Overview of the Style Change Detection Task at PAN 2022

Eva Zangerle[1], Maximilian Mayerl[1], Martin Potthast[2] and Benno Stein[3]

[1]Universität Innsbruck
[2]Leipzig University
[3]Bauhaus-Universität Weimar

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

Abstract

Style change detection means to identify positions at which the authorship in a multi-author document changes. Reliably detecting these positions is key for multi-author document analyses, and it is a preliminary step for authorship identification. This year style change detection task at PAN features three connected subtasks: (1) For a text written by two authors, which 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) 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. (3) For a text written by two or more authors, find all positions of writing style change. In particular, style changes may occur both between paragraphs but also at sentence level. The task is evaluated on a dataset compiled from an English Q&A platform. The paper in hand introduces the style change detection task, the underlying dataset, the approaches employed by the participants, and the achieved results.

1. Introduction

The goal of the style change detection task is to identify the positions in a document where authorship changes. Previous editions of the style change detection task at PAN included certain variants of this task: In 2016, the identification and clustering of text segments by author [1]. In 2017, to first detect whether a given document was written by multiple authors [2], and, given the case, to identify the exact positions at which authorship changes. The (weak) results showed that this task was beyond the state-of-the-art at that time. Hence, at PAN 2018, the task was relaxed to a binary classification task, namely, to distinguish single-author from multi-author documents [3]. PAN 2019 extended the task and asked participants to also predict the number of authors for all detected multi-author documents [4]. Similarly, PAN 2020 focused on binary classification (single versus multiple authors) and to determine the positions of style changes at paragraph level [5]. At PAN 2021, the participants were asked to determine whether a given document was written by multiple authors and, given the case, to detect the style changes at paragraph level and to assign authors to paragraphs [6].

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

CEUR Workshop Proceedings (CEUR-WS.org)

This year, we again advanced the field of multi-author analysis by increasing the complexity of the identification problem. Specifically, participants were asked (1) to find the position of a style change in documents with a single style change (at the paragraph level), (2) to find all style changes in a document written by up to five authors at the paragraph level, and (3) to find the positions of style changes at the sentence level.

The remainder of this paper is structured as follows. Section 2 discusses previous style change detection approaches. Section 3 introduces this year's style change detection task and its subtasks, along with the datasets and the used evaluation (performance) measures. Section 4 surveys the participants' submissions. Section 5 analyzes and compares the achieved results and Section 6 concludes the paper.

2. Related Work

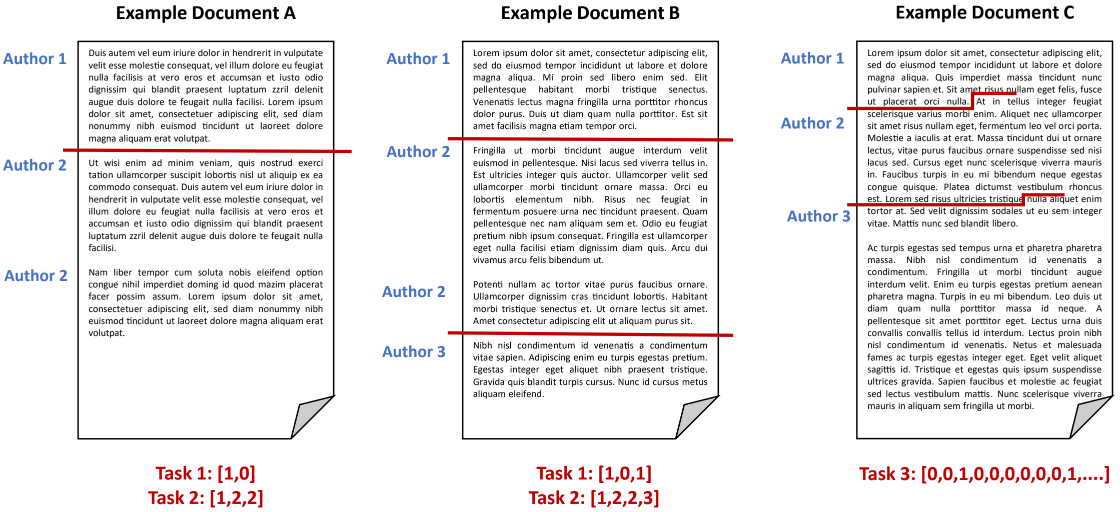
Style change detection requires an "intrinsic document analysis", in contrast to an analysis that can use knowledge from other corpora. An intrinsic analysis includes the computation of a stylistic profile for each sentence or paragraph, which is used to spot style change positions by either comparing similarities [7, 8] or by an outlier detection analysis [9]. Stylistic profiles may comprise lexical features such as character n-grams (e.g., [10, 11]), word frequencies (e.g., [12]) and average word or sentence lengths (e.g., [13]), syntactic features such as part-of-speech tag frequencies and structures (e.g., [14]), grammar trees (e.g., [15]), or structural features such as indentation usage (e.g., [13]).

