Computational Linguistics in the Netherlands Journal 11 (2021) 71-84 Submitted 10/2021; Published 12/2021

Evaluating the Impact of Word Classes on Cross-Domain Age Detection Models' Performance

Jens Van Nooten* JENS.VANNOOTEN@UANTWERPEN.BE
Ilia Markov* ILIA.MARKOV@UANTWERPEN.BE
Walter Daelemans* WALTER.DAELEMANS@UANTWERPEN.BE

* *University of Antwerp (CLiPS),*
Lange Winkelstraat 40,
2000 Antwerp,
Belgium

Abstract

In this paper, we examine the importance of word category information for the age detection task – the task of identifying the age of a person based on their writing – both under in-domain and cross-domain conditions. We remove entire word classes and study its effect using both Support Vector Machines (SVM) and pre-trained contextual word embeddings (BERT). By conducting these experiments, we aim to gain insight into how both approaches handle cross-domain conditions. Our experiments show that, on the one hand, SVM mainly relies on content words in the in-domain settings, while function words are the most indicative features in the cross-domain setup. BERT, on the other hand, mainly relies on highly-frequent word classes, such as nouns and punctuation, to make predictions both under in-domain and cross-domain age detection conditions.

1. Introduction

Age detection is a sub-task of author profiling that comprises the inference of an author's age based on text that they have written. The task keeps generating interest and finding more applications in real life, such as in forensic linguistics and marketing. An important application of automatic age detection in forensic linguistics is, for example, the detection of online predatory acts and sexually transgressive behaviour (van de Loo et al. 2016). In marketing, companies could implement author profiling tools to market their products to specific target age groups.

From a machine learning perspective, age detection can be approached as a multiclass categorisation problem, where the input is text and the output is one item of a pre-defined label set (age categories in this case), e.g., 18-24, 25-34, and so on. Age detection can also be approached as a regression problem, where the exact age of the author of a text is predicted. Throughout the years, a large amount of distinct (mainly machine learning) techniques have been proposed to reveal a user's age.

The field of Natural Language Processing (NLP) was greatly influenced by the introduction of BERT (Bidirectional Encoder Representations from Transformers) in 2018 (Devlin et al. 2019), a pre-trained language model that employs a transformer-based architecture based on bidirectional attention. This pre-trained language model achieves state-of-the-art results on various NLP tasks and has also found its way into author profiling, e.g., (Abdul-Mageed et al. 2019, Zhang and Abdul-Mageed 2019, Polignano et al. 2020, Suman et al. 2021).

Despite these recent advances in the field of NLP, one of the remaining challenges in age detection and related tasks is the cross-domain performance of automated approaches. When the sources of the training and test corpora are different, models that rely on content words may overfit on the training data, since these features are highly domain-dependent (Nguyen et al. 2011). Such models provide good results on the data belonging to the same domain, but yield lower results on out-of-

domain data, while in real-word applications cross-domain scalability is essential, that is, models have to generalise well across different domains and topics, since test data is often from a different source than the data that the models are trained on.

In this paper, we examine the importance of word classes under in-domain and cross-domain age detection conditions by conducting part-of-speech (POS) ablation experiments (removing entire word classes) and evaluating the performance of Support Vector Machines (SVM), on the one hand, and a pre-trained language model (BERT), on the other hand, on the modified data. Though we realise that a direct comparison of the two models is not feasible due to the different nature of these models: SVM builds feature vectors directly from the manipulated data, while BERT is pre-trained on a large amount of texts and is only confronted with manipulated data during fine-tuning, by removing entire word classes and analyzing the most indicative word classes each of the models relies on when making predictions, we attempt to gain insight into how the two approaches handle cross-domain conditions. Since the lexical overlap between datasets in a cross-domain condition is likely to be quite low, our expectation is that removing content word POS classes will decrease in-domain but increase cross-domain performance of the models.

2. Related Work

2.1 Conventional machine learning approaches for age detection

Age-related language variation has been discussed in numerous studies for several languages in the field of sociolinguistics; e.g., Rickford and Mackenzie (2013) researched the usage of vernacular English in several subjects over time. The researchers found that the habitual usage of vernacular English decreased as the subjects grew older. Additionally, in (Barbieri 2008) a number of discriminative features for younger people are derived from a corpus of spoken American English. One of the observations included that young people tended to use more discourse markers, such as 'like', 'just', 'okay', 'yeah', 'I mean', 'I guess', 'kinda' and 'sorta'. Moreover, young people were found to use more personal pronouns in discourse. In contrast, older people tended to use fewer discourse markers, fewer personal pronouns, more positive emotion words, more future tense and fewer past tense verbs.

