Systematic Errors and the Stability of Feature Importance

An Assessment for the Social Scientists

Marcel H. Schubert[1]

[1]Max-Planck Institute for Research on Collective Goods

Machine-learning is used ever more widely, particularly in the social sciences and in law. Here, Natural Language Processing (NLP) finds many cases of use as it enables the processing of large-scale online text data. However, often the focus only lies on a model's performance and not on its transparency. That lack of transparency is especially troubling as the users of those models may often not be its designers. As such, they have no way of assessing whether there are systematic errors or whether the performance hinges on unknown, possibly unstable, factors. Here, we offer an in-depth analysis of the effect small variations in the input have on systematic errors and feature stability. Our aim is to enable social scientists and practitioners using the technology to assess whether they may invite either normative or unwanted systematic errors into their results when using current technologies.

1 Introduction

In traditional studies within the social sciences, characteristics such as age or gender are key traits as they have proven to be central to understanding and modeling human behavior (e.g., Lahey et al., 2000; Gneezy and Rustichini, 2004; Charness and Gneezy, 2012; Booth and Nolen, 2012; Sutter and Glätzle-Rützler, 2014; Bian, Leslie, and Cimpian, 2017). While traditionally studies within the field focused on lab experiments as well as questionnaires, large-scale datasets have, until recently, been of limited availability. Especially when it comes to text data, the analysis often proved resource-intensive and the results are difficult to assess. However, in general, that data is a wealth of information, even more so since the amount of text data generated by individuals increased massively with the advent of services such as Facebook or Twitter (InternetLiveStats, 2019). With such large amounts of data, new methods in machine-learning and natural language processing (NLP) gained great popularity, especially within the social sciences. In order to mine that treasure trove, researchers increasingly turn to the application of NLP to answer prevailing questions in the social sciences (Bail, 2016; Pavlick et al., 2016; Costa-jussà, 2019; Burley et al., 2020, e.g.,). The technology as such is, depending on the algorithms, comparatively simple to use in terms of know-how when one considers the textual features for identification as well as the difficulty of setting up and training the respective algorithms (Narayanan et al., 2012). By reducing the entrance cost in such a way, this technology certainly has great potential. Here, the section of authorship analysis is becoming an area of special relevance to the social scientists. The reason is simply that, while a large amount of text data is available, characteristics of the individual, such as age and gender, but also their identity, often are not. The main goal of that particular area is therefore to profile characteristics such as age and gender, but also political orientation or even the author's identity from texts written by the individual. As such, while studies within the social sciences, in making use of that new type of data, continue to include these characteristics due to their proven relevance in past research (e.g. Bail, 2016; Colleoni, Rozza, and Arvidsson, 2014), authorship analysis is used to compensate for the lack of ground truth. That is often done by using a layered approach, first inferring the missing characteristics with a trained classifier, and then in turn using that as input to their research approach (Barberá and Rivero, 2015; Huang, Su, and Iwaihara, 2020). Moreover, these characteristics are not only important as an input for further research, but also as identifying information. In the adjacent field where NLP and behavioral sciences intersect, the law community is making use of such analysis for the targeting and researching of incriminating online behavior, e.g., hate speech (Djuric et al., 2015; Laub, 2019; Zufall et al., 2020). The practical application of authorship analysis also becomes relevant when pursuing offenses. Often, online users do not use their real names, and finding out their identities becomes difficult either because it is not available or because the companies having access to the information are not willing to share it (Stuttgarter Nachrichten, 2017). Consequently, often during such an investigation, *forensic* authorship analysis is employed to gain additional information on the potential offender. Due to the high volume of data, these processes become increasingly automated and as such the research field of automated forensic authorship analysis is

well established (Rocha et al., 2017).