By analyzing stylometric features, Glover and Hirst [16] identify inconsistencies in writing style in collaborative documents by detecting author boundaries. Meyer zu Eißen and Stein [17, 18, 19] analyze intrinsic plagiarism detection based on style change detection using word frequency classes. Koppel et al. [20, 21], as well as Akiva and Koppel [22, 23] use clustering based on lexical features to decompose multi-authors into authorial threads. The approach by Tschuggnall et al. [15] relies on grammar tree features for an unsupervised decomposition approach. Rexha et al. [24] use stylistic features to predict the number of authors who wrote a text. Bensalem et al. [25] use n-grams to identify author style changes, while Gianella [26] employs Bayesian modeling to decompose a document by author.

To this end, we observe at PAN a shift from the use of traditional stylistic features to pre-trained language models for characterizing paragraphs or sentences. For instance, in 2018, the best binary classification results (distinguish between single- or multi-authored document) were obtained by a stacking ensemble classifier based on lexical and syntactical features extracted via multiple sliding window approaches [27]. In 2020 and 2021, pre-trained BERT models that were fine-tuned on the training dataset have shown to achieve the best results [28, 29].

3. Style Change Detection Task

This section presents the style change detection task and its subtasks, the dataset underlying the task, and the used evaluation (performance) measures.



Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Mi proin sed libero enim sed. Elit pellentesque habitant morbi tristique senectus. Venenatis lectus magna fringilla urna porttitor rhoncus dolor purus. Duis ut diam quam nulla porttitor. Est sit amet facilisis magna etiam tempor orci.

Fringilla ut morbi tincidunt augue interdum velit euismod in pellentesque. Nisi lacus sed viverra tellus in. Est ultricies integer quis auctor. Ullamcorper velit sed ullamcorper morbi tincidunt ornare massa. Orci eu lobortis elementum nibh. Risus nec feugiat in fermentum posuere urna nec tincidunt praesent. Quam pellentesque nec nam aliquam sem et. Odio eu feugiat pretium nibh ipsum consequat. Fringilla est ullamcorper eget nulla facilisi etiam dignissim diam quis. Arcu dui vivamus arcu felis bibendum ut.

Potenti nullam ac tortor vitae purus faucibus ornare. Ullamcorper dignissim cras tincidunt lobortis. Habitant morbi tristique senectus et. Ut ornare lectus sit amet. Amet consectetur adipiscing elit ut aliquam purus sit.

