Increasing In-Class Similarity by Retrofitting Embeddings with Demographic Information

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Tommaso Fornaciari
Bocconi University
fornaciari@unibocconi.it

Abstract

Most text-classification approaches represent the input based on textual features, either feature-based or continuous. However, this ignores strong non-linguistic similarities like homophily: people within a demographic group use language more similar to each other than to non-group members. We use homophily cues to retrofit text-based author representations with non-linguistic information, and introduce a trade-off parameter. This approach increases in-class similarity between authors, and improves classification performance by making classes more linearly separable. We evaluate the effect of our method on two author-attribute prediction tasks with various training-set sizes and parameter settings. We find that our method can significantly improve classification performance, especially when the number of labels is large and limited labeled data is available. It is potentially applicable as preprocessing step to any text-classification task.

1 Introduction

Predicting socio-demographic author characteristics is becoming ever more relevant with the pervasive use of user-generated content. Classifying user attributes such as age and gender is useful for a number of applications both in the public sector, where it can support the investigation of crime (in forensic linguistics) or the determination of social policies, and in the private sector, where companies want to profile a potential consumer market, targeting communication strategies and advertising to specific communities. Furthermore, recent work in NLP has shown that incorporating author attributes in various NLP tasks can also improve performance (Volkova et al., 2013; Hovy, 2015; Hovy and Søgaard, 2015; Lynn et al., 2017; Preotiuc-Pietro et al., 2016).

In these tasks, authors are typically represented via their linguistic profiles, i.e., information available in the text. This includes both word-based features as well as continuous representations (embeddings). Generally, linguistic features are divided into content-related and strictly stylistic features. While the first can be effectively represented by (n-grams of) words which capture the topic and meaning of a text, the second ones focus on the use of function words, expressions, pronouns, syntactic structures, etc. There is evidence in the literature that content-related text characteristics are more effective than stylistic features for gender and age prediction (Fatima et al., 2017; Rosenthal and McKeown, 2011). This effect is the consequence of a non-linguistic auto-selection process known as *homophily*: people within in a demographic group tend to be more similar to each other than to other groups, and subjects belonging to different groups are therefore naturally more prone to discuss different topics.

Despite the large amount of available social media data (in April 2018, Facebook had more than two billion active users, YouTube and WhatsApp each one-and-a-half billion, and Twitter 330 million, see statista.com), we often encounter scenarios with limited availability of ground-truth user attributes, leading to remarkable performance differences to, say, blogs. This difference is due to the shortness of social media texts and the wider range of topics (Rangel et al., 2016), which weaken linguistic profile features. In such cases, improving author representations beyond the linguistic profiles can be especially useful.

We implement this intuition of leveraging demographic homophily by using retrofitting (Faruqui et al., 2015), a method introduced to refine word vectors to reflect semantic similarity information from lexicons. In our case, we increase the similarity between the (linguistically-based) continuous authors representations within each class (here: age or gender). Authors who

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 671–677
Brussels, Belgium, October 31 - November 4, 2018.

share the same gender or age therefore get more similar vector representations (see section 3.2). This effectively increases class-separability and can thereby improve classification performance.

We also experiment with a trade-off parameter α, which controls the relative influence of the retrofitting process vs. the original embedding vector on the retrofit representation, allowing us to explore the effect of both factors on the final prediction outcome.

In order to extend the in-class homophily information to unlabeled data, we induce a transformation matrix to translate between the original and retrofitted embedding space. This matrix can be applied to unlabeled data to transform the author representations in the test set.

We use a set of almost 100K authors to predict age and gender. In order to explore limited-resource scenarios, we experiment with a range of training set sizes. Our results indicate that demographic retrofitting of linguistic representations substantially increases classification performance for age and gender prediction, especially in low-resource scenarios.

It is an easy, fast, and efficient preprocessing step that can substantially improve classification performance. We show our method for author-attribute prediction, but believe it can potentially be applied to *any* text-classification task.

Contributions In this paper, we introduce demographic retrofitting based on in-class homophily, and make the following contributions:

1. we present a substantial expansion of the original retrofitting algorithm (Faruqui et al., 2015). In contrast to prior work, which relies on external ontologies, our method relies solely on the information contained within the training data.

2. We show how to generalize the transformation from training data to unlabeled data, using a translation matrix.

3. We publicly release all our data and models on our GitHub page, https://github.com/Bocconi-NLPLab/retrofit_attributes.