However, while state-of-the-art methods regularly manage to achieve a high accuracy for authorship analysis (Rocha et al., 2017), some call the field's scientific character into question (Chaski, 2001). This is due to the fact that current research of automated authorship analysis mostly focuses on correct predictions, using a wide set of features which often varies between different papers (Rocha et al., 2017). The focus seems to lie on achieving the best results, showing the viability of automated authorship analysis often to the detriment of rigorous explainability and transparency (Chaski, 2012). This not only generally concerns research employing such models but, when used by law enforcement, it is also directly related to the admissibility of findings from automated authorship analysis in the courts, as transparency and rigor might not satisfy the demands before the law (Chaski, 2012). As such, this topic is very much of normative interest. Explainability and transparency are the central aspects when a decision made by some model affects individuals. Within the field of law, as outlined above, that may affect the individual directly as she comes under suspicion when identified by the machine. In social sciences, they may be more indirectly affected. One way would be when predicted labels on a datasset are used for further analysis. The result from such an analysis may inform policy decisions. However, the error will propagate. As the underlying labels assigned by the machine were systematically faulty, the policy design will be as well. On the one hand, it is therefore paramount that, when exposed to such machine-informed decisions and label assignments, there are no systematic patterns of errors. On the other hand, the things driving the prediction result should be more than just spurious correlations, present in one dataset, absent in the next. Consequently, the technology is in need of further assessment before being used even more widely than is already the case.

Besides understanding the algorithm, such an explainability would require two things: One concerns the topic independence, which means that the features predictive of an author should not depend on the content of the text (Narayanan et al., 2012). The second aspect is rarely mentioned. Most models are trained, at one point in time, on one particular text corpus related to one domain. Using the trained model at a later point in time, however, assumes that in the meantime there was no shift in the underlying features used or shifts in the set used by a particular author. If one looks at age, for example, it may be that older people are currently using more grammatically correct language in the online environment compared to younger people (Flekova, Preoţiuc-Pietro, and Ungar, 2016). Naturally, that is merely a snapshot of the current environment. It does not imply that the next generation features the same pattern, and consequently, any model trained on the old pattern might inadvertently misclassify when confronted with the new pattern. For that reason, it is necessary to assess the stability of individual features used by the classifier, when the underlying data and thus the patterns change slightly. There are only very few forays seeking to address such problems, for example as Azarbonyad et al. (2015) do with their temporal weighting of features. As stylometry is rooted in the humanities and thus the social sciences (Neal et al., 2017), it is surprising that not more efforts have been made so far to see whether some author characteristics

result in some stable topic- and domain-independent features and feature importance.

Therefore, the central aspect of our approach is helping to answer that question about the stability. We assess the stability in terms of predictions and in terms of feature importance. We seek to extend the understanding about the fundamentals of language and communication. We think that such a stability analysis would aid immensely in assessing the rigor of predictions, therefore making them safer to use in the legal context. Moreover, it would also help us to better establish the boundaries of transferability and stability of models and their predictions, which is needed predictions of such models are used as input for further research. This contribution is therefore interdisciplinary in nature, as it tries to address an issue affecting multiple fields. While the lack of contributions has already been pointed out Rocha et al. (2017), only recently, have there been any notable forays. In general, there has been an effort to make model predictions more explainable (Ribeiro, Singh, and Guestrin, 2016; Samek et al., 2019). A systematic approach , however, looking at changes when features are systematically varied, is still limited. The study by Koppel, Schler, and Argamon (2011) looks at authorship analysis "in the wild" and systemically varies the number of authors as well as the number of features to assess and quantify gains and losses in performance. However, the authors do not focus on feature types and do not extend their study towards analyzing the changes in within the model. Recently, Boenninghoff et al. (2019) showed a method to make a complex model based on a neuronal net explainable. Their approach is limited to their specific model and does not analyze either what the decisive features correspond to, i.e., how much context they encode. In that vein, Sanchez-Perez et al. (2017) is closer to our approach. They also seek to limit topic dependency and focus on feature types. However, their goal is to find a good subset of n-grams for their feature set with high predictive power. The paper closest to ours is the one by Sage et al. (2020). Their analysis is focused on different feature types and the influence of varying n-gram lengths. They systematically vary both in order to find the impact on performance. However, they do not extend their analysis to different input sets and also focus on longer news articles instead of the more common data of microblog texts. Moreover, we also extend that analysis into the domain of stability, assessing whether there are shifts in feature importance.