Nibh nisl condimentum id venenatis a condimentum vitae sapien. Adipiscing enim eu turpis egestas pretium. Egestas integer eget aliquet nibh praesent tristique. Gravida quis blandit turpis cursus. Nunc id cursus metus aliquam eleifend.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Quis imperdiet massa tincidunt nunc pulvinar sapien et. Sit amet risus nullam eget felis, fusce ut placerat orci nulla. At in tellus integer feugiat scelerisque varius morbi enim. Aliquet nec ullamcorper sit amet risus nullam eget, fermentum leo vel orci porta. Molestie a iaculis at erat. Massa tincidunt dui ut ornare lectus, vitae purus faucibus ornare suspendisse sed nisi lacus sed. Cursus eget nunc scelerisque viverra mauris in. Faucibus turpis in eu mi bibendum neque egestas congue quisque. Platea dictumst vestibulum rhoncus est. Lorem sed risus ultricies tristique nulla aliquet enim tortor at. Sed velit dignissim sodales ut eu sem integer vitae. Mattis nunc sed blandit libero.

Ac turpis egestas sed tempus urna et pharetra pharetra massa. Nibh nisl condimentum id venenatis a condimentum. Fringilla ut morbi tincidunt augue interdum velit. Enim eu turpis egestas pretium aenean pharetra magna. Turpis in eu mi bibendum. Leo duis ut diam quam nulla porttitor massa id neque. A pellentesque sit amet porttitor eget. Lectus urna duis convallis convallis tellus id interdum. Lectus proin nibh nisl condimentum id venenatis. Netus et malesuada fames ac turpis egestas integer eget. Eget velit aliquet sagittis id. Tristique et egestas quis ipsum suspendisse ultrices gravida. Sapien faucibus et molestie ac feugiat sed lectus vestibulum mattis. Nunc scelerisque viverra mauris in aliquam sem fringilla ut morbi.

Figure 1: Sample documents that illustrate different style change situations and the expected solution. From left to right: single style change (Subtask 1), multiple style changes and attribution (Subtask 2), and multiple style changes on the sentence level (Subtask 3).

3.1. Task Definition

The goal of the style change detection task is to identify positions at which the authorship of a multi-author document changes. We study the following subtasks in this regard:

Style Change Basic 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 at the paragraph-level.

Style Change Advanced 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.

Style Change Real-World 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.

Figure 1 illustrates three example documents and the expected outcome for the three subtasks. Document A has a single style change between the first and second paragraph, Document B contains two style changes on the paragraph level and was authored by three different authors, and Document C contains two style changes on the sentence level and was authored by three authors.

Participants either deployed their software on the TIRA platform [30] or uploaded their predictions. TIRA allows participants to tune their approaches on the training and validation dataset, as well as to self-evaluate their software on the unseen test dataset. By enabling blind evaluation, TIRA prevents optimization against test data.

3.2. Dataset

The datasets that we provided for this task have been created from posts on 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 broad topics (so-called sites).[1] The cleansing of this raw data included the removal of questions and answers that were edited after they were originally posted, as well as the removal of images, URLs, code snippets, block quotes, and bullet lists.

The procedure for forming datasets works as follows. All questions and answers are split into paragraphs, where paragraphs with less than 100 characters are discarded. Then, artificial documents are synthesized by drawing paragraphs from a single question thread to ensure that topic changes cannot be exploited for detecting style changes. The number of authors for each artificial document is picked randomly between one and five. We randomly chose a corresponding number of authors from the set of authors who contributed to the question thread we are 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 shorter than 1,000 characters or longer than 10,000 characters, 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 of Subtask 1, we ensured that every generated document features exactly one style change. For the dataset of Subtask 2, we used the procedure as outlined above. For the dataset of Subtask 3, we changed the procedure to operate on sentences instead of on paragraphs. The three datasets that we obtained in this way contain $2,000$, $10,000$, and $10,000$ documents respectively. All datasets are split into training, test, and validation sets in the ratio 70:15:15.