2 Data

We use data from Hovy et al. (2015), a collection of reviews of online companies from various countries, including author information. We select all reviews written in English from American and British sources, if they include both age and gender of the author, and if the review is at least 10 tokens long (shorter reviews tend to be mis-tokenized URLs or replies). We aggregate the reviews by users, so that each instance is a collection of texts' from a unique user. This leaves us with 98,608 individual users, and about 8M words (roughly 80 words per instance). For each user, we use the age (discretized by decade) and gender (self-stated as binary, and augmented by Hovy et al. (2015) based on the users' first name) as target variables. We minimally preprocess the text data, collapsing all numbers into 0s, and tokenizing via spacy (Honnibal and Johnson, 2015).

3 Methodology

In our experiments, we are interested in the effect of homophily-inducing retrofitting on author-attribute prediction. In order to evaluate the effect, we compare the performance of author representations based on linguistic input to the performance of the same representation retrofitted to the author attribute class in question. In this section, we outline the details for the different steps.

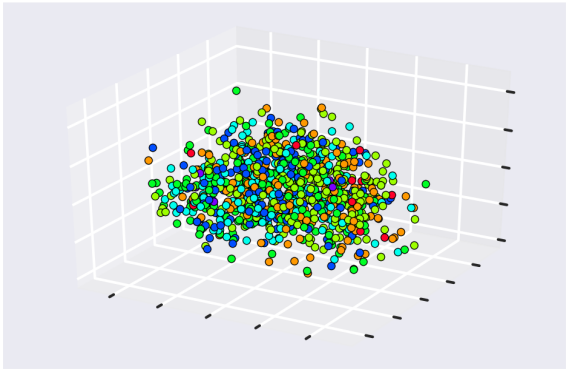
3.1 Linguistic author representations

We train Doc2Vec, a paragraph2vec (Le and Mikolov, 2014) implementation, on the corpus, inducing a 98K-by-300 matrix D, where each row represents an author. We follow the parametrization suggested in Lau and Baldwin (2016), setting the window size to 15, minimum word-frequency to 10, negative samples to 5, downsampling rate to 0.00001, and run for 1000 iterations. We use the resulting author embeddings as input to the author-attribute classifier (see 3.4). We induce the author embeddings over the entire corpus of 98K authors, without recurrence to age or gender information.

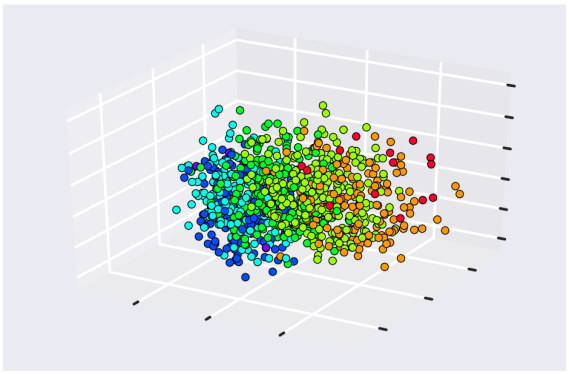
As comparison, we also create a bag-of-words (BOW) representation with the same dimensionality. We use χ^2 as selection criterion to find the top 300 words in the training data, separately for both age and gender classification.

3.2 Retrofitting

Our goal is to enhance the author representations, which are based on linguistic similarity, with demographic information about the target variable (say, age). In order to introduce this information into the vector space, we rely on *retrofitting*, by in-



(a) Non-retrofitted



(b) Retrofitted

Figure 1: Schematic representation of 500 authors colored by age group, without (top) and with (bottom) retrofitting

creasing the similarity of authors within the same target group (say, people in their 20s). We thereby separate the target classes in embedding space, making them easier to differentiate by a classifier.

Faruqui et al. (2015) introduced retrofitting of word vectors based on external ontologies, such as WordNet (Miller, 1995) or PPDB (Ganitkevitch et al., 2013). Instead of these resources, we map each labeled author to the list of all other authors with the same label in the training data. Formally, we create a set Ω containing tuples of authors $(d_i, d_j | y_i = y_j)$. We do this separately for each demographic dimension - age and gender.