2 Experimental Design and Data

In order to conduct our stability analysis, we conduct an experiment as used in the field of machine-learning by introducing controlled variations to an underlying, given dataset. To that end, we use a fixed setup of machine-learning (ML) models and test their internal stability, when they are exposed to these controlled variations.

Synopsis The dataset used for the experiment is the PAN @CLEF 2019 Celebrity Profiling (PAN2019) dataset.[1] As our goal is to assess te performance and the stability

[1]The dataset may be downloaded from the website of the PAN challenge: PAN Challenge 2019.

of importance in regards to single features, we try to reduce variation present directly within the authors as much as possible. Therefore, we focus only on authors from one category, namely those dubbed "creator". As suggested in the guidelines of the original PAN challenge, we change the age from a numerical variable to their categorical one consisting of five age brackets. As a prediction target, we select "age" and "gender", two commonly used characteristics in the social sciences. Moreover, in order to exclude further any variation introduced by an imbalanced dataset, we undersample the data in such a way that the genders as well as the age groups are balanced. For the comparison in the author dimension, we create four subsets consisting of 50, 150, 500, and 1000 authors, respectively. The upper limit of 1000 authors reflects the maximum number of authors for whom it is still possible to balance the dataset. Furthermore, we repeat the experiment three times for different minimum lenghts per training instance, as the text lengths were shown to impact classifier performance; in doing this, we hold the model setup constant (Custódio and Paraboni, 2021). The minimum lengths are 100, 250, and 500 characters, respectively. In order to achieve these minimum lengths, tweets from the same author were concatenated. As feature types we use the following ones, sorted in ascending order in terms of encoded context information: DIST, CHAR, ASIS, POS, TAG, DEP, LEMMA, WORD, NUM. For all types, we apply the n-gram ranges found to be useful by prior research (Custódio and Paraboni, 2021). The evaluation is conducted using the 500-score as the performance measure. To assess the stability of features, we use Spearman's Rho for rank order correlation. All evaluations were done on a separate hold-out dataset, the test set. That data was not used during training at any point. In the following paragraphs, we outline our design choices in detail.

2.1 Data

While the literature for authorship analysis is abundant and only increased during recent years, there are no easily identified commonly used datasets across a wide range of studies. Comparison between studies is therefore difficult. This is well illustrated by Neal et al. (2017) who list 13 datasets used more than once and a multitude of others used less frequently. However, for Twitter, they list only two. This means that most studies on authorship analysis additionally suffer from at least one of two limitations. Either they focus on authorship analysis while using traditional, longer texts, such as articles or blog posts, or they make use of custom datasets (Neal et al., 2017). The latter are sometimes described as a great challenge in the field of authorship analysis, making replication as all well verification of results difficult (Halvani, Winter, and Pflug, 2016). The former implies that past studies not focusing on online short text messages are looking at a fundamentally different research problem compared to Twitter texts. At the same time, most automated authorship analysis research acknowledges the fact that use cases for these tools will consist of attributing micro-blog texts to an author (see, for example, Narayanan et al., 2012; Rocha et al., 2017; Spitters et al., 2016). Consequently, the dataset used is from that platform, as its prevalence makes it especially relevant. Moreover, it reflects the text data, in style and characteristics commonly found for chat messages. Especially in terms of length, it is also similar to text data generated during

studies and experiments within the social sciences.