Due to the wide-spread popularity of blogging and social media, on the one hand, and computational developments, on the other hand, the research into age-related stylistic variation gained a significant boost in insights and techniques. Argamon et al. (2009) discern two different feature types for the author profiling task, namely content-based features (individual words related to topics) and style-based features (e.g., function words). The importance of function words in day-to-day conversation and social interactions has been researched in (Chung and Pennebaker 2007). The authors concluded that function words "carry an array of psychological meanings and set the tone for social interactions".

Early attempts at age prediction, as in (Schler et al. 2006), already consisted of extracting function words and POS tags in combination with content words. Later approaches also included several other feature types, such as subtrees of syntactic trees (Johannsen et al. 2015), bleached text (van der Goot et al. 2018) and relative frequencies of function words (Rangel et al. 2015, Pardo et al. 2016).

However, a great number of approaches for age detection mainly focused on content-based features such as word and lemma n-grams (Rangel et al. 2015, Pardo et al. 2016, Shrestha et al. 2016). Though these features yield high results for the in-domain age prediction task, they are prone to overfitting and limit the model's applicability in cross-domain conditions (Argamon et al. 2009, Nguyen et al. 2011).

Several researchers focused on age detection in a cross-domain condition for the PAN'16 author profiling competition (Pardo et al. 2016). For example, Modaresi et al. (2016) used word unigrams, word bigrams, character 4-grams, the average amount of spelling errors and punctuation features,

and reached an accuracy of 51.28% for six classes. For the same competition, Bilan and Zhekova (2016) constructed a model based on more abstract features, such as dictionary-based (e.g., stop words, connective words and emotion words), POS-based, text-structure-based (e.g., type/token ratio, average word length) and stylistic features (e.g., frequency of different adjectival and adverbial suffixes). Using this approach, they reached an accuracy score of 44.87% for English.

2.2 Neural network approaches for age detection

In more recent years, researchers shifted their attention to applying neural networks for author profiling. For example, Chen et al. (2019) implemented LSTM (Long-Short-Term-Memory) neural networks (Hochreiter and Schmidhuber 1997) to jointly learn age classification and regression models. Additionally, with the rise of BERT in 2018 (Devlin et al. 2019), this state-of-the-art language model has also been employed for age prediction, e.g., (Abdul-Mageed et al. 2019, Zhang and Abdul-Mageed 2019, Polignano et al. 2020, Suman et al. 2021).

Abdul-Mageed et al. (2019) used sentence-level representations from BERT for multi-task learning, i.e., jointly predicting the age (three classes) and gender of an author. Using BERT, they achieved an accuracy score of 50.95% for the age prediction task. Polignano et al. (2020) also used contextualised sentence embeddings from BERT to predict an author's age (besides gender, fame and occupation). Age prediction was approached as a linear regression problem, where the birth year was to be predicted. BERT outperformed SVM on all of the tasks, reaching an accuracy of 83% for age detection.

In order to assess the quality of datasets for natural language inference, Talman et al. (2021) corrupted the data in several benchmark datasets. By removing entire word classes, the authors showed that the performance of the BERT model is the lowest when content-bearing words, such as nouns and verbs, were removed from the datasets. It should also be noted that the performance of the models did not drop significantly, which as the researchers suggest, might be due to "other clues and biases" left in the data.

The role of pre-processing has also been explored for gender prediction by Alzahrani and Jololian (2021). The researchers examined different commonly used pre-processing techniques from the PAN'16 competition and found that the BERT model that was trained on data without any pre-processing (e.g., removal of punctuation and stop words) achieved the highest result.

In this paper, we zoom in on different part-of-speech categories and examine their importance in the in-domain and cross-domain age detection settings. We remove entire word classes and evaluate its impact on the performance of SVM and BERT models.

2.3 Datasets

In order to evaluate the importance of features based on their grammatical categories, we conducted experiments on four datasets designed for the age detection task, namely the PAN'15 dataset (Rangel et al. 2015), the PAN'16 dataset (Pardo et al. 2016), the WebMD dataset[1] and the Blog Authorship Corpus (Schler et al. 2006). For the PAN'15 and PAN'16 datasets, only the English subsets were used. More information about the genres, number of entries, number of authors, and number of tokens in the datasets can be found in Table 1.

In order to limit user bias in the in-domain conditions, authors from the training sets in the PAN 15 and 16 datasets, and in the Blog Authorship Corpus did not appear in the test set. We used 15% of the dataset as test data for the in-domain experiments[2], and all available age categories in each dataset (see Table 2).

In order to conduct cross-domain experiments with as many age categories as possible, all age categories in the datasets were converted to 18-24, 25-34 and 35-xx. The transformed classes are

1. More information about this dataset can be found on www.kaggle.com/rohanharode07/webmd-drug-reviews-dataset.
2. The gold labels for the test sets used in the PAN'15 and PAN'16 competitions are not made available.