3.3. Performance Measures

The three subtasks are evaluated independently. As a primary evaluation metric, we compute the macro-averaged F1-score across all documents for all three subtasks. To get a deeper understanding of the performance of the authorship attribution in Subtask 2, we employ two additional measures, borrowed from the field of text transcription and speaker diarization. Transferred to the style change detection task, these measures essentially capture the fraction of text that is not correctly attributed to an author. The Diarization Error Rate (DER) measure [31] captures the fraction of wrongly attributed segments. The Jaccard Error Rate (JER) [32] gives equal weight to each author. For each reference author ref_a, we compare the set of segments authored by ref_a (either paragraph or sentence, depending on the subtask) against the set of predicted authors $pred_a$ for these segments. Based on the Jaccard Error Rate, we compute the ratio between the sizes of the intersections and unions of the two sets of segments (see Equation 1). The

[1] 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.

final JER results from the average of the author-specific Jaccard Error Rates.

$$JER(a) = 1.0 - \frac{|ref_a \cap pred_a|}{|ref_a \cup pred_a|} \tag{1}$$

4. Survey of Submissions

For the 2022 edition of the style change detection task, we received nine submissions; eight of which used intrinsic approaches and one used an extrinsic approach. We briefly describe the approaches proposed by the participants in the following.

Alshamasi and Menai [33] rely on a set of lexical and syntactic features that are extracted on the sentence- and on the paragraph level. Based on these text representations, the authors apply k-means clustering, where the number of clusters k is evaluated using the within-cluster sum-of-squares error. This method aims to assign all paragraphs (sentences) of an author to a single cluster. For Subtask 1, where only a single style change is contained in the input documents, k is set to 2 in order to derive the potential authors of individual paragraphs and detect candidate positions for the switches. Finally, the pair of paragraphs with the highest cosine distance is chosen. For Subtask 2 and Subtask 3, the assignments of the individual paragraphs to the k clusters is used for the prediction based on the extracted paragraph or sentence vectors.

Lao et al. [34] rely on a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [35] model that is fine-tuned based on the provided dataset. The output of the BERT model is then fed into a one-dimensional convolution that allows obtaining dense feature representations of individual sentences/paragraphs. These representations are then the input to a max-pooling layer to arrive at a binary class capturing whether there is a style change between two paragraphs (sentences) or not. This binary output can directly be used for the subtasks 1 and 3. For Subtask 2, the paragraph representations are used to compute pair-wise similarities among paragraphs to compute the number of authors and the assignments of authors.

Jiang et al. [36] also apply transformer-based neural networks. However, they relied on the Electra model [37], which has been shown to be more efficient in training than masked-language modeling training as used in BERT. The authors chose to utilize three pre-trained Electra models for the three subtasks, depending on subtask complexity and the amount of data available.

Zi et al. [38] also apply BERT to compute word representations using masked-language modeling (MLM) training. These representations are fed into a Bi-LSTM (Bidirectional Long Short-term Memory) to enhance the representations with context information. The representations are then fed to a convolution and a max-pooling layer to compute a more dense representation. A fully connected layer is used to compute the final predictions.

Rodríguez-Losada and Castro [39] apply a mixture of approaches for the three subtasks. To represent the given texts as features, they use transformer models, the frequency of punctuation marks, and the frequency of discourse markers. For Subtask 1, they compute the similarity between all consecutive paragraphs in a given document and predict a style change at the paragraph boundary with the lowest similarity. For the subtasks 2 and 3, they compute similarities for each of the three feature categories and define a similarity threshold for each category. They predict a style change if either (a) all similarities are under the threshold or (b) if two of the three similarities were under the threshold.

Zhang et al. [40] utilize a prompt-based model to determine the similarity in writing style between two adjacent paragraphs or sentences. They train a BERT model to learn how to fill in the blank in a text of the form 'They are the [blank] writing style: {First Paragraph} and {Second Paragraph}'. For '[blank]' they use a vocabulary of possible terms that cover a range of similarities, including terms such as 'same', 'equal', 'different', 'unlike' etc. These predictions are then used to solve all three tasks.

Lin et al. [41] apply an ensemble model to solve this year's tasks. They trained three separate classifiers for determining whether a given pair of paragraphs or sentences is written by the same author or not. Each of these classifiers is based on a different pretrained language model for feature extraction: one on BERT [35], one on RoBERTa [42], and one on ALBERT [43]. The three classifiers are combined in a majority voting ensemble to make a final prediction.

