Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection*

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

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

pan@webis.de https://pan.webis.de

Abstract We briefly report on the four shared tasks organized as part of the PAN 2020 evaluation lab on digital text forensics and authorship analysis. Each tasks is introduced, motivated, and the results obtained are presented. Altogether, the four tasks attracted 228 registrations, yielding 82 successful submissions. This, and the fact that we continue to invite the submissions of software rather than its run output using the TIRA experimentation platform, marks for a good start into the second decade of PAN evaluations labs.

1 Introduction

The PAN 2020 evaluation lab organized four shared tasks related to authorship analysis, i.e., the analysis of authors based on their writing style. Two of the tasks addressed the profiling of authors with respect to traditional demographics as well as new ones from two perspectives: whether the authors are inclined to spread fake news, and whether the stylometric properties of demographic are also represented in their followers' text. The third task started a new evaluation cycle on authorship verification as the core authorship analysis discipline, starting with closed-set attribution on a significantly improved dataset. The fourth task addressed the important, yet exceedingly difficult task of handling multi-author documents and the detection of style changes within a given text written by more than one author.

In this paper, each of the following sections gives a brief, condensed overview of the four aforementioned tasks, including their motivation and the results obtained.

* Authors are listed in alphabetical order.

2 Authorship Verification

From the very beginning onward, authorship analysis tasks have played a key role in the PAN series [1]. Many task variations have been devised over the last decade, including the development of the respective corpora for authorship attribution, authorship clustering, and authorship verification, both within and across genres, and within and across languages. This year we opted for a task in the domain of authorship verification, that fits in a renewed three-year strategy, via which we aim to contribute tasks of an increasing difficulty and realism. In this endeavour, special attention will go out to open challenges in the field, such as topical shifts (author-topic orthogonality), text varieties (cross-genre authorship) and limited text length.

2.1 Dataset

This year, two training datasets of different magnitudes ("small" and "large") are provided with text pairs, crawled from `fanfiction.net`, a sharing platform for fanfiction that comes from various topical domains (or 'fandoms') and with rich, user-contributed metadata [7]. Participants were allowed to submit systems calibrated on either dataset (or both). All texts were heavily preprocessed to avoid textual artifacts [2] and have a length of $\approx$ 21,000 characters. To construct the dataset, we bucketed the texts by author and fandom to ensure a good mix of the two and, despite the very uneven popularity of fandoms and activity of authors, prevent gross overrepresentation of individual fandoms and authors. For the large dataset, 148,000 same-author (SA) and 128,000 different-authors (DA) pairs were drawn from the fan fiction crawl. The SA pairs encompass 41,000 authors of which at least 4 and not more than 400 have written in the same fandom (median: 29). In total, 1,600 fandoms were selected and each single author has written in at least 2, but not more than 6 fandoms (median: 2). The pairs were assembled by building all possible $\binom{n}{2}$ pairings of author texts (n being the actual number of texts from this author) without allowing two pairs with the same author *and* fandom. The small training set is a subset of the large training set with 28,000 same-author and 25,000 different-authors pairs from the same 1,600 fandoms, but with a reduced author number of 6,400 (4–68 per fandom, median: 7) and 48,500 (2–63 per fandom, median: 38), respectively. The test dataset contains 10,000 same-author and 6,900 different-authors pairs from 400 fandoms and 3,500 / 12,000 authors which are guaranteed to exist in the training sets, but either in a different author-fandom relation or in the same author-fandom relation, but with a previously unseen text. This creates a closed-set authorship identification scenario, a condition which will be broken in the next year with unseen fandoms and authors.