Dataset	Genre	Number of entries	Authors	Avg. number of tokens per entry	Total number of tokens
PAN'15	Tweets	13,446	152	15	201,264
PAN'16	Tweets	174,565	402	17	2,936,540
WebMD	User reviews	340,546	N/A	67	22,831,929
Blog Authorship Corpus	Blogs	609,164	18,978	251	152,643,858

Table 1: Statistics of the datasets used for the experiments.

summarised in Table 2. For the cross-domain experiments when PAN data was used, the training partitions of the PAN'15 and PAN'16 datasets were merged to obtain more data for training the supervised approaches described below. For the cross-domain experiment where the Blog Authorship Corpus and the WebMD dataset were used, all authors below the age of 18 were removed. The statistics for the datasets used for the cross-domain experiments in terms of the number of entries for each age category can be found in Table 3.

PAN'15	PAN'16	WebMD	Blog Authorship Corpus	Combined age category
/	/	xx-18 (10,024)	xx-18 (203,935)	/
18-24 (5,363)	18-24 (11,546)	19-24 (24,230)	18-24 (138,124)	18-24
25-34 (5,250)	25-34 (58,940)	25-34 (49,705)	25-34 (187,204)	25-34
35-49 (1,840)	35-49 (69,802)	35-44 (55,010)	35-49 (79,901)	35-xx
50-xx (993)	50-64 (32,927)	45-54 (80,032)	/	
/	65-xx (1,350)	55-64 (75,129)	/	
/	/	65-74 (41215)	/	
/	/	75-xx (15,225)	/	

Table 2: Number of entries (provided in parentheses) per age category used for the in-domain experiments and the transformed classes for the cross-domain experiments.

Dataset	# 18-24	# 25-34	# 35+	Total
PAN '15	5,428	5,283	2,878	13,589
PAN '16	11,869	60,202	106,341	178,412
WebMD	24,230	49,705	266,611	340,546
Blog Authorship Corpus	152,051	206,226	84,064	442,341

Table 3: Number of entries per age category in the datasets used for the cross-domain experiments.

3. Methodology

3.1 Data pre-processing

Before conducting the experiments, all duplicate entries (and retweets, if applicable) in all the datasets were removed. For the experiments with Support Vector Machines, all entries were lowercased. Additionally, in the PAN'15 and PAN'16 datasets, all usernames, hashtags and URLs were replaced with placeholders. After this, we extracted POS tags using spaCy[3].

3.2 Experimental setup

We developed a suite of experiments in order to evaluate the importance of individual word classes both under in-domain and cross-domain age detection conditions. We conduct in-domain experiments on all four datasets described above, as well as carry out three cross-domain experiments (training on one dataset and testing on out-of-domain data), as summarised in Table 4.

Condition	**Training data**	**Test data**
In-domain	PAN'15 (tweets)	PAN'15 (tweets)
In-domain	PAN'16 (tweets)	PAN'16 (tweets)
In-domain	WebMD (reviews)	WebMD (reviews)
In-domain	Blog Corpus (blogs)	Blog Corpus(blogs)
Cross-domain	PAN-data (tweets)	Blog Corpus (blogs)
Cross-domain	PAN-data (tweets)	WebMD (reviews)
Cross-domain	Blog Corpus (blogs)	WebMD (reviews)

Table 4: Overview of the in-domain and cross-domain experimental setups. 'PAN-data' refers to a combination of the PAN'15 and PAN'16 datasets.

3.2.1 SVM experiments

In order to examine the importance of individual word classes, POS ablation experiments were conducted. These experiments consisted of first training the SVM model using word n-gram features (n = 1–3) and evaluating its performance when all word classes are included and secondly redoing the experiments with all words of a particular POS tag (e.g., nouns) removed both from the training and test sets. The latter result was then subtracted from the former result. This was repeated for each of the 17 universal POS tags shown in Table 5. This experiment was conducted for all the in-domain and cross-domain experimental settings described above.

We used a Scikit-learn (version 0.24.1) implementation of the linear SVM classifier with the $tf-idf$ weighting scheme. As the evaluation metric, we used F1-macro, since the class distributions in the datasets used are imbalanced.

3.2.2 BERT experiments

For the experiments with BERT, we adopt the methodology proposed in (Talman et al. 2021), that is, removing entire word classes before fine-tuning (i.e., in the pre-processing stage), and subsequently fine-tuning and testing the model on the modified versions of the datasets. We used the BERT base model (cased) from the Hugging Face transformers library[4]. The model was fine-tuned for one epoch with a learning rate of 2×10^{-5} and a batch size of 8. This was repeated for each of the 17 universal POS tags. For these experiments, we also used the F1-score macro.