Alvi et al. [44] apply a set of handcrafted discourse markers to characterize the writing style for a paragraph or sentence. For Sutask 1, they identify conversational patterns to predict the position of the author change. For the Sutasks 2 and 3, they extract occurrence counts for their set of discourse markers to first determine the number of authors in a document via a random forest model, and then use *k*-means clustering to cluster paragraphs or sentences into author clusters.

5. Evaluation Results

The summary of the evaluation results for the nine submissions to the Style Change Detection task at PAN 2022, as well as a baseline, is shown in Table 1.

The baseline approach uses uniformly distributed random predictions to assign paragraphs to authors, and then infers style change locations based on these, predicting a style change between all paragraphs or sentences that have a different author label. The random author assignments take into account that authors must be labelled with increasing identifiers depending on the order in which they first appear in a document. In other words, the first author appearing in the document is assigned label 1, the second author label 2, etc. As can be seen in Table 1, all submitted approaches outperformed the baseline for Subtask 1 and Subtask 3 in terms of F_1 score. For Subtask 2, only the approach submitted by Al-Shamasi and Menai [33] achieved a lower F_1 score than the baseline. A similar picture emerges for the JER and DER scores, where the same submission is the only one that has a higher JER than the baseline.

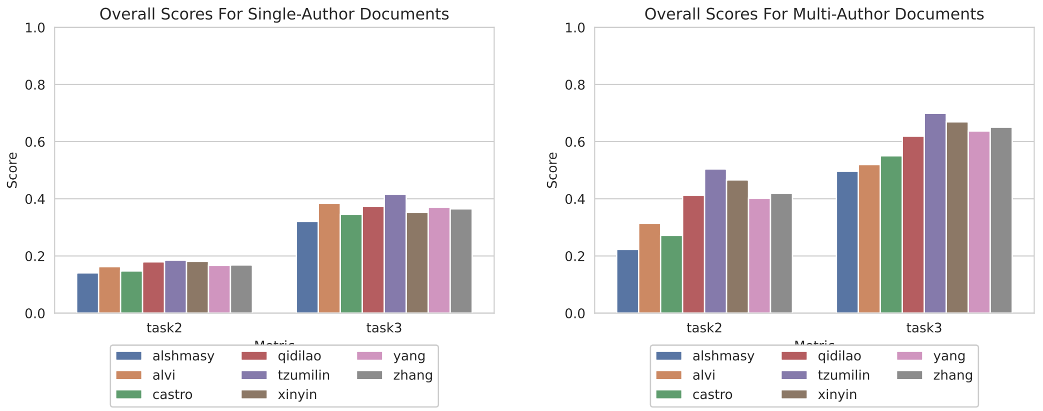
In terms of submitted approaches, this year (and for the first time) we received not only submissions with intrinsic approaches, but also a single submission with an extrinsic approach to style change detection. In Table 1, these approaches are shown separated since they are inherently not comparable. The extrinsic approach, graner22, clearly outperforms all other approaches due to its use of external information. In the following, we will therefore focus our discussions on the submitted intrinsic approaches. For these,

Table 1
Overall results for the style change detection task, ranked by average F_1 performance across all three subtasks (ST). The best (intrinsic) score for each metric is highlighted in bold.

Participant	**ST1 F_1**	**ST2 F_1**	**ST3 F_1**	**ST3 DER**	**ST3 JER**
		Intrinsic Approaches			
tzumilin22	**0.7540**	**0.5100**	**0.7156**	**0.1941**	**0.3095**
xinyin22	0.7346	0.4687	0.6720	0.2380	0.3138
qidilao22	0.7471	0.4170	0.6314	0.2636	0.3641
zhang22	0.7162	0.4174	0.6581	0.2886	0.3556
yang22	0.6690	0.4011	0.6483	0.2964	0.3677
alvi22	0.7052	0.3213	0.5636	0.3924	0.5218
castro22a	0.5661	0.2735	0.5565	0.4035	0.5771
alshmasy22	0.5272	0.2207	0.4995	0.4240	0.6444
		Extrinsic Approaches			
graner22	0.9932	0.9855	0.9929	0.0040	0.0040
		Baseline			
Random	0.3222	0.2651	0.4809	0.4568	0.5938

the overall best results were achieved by Lin et al. [41], whose approach achieved the best scores across all the metrics we considered. Some of the other approaches achieved a similar, albeit slightly reduced performance, especially those of Jiang et al. [36] and Lao et al. [34].