No. of Characters	Target No. of Authors	avg_instance age	avg_instance gender	avg_tweet age	avg_tweet gender	avg_tweet_per_instance age	avg_tweet_per_instance gender
100	50	160.30	162.94	109.58	111.92	1.46	1.46
	150	160.52	162.37	107.38	112.94	1.49	1.44
	500	160.78	161.63	109.21	111.86	1.47	1.44
	1000	160.07	160.99	109.28	111.84	1.46	1.44
250	50	313.16	315.73	109.05	112.02	2.87	2.82
	150	313.25	315.58	107.43	112.84	2.92	2.80
	500	313.26	314.66	109.09	111.67	2.87	2.82
	1000	313.30	314.34	109.18	111.84	2.87	2.81
500	50	565.60	568.76	109.12	111.87	5.18	5.08
	150	565.89	568.71	107.48	112.84	5.27	5.04
	500	566.13	567.54	109.15	111.70	5.19	5.08
	1000	566.10	567.38	109.17	111.85	5.19	5.07

Table 1: Statistics of the Dataset

Most studies focusing on short-text online media such as Twitter use different data sets. This is due to the fact that the user agreement for the API of this particular platform does not generally give permission to publish a scraped data set online (Theophilo, Pereira, and Rocha, 2019). At this point in time, we know of three public datasets: Twisty (Verhoeven, Daelemans, and Plank, 2016), ISOT (Brocardo, Traore, and Woungang, 2015), and PAN (Stamatatos et al., 2015), a yearly challenge tackling different aspects of authorship analysis. The Twisty dataset includes a multitude of languages, making it unusable for this task as it was shown that language has major impact on the results (Halvani, Winter, and Pflug, 2016). Another problem mentioned before concerns the high number of troll profiles, as well as potential alias accounts (Varol et al., 2017) in an arbitrarily captured data set. This is sometimes referred to as the ground truth problem (Narayanan et al., 2012). As the research question is focused on characteristics of individual people, this is particularly problematic. For this reason, the ISOT dataset, too, is unusable, as neither the problematic of troll profiles nor the problem of double accounts for a single user can be addressed.

To overcome this, a special version of the PAN dataset focusing on profiling celebrities (PAN, 2019) is used. For this dataset it can at least be established that the accounts relate to a real, individual human. Naturally, there may be new limitations, e.g., it may not be guaranteed that celebrities always write their own posts. However, Twitter is more and more considered to be a medium offering the possibility of interacting directly with followers by circumventing the filter, interpretation, and comments of traditional media (thus enabling "authenticity") (Schmidt, 2014). Consequently, the problem of other people messaging instead of the celebrities themselves is considered minor by the authors of the dataset when compared to the problem of having unknown fake profiles.

2.2 Feature Engineering

In order to use text input for machine-learning models, the text has to be transformed into a numerical representation. The chosen representation we call *feature type* here. Within one feature type, there may be many features. For an example of two words, each may be mapped to a number, so there would be two features.
For automated authorship analysis, one may in principle choose from or combine a wide range of possible features for prediction. The natural approach would be to use word-based features. However, this comes with the limitation that rather than finding features predictive of a certain gender or age, it is more likely that the topic is a latent variable driving the result. As our goal is to control most of the information from outside the feature itself, e.g., topic or other context, the selection has to be more nuanced. This brings us to character-based features. Character n-grams, are based on concatenating characters; in the form of 1-grams they equal uni-grams, i.e., single characters. They are maybe one of the most commonly used feature sets within the literature (Rocha et al., 2017). Spitters et al. (2016) find in their exhaustive review of the literature that most studies employ them in one form or another. This is due to the fact that such character n-grams were shown in multiple studies to perform robustly (Kešelj et al., 2003; Stamatatos, 2009; Peng et al., 2003). Some authors like Forstall and Scheirer (2010) link this performance to the fact that n-grams are very closely related to pronunciation. Moreover, due to the fact that the n-gram length may be reduced, many outside influences which introduce context in terms of topics, text type, and even language as a whole may be removed. For example, the cross-domain analysis by Stamatatos (2013) shows that, compared to traditional word-based features, character n-grams outperform them in terms of cross-domain stability. N-grams were also shown to capture many different features such as punctuation or spelling mistakes. Regarding the length of n-grams, it must be noted that for English, those with a length of three and above are shown to capture content again partially, thus becoming topic-dependent (Narayanan et al., 2012; Spitters et al., 2016). Therefore, not only is the type of feature important when controlling for the relevance of topic and content but also the n-grams themselves are crucial. As such, we have a layered approach, controlling for the feature type, while also varying the n-grams employed within one feature type. In the following, we construct a hierarchy ranging from the type of features mostly removed from content to the ones which partially capture content. In between, we can place those features which are still related to style and structure, but which necessitate a certain amount of text. In terms of the actual feature types as well as range of the n-grams, this study mainly follows Custódio and Paraboni (2021) with the numerical features taken from Huang, Su, and Iwaihara (2020). It gives us the following types as input in ascending order, when compared on their context-dependency.