3. https://spacy.io/usage/linguistic-features
4. https://huggingface.co/bert-base-cased

Tag	Part Of Speech
ADJ	adjective
ADV	adverb
ADP	adposition
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	number
PART	participle
PROPN	propernoun
PRON	pronoun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

Table 5: Universal POS tags used for the ablation experiments.

4. Results and Discussion

In this section, we discuss our experimental results. Tables with the detailed results are provided in the Appendix.

The in-domain experimental results show that discarding content words (such as nouns and proper nouns) and punctuation led to the highest score drop in the majority of cases both for the SVM and BERT models across all the datasets. While punctuation features are considered indicative both for age detection (Markov et al. 2016) and for other stylometry-related tasks, such as authorship attribution (Markov et al. 2018a) and native language identification (Markov et al. 2018b, Markov et al. 2020), the high score drops in terms of F1-score that are observed when nouns and proper nouns are removed may be caused by topic bias, that is, certain age groups in the in-domain datasets may discuss certain specific topics that other age groups do not discuss.

An anomaly across the datasets in terms of informative word classes is the WebMD dataset, where numbers were quite informative. We used the ELI5 library[5] to interpret the features used in the SVM model for predicting age labels. This revealed that the most informative features for each age category mention the author's actual age, e.g. 'am 28', 'am 23', '51', etc. Additional informative nouns also tended to refer to the age group that the user belongs to, e.g., 'college' (18-24), 'menopause' (45-54) and 'retired' (65-74).

In two cross-domain conditions (see Tables 16 and 18), we observed that for the SVM model, determiners, participles, auxiliaries, and pronouns, i.e., function words, were the most informative features, while the content-based features (i.e., proper nouns and nouns) were the least informative. The low importance of content words (proper nouns, nouns, etc.) in the cross-domain experiments with SVM (see Tables 16, 17 and 18) could be attributed to the low lexical overlap between the datasets used in the cross-domain experiments. We determined the lexical overlap between the lemmas of content words by selecting non-overlapping subsets of the datasets of the same size and calculating the Jaccard similarity scores. It was observed that the lexical overlap between the datasets of different domains was significantly lower than the lexical overlap between two partitions

5. ELI5 is a library designed for debugging machine learning algorithms and retrieving the most informative features. More information about this package can be found on https://eli5.readthedocs.io/en/latest/overview.html.

	PAN (1)	WebMD (1)	Blogs (1)
PAN (2)	**0.44**	0.15	0.13
WebMD (2)	0.15	**0.30**	0.09
Blogs (2)	0.12	0.09	**0.28**

Table 6: Jaccard similarity scores between data partitions. The partitions - (1) and (2) - are all equal in size (94,006) and non-overlapping.

of the same dataset. Low lexical overlap could be one of the reasons why content words lose their importance in the cross-domain condition.

These results seem to correlate with the performance of both the SVM and BERT models in the cross-domain settings: the higher the lexical overlap between the datasets, the higher the performance of the models. For example, the lexical overlap between the PAN data and the WebMD dataset (0.15) is higher than the overlap between the PAN data and Blog Authorship Corpus (0.12). Consequently, the performance of the SVM and BERT models when they are trained on the PAN data and tested on the WebMD data (Tables 16 and 19) is higher than when the models are trained on the PAN data and tested on the Blog Authorship corpus (Tables 18 and 21).

For the experiments with BERT in the cross-domain settings, we observe a substantial drop in performance on all the datasets when content-bearing words (i.e., nouns, adverbs) are removed and a relatively small drop when function words are omitted. These cross-domain results are in line with the in-domain observations by Talman et al. (2021) for a non-stylometric task (natural language inference), who report that the performance of the BERT model drops the most when content-bearing words are removed from the data. We can also observe that removing punctuation leads to a high score drop as well, which is the second most frequent word class in our data (Table 7 provides the relative frequency of the word classes in the datasets used).

	PAN '15	PAN '16	PAN data	WebMD	Blogs
ADJ	5.02	5.94	5.89	6.40	5.57
ADP	6.18	7.78	7.67	8.25	8.30
ADV	4.09	3.81	3.82	6.35	6.25
AUX	4.06	3.72	3.74	6.83	5.12
CCONJ	1.60	1.74	1.73	3.66	3.26
DET	6.25	6.68	6.65	9.38	7.67
INTJ	0.65	0.56	0.56	0.14	1.01
NOUN	15.91	**19.86**	**19.6**	**15.89**	**14.61**
NUM	1.38	1.72	1.7	2.3	1.23
PART	2.17	2.13	2.13	2.72	1.35
PRON	6.33	4.55	4.67	8.82	11.03
PROPN	12.40	12.85	12.82	3.02	4.42
PUNCT	**18.56**	13.41	13.74	10.06	13.61
SCONJ	0.94	0.84	0.84	1.72	1.35
SYM	1.80	2.16	2.14	0.22	0.13
VERB	9.37	9.47	9.47	12.30	13.04
X	3.29	2.79	2.82	0.06	0.36

Table 7: Relative frequency of the POS tags in the datasets.