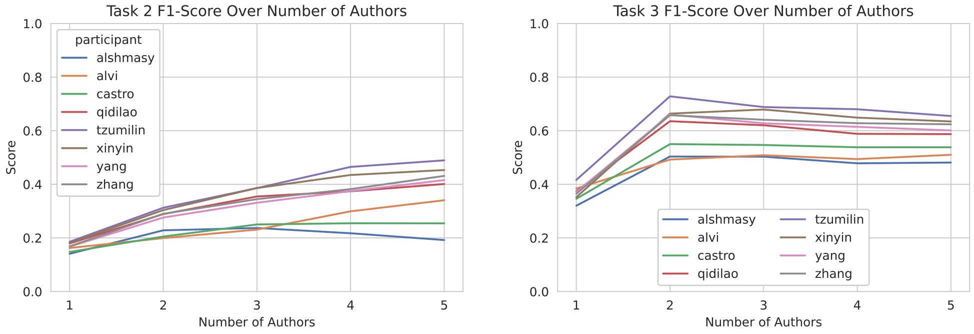
In addition to the overall scores given in Table 1, we also analyzed how the participant's systems performed depending on the true number of authors in a document. This is shown only done for Subtask 2 and Subtask 3, since all documents for Subtask 1 were written by exactly two authors. First, we considered single-author versus multi-author



(a) Single-author documents

(b) Multi-author documents

Figure 2: Scores (F_1) for the subtasks 2 and 3 separately for single-author (left) and multi-author documents (righ).



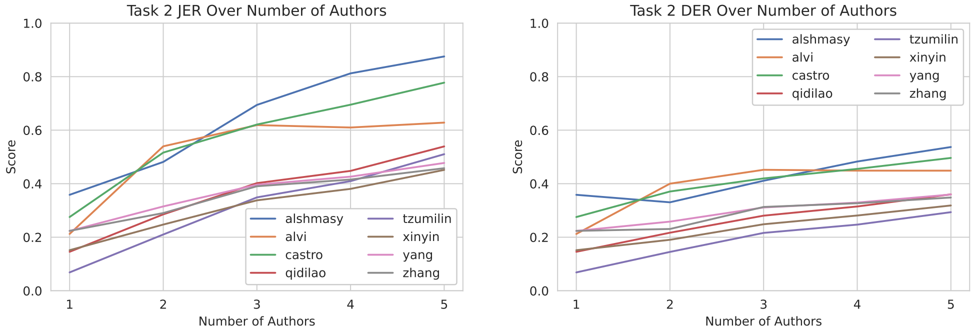
(a) Task 2

(b) Task 3

Figure 3: Scores (F_1) for Task 2 and Task 3, depending on the true number of authors in a document.

documents. The results of this analysis are shown in Figure 2. The first interesting observation that we can make here is that all systems perform better if a document was written by multiple authors. We can also see that the ranking of the systems is pretty stable, with only small changes, such as the approaches by Rodríguez-Losada and Castro [39] and Lao et al. [34] outperforming the approach submitted by Alvi et al. [44] for multi-author documents. We also looked at how the performance of all systems changed with the concrete number of authors (see Figure 3). For Subtask 2, we observe an performance increase for all systems until the number of three authors is reached in a document. When confronted with more authors, the systems behave differently: Some systems show a continued performance increase up to five authors (which is the maximum number of authors in a document in our datasets), while other systems show a drop in performance. For Subtask 3, we can see that almost all systems have a performance peak at two authors, with a sharp increase when going from single-author documents to two-author documents, followed by a slow decline as the number of authors grows.