During retrofitting, we iteratively update the author representation in the training data (initially linguistically-based) to increase the cosine similarity between authors within the same class (as defined in Ω). This creates a retrofitted matrix $\hat{D}_{train}$ of the original author matrix D_{train}. The update for an author representation d_i is a weighted combination of the original embedding and the average over all its current neighbors:

$$\hat{d}_i = \alpha d_i + \beta \frac{\sum_{j:(i,j)\in\Omega} \hat{d}_j}{N}$$

where d_i is the original linguistic representation vector, $N = |\{\forall j : (i,j) \in \Omega\}|$ is the set of all embeddings in the same label group, and α and β are hyper-parameters that control the trade-off between the original representation and the updates from the neighboring embeddings during retrofitting. In Faruqui et al. (2015), $\alpha = \beta$. In contrast, we define

$$\beta = 1 - \alpha$$

By varying α from 0 to 1, we can control the strength of the retrofitting process. $\alpha = 1$ simply reproduces the original matrix, i.e., $\hat{D} = D$, whereas $\alpha = 0$ only relies on the neighborhood updates after the initialization. Figure 1 shows a sample of 500 users in a non-retrofitted (1a) and retrofitted (1b) 3D embedding space, colored by class. The color distribution shows how people belonging to the same group get drawn closer to each other in embeddings space when using retrofitting.

3.3 Translation

We can only retrofit the embeddings of authors in the training set D_{train}, since we need information about the class label in order to construct Ω. However, the retrofitting process changes the configuration of the embedding space (into $\hat{D_{train}}$), so a separating hyperplane learned on $\hat{D}_{train}$ will not be applicable to a test set D_{test} in the original embedding space.

In order to extend the homophily information to authors in the test set, we use a *translation matrix* T (a 300×300 matrix), which approximates the transformation from the original training data matrix D_{train} into the retrofitted matrix $\hat{D}_{train}$. We obtain T by minimizing the least-square difference in $D_{train} \cdot T = \hat{D}_{train}$.

T captures the retrofitting operation, and allows us to modify the test subjects' representations *as if* age and gender were known, despite the absence of class information. In particular, by applying T to the matrix of the author embeddings in the unlabeled test set D_{test}, we obtain a retrofitted version $\hat{D_{test}}$ that preserves the transformation learned on the training data. Since the least-square approximation is not perfect, we find that in practice fitting a classifier on the approximation $D_{train} \cdot T$

works better than using $\hat{D}_{train}$, acting as a regularizer.

3.4 Classification

We retrofit the author embeddings in the training set (see 3.2) and learn a translation matrix to transform the representations of the remaining authors in the test set. We train three separate Logistic Regression classifiers: one on author embeddings, one on the retrofit embeddings, and one on BOW features. It is technically possible to retrofit BOW representations as well, but in practice, the classifier does not converge, as word count-based vectors do not represent a continuous space that captures latent similarities.

We then use the three classifiers to predict the author attributes of the remaining authors in the test data set. We evaluate the results via micro-F1 score (averaged over 100 runs), since our tasks include imbalanced multi-class scenarios: micro-F1 weights the contribution of each class according to their relative size and is therefore more informative than accuracy.

Since we are interested in the effect of the training set size on performance, we vary the number of available training examples from 1000 to 10,000, using the remaining authors as test set. For each training set size, we collect 100 random subsamples and average over them.

4 Results

The learning curves in Figure 2 show that retrofitting outperforms both the original author embeddings and BOW representations for age (top) and gender (bottom) prediction in terms of F1. The effect is stronger when little training data is available. We evaluate the statistical significance of the difference between results with retrofitting and original embeddings via a bootstrap sampling test. We do not test against BOW, since this is consistently lower than embeddings. The resulting p-value are given in the respective figures. For gender classification, there are small, but not significant improvements with retrofitting. By contrast, for age classification, small values of α (0.01, 0.1, and 0.25) result in significantly better classification than when using any other method.

The performance differences between the methods are generally more pronounced for age-prediction, which has 10 possible labels, than for gender prediction (two labels). The difference in optimal α value suggests a relation between α and label space.

In both tasks, the best result is achieved by choosing a low α, i.e., by giving more weight to the demographic association of the users than to their linguistic feature representations. In practice, this value should be determined via cross-validation: here we show different levels of α in order to give some intuition of its on performance. Note that the curves for the original embeddings and BOW are unaffected by α and do not change. We repeat them at each figure for comparison. Increasing α eventually converges with the original embeddings, but we see that even intermediate values can be close to the original embeddings.