While for the SVM approach features are extracted directly from the modified data, due to the nature of the BERT model that already has prior knowledge from pre-training, removing highly-frequent word classes essential for the meaning representation (content-bearing words), and therefore, fine-tuning and testing the model on meaningless data, affects the transfer learning and therefore its performance both under in-domain and cross-domain age detection conditions. In other words, these results indicate that BERT does not only rely on stylometric information (for example, function words and punctuation mark usage) when predicting the age of the author, but also on other cues

that might be related to topical and statistical bias in the data, as shown by a large drop in the cross-domain setup when content words are omitted.

Considering the overall effect of the POS ablation, the BERT model seems to be more sensitive to the removal of individual POS tags than the SVM model. For example, when comparing the performance of the SVM and BERT models trained on the Blog Authorship Corpus and tested on the WebMD dataset (Tables 17 and 20), the performance drop for the BERT model is considerably higher than from the SVM model. For example, removing nouns lowers the performance of the SVM model by 2.84 F1 points, while removing the same word class lowers the performance of the BERT model by 5.12 F1 points.

Furthermore, BERT performs better in almost all the examined experimental settings, with the exception of the results on the WebMD dataset (see Table 10 and 14), where SVM performs better (29.85% versus 50.06% F1-macro). This could be related to the relatively higher number of classes in this dataset (seven).

In the cross-domain condition, the performance of both SVM and BERT models drops substantially compared to the in-domain condition, despite the smaller number of classes, which is generally the case for cross-domain experiments.

5. Conclusion

In this paper, the importance of word classes for the in-domain and cross-domain age detection task was explored using both SVM and BERT. The experimental results with SVM showed that in the in-domain settings, content words are the most informative features, while in the cross-domain setup, function words are the most informative features in the majority of cases. This indicates that more attention should be paid to avoiding overfitting on content-based or topical words when developing machine learning approaches for the cross-domain age detection task.

The results for the BERT model, however, are different. BERT relies on highly-frequent word classes such as nouns and punctuation when making predictions. The observed behavior can be attributed to the different nature of BERT: it is a pre-trained language model, while SVM builds feature vectors directly from the modified data. This however may also indicate that BERT relies not only on stylometric information, such as punctuation mark usage, when predicting the age of the author, but also on other cues related to topical and statistical bias present in the data, as evidenced by the high importance of content-carrying words in the cross-domain setup.

In future work, we will conduct a detailed analysis of the topics present in the datasets in order to investigate whether the importance of content words for BERT is partially related to the topic bias present in the data. We will also examine other pre-trained language models in order to better generalize the findings.

6. Acknowledgement

This research has been supported by the Flemish Research Foundation through the bilateral research project FWO G070619N "The Linguistic Landscape of Hate Speech on Social Media". The research also received funding from the Flemish Government (AI Research Program).

References

Abdul-Mageed, Muhammad, Chiyu Zhang, Arun Rajendran, AbdelRahim Elmadany, Michael Przystupa, and Lyle Ungar (2019), Sentence-Level BERT and Multi-Task Learning of Age and Gender in Social Media, *CoRR*.

Alzahrani, Esam and Leon Jololian (2021), How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors, *CS & IT Conference Proceedings*, Vol. 11, pp. 1–8.

Argamon, Shlomo, Moshe Koppel, James Pennebaker, and Jonathan Schler (2009), Automatically Profiling the Author of an Anonymous Text, *Communications of the ACM* **52** (2), pp. 119–123, Association for Computing Machinery.

Barbieri, Federica (2008), Patterns of age-based linguistic variation in American English, *Journal of Sociolinguistics* **12** (1), pp. 58 – 88, John Wiley & Sons.

Bilan, Ivan and Desislava Zhekova (2016), Caps: A cross-genre author profiling system., *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CLEF and CEUR-WS.org, Évora, Portugal, pp. 824–835.

Chen, Jing, Long Cheng, Xi Yang, Jun Liang, Bing Quan, and Shoushan Li (2019), Joint Learning with both Classification and Regression Models for Age Prediction, *Journal of Physics: Conference Series*, Vol. 1168, IOP Publishing, pp. 1–12.

Chung, Cindy and James W Pennebaker (2007), The Psychological Functions of Function Words, *Social Communication* **1**, pp. 343–359, Psychology Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, **1**, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota.

Hochreiter, Sepp and Jürgen Schmidhuber (1997), Long Short-Term Memory, *Neural Computation* **9** (8), pp. 1735–1780, Neural Computation.