Text Distortion Symbols within text are usually disregarded for the standard approaches of text-based models. However, past research has shown that these features serve as valuable information in the context of authorship analysis (Stamatatos, 2017). For this feature type, all a-z characters are mapped to "*", which only leaves punctuation

and other markers. We call this type of feature *DIST* and apply n-grams $\in [2, 5]$

Character Character n-grams were shown to capture many idiosyncrasies present in text, while yielding a stable performance (Stamatatos, 2013). Moreover, the amount of context present in the n-grams can easily be adjusted by their range (Rocha et al., 2017). Thus, we include them in the range of [2,5], referring to them as *CHAR*.

Unprocessed Text In essence, this feature type is a combination of text distortion and character n-grams. As input, the unprocessed text, including all special characters and punctuation, is transformed into n-grams. The n-gram range is also [2,5]. This feature type we refer to as *ASIS*.

Part-of-Speech Part-of-Speech tags capture linguistic style patters and general information such as grammatical classes, e.g. "noun" or "verb". We tag the text by employing the SpaCy[2] tagger. This feature type is called *POS* and the n-gram range is [1,3].

Language-specific morphological features Within a language, one is also able to find more fine-grained features related to morphology. Such features concern, for example, the gender of a word, tenses and others. To extract these, the tags generated by SpaCy on its *TAG* level are used. Following this, we call this feature type *TAG* and employ the n-gram range [1,3].

Syntactic Dependencies This type of feature captures structural information, e.g., the use of the passive over the active voice. The dependencies are generated using SpaCy's dependency parser. We refer to it as *DEP* and the n-gram range is [1,3] as well.

Lemma Lemmas are essentially word-like features, although the words are reduced to a common, lowercase form. For example, "I'm" would be converted to "i" and "am", while "played" and "playing" would both be mapped onto "play". In such a way, words are captured but not their transformations. We call this feature *LEMMA* and include it with [1,2]-grams.

Words This feature type is created by forming the n-grams directly from the words without any preprocessing besides lowercasing, and removing all characters that are not within the A-Z range. Again, we employ [1,2]-grams, the feature type is called *WORD*.

Numerical Features Huang, Su, and Iwaihara (2020) additionally suggest numerical features describing the content and the form of the tweet. The feature type *NUM* is thus comprised of the following attributes: average tweet length, number of URLs, number of dates and times, number of emoticons, number of emojis, as well as polarity and subjectivity.

[2]https://spacy.io

Preprocessing, Models, and Targets For the preprocessing, we apply the specific ones outlined above to each feature type. In general, emojis and emoticons were always counted as one singular feature and marked by an $< EMOJI >$ or $< EMOTICON >$ token in the beginning and end. Furthermore, we replaced the unicode string by the textual description using the package demoji.[3] For all text-based features, we apply count vectorization and tf-idf scaling. The individual features were kept when they appeared in more than 1% of the samples. For the feature type NUM, we apply scaling and centering. For the model, there is the option of linear and non-linear models. While neuronal nets and transfer-learning models gain huge popularity, for our case of authorship analysis, it turns out that simple linear models regularly outperform the more complex ones (Rocha et al., 2017; Custódio and Paraboni, 2021). Moreover, as we also address the social sciences as well as the law community, interpretability and transparency are key aspects. Thus, we focus here on well-researched models which also enable a mathematically global interpretation, as well as attribution of outcome to individual features of the input. While there is a wide range of models employed, the most common ones are a SVM, a logistic classifier, and Naive Bayes Classifiers (Rocha et al., 2017). We test all three of them and use the overall best-performing one for the evaluation. For the SVM, the analysis is limited to a linear kernel as only this type enables us to interpret the weight matrix directly in terms of feature importance. As target, we selected two author characteristics of high relevance for the social sciences, namely *age* and *gender*.