2.2 Evaluation

Metrics Because of the considerable size of the data sets, we opted for a combination of 4 evaluation metrics that each focus on different aspects. For each problem (i.e. individual text pair) in the test set, the participating systems submitted a scalar in the [0,1] range, indicating the probability of this being a SA pair. For a small number of difficult cases, the systems could submit a score of exactly 0.5, which was equivalent to

Submission	AUC	c@1	F0.5u	F1-score	Overall
boenninghoff20-large	**0.969**	**0.928**	**0.907**	**0.936**	**0.935**
weerasinghe20-large	0.953	0.880	0.882	0.891	0.902
boenninghoff20-small	0.940	0.889	0.853	0.906	0.897
weerasinghe20-small	0.939	0.833	0.817	0.860	0.862
halvani20b-small	0.878	0.796	0.819	0.807	0.825
kipnis20-small	0.866	0.801	0.815	0.809	0.823
araujo20-small	0.874	0.770	0.762	0.811	0.804
niven20	0.795	0.786	0.842	0.778	0.800
gagala20-small	0.786	0.786	0.809	0.800	0.796
araujo20-large	0.859	0.751	0.745	0.800	0.789
baseline (naive)	0.780	0.723	0.716	0.767	0.747
baseline (compression)	0.778	0.719	0.703	0.770	0.742
ordonez20-large	0.696	0.640	0.655	0.748	0.685
faber20-small	0.293	0.331	0.314	0.262	0.300

Table 1. Evaluation results for authorship verification at PAN-2020 in terms of area under the curve (AUC) of the receiver operating characteristic (ROC), c@1, F0.5u, F1-score and overall score (sorted by overall score). Large stands for training on the large dataset; small stands for training on the small dataset.

a non-response [9]. The following metrics were used to score the submissions: (1) **AUC**: the conventional area-under-the-curve score, in a reference implementation [10]; (2) **F1-score**: the well-known performance measure (*not* taking into account non-answers), in a reference implementation [10]; (3) **c@1**: a variant of the conventional F1-score, which rewards systems that leave difficult problems unanswered [9]; (4) **F0.5u**: a newly proposed measure that puts more emphasis on deciding same-author cases correctly [3]. The overall score is the mean of the scores of all the evaluation metrics.

Baselines We applied two baseline systems (calibrated on the small training set). (1) The first method calculates the cosine similarities between TFIDF-normalized tetragram representations of the texts in a pair. The resulting scores are shifted using a grid search on the calibration data (naive, distance-based baseline). (2) Secondly, we applied a text compression method that, given a pair of texts, calculates the cross-entropy of text2 using the Prediction by Partial Matching model of text1 and vice-versa. The mean and absolute difference of the two cross-entropies are used by a logistic regression model to estimate a score in [0,1].

2.3 Results

The authorship verification task received submissions from nine participating teams. A detailed evaluation results can be found in Table 1. A pairwise significance comparison of the F1-scores (according to approximate randomization test [15]) is shown in Table 2. The symbolic notation is based on the following thresholds: '=' (not significantly different: $p > 0.5$), '*' (significantly different: $p < 0.05$), '**' (very significantly different: $p < 0.01$), '***' (highly significantly different: $p < 0.001$). These comparisons highlight how, compared to recent editions, the received submissions used a variety of learning approaches and feature extractors. Consequently, the reported scores lie in a wide range.

	boenninghoff20-large	weerasinghe20-large	boenninghoff20-small	weerasinghe20-small	halvani20b-small	kipnis20-small	araujo20-small	niven20	gagala20-small	araujo20-large	baseline (naive)	baseline (compression)	ordonez20-large	faber20-small
boenninghoff20-large		***	***	***	***	***	***	***	***	***	***	***	***	***
weerasinghe20-large			***	***	***	***	***	***	***	***	***	***	***	***
boenninghoff20-small				***	***	***	***	***	***	***	***	***	***	***
weerasinghe20-small					***	***	***	***	***	***	***	***	***	***
halvani20b-small						=	=	***	=	=	***	***	***	***
kipnis20-small							***	***	=	=	***	***	***	***
araujo20-small								***	**	***	***	***	***	***
niven20									***	***	***	***	***	***
gagala20-small										=	***	***	***	***
araujo20-large											***	***	***	***
baseline (naive)												=	=	***
baseline (compression)													=	***
ordonez20-large														***

Table 2. Significance of pairwise differences in output between submissions (using F1-score as the reference metric).