Johannsen, Anders, Dirk Hovy, and Anders Søgaard (2015), Cross-lingual syntactic variation over age and gender, *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Beijing, China, pp. 103–112.

Markov, Ilia, Efstathios Stamatatos, and Grigori Sidorov (2018a), Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing, *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Vol. 10762, Springer, Budapest, Hungary, pp. 289–302.

Markov, Ilia, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh (2016), Adapting Cross-Genre Author Profiling to Language and Corpus, *Working Notes Papers of the CLEF 2016 Evaluation Labs*, Vol. 1609, CLEF and CEUR-WS.org, Évora, Portugal, pp. 947–955.

Markov, Ilia, Vivi Nastase, and Carlo Strapparava (2018b), Punctuation as Native Language Interference, *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3456–3466.

Markov, Ilia, Vivi Nastase, and Carlo Strapparava (2020), Exploiting Native Language Interference for Native Language Identification, *Natural Language Engineering* **26**, pp. 1–31, Cambridge University Press.

Modaresi, Pashutan, Matthias Liebeck, and Stefan Conrad (2016), Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016., *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CLEF and CEUR-WS.org, Évora, Portugal, pp. 970–977.

Nguyen, Dong, Noah A. Smith, and Carolyn P. Rosé (2011), Author Age Prediction from Text using Linear Regression, *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 115–123.

Pardo, F., P. Rosso, B. Verhoeven, W. Daelemans, Martin Potthast, and Benno Stein (2016), Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations, *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CLEF and CEUR-WS.org, Évora, Portugal, pp. 750–784.

Polignano, Marco, Marco de Gemmis, and Giovanni Semeraro (2020), *Contextualized BERT Sentence Embeddings for Author Profiling: The Cost of Performances*, Springer International Publishing, Online, pp. 135–149.

Rangel, Francisco, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans (2015), Overview of the 3rd Author Profiling Task at PAN 2015, *in* Cappellato, Linda, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, CEUR-WS.org, Toulouse, France, pp. 1–8.

Rickford, John and Price Mackenzie (2013), Girlz II women: Age-grading, language change and stylistic variation, *Journal of Sociolinguistics* **17** (2), pp. 143–179, John Wiley & Sons.

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James Pennebaker (2006), Effects of Age and Gender on Blogging., *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6, Stanford, California, USA, pp. 199–205.

Shrestha, Prasha, Nicolas Rey-Villamizar, Farig Sadeque, Ted Pedersen, Steven Bethard, and Thamar Solorio (2016), Age and Gender Prediction on Health Forum Data, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portoroz, Slovenia, pp. 3394–3401.

Suman, Chanchal, Anugunj Naman, Sriparna Saha, and Pushpak Bhattacharyya (2021), A Multimodal Author Profiling System for Tweets, *IEEE Transactions on Computational Social Systems* **8** (6), pp. 1407–1416, IEEE.

Talman, Aarne, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann (2021), NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance, *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Linköping University Electronic Press, Reykjavik, Iceland (Online), p. 276–287.

van de Loo, Janneke, Guy De Pauw, and Walter Daelemans (2016), Text-Based Age and Gender Prediction for Online Safety Monitoring, *International Journal of Cyber-Security and Digital Forensics* **5**, pp. 46–60, The Society of Digital Information and Wireless Communications.

van der Goot, Rob, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank (2018), Bleaching Text: Abstract Features for Cross-lingual Gender Prediction, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, Association for Computational Linguistics, Melbourne, Australia, pp. 383–389.

Zhang, Chiyu and Muhammad Abdul-Mageed (2019), BERT-Based Arabic Social Media Author Profiling, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, Vol. 2517, Kolkata, India, pp. 84–91.

7. Appendix

Removed tag	F1-macro	Score change
None	58.47	0.00
DET	58.65	+0.18
SYM	58.76	+0.29
PART	58.30	-0.17
CCONJ	58.20	-0.27
PRON	58.13	-0.34
ADV	58.11	-0.36
SCONJ	58.09	-0.39
ADP	58.00	-0.47
INTJ	57.93	-0.54
NUM	57.87	-0.61
AUX	57.53	-0.94
VERB	57.43	-1.04
ADJ	56.89	-1.58
PROPN	55.42	-3.05
PUNCT	55.34	-3.13
X	55.04	-3.43
NOUN	54.22	-4.25

Table 8: In-domain results for the ablation experiments with SVM on the PAN '15 dataset (4 classes).