Finally, we took a look at how the JER and DER scores for task 2 changed depending



(a) JER

(b) DER

Figure 4: Scores (JER and DER) for Task 2, depending on the true number of authors in a document.

on the number of authors. The result for this are given in Figure 4. Here, performance generally seems to decrease the more authors are in a document.

6. Conclusion

In the 2022 edition of the Style Change Detection task at PAN, we asked participants to detect style changes on the paragraph and sentence level (subtasks 1 and 3) and to assign paragraphs to authors based on the detected style changes (Subtask 2). We have received nine submissions by participants. The best results were obtained by utilizing pre-trained language models (BERT or Electra) to compute semantic representations of the texts across all three tasks. Altogether, we consider the achieved performance values as solid and promising.

References

[1] E. Stamatatos, M. Tschuggnall, B. Verhoeven, W. Daelemans, G. Specht, B. Stein, M. Potthast, Clustering by Authorship Within and Across Documents, in: Working Notes Papers of the CLEF 2016 Evaluation Labs, CEUR Workshop Proceedings, CLEF and CEUR-WS.org, 2016. URL: http://ceur-ws.org/Vol-1609/.

[2] M. Tschuggnall, E. Stamatatos, B. Verhoeven, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), Working Notes Papers of the CLEF 2017 Evaluation Labs, volume 1866 of CEUR Workshop Proceedings, CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-1866/.

[3] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), Working Notes Papers of the CLEF 2018 Evaluation Labs, volume 2125 of CEUR Workshop Proceedings, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[4] E. Zangerle, M. Tschuggnall, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2019, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[5] E. Zangerle, M. Mayerl, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[6] E. Zangerle, M. Mayerl, , M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[7] D. Karaś, M. Śpiewak, P. Sobecki, OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection—Notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-1866/.

[8] J. Khan, Style Breach Detection: An Unsupervised Detection Model—Notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-1866/.

[9] K. Safin, R. Kuznetsova, Style Breach Detection with Neural Sentence Embeddings—Notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-1866/.

[10] E. Stamatatos, Intrinsic Plagiarism Detection Using Character n-gram Profiles, in: B. Stein, P. Rosso, E. Stamatatos, M. Koppel, E. Agirre (Eds.), SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), Universidad Politécnica de Valencia and CEUR-WS.org, 2009, pp. 38–46. URL: http://ceur-ws.org/Vol-502.

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

[12] D. I. Holmes, The Evolution of Stylometry in Humanities Scholarship, Literary and Linguistic Computing 13 (1998) 111–117.

[13] R. Zheng, J. Li, H. Chen, Z. Huang, A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques, Journal of the American Society for Information Science and Technology 57 (2006) 378–393.

[14] M. Tschuggnall, G. Specht, Countering Plagiarism by Exposing Irregularities in Authors' Grammar, in: Proceedings of the European Intelligence and Security Informatics Conference (EISIC), IEEE, Uppsala, Sweden, 2013, pp. 15–22.

[15] M. Tschuggnall, G. Specht, Automatic decomposition of multi-author documents using grammar analysis, in: F. Klan, G. Specht, H. Gamper (Eds.), Proceedings of the 26th GI-Workshop Grundlagen von Datenbanken, volume 1313 of CEUR Workshop Proceedings, CEUR-WS.org, 2014, pp. 17–22. URL: http://ceur-ws.org/Vol-1313.

[16] A. Glover, G. Hirst, Detecting Stylistic Inconsistencies in Collaborative Writing, Springer London, London, 1996, pp. 147–168. doi:10.1007/978-1-4471-1482-6_12.

[17] S. Meyer zu Eißen, B. Stein, Intrinsic Plagiarism Detection, in: M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval. 28th European Conference on IR Research (ECIR 2006), volume 3936 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2006, pp. 565–569. doi:10.1007/11735106_66.