2.3 Experimental Setup

In order to assess how different combinations of feature types impact the outcome, we use three different approaches to feed them into a classifier.

1. Baseline: Here, the model gets only one feature type (although with varying n-gram ranges). Thus, it enables us to compare the performance of individual feature types against one another.

2. Cumulated: For this approach, we feed the classifier combinations of feature types such that we combine them in an ascending order in terms of context-content.

3. Stacked: Here, too, we use different feature types as input. However, we first make predictions using individual feature types (as in the baseline setup) and then apply a second classifier, a logistic one, on top, using the predictions as input to predict the target again. That, in essence, is an ensemble approach (Dietterich, 2000) and a variation of the successful DynAA model by Custódio and Paraboni (2021).

To assess the stability in performance as well as relevance, we compare the different feature types we introduce a small, controlled variation on the input data. In order to simulate (possible) shifting variations in the patterns of feature use, we vary the number of authors within the dataset. To that end, we construct four subsets from our dataset.

[3] `https://pypi.org/project/demoji/`

Each subset is comprised of a different number of authors (50, 150, 500, 1000). Furthermore, the sets are constructed in such a way that all authors present in the smaller set are also present in all the larger ones. That means the 50 authors from the smallest set are part of all three larger sets as well. We chose that approach in order to increase the number, and thus the potential variation, while at the same time keeping prior information. Moreover, the authors are balanced in gender as well as in age. That is necessary so that the model has no advantage by focusing on one class to the detriment of others. That gives us a cleaner result when analyzing the impact of the individual feature types as well as n-grams.

We conduct the whole experiment three times, varying the input length of the individual text instances each time. As previous research has shown that text length greatly influences the outcome (Custódio and Paraboni, 2021), we construct input instances of different minimum lengths. We do so by concatenating different tweets by the same author together until the minimum length is reached. No n-grams are constructed in such a way that they would contain information from two different tweets. Naturally, when we increase the minimum length, the number of individual training instances declines, as more tweets are needed to form one training instance. As minimum lengths we use 150, 250, and 500 characters. The summary statistics for the dataset may be found in Table 1.

In order to have no spillover of information between evaluation and training, we split the dataset into two subsets, training and testing. All training was done on the training dataset, while all evaluations shown here are done on the hold-out test set. For the stacked approach, we split the test again, this time into a validation and test set. Here, the first layer of the model is trained with the training set, which is the same for all classifiers. The second layer is then trained on the validation set. Finally the evaluations are conducted on the hold-out test set. Due to this setup, all classifiers are trained on exactly the same first-layer input in order to increase comparability.

2.4 Evaluation Measures

We seek to answer the question of stability in predictions as well as stability on the level of features. For the former, we test the predictive power of the classifier for both targets on different inputs and sets varying in the number of authors. Our analysis compares different feature types and their performance against each other. Their difference lies in what they capture, especially in terms of the amount of context. We also analyze by how much their inclusion improves the model's performance. As the evaluation metric of choice, we use the *macro* F1-score. The score is an equally weighted mean of precision[4] and recall[5]. Moreover, on a balanced dataset, the score is nearly equal to the accuracy. The score is bounded between 0 and 1, with 1 being the optimum.
Moreover, we analyze the performance on the author level. That helps us to test whether the models make systematic errors for specific authors. That is indeed important as it

[4] $TruePositives * (PredictedPositives)^{-1}$.
[5] $TruePositives * Positives^{-1}$.

tells us something about how patterns, found to be predictive for target categories, may systematically disadvantage some individuals compared to others. For that part of the analysis, we look at author level accuracy as well as the stability in classifications patterns. For the latter, we evaluate confusion matrices. The results for that analysis may be found in Section 3.