3 Celebrity profiling

In 2019, we introduced the task of celebrity profiling [17] and organized the first competition on this task [18] with the goal of predicting the demographics age, gender, fame, and occupation of a celebrity from the matching Twitter timeline. For the continuation of the celebrity profiling task at PAN, we utilize the unique position of celebrities highly influential hubs of their communities to explore the idea of distributional author profiling: If the stylometric features of a demographic are consistent within a community, then we can profile an author from the texts of his followers. For this task, we compiled the Twitter timelines of 10 followers for 2,320 celebrities and asked participants to determine age, gender, and occupation of each celebrity by profiling the Tweets of the followers. We received submissions by 3 teams, all beating the baselines and demonstrating with a healthy margin above random that the task can be solved.

3.1 Dataset

We compiled the dataset based on the PAN19 Celebrity Profiling dataset by extracting all celebrities with an annotated birthyear between 1940 and 1999, a binary gender, and an occupation of either sports, performer, creator, politics. We discarded all celebrities with less than 1,000 followers, which left 10,585 complete celebrity profiles. For this initial set of celebrities, we compiled the follower network and collected the timelines of all followers, discarding all followers with less than 10 English tweets excluding retweets, more than 100,000 or less than 10 followers, and more than 1,000 or less than 10 followees, yielding reasonably active and well-connected followers. From the remaining list of followers, we randomly selected 10 followers for each celebrity.

Participant	cRank	Age	Gender	Occupation
hodge20	0.577	0.432	0.681	0.707
koloski20	0.521	0.407	0.616	0.597
tuksa20	0.477	0.315	0.696	0.598
baseline-oracle	0.631	0.500	0.753	0.700
baseline-ngram	0.469	0.362	0.584	0.521
expectation	0.333	0.333	0.500	0.250

Table 3. Overall results for the celebrity profiling task.

From the selected timelines, we removed retweets and non-English tweets and sampled a 2,320 celebrity dataset that is balanced by occupation and by gender, leaving 8,265 celebrities for an unbalanced, supplemental dataset. We split the 2,320 celebrity dataset roughly 80:20 into a 1,920 author training dataset and a 400 author test dataset test. We handed out the training and supplemental datasets to the participants and kept the test dataset hidden for evaluation on TIRA.

3.2 Evaluation

As in 2019, the decisive performance metric for this task is the harmonic mean of the minor metrics for each demographic:

$$cRank = \frac{3}{\frac{1}{\mathrm{F}_{1,\mathrm{age}}} + \frac{1}{\mathrm{F}_{1,\mathrm{gender}}} + \frac{1}{\mathrm{F}_{1,\mathrm{occupation}}}} \tag{1}$$

The performances of the gender and occupation predictions are evaluated as micro-averaged, multi-class F_1, which is consistent with the 2019 task on celebrity profiling. Since we commit to precisely predicting age instead of bucketing age-groups, the performance of the age predictions is evaluated with a variable-bucket strategy, where the predicted age of an author is correct if it is within an m-window of the truth. The window size m is between 2 and 9 years, increasing linearly with the true age of the author.

We released the results of three baselines at the beginning of the evaluation cycle: (1) the expected random values, (2) baseline-ngram, a logistic regression classifier using tf-idf weighted word 3-grams on the concatenated follower tweets, and (3) baseline-oracle, which is identical to baseline-ngram but uses the celebrities' timelines instead of the follower timelines.

3.3 Results

Table 3.2 shows the results of the participants with successful submissions as well as the baseline performance. All participants managed to surpass the random expectation and improve on the baseline by a healthy margin. The peak performance of the submitted solutions already closes in on the oracle-baseline, which shows that the followers' texts contain noticeable hints about the demographics of the followee. The details of the submitted solutions are discussed in the overview paper of this task [16].