5 Related Work

The first studies to apply statistical NLP techniques to author attribute prediction are Koppel et al. (2002); Argamon et al. (2003), using the British National Corpus (BNC). The same authors also introduced the use of blogs as data source (Koppel et al., 2006).

In recent years, predicting socio-demographic variables from text has seen increased interest, with several corpora for the classification of age and gender, covering various languages, such as English (Schler et al., 2006; Rosenthal and McKeown, 2011), Spanish, French, German, Dutch (Company and Wanner, 2015), Greek (Mikros, 2012), Chinese (Zhang et al., 2016), and Vietnamese (Pham et al., 2009).

A big contribution in this field, however, was the shared tasks of the PAN workshops (Rangel et al., 2013, 2014, 2015, 2016).

Research has identified a variety of linguistic features, ranging from "stylistic features with n-grams models, parts-of-speech, collocations, LDA, different readability indexes, vocabulary richness, correctness or verbosity" (Rangel et al., 2016). However, none of these papers used demographic information directly in the author representations.

Closest to our method are Lopez-Monroy et al. (2013), who propose the use of second-order representations. They created specific profiles for the target classes, and exploited them for the creation of the profile of each document. In both cases, the linguistic representation of the documents passes through a class-related profile.

The methods applied in the PAN workshops also reflect the recent research trend towards word embeddings, which we explore in this pa-

