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

Janek Bevendorff,[1] Berta Chulvi,[2] Elisabetta Fersini,[3] Annina Heini,[4] Mike Kestemont,[5] Krzysztof Kredens,[4] Maximilian Mayerl,[6] Reynier 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, Germany
[2]Universitat Politècnica de València, Spain
[3]Universitty Milano-Bicocca, Italy
[4]Aston University, UK
[5]University of Antwerp, Belgium
[6]University of Innsbruck, Austria
[7]Leipzig University, Germany
[8]Symanto Research, Germany
[9]University of the Aegean, Greece

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

Abstract The paper gives a brief overview of three shared tasks which have been 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 and author 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). 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. 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

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

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 continued 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 ask for software submissions on our TIRA platform [11]. We briefly outline the 2022 tasks and results 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 [9]. Previous editions of PAN included across-domain authorship verification tasks where texts of known and unknown authorship come from different domains [28, 3, 2]. 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 [28]. 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.

In the current edition of the authorship verification task we adopt the simplified version used in the most recent PAN editions [2, 3] where text pairs are considered. Formally, one has to approximate the target function $\phi : (d_k, d_u) \rightarrow \{T, F\}$, d_k being a text of known authorship and d_u being a text of unknown or disputed 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 as the author of d_u. The main novelty of the current edition is that d_k and d_u belong to different discourse types.

Dataset

A new dataset has been created based on the recent Aston 100 Idiolects Corpus in English [2] including a rich set of DTs written by around 100 individuals. We

[2] https://fold.aston.ac.uk/handle/123456789/17

used the following DTs: emails, essays, 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).

First, we split available individuals into two equal and non-overlapping sets, one to be used for the training dataset and the other for the test dataset. That way, it is ensured that any kind of particularities among the training authors will not affect the performance on the test dataset. In addition, we took advantage of available demographic metadata and used a similar gender distribution of individuals in both training and test datasets.

The dataset comprises a set of text pairs and in each pair the two texts belong to two different DTs. All six combinations of the four available DTs are taken into account. However, the distribution of text pairs over the combination of DTs is not homogeneous since it depends on the available texts belonging to each DT. For example, the corpus comprises only one business memo and multiple email messages per individual. Anyway, the distribution of verification instances per DT combination is similar in both training and test datasets as can be seen in Table 1. Similar, both training and test datasets have balanced distribution of positive/negative verification cases. This is also valid for each combination of DTs (e.g., half of the pairs belonging to the combination essay-email is positive and the other half is negative).

Since the length of texts belonging to certain DTs is very small, we concatenated multiple texts of the same DT to produce longer text samples that are used in the text pairs of authorship verification instances. In more detail, email messages are concatenated so that a text sample of at least 2,000 characters is obtained. The date of email messages is taken into account so that consecutive messages are concatenated. In the case of text messages, we concatenate messages sent either to friends or family so that text samples of at least 500 characters are obtained. The text length information provided in Table 1 for email and text messages refers to text samples produced as explained above.

Evaluation Setup and Results

The evaluation framework is similar to the one used in recent shared tasks at PAN. For each AV instance (a text pair) of the test dataset, participants have to produce a scalar score a_i (in the $[0, 1]$ range) indicating the probability that the pair was written by the same author. It is possible for participants to leave text pairs unanswered by submitting a score of precisely $a_i = 0.5$. As concerns the experimental setup, the set of evaluation measures used in the last edition of PAN is also adopted. These include the area under ROC (AUROC), c@1 that rewards unanswered cases over wrong predictions, F_1, $F_{0.5u}$, and the complement of Brier score (so that higher scores correspond to better performance) [2]. The average of these diverse measures is used as final score to rank participants.

Two baseline approaches were made available to the participants: a compression-based approach based on Prediction by Partial Matching

Table 1. Statistics of the new dataset used in the authorship verification task.

	Training	Test
Text pairs		
Positive	6,132 (50.0%)	5,239 (50.0%)
Negative	6,132 (50.0%)	5,239 (50.0%)
Email - Text message	7,484 (61.0%)	6,092 (58.1%)
Essay - Email	1,618 (13.2%)	1,454 (13.9%)
Essay - Text message	1,182 (9.6%)	1,128 (10.8%)
Business memo - Email	1,014 (8.3%)	900 (8.6%)
Business memo - Text message	780 (6.4%)	718 (6.9%)
Essay - Business memo	186 (1.5%)	186 (1.8%)
Text length (avg. chars)		
Essay	11,098	10,117
Email	2,385	2,323
Business memo	1,255	1,042
Text message	611	601

Table 2. Final results for the cross-discourse-type authorship verification task at PAN'22. Submitted systems are ranked by their mean performance across five evaluation metrics. Best result per column is shown in bold.