4 Profiling Fake News Spreaders on Twitter

Although the detection of fake news, and credibility in general, has received a lot of research attention [6], there are only few studies that have addressed the problem from a user or author profiling perspective. For example, Shu et al. [13] analyzed different features, such as registration time, and found that users that share fake news have more recent accounts than users who share real news. Vo and Lee [14] analyzed the linguistic characteristics (e.g., use of tenses, number of pronouns) of fact-checking tweets and proposed a deep learning framework to generate responses with fact-checking intention. Recently, Giachanou et al. [5] employed a model based on a Convolutional Neural Network that combines word embeddings with features that represent users' personality traits and linguistic patterns, to discriminate between fake news spreaders and fact-checkers.

We believe that fact-checkers are likely to have a set of different characteristics compared to fake news spreaders. For example, fact-checkers may use different linguistic patterns when they share posts compared to fake news spreaders. This is what we aim at investigating in this year's author profiling shared task where we address the problem of fake news detection from the author profiling perspective. The final goal is profiling those authors that have shared some fake news in the past. This will allow for identifying possible fake news spreaders on Twitter as a first step towards preventing fake news from being propagated among social media users. This should help for their early detection and, therefore, for preventing their further dissemination.

4.1 Dataset and Evaluation

We built a dataset of fake and real news spreaders, i.e. discriminating authors that have shared some fake news in the past from those that, to the best of our knowledge, have never done it. Table 4.1 presents the statistics of the dataset that consists of 500 authors for each of the two languages, English and Spanish. For each author, we retrieved via the Twitter API her last 100 Tweets. The dataset for each language is balanced, with 250 authors for each class (fake and real news spreaders).

Therefore, the performance of the systems has been ranked by accuracy. For each language, we calculated individual accuracy in discriminating between the two classes. Finally, we averaged the accuracy values per language to obtain the final ranking.

Language	Training	Test	Total
English	300	200	500
Spanish	300	200	500

Table 4. Number of authors in the PAN-AP-20 dataset created for this task.

4.2 Results

We represent each author in the dataset by concatenating her tweets into one document and then we feed this document to the models.