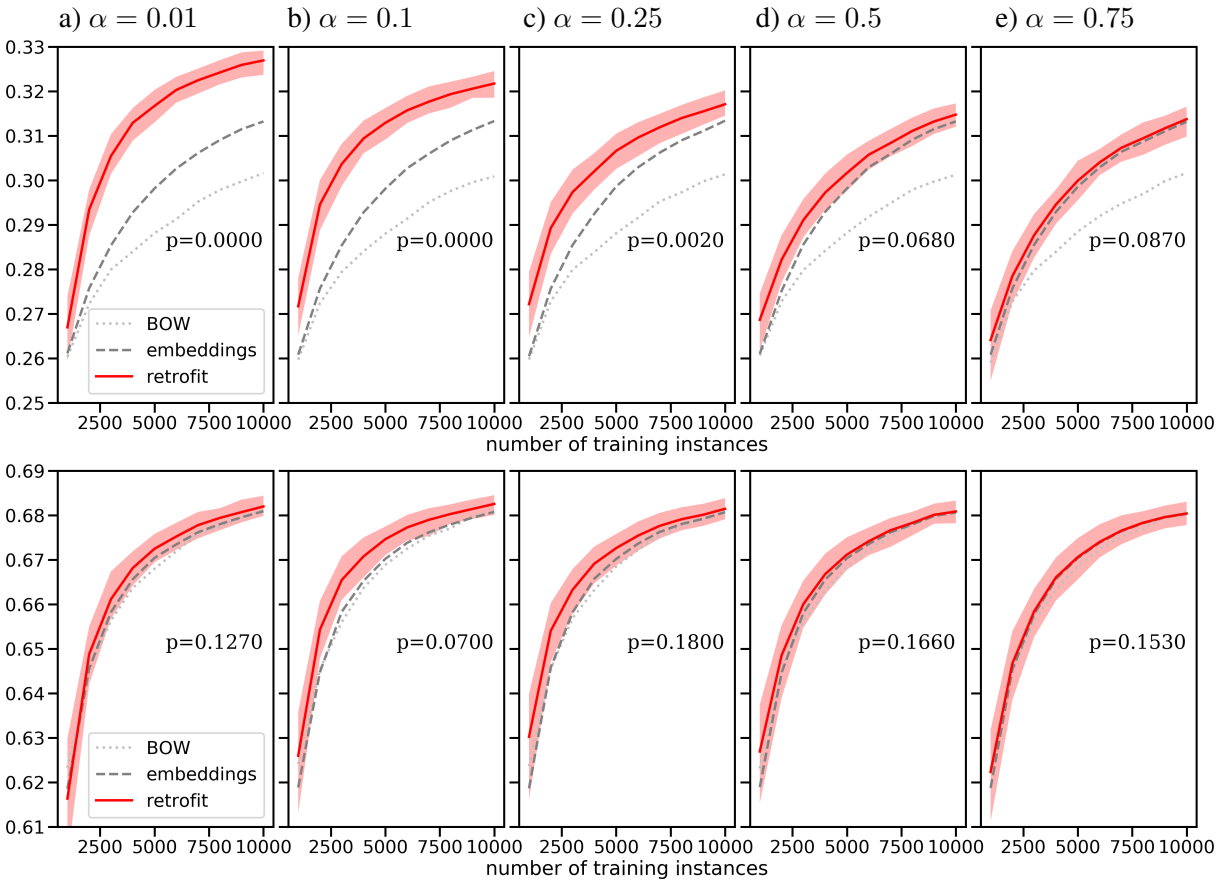


Figure 2: Learning curves (micro-F1) for 3 classifiers on age (top) and gender prediction (bottom) for different values of α. Retrofitting influence decreases from left to right. All data points averaged over 100 runs. Shaded area is 95%-confidence interval for retrofitting. p-values denote statistical difference between original and retrofit embeddings according to bootstrap test.

per. Bayot and Gonçalves (2016) first used `word2vec` embeddings as input features to a SVM classifier, followed by the use of convolutional (CNN) and recurrent neural networks (RNN) by Miura et al. (2017). Markov et al. (2016) also created document representations through word2vec, using a Logistic Regression classifier.

6 Conclusion

We explore retrofitting text-based author embeddings with non-linguistic demographic information to increase in-class similarity. This method increases class-separability to improve classification performance. We use a corpus of almost 100K users, and evaluate the effect of our method on age and gender prediction for various levels of available training data. We find that aggressive retrofitting (prioritizing homophily over linguistic embeddings) is beneficial for prediction performance, especially when the available amount of training data is limited. While the effect diminishes with increased training data size, our approach provides a simple method to incorporate non-linguistic knowledge into author representations. For another application (introducing geographic information into city representations), see Hovy and Purschke (2018). Our method is fast, simple, and applicable to any problem represented in embedding space. It is therefore a viable preprocessing step to any text-classification task.

Acknowledgements

We would like to thank the anonymous reviewers, whose constructive feedback improved the paper, as well as Carlo Baldassi for many enlightening discussions, as well as Barbara Plank, Noah Smith, and Afshin Rahimi, whose feedback helped our conceptualization of the problem.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TextInterdisciplinary Journal for the Study of Discourse*, 23(3):321–346.

Roy Bayot and Teresa Gonçalves. 2016. Author Profiling using SVMs and Word Embedding Averages-Notebook for PAN at CLEF 2016.

Juan Soler Company and Leo Wanner. 2015. Multiple Language Gender Identification for Blog Posts. In *CogSci*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Mehwish Fatima, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab. 2017. Multilingual author profiling on Facebook. *Information Processing & Management*, 53(4):886–904.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, pages 452–461. International World Wide Web Conferences Steering Committee.

Dirk Hovy and Christoph Purschke. 2018. Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. In *Proceedings of EMNLP*.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 483–488.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. page 78.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

A Pastor Lopez-Monroy, Manuel Montes-Y-Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esau Villatoro-Tello. 2013. INAOEs participation at PAN13: Author profiling task. In *CLEF 2013 evaluation labs and workshop*.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. 2017. Human Centered NLP with User-Factor Adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155.

Ilia Markov, Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and Alexander Gelbukh. 2016. Author profiling with doc2vec neural network-based document embeddings. In *Mexican International Conference on Artificial Intelligence*, pages 117–131. Springer.

George K Mikros. 2012. Authorship attribution and gender identification in Greek blogs. *Methods and Applications of Quantitative Linguistics*, 21:21–32.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma. 2017. Author profiling with word+ character neural attention network. *Cappellato et al.[13]*.

Dang Duc Pham, Giang Binh Tran, and Son Bao Pham. 2009. Author profiling for Vietnamese blogs. In *Asian Language Processing, 2009. IALP'09. International Conference on*, pages 190–194. IEEE.

Daniel Preotiuc-Pietro, Wei Xu, and Lyle H Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *AAAI*, pages 3030–3037.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–30.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF*, page 2015. sn.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.*, pages 750–784.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, pages 1815–1827.

Wanru Zhang, Andrew Caines, Dimitrios Alikaniotis, and Paula Buttery. 2016. Predicting Author Age from Weibo Microblog Posts. In *LREC*.