System	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
BASELINE-cngdist	0.546	0.496	0.669	0.542	0.749	**0.600**
najafi22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
galicia22	0.512	0.499	0.628	0.544	0.741	0.585
jinli22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-compressor	0.541	0.493	0.570	0.478	**0.750**	0.566
lei22	0.539	0.539	0.399	0.488	0.539	0.501
yihuiye22	0.542	0.526	0.398	0.461	0.565	0.499
huang22	0.519	0.519	0.196	0.328	0.519	0.416
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(PPM) [30] and a naive distance-based character n-gram model [7]. We received 7 submissions and evaluated their performance using the TIRA experimentation framework. The overall results of all participants and the baselines can be found in Table 2.

As can be seen, the general performance of all submissions is quite low reflecting the difficulty of the task. It is surprising that a naive baseline achieved the best overall score despite the fact that most participant models are quite sophisticated. On the other hand, the most effective submitted method (najafi22) outperforms all other submissions and baselines in three out of five evaluation measures indicating a promising potential. More details on the evaluation results and the submissions will be available in the task overview paper [27].

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.[3] In 2013 the aim was to identify gender and age in social media texts for English and Spanish [18]. 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 [16]. In 2015, apart from age and gender identification, we addressed also personality recognition on Twitter in English, Spanish, Dutch, and Italian [20]. 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 [21]. In 2017, we addressed gender and language variety identification in Twitter in English, Spanish, Portuguese, and Arabic [19]. 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 [17]. In 2019 the focus was on profiling bots and discriminating bots from humans on the basis of textual data only [15]. 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 [13]. 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 [12]. 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.

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 [22]. 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 [6]. Stereotypes are often used, especially in discussions about controversial issues such as immigration [29] or sexism [23] and misogyny [1]. At PAN 2022 we focused on profiling ironic authors in Twitter. Special emphasis was given to those authors that employ irony to spread stereotypes. The goal was to classify authors as ironic or not depending on their number of tweets with ironic content. Among those authors we considered 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

[3] To generate the datasets, we have followed a methodology that complies with the EU General Data Protection Regulation [14].

tweets, the goal was 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 created a collection that contains tweets posted by users in Twitter. One document consisted of a feed of tweets written by the same user.

Taxonomy of Stereotype Categories

Recently [26] developed the Social Bias Frame, a new conceptual formalism that aims to model the pragmatic frames in which people project social biases and stereotypes onto others. To support this research they developed the Social Bias Inference Corpus (SBIC) with 150K structured annotations of social media posts covering 34k implications about social groups. For example: "If cameras do really add ten pounds, do Africans really exist?". For each post, annotators from Amazon Mechanical Turk indicate whether or not: (i) the post is offensive, (ii) the intent is to offend, and (iii) it contains lewd or sexual content. Only if annotators indicate potential offensiveness they answer the group implication question: who is referred to/targeted by this post? Two possible answers were: (i) yes, this could be offensive to a group and (ii) no, this is just an insult to an individual or a non-identity-related group of people. If the post targets or references a group or demographic, annotators select or write which group is referenced. For each selected group, they then write two to four stereotypes that are used in this post; for the given example, annotators write as stereotype: "Africans are all starving". Finally, workers are asked whether they think the speaker is part of one of the minority groups referenced by the post. From 16,739 instances in SBIC, 8,167 refer to a group of people in the field of "target minority".

To build the IROSTEREO corpus we examine the "target minority" field of SBIC and we identify 600 unique labels that could be considered a social group or a social category. We define a social category following a long tradition of research in Social Psychology [4] which considers that a social group exists when two or more persons define themselves as members of the group and when their existence is recognised by at least one other person. [26] classify the groups referenced in seven categories: (1) body (2) culture (3) disabled (4) gender (5) race (6) social and (7) victims. In order to focus specifically on stereotypes as the expression of a prejudice against certain groups or social categories that are often the object of an ironic and hurtful discourse we create a more granular taxonomy to classify the 600 labels in 17 categories: (1) national majority groups, (2) illness/health groups, (3) age and role family groups, (4) victims, (5) political groups, (6) ethnic/racial minorities, (7) immigration/national minorities (8) professional and class groups, (9) sexual orientation groups, (10) women, (11) physical appearance groups, (12) religious groups, (13) style of life groups, (14) non-normative behaviour groups, (15) man/male groups, (16) minorities expressed in generic terms and (17) white people. As keywords to retrieve the tweets we use the labels associated to groups only from categories 5 to 14 of the taxonomy.

Dataset and Annotation Process

The Twitter API was used to retrieve tweets with two conditions: (i) tweets that contain the hashtag #irony or #sarcasm and at least one of the labels included in categories 5 to 14 of the taxonomy and (ii) the same labels about social groups but without #irony or #sarcasm. Users with more cases in classes 1 and 2 were identified and the tweets that accomplish these two conditions were downloaded. The annotators had to identify ironic tweets and tweets that use stereotypes among this set of users. To identify irony, the annotators were asked to mark the tweets where the user "expresses the opposite of what was saying as a disguised mockery". If a user had more than five ironic tweets it was labelled as ironic.

