance, for fighting against misogyny and xenophobia [1]. Having previously profiled bots and fake news spreaders, at PAN'21 we will focus on PROFILING HATE SPEECH SPREADERS in social media, more specifically on Twitter. We will address the problem both in English and Spanish, as we did in the previous author profiling tasks. The goal will be to identify those Twitter users that can be considered haters, depending on the number of tweets with hateful content that they had spread (tweets will be manually

[2] To generate the datasets, we have followed a methodology that complies with the EU General Data Protection Regulation [12].

annotated). As an evaluation setup, we will create a collection that contains Spanish and English tweets posted by users on Twitter. One document will consist of a feed of tweets written by the same user. The goal will then be to classify the user as hater or not hater (binary classification). Given that we plan to create a balanced dataset (although this is not a realistic scenario,[3] we balance the dataset to reinforce the haters' view and to prevent machine/deep learning models from being skewed towards tweets), we will use accuracy as the evaluation metric for the binary classification.

3 Author Identification

Authentication is a major safety issue in today's digital world and in this sense it is unsurprising that (computational) author identification has been a long-standing task at PAN. Author identification still poses a challenging empirical problem in fields related to Information and Computer Science, but the underlying techniques are nowadays also frequently adopted as an auxiliary component in other application domains, such as literary studies or forensic linguistics. These scholarly communities are strongly dependent on reliable and transparent benchmark initiatives that closely monitor the state of the art in the field and enable progress [9]. 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 [21, 4, 7]. At different editions of PAN since 2007, author identification has been studied in multiple incarnations: as authorship attribution (given a document and a set of candidate authors, determine which of them wrote the document; 2011-2012 and 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), and obfuscation evaluation (devise and implement performance measures that quantify the safeness, soundness, and/or sensibleness of obfuscation software; 2016-2018).

For the next edition, we will continue to capitalize on so-called fanfiction, as we did in previous years [6, 5]. 'Fanfiction' or 'transformative literature' refers to the worldwide cultural phenomenon of (non-professional) writers producing (largely unauthorized) literary fiction in the tradition of well-known, influential domains in culture, called 'fandoms', such as J.K. Rowling's Harry Potter or Sherlock Holmes [3]. Fanfiction is nowadays estimated to be the fastest growing form of online writing [2] and the abundance of data in this area is a major asset. Typically, fan writers actively aim to attract more readers and on most platforms (e.g. archiveofourown.org or fanfiction.net) the bulk of their writings can be openly accessed although the intellectual rights relating to these texts are convoluted [23]. Multilinguality of the phenomenon is another asset since fanfiction extends far beyond the Indo-European language area that is the traditional focus of shared tasks. Finally, fanfiction is characterized by a remarkable

[3] In a realistic scenario, we would need to know a priori the distribution of haters vs non-haters: this information is unknown and impossible to calculate manually; one of the aims of this shared task is to foster research on profiling haters in order to address this problem automatically.

wealth of author-provided metadata related to the textual domain (the fandom), popularity (e.g. number of 'kudos'), time of publication, and even intended audience (e.g. maturity ratings).

Cross-domain Authorship Verification at PAN'21

Fanfiction provides an excellent source of material to study cross-domain attribution scenarios since users usually publish narratives that range over multiple domains, the previously-mentioned 'fandoms': Harry Potter, Twilight, Marvel comics, for instance. Previous editions of PAN, in particular the last one, have already included a cross-domain authorship attribution task set in the context of fanfiction. Two basic cross-domain setups specific to fanfiction (training and test documents from disjoint fandoms) were examined: closed-set attribution (the true author of a test document belongs to the set of candidates) and open-set attribution (the true author of a test document could not be one of the candidates). For the 2021 edition, we will focus on the (OPEN) AUTHORSHIP VERIFICATION scenario: given two documents belonging to different fandoms, determine whether they are written by the same, previously unseen author. This is a fundamental task in author identification and all cases, be it closed-set or open-set ones, and can be decomposed into a series of verification instances. Again exploiting fanfiction – where the topic is easily controlled and a larger volume (on the order of thousands) of verification instances can be produced covering multiple languages – we will also attempt to mitigate the effect of certain weaknesses identified in the evaluation framework of previous authorship verification evaluations (e.g., ensuring that each verification instance is handled separately).

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 [22]. 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 [6]. 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 [25]. 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 thee 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 [24].

Style Change Detection at PAN'21

In today's scientific practice, usually a team of researchers is involved in writing a paper and conducting the underlying research—research work is teamwork. Hence, a fundamental question is the following: If multiple authors together have written a text, can we find evidence for this fact, e.g., do we have a means to detect variations in the writing style? Answering this question belongs to the most difficult and most interesting challenges in author identification and is the only means to detect plagiarism in a document if no comparison texts are given; likewise, it can help to uncover gift authorships, to verify a claimed authorship, and to develop new technology for writing support. We tackle this challenge by providing STYLE CHANGE DETECTION tasks of increasing difficulty which will attract both novices and experts in the field of authorship analytics: (1) Single vs. Multiple authors: given a text, find out whether the text is written by a single author or by multiple authors, (2) Style Change Basic: given a text written by two authors that contains a single style change only, find the position of this change, i.e., cut the text into the two authors' texts; note that this task corresponds to authorship verification where the two authors are responsible for the beginning and the end of the text respectively, (3) Style Change Real-World: 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 you assume for the multi-author document. For this year's edition, we will introduce a new type of corpus which is based on a publicly available dump of a Q&A platform and which is particularly suited for these tasks because of its topic homogeneity. For all three task variants, we will guarantee that each paragraph in a text is authored by a single author, in other words, a style change may be observed only at the beginning of a paragraph.

Acknowledgments

The work of the researchers from 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).

Bibliography

[1] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., Sanguinetti, M.: SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In: Proc. of the 13th Int. Workshop on Semantic Evaluation (SemEval-2019), co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019) (2019)

[2] Fathallah, J.: Fanfiction and the Author. How FanFic Changes Popular Cultural Texts. Amsterdam University Press (2017)
[3] Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press (2014)
[4] Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval **1**(3), 233–334 (2006)
[5] Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)
[6] 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)
[7] 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)
[8] Nockleby, J.T.: Hate speech. In: Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan), pp. 1277–1279 (2000)
[9] 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, Springer International Publishing (2016)
[10] 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
[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] 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)

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

[22] 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)

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

[24] 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)

[25] 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)