Removed tag	F1-macro	Score change
None	59.67	0.00
SCONJ	59.73	+0.07
NUM	59.65	-0.02
CCONJ	59.59	-0.08
X	59.57	-0.09
AUX	59.52	-0.15
INTJ	59.50	-0.16
SYM	59.41	-0.26
PRON	59.38	-0.29
ADV	59.26	-0.41
PART	59.18	-0.49
VERB	59.05	-0.62
DET	58.67	-1.00
ADP	58.54	-1.13
ADJ	58.48	-1.19
PUNCT	58.14	-1.53
PROPN	55.30	-4.36
NOUN	54.82	-4.84

Table 9: In-domain results for the ablation experiments with SVM on the PAN '16 dataset (5 classes).

Removed tag	F1-macro	Score change
None	50.06	0.00
X	50.08	+0.02
PUNCT	50.08	+0.02
PART	50.13	+0.07
CCONJ	50.19	+0.13
PRON	50.19	+0.13
ADP	50.20	+0.14
DET	50.27	+0.21
SYM	50.05	+0.00
INTJ	50.05	-0.01
SCONJ	49.99	-0.06
AUX	49.93	-0.13
ADV	49.92	-0.14
ADJ	49.79	-0.27
VERB	49.79	-0.27
PROPN	49.50	-0.56
NUM	49.13	-0.93
NOUN	48.54	-1.51

Table 10: In-domain results for the ablation experiments with SVM on the WebMD dataset (7 classes).

Removed tag	F1-macro	Score change
None	43.45	0.00
DET	43.45	+0.00
SYM	43.45	+0.00
PART	43.45	+0.01
CCONJ	43.50	+0.05
PUNCT	43.44	-0.01
X	43.43	-0.02
AUX	43.33	-0.12
SCONJ	43.32	-0.12
ADV	43.29	-0.16
ADP	43.22	-0.23
NUM	43.21	-0.24
PRON	43.18	-0.27
INTJ	43.18	-0.27
ADJ	43.03	-0.42
VERB	42.91	-0.54
PROPN	42.61	-0.84
NOUN	41.14	-2.31

Table 11: In-domain results for the ablation experiments with SVM on the Blog Authorship Corpus (4 classes).

Removed tag	F1-macro	Score change
None	68.09	0.00
CCONJ	68.43	+0.34
PART	68.25	+0.16
SCONJ	67.44	-0.65
PRON	67.37	-0.73
NUM	67.32	-0.77
INTJ	67.21	-0.88
VERB	67.12	-0.97
ADJ	66.97	-1.12
SYM	66.87	-1.22
DET	66.65	-1.44
AUX	66.64	-1.46
X	66.33	-1.76
ADP	66.30	-1.79
ADV	65.96	-2.13
NOUN	64.55	-3.55
PROPN	63.54	-4.56
PUNCT	62.68	-5.41

Table 12: In-domain results for the ablation experiments with BERT on the PAN '15 dataset (4 classes).

Removed tag	F1-macro	Score change
None	66.33	0.00
CCONJ	66.17	-0.17
SYM	66.03	-0.30
ADV	65.98	-0.36
PRON	65.80	-0.54
PART	65.76	-0.57
SCONJ	65.71	-0.62
DET	65.70	-0.63
NUM	65.62	-0.71
VERB	65.43	-0.90
AUX	65.42	-0.91
X	65.32	-1.01
INTJ	65.31	-1.02
ADP	64.93	-1.40
ADJ	64.12	-2.21
PUNCT	62.80	-3.54
PROPN	60.78	-5.55
NOUN	59.85	-6.48

Table 13: In-domain results for the ablation experiments with BERT on the PAN '16 dataset (5 classes).

Removed tag	F1-macro	Score change
None	29.85	0.00
INTJ	30.26	+0.41
X	30.19	+0.34
SYM	30.16	+0.30
PRON	30.11	+0.26
PART	30.09	+0.24
ADV	29.94	+0.09
SCONJ	29.81	-0.04
ADP	29.76	-0.09
ADJ	29.67	-0.18
DET	29.64	-0.21
CCONJ	29.63	-0.23
AUX	29.42	-0.44
PUNCT	29.34	-0.52
PROPN	29.28	-0.57
VERB	29.02	-0.83
NUM	28.72	-1.13
NOUN	28.49	-1.36

Table 14: In-domain results for the ablation experiments with BERT on the WebMD dataset (7 classes).

Removed tag	F1-macro	Score change
NONE	46.87	0.00
AUX	46.97	+0.11
SCONJ	46.86	-0.01
PART	46.86	-0.01
DET	46.81	-0.06
ADP	46.67	-0.20
CCONJ	46.66	-0.21
SYM	46.63	-0.23
INTJ	46.63	-0.24
VERB	46.56	-0.31
X	46.51	-0.36
ADV	46.42	-0.45
ADJ	46.26	-0.61
PROPN	46.04	-0.83
PUNCT	45.86	-1.01
PRON	45.82	-1.05
NUM	45.68	-1.19
NOUN	43.82	-3.05

Table 15: In-domain results for the ablation experiments with BERT on the Blog Authorship Corpus (4 classes).