Positive examples of classes 1 (users that express irony without stereotypes), 2 (non-ironic users that use stereotypes) and 3 (users that express irony and use stereotypes) were selected and 200 tweets from their timeline were downloaded. To find the non-ironic and non-stereotype class (4) the lexicon used in the three previous classes was analysed in order to reduce topic bias. Moreover, tweets should not contain the labels of social groups associated to stereotypes. A second annotation was done to check that class 4 does not contain irony.

Table 3 presents the statistics of the corpus that consists of 600 authors for English language, completely balance between the two classes (ironic and non ironic), and with a 66/33 balance between users using stereotypes or not for each class. For each author, we retrieved via the Twitter API their last 200 tweets. We have split the corpus into training and test sets, following a proportion of 70/30 for training and testing respectively.

Table 3. Number of authors in the PAN-AP-22 corpus distributed between the two classes, Ironic vs Non-Ironic, and within each class, distributed between users who use stereotypes vs. users who do not use stereotypes.

	Ironic			**Non Ironic**			
Set	**Stereotypes**	**Non stereo.**	**Total**	**Stereotypes**	**Non stereo.**	**Total**	**Total**
Training	140	70	210	140	70	210	420
Test	60	30	90	60	30	90	180
Total	200	100	300	200	100	300	600

Evaluation Setup

Since the dataset is completely balanced for the two target classes, ironic vs. non ironic, we have used the accuracy measure and ranked the performance of the systems by that metric. More than 60 teams participated in the IROSTEREO author profiling task. At the moment of the writing-up of this overview paper, we are still evaluating the last submissions. The results will be presented in the IROSTEREO overview paper [10].

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 at which the author switches in a multi-author document. Style change detection is a crucial part of the authorship identification process and multi-author document analysis. 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 [24]. In 2017, the task was to detect whether a given document is multi-authored. If the document was indeed multi-authored, participants were asked to determine the positions at which authorship changes [31]. Since this task was deemed highly complex, we reduced the complexity of the task in 2018 and asked participants to predict whether a given document is single- or multi-authored [8], which has to lead promising results. In 2019, participants were asked first to detect whether a document was single- or multi-authored and to predict the number of authors if it was indeed written by multiple authors [35]. In 2020, we steered the task back to its original definition, i.e., to find the positions at which authorship changes. We asked participants to first determine whether a document was written by one or by multiple authors and, for multi-author documents, they had to detect between which paragraphs the authors change [34]. Continuing these efforts, in the 2021 edition, we asked participants to first 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 and identify which author wrote any given paragraph [32].

Multi-Author Writing Style Analysis at PAN'22

The analysis of author writing styles is the foundation for author identification. 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.

Based on the progress made towards this goal in previous years and to entice novices and experts, we extend the set of challenges. Therefore, the style change detection task at PAN'22 involves three subtasks in increasing difficulty: (1) Style Change Basic (subtask1): for 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 on the paragraph-level), (2) Style Change Advanced (subtask2): for 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), and (3) Style Change Real-World (subtask3): for 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.

Data set and evaluation

The datasets underlying this task were created from posts of the popular StackExchange network of Q&A sites. Based on a dump of questions and answers from the StackExchange network, we extracted a subset of topics (so-called sites)[4]. Initial data cleaning involved removing questions and answers that were edited after they were originally posted and removing images, URLs, code snippets, block quotes, and bullet lists from all questions and answers. The general procedure for generating one of our datasets then works as follows. All questions and answers were split into paragraphs; we removed paragraphs of less than 100 characters. Based on these paragraphs, we create documents by drawing paragraphs from a single question thread to ensure that topic changes cannot be leveraged for detecting style changes. We randomly pick the number of authors per document between one and five. Following that, we randomly choose a corresponding number of authors from the authors who contributed to the question thread we were drawing paragraphs from. In the next step, we take the paragraphs written by the selected authors and shuffle them to obtain the final documents. If a resulting document has fewer than two paragraphs or is fewer than 1,000 or more than 10,000 characters long, we discard it.

We applied this procedure, with slightly different parameters, to generate a separate dataset for each of this year's three subtasks. For the dataset for subtask 1, we ensured that every generated document has exactly one style change in it. For subtask 2, we used the procedure exactly as outlined above. For subtask 3, we changed the procedure to operating on sentences instead of paragraphs. The three datasets we obtained in this way contain a total of $2,000$, $10,000$, and $10,000$ documents, respectively, and were then all split into training, validation, and test sets. The training sets consist of 70% of all generated documents for a given dataset, whereas the test and validation set each consist of 15% of the documents.