[18] B. Stein, S. Meyer zu Eißen, Intrinsic Plagiarism Analysis with Meta Learning, in: B. Stein, M. Koppel, E. Stamatatos (Eds.), 1st Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007) at SIGIR, 2007, pp. 45–50. URL: http://ceur-ws.org/Vol-276.

[19] B. Stein, N. Lipka, P. Prettenhofer, Intrinsic Plagiarism Analysis, Language

Resources and Evaluation (LRE) 45 (2011) 63–82. doi:10.1007/s10579-010-9115-y.

[20] M. Koppel, N. Akiva, I. Dershowitz, N. Dershowitz, Unsupervised decomposition of a document into authorial components, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 1356–1364. URL: https://www.aclweb.org/anthology/P11-1136.

[21] M. Koppel, N. Akiva, I. Dershowitz, N. Dershowitz, Unsupervised decomposition of a document into authorial components, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 1356–1364. URL: http://www.aclweb.org/anthology/P11-1136.

[22] N. Akiva, M. Koppel, Identifying Distinct Components of a Multi-author Document, in: N. Memon, D. Zeng (Eds.), 2012 European Intelligence and Security Informatics Conference, EISIC 2012, IEEE Computer Society, 2012, pp. 205–209. URL: https://doi.org/10.1109/EISIC.2012.16. doi:10.1109/EISIC.2012.16.

[23] N. Akiva, M. Koppel, A Generic Unsupervised Method for Decomposing Multi-Author Documents, JASIST 64 (2013) 2256–2264. URL: https://doi.org/10.1002/asi.22924. doi:10.1002/asi.22924.

[24] A. Rexha, S. Klampfl, M. Kröll, R. Kern, Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features, in: P. Mayr, I. Frommholz, G. Cabanac (Eds.), Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016), volume 1567 of CEUR Workshop Proceedings, CEUR-WS.org, 2016, pp. 26–31. URL: http://ceur-ws.org/Vol-1567.

[25] I. Bensalem, P. Rosso, S. Chikhi, Intrinsic plagiarism detection using n-gram classes, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1459–1464. URL: https://www.aclweb.org/anthology/D14-1153. doi:10.3115/v1/D14-1153.

[26] C. Giannella, An Improved Algorithm for Unsupervised Decomposition of a Multi-Author Document, JASIST 67 (2016) 400–411. URL: https://doi.org/10.1002/asi.23375. doi:10.1002/asi.23375.

[27] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/.

[28] A. Iyer, S. Vosoughi, Style Change Detection Using BERT, in: L. Cappellato, N. Ferro, A. Névéol, C. Eickhoff (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020.

[29] Z. Zhang, X. Miao, Z. Peng, J. Zeng, H. Cao, J. Zhang, Z. Xiao, X. Peng, Z. Chen, Using Single BERT For Three Tasks Of Style Change Detection, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops,

Notebook Papers, CEUR-WS.org, 2021.

[30] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.

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

[32] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The second dihard diarization challenge: Dataset, task, and baselines, arXiv preprint arXiv:1906.07839 (2019).

[33] S. Al-Shamasi, M. Menai, Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[34] Q. Lao, L. Ma, W. Yang, Y. Zexian, D. Yuan, Z. Tan, L. Liang, Style Change Detection Based On Bert And Conv1d, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[36] X. Jiang, H. Qi, Z. Zhang, Style Change Detection: Method Based On Pre-trained Model And Similarity Recognition, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[37] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[38] J. Zi, L. Zhou, Z. Liu, Style Change Detection Based On Bi-LSTM And Bert, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[39] C. A. R. Losada, D. C. Castro, Three style similarity: sentence-embedding, auxiliary words, punctuation, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[40] Z. Zhang, Z. Han, L. Kong, Style Change Detection based on Prompt, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[41] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble Pre-trained Transformer Models for Writing Style Change Detection, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[42] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).

[43] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942 (2019).

[44] F. Alvi, H. Algafri, N. Alqahtani, Style Change Detection using Discourse Markers, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.