	Participant	En	Es	Avg
1	bolonyai20	**0.750**	0.805	0.7775
1	pizarro20	0.735	**0.820**	0.7775
	SYMANTO (LDSE)	*0.745*	*0.790*	*0.7675*
3	koloski20	0.715	0.795	0.7550
3	deborjavalero20	0.730	0.780	0.7550
3	vogel20	0.725	0.785	0.7550
6	higueraporras20	0.725	0.775	0.7500
6	tarela20	0.725	0.775	0.7500
8	babaei20	0.725	0.765	0.7450
9	staykovski20	0.705	0.775	0.7400
9	hashemi20	0.695	0.785	0.7400
11	estevecasademunt20	0.710	0.765	0.7375
	SVM + c nGrams	*0.680*	*0.790*	*0.7350*
12	castellanospellecer20	0.710	0.760	0.7350
13	shrestha20	0.710	0.755	0.7325
13	tommasel20	0.690	0.775	0.7325
15	johansson20	0.720	0.735	0.7275
15	murauer20	0.685	0.770	0.7275
17	espinosagonzales20	0.690	0.760	0.7250
17	ikae20	0.725	0.725	0.7250
19	morenosandoval20	0.715	0.730	0.7225
20	majumder20	0.640	0.800	0.7200
20	sanchezromero20	0.685	0.755	0.7200
22	lopezchilet20	0.680	0.755	0.7175
22	nadalalmela20	0.680	0.755	0.7175
22	carrodve20	0.710	0.725	0.7175
25	gil20	0.695	0.735	0.7150
26	elexpuruortiz20	0.680	0.745	0.7125
26	labadietamayo20	0.705	0.720	0.7125
28	grafiaperez20	0.675	0.745	0.7100
28	jilka20	0.665	0.755	0.7100
28	lopezfernandez20	0.685	0.735	0.7100
31	pinnaparaju20	0.715	0.700	0.7075
31	aguirrezabal20	0.690	0.725	0.7075
33	kengyi20	0.655	0.755	0.7050
33	gowda20	0.675	0.735	0.7050
33	jakers20	0.675	0.735	0.7050
33	cosin20	0.705	0.705	0.7050
37	navarromartinez20	0.660	0.745	0.7025
38	cardaioli20	0.675	0.715	0.6950
38	females20	0.605	0.785	0.6950
	NN + w nGrams	*0.690*	*0.700*	*0.6950*
38	kaushikamardas20	0.700	0.690	0.6950
41	monteroceballos20	0.630	0.745	0.6875
42	ogaltsov20	0.695	0.665	0.6800
43	botticebria20	0.625	0.720	0.6725
43	lichouri20	0.585	0.760	0.6725
45	manna20	0.595	0.725	0.6600
46	fersini20	0.600	0.715	0.6575
47	jardon20	0.545	0.750	0.6475
	EIN	*0.640*	*0.640*	*0.6400*
48	shashirekha20	0.620	0.645	0.6325
49	datatontos20	0.725	0.530	0.6275
50	soleramo20	0.610	0.615	0.6125
	LSTM	*0.560*	*0.600*	*0.5800*
51	russo20	0.580	0.515	0.5475
52	igualadamoraga20	0.525	0.505	0.5150
	RANDOM	*0.510*	*0.500*	*0.5050*

	Participant	En
53	hoertenhuemer20	0.725
54	duan20	0.720
54	andmangenix20	0.720
56	saeed20	0.700
57	baruah20	0.690
58	anthonio20	0.685
59	zhang20	0.670
60	espinosaruiz20	0.665
61	shen20	0.650
62	suareztrashorras20	0.640
63	niven20	0.610
64	margoes20	0.570
65	wu20	0.560

Table 5. Overall accuracy of the submission to the task on profiling fake news spreaders on Twitter: The teams that participated in both languages (English and Spanish) are ranked by the average accuracy between both languages, teams that participated only in English (bottom right) are ranked by the accuracy on English. The best results for each language are printed in bold.

In total 65 teams participated in this year's author profiling task on profiling fake news spreaders on Twitter (record in terms of participants at PAN Lab). In Table 4.2 we present the results in terms of accuracy of the teams that participated in both languages and the results of the teams that addressed the problem only in English.

As baselines to compare the performance of the participants with, we have selected: (1) an LSTM that uses fastText[1] embeddings to represent texts; (2) a Neural Network (NN) with word n-grams (size 1-3) and (3) a Support Vector Machine (SVM) with char n-grams (size 2-6); (4) an SVM with Low Dimensionality Statistical Embeddings (LDSE) [12] to represent texts; (5) the Emotionally-Infused Neural (EIN) network [4] with word embedding and emotional features as the input of an LSTM, and (6) a Random prediction.

The description of the models of the participating teams and the detailed analysis of the results are presented in the shared task overview paper [11].

[1] https://fasttext.cc/docs/en/crawl-vectors.html

5 Style Change Detection

In previous editions, the style change detection task aimed at detecting whether a document is single- or multi-authored [20] or predicting the actual number of authors within a document [8]. Considering the promising results achieved in the last years, we steer the task back to its original goal: detecting the exact position of authorship changes. Therefore, the goal is to determine whether the given document contains style changes and if it indeed does, we aim to find the position of the change in the document (between paragraphs). 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. Consequently, we ask participants to answer the following two questions for a given document: (1) Task 1: Was the given document written by multiple authors? (2) Task 2: For each pair of consecutive paragraphs in the given document: is there a style change between these paragraphs?.