Removed tag	F1-macro	Score change
None	34.39	0.00
X	34.97	+0.58
PROPN	34.84	+0.45
CCONJ	34.74	+0.35
ADP	34.63	+0.24
NOUN	34.57	+0.17
ADJ	34.55	+0.16
VERB	34.54	+0.15
INTJ	34.51	+0.11
NUM	34.45	+0.06
SCONJ	34.41	+0.02
SYM	34.38	-0.01
PUNCT	34.33	-0.06
PRON	34.28	-0.11
ADV	34.29	-0.11
AUX	34.25	-0.14
PART	34.19	-0.20
DET	34.18	-0.21

Table 16: Cross-domain results for the ablation experiments with SVM: training on PAN data, testing on WebMD (3 classes).

Removed tag	F1-macro	Score change
None	25.50	0.00
SYM	25.50	+0.00
PROPN	25.53	+0.03
ADV	25.56	+0.06
NUM	25.62	+0.12
CCONJ	25.76	+0.26
DET	25.90	+0.40
PUNCT	25.49	-0.01
INTJ	25.46	-0.04
X	25.41	-0.09
VERB	25.40	-0.10
SCONJ	25.36	-0.14
PART	25.35	-0.15
ADP	25.27	-0.23
ADJ	24.85	-0.65
AUX	24.70	-0.80
PRON	24.51	-0.99
NOUN	22.66	-2.84

Table 17: Cross-domain results for the ablation experiments with SVM: training on the Blog Authorship Corpus, testing on WebMD (3 classes).

Removed tag	F1-macro	Score change
None	24.06	0.00
X	27.47	+3.41
PROPN	26.79	+2.73
ADP	25.95	+1.89
NOUN	25.06	+1.00
ADJ	24.34	+0.28
ADV	24.34	+0.28
SCONJ	24.33	+0.27
NUM	24.32	+0.26
AUX	24.27	+0.21
INTJ	24.11	+0.05
SYM	24.06	0.00
PUNCT	23.78	-0.28
PART	23.33	-0.73
VERB	23.28	-0.78
DET	21.01	-3.05
CCONJ	20.22	-3.84
PRON	20.07	-3.99

Table 18: Cross-domain results for the ablation experiments with SVM: training on PAN data, testing on the Blog Authorship Corpus (3 classes).

Removed tag	F1-macro	Score change
None	35.38	0.00
CCONJ	36.26	+0.88
SYM	35.98	+0.60
INTJ	35.96	+0.57
ADP	35.71	+0.33
PRON	35.70	+0.32
SCONJ	35.66	+0.28
PART	35.52	+0.14
AUX	35.34	-0.04
VERB	35.32	-0.06
ADJ	35.25	-0.14
NUM	35.23	-0.15
PROPN	35.19	-0.19
X	34.52	-0.86
DET	34.45	-0.93
ADV	34.32	-1.06
NOUN	33.60	-1.79
PUNCT	33.50	-1.88

Table 19: Cross-domain results for the ablation experiments with BERT: training on PAN data, testing on the WebMD dataset (3 classes).

Removed tag	F1-macro	Score change
NONE	35.44	0.00
ADV	35.37	-0.07
ADP	35.29	-0.15
INTJ	34.88	-0.56
DET	34.83	-0.62
PART	34.42	-1.02
SYM	33.93	-1.51
PROPN	33.87	-1.57
AUX	33.52	-1.92
ADJ	33.14	-2.31
X	32.66	-2.78
NUM	32.63	-2.81
PRON	32.50	-2.94
SCONJ	32.40	-3.04
CCONJ	32.28	-3.17
VERB	31.28	-4.16
NOUN	30.32	-5.12
PUNCT	28.02	-7.43

Table 20: Cross-domain results for the ablation experiments with BERT: training on the Blog Authorship Corpus, testing on the WebMD dataset (3 classes).

Removed tag	F1-macro	Score change
None	27.39	0.00
PRON	26.97	-0.42
ADP	26.39	-1.00
NUM	26.37	-1.02
VERB	26.26	-1.13
CCONJ	26.24	-1.15
SCONJ	26.18	-1.21
SYM	26.15	-1.24
INTJ	25.72	-1.67
AUX	25.72	-1.67
PART	25.63	-1.76
X	24.99	-2.40
ADJ	24.77	-2.62
DET	24.48	-2.91
NOUN	24.20	-3.19
PROPN	23.98	-3.41
ADV	23.70	-3.69
PUNCT	23.05	-4.34

Table 21: Cross-domain results for the ablation experiments with BERT: training on PAN data, testing on the Blog Authorship Corpus (3 classes).