The three subtasks are evaluated independently. As primary evaluation metric, we compute the macro-averaged F1-score value across all documents. To add a further perspective on the results obtained, we evaluate two further measures for subtask 2: Diarization Error Rate (DER) [5] and Jaccard Error Rate (JER) [25]. These measures essentially capture the fraction of text that is not correctly attributed to an author and are borrowed from the field of text transcription.

[4]The following StackExchange sites were used: Code Review, Computer Graphics, CS Educators, CS Theory, Data Science, DBA, DevOps, GameDev, Network Engineering, Raspberry Pi, Superuser, and Server Fault.

Table 4. Overall results for the style change detection task, ranked by average F_1 performance across all three subtasks (ST).

Participant	ST1 F_1	ST2 F_1	ST3 F_1	ST3 DER	ST3 JER
Intrinsic Approaches					
tzumilin22	0.7540	0.5100	0.7156	0.8059	0.6905
xinyin22	0.7346	0.4687	0.6720	0.7620	0.6862
qidilao22	0.7471	0.4170	0.6314	0.7364	0.6359
zhang22	0.7162	0.4174	0.6581	0.7114	0.6444
yang22	0.6690	0.4011	0.6483	0.7036	0.6323
alvi22	0.7052	0.3213	0.5636	0.6076	0.4782
castro22a	0.5661	0.2735	0.5565	0.5965	0.4229
alshmasy22	0.5272	0.2207	0.4995	0.5760	0.3557
Extrinsic Approaches					
graner22	0.9932	0.9855	0.9929	0.9960	0.9960

Results

The style change detection task received nine software submissions, eight of which used intrinsic approaches and one used an extrinsic approach. The individual results achieved by the participants are presented in Table 4. For the intrinsic approaches, the best results were achieved by *tzumilin22*, who obtained the highest score for every subtask and evaluation metric. Further details on the approaches taken can be found in the overview paper [33].

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).

Bibliography

[1] Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In: Proc. 23rd Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2018, Springer-Verlag, LNCS(10859), pp. 57-64 (2018)

[2] 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: Candan, K.S., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, Lecture Notes in Computer Science, vol. 12880, pp. 419–431, Springer (2021)

[3] 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: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, Lecture Notes in Computer Science, vol. 12260, pp. 372–383, Springer (2020)

[4] Brown, R.: Prejudice: Its social psychology. John Wiley & Sons (2011)

[5] Fiscus, J.G., Ajot, J., Michel, M., Garofolo, J.S.: The rich transcription 2006 spring meeting recognition evaluation. In: International Workshop on Machine Learning for Multimodal Interaction, pp. 309–322, Springer (2006)

[6] Frenda, S., Cignarella, A., Basile, V., Bosco, C., Patti, V., Rosso, P.: The unbearable hurtfulness of sarcasm. In: Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2021.116398 (2022)

[7] Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., Daelemans, W.: Authenticating the writings of julius caesar. Expert Systems with Applications **63**, 86–96 (2016)

[8] 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)

[9] 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)

[10] Ortega-Bueno, R., Chulvi, B., Rangel, F., Rosso, P., Fersini, E.: Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)

[11] 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

[12] 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)
[13] 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)
[14] 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)
[15] 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)
[16] 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)
[17] 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)
[18] 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)
[19] 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)
[20] 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)
[21] 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
[22] Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: Beyond a simple case of negation. In: Knowledge and Information Systems, vol. 40, issue 3, pp. 595-614 (2014)
[23] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of exist 2021: Sexism identification in social networks. In: Procesamiento del Lenguaje Natural (SEPLN), num. 67, pp. 195-207 (2021)
[24] 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)
[25] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., Liberman, M.: The second dihard diarization challenge: Dataset, task, and baselines. arXiv preprint arXiv:1906.07839 (2019)
[26] Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A., Choi, Y.: Social bias frames: Reasoning about social and power implications of language. In:

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5477–5490, Association for Computational Linguistics, Online (Jul 2020), https://doi.org/10.18653/v1/2020.acl-main.486, URL https://aclanthology.org/2020.acl-main.486

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

[28] Stamatatos, E., Potthast, M., Pardo, F.M.R., Rosso, P., Stein, B.: Overview of the PAN/CLEF 2015 evaluation lab. In: Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G.J.F., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, Lecture Notes in Computer Science, vol. 9283, pp. 518–538, Springer (2015)

[29] Sánchez-Junquera, J., Chulvi, B., P, P.R., Ponzetto, S.: How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. In: Applied Science, 11(8), 3610 (2021)

[30] Teahan, W.J., Harper, D.J.: Using Compression-Based Language Models for Text Categorization, pp. 141–165. Springer Netherlands (2003)

[31] 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)

[32] 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)

[33] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2022. In: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (2022)

[34] 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)

[35] 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)