5.1 Dataset

For this year's style change detection task, we prepared two datasets. Both datasets were extracted from the StackExchange network of Q&A sites; nonetheless, they differ in the number and topical variety of sites included in the dataset. The first dataset, *dataset-narrow*, includes texts from StackExchange sites dealing with topics related to computer technology. The second dataset, *dataset-wide*, includes texts from a broader and larger selection of StackExchange sites, and therefore covers a broader range of topics. The goal behind using those two different datasets was to see how the topical range of texts impacts the performance of the submitted approaches.

Aside from the specific sites that were included, both datasets were generated in the same way. We used a dump of questions and answers on the StackExchange network as our data source, which we cleaned by removing questions and answers that contain fewer than 30 characters, or that were edited by a different user than the original author. We also removed images, URLs, code snippets, blockquotes, and bullet lists from all questions and answers. We then took all the questions and answers written by the same user and split them into paragraphs, dropping all paragraphs with fewer than 100 characters. This gave us a list of paragraphs for every user on a single StackExchange site. We constructed documents by drawing paragraphs from those lists. We generated an equal number of single-author and multi-author documents for our datasets. For single-author documents, the paragraphs making up the document are drawn from the paragraph list of a single user of a single StackExchange site. For multi-author documents, we combine paragraphs from the paragraph lists of two or three users, in a way that leads to the author changing between paragraphs between one and ten times for a single document; again, combining only paragraphs of the same StackExchange site. A more detailed description of the dataset generation can be found in the task overview.

Both datasets were then split into training, validation, and test sets, with 50% of the documents going into the training set and 25% each going into the validation and test set. Table 6 summarizes the properties of the documents in our datasets, and the exact composition of both the narrow and the wide dataset, showing the number of documents written by one, two, and three authors in the training, validation, and test sets of both.

Parameter	Configurations
Number of collaborators	1–3
Number of style changes	0–10
Document length	1,000–3,000
Change positions	Between paragraphs
Document language	English

Dataset	Training Set			Validation Set			Test Set		
	1	2	3	1	2	3	1	2	3
narrow	1,709	854	855	855	415	443	852	426	423
wide	4,025	1,990	2,015	2,018	969	1032	2,014	987	1,004

Table 6. Left. Properties for the documents in the style change detection datasets. **Right.** Overview of the datasets, listing the number of documents per dataset (narrow and wide) for the training, validation, and test sets split by the number of authors per document.

5.2 Evaluation

For the comparison of the submitted approaches, we report both the achieved performances for the subtasks in isolation and their combination as a staged task. Furthermore, we evaluate the approaches on both datasets individually.

Submissions are evaluated by the F_α-Measure for each document, where we set α to 1. For task 1, we compute the average F_1 measure across all documents, and for task 2, we use the micro-averaged F_1 measure across all documents. The submissions for the two datasets are evaluated independently and the resulting F_1 measures for the two tasks will be averaged across the two datasets.

5.3 Results

The style change detection task received three software submissions, which were evaluated on the TIRA experimentation platform. Table 7 depicts the results of the individual submissions for both tasks independently and the average of the two task results per participant. We also include a random baseline, which predicts a document being single- vs multi-authored as well as author changes occurring between every two paragraphs at random, with equal probabilities. As can be seen, iyer20 achieved the highest scores in both tasks, whereas the other two participants achieved comparable results in both tasks. Every approach managed to beat the baseline on both tasks, with the differences between the baseline and the participants' approaches being particularly noteworthy for task 2. More details on the approaches taken can be found in the task overview paper [19].

Participant	Task1 F_1	Task2 F_1	Avg. F_1
iyer20	0.6401	0.8567	0.7484
castro20	0.5399	0.7579	0.6489
nath20	0.5204	0.7526	0.6365
baseline (random)	0.5007	0.5001	0.5004

Table 7. Overall results for the style change detection task ranked by average F_1.

6 Summary and Outlook

Despite the generally bleak circumstances, this year's PAN lab has succeeded in both retaining the core community and in expanding beyond it. Although we had far fewer registrations than in 2019, we managed to increase the turnout and thus the number of submissions from 72 last year to 81 in 2020. The increasing participation can mostly be attributed to the tireless effort of PAN's largest task with 64 participants, Profiling Fake News Spreaders on Twitter, while the other recurring tasks addressed their core community and retained consistent participation.

Going into PAN 2020, we continued to tackle long-standing authorship issues and scrutinize societal problems through the lense of stylometry, improved our datasets, and re-invented our task design. As a larger innovation, we experimented with the design of evaluation episodes as multi-year series of shared tasks on difficult problems: At the start of this years' evaluation cycle we announced the future questions for some tasks two years in advance, not only to provide necessary context but to create the stability needed for participants to invest in difficult challenges. Besides positive feedback from the community, we already noticed significant improvements in the quality of the submitted approaches and aim to expand this strategy to our other tasks and nurture the idea of organizing evaluation episodes over mere evaluation cycles.

7 Acknowledgments

We thank Symanto for sponsoring the ex aequo award for the two best performing systems at the author profiling shared task of this year on Profiling fake news spreaders on Twitter. 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). The work of Anastasia Giachanou is supported by the SNSF Early Postdoc Mobility grant under the project Early Fake News Detection on Social Media, Switzerland (P2TIP2_181441).

Bibliography

[1] Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., Pardo, F.M.R., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Wiegmann, M., Zangerle, E.: Shared tasks on authorship analysis at PAN 2020. 42nd European Conference on IR Research (ECIR). 2020.

[2] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Bias Analysis and Mitigation in the Evaluation of Authorship Verification. 57th Annual Meeting of the Association for Computational Linguistics (ACL). 2019.

[3] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Generalizing unmasking for short texts. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. 2019.

[4] Ghanem, B., Rosso, P., Rangel, F.: An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology (TOIT). 2020.

[5] Giachanou, A., Ríssola, E.A., Ghanem, B., Crestani, F., Rosso, P.: The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers. In: International Conference on Applications of Natural Language to Information Systems. 2020.
[6] 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. 2019.
[7] Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019. Working Notes Papers of the CLEF 2019 Evaluation Labs. CEUR Workshop Proceedings. 2019.
[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. Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings. 2018.
[9] Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.
[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011.
[11] Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. CLEF 2020 Labs and Workshops, Notebook Papers. 2020.
[12] Rangel, F., Rosso, P., Franco-Salvador, M.: A Low Dimensionality Representation for Language Variety Identification. 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLingâĂŹ16. 2018.
[13] Shu, K., Wang, S., Liu, H.: Understanding User Profiles on Social Media for Fake News Detection. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). 2018.
[14] Vo, N., Lee, K.: Learning from Fact-checkers: Analysis and Generation of Fact-checking Language. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019.
[15] W. Noreen, E.: Computer-Intensive Methods for Testing Hypotheses: An Introduction. A Wiley-Interscience publication (1989)
[16] Wiegmann, M., Potthast, M., Stein, B.: Overview of the Celebrity Profiling Task at PAN 2020. CLEF 2020 Labs and Workshops, Notebook Papers. 2020.
[17] Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics. 2019.
[18] Wiegmann, M., Stein, B., Potthast, M.: Overview of the Celebrity Profiling Task at PAN 2019. CLEF 2019 Labs and Workshops, Notebook Papers. 2019.
[19] Zangerle, E., Mayerl, M., Specht, G., Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2020. CLEF 2020 Labs and Workshops, Notebook Papers. 2020.
[20] Zangerle, E., Tschuggnall, M., Specht, G., Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2019. CLEF 2019 Labs and Workshops, Notebook Papers. 2019.