The second, central aspect in this study is that of stability on the feature level. The question we try to answer here is by which degree stays the importance assigned to single features constant, when the classifier input-set used for training is slightly changed. The change introduced here is the increase in the number of authors. What we want to assess is by how much the importance of individual features shifts when such a change occurs. We developed the following approach: First, we extract the weight matrix of the two relevant models. When all input features are scaled to the same range as well as centered, the matrix contains the information about the relative importance of each feature when predicting the outcome. As we are using linear models, these weights are global, i.e., the importance assigned to a feature is the same no matter which individual instance is assessed.

To assess the potential shift in importance, we rank the individual features in terms of weights assigned. In a second step, we then calculate Spearman's ρ in order to assess by how much the importance placed on individual input features shifts when introducing a small variation in the underlying data. The reason why this works is because the ranking of a feature directly reflects the weight the classifier places on it. Within our linear models, this is a direct mapping on its importance to the prediction result. Thus, when this coefficient is 1, the distribution of importance across features is completely identical; if it were -1, it would be completely inverse. Consequently, we say the stability is high for values going towards 1, while values close to 0 imply that there is no recognizable relationship, and thus very high instability. Moreover, we selected an ordinal measure, as it allows for more latitude. While the absolute values of the coefficients might change (and indeed have to when more features are included), their ordering may still be constant. Consequently, their importance when compared to each other can still be stable. That means that any stability found here is to be considered the upper bound.

However, when increasing the number of authors, the underlying feature set might also increase and thus the two matrices do not have the same dimensionality anymore. To tackle that problem, we follow two approaches: The first is to expand the smaller matrix by adding columns of ∞ for the missing features. This ensures, that these will always be assigned the highest possible rank in the smaller matrix (the rankings are sorted in ascending order during comparison). We call this the *extended* Spearman correlation and it enables us to assess the absolute feature importance ranking. The second option is to assess only the features present in both matrices, and therefore to reduce the dimensionality of the larger one. This ensures that we assess relative importance, but ignore that additional features might have great influence on the outcome. That we refer to as the *reduced* Spearman correlation. The result for the analysis of the feature

importance and its stability may be found in Section 4.

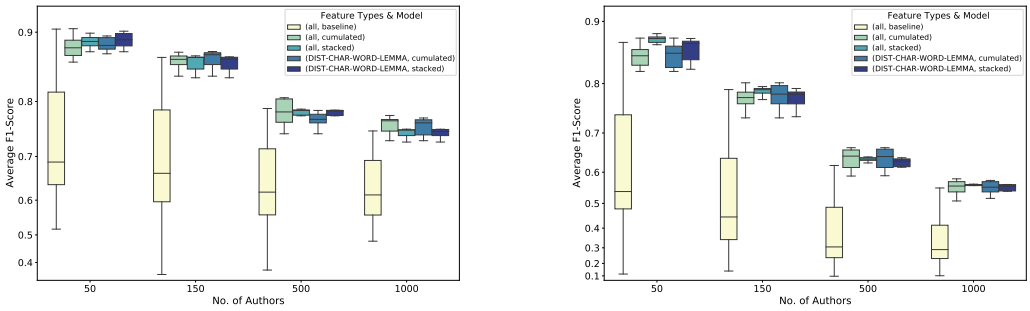
3 Stability of Predictions

The results were computed on HPC system, making use of 72 cores with 256GB of RAM.[6] Of our classifiers, the SVM outperformed the logistic classifier as well as the Naive Bayes Classifier as a first-layer model. Thus, for all results a SVM was used as the first-layer model. For the stacked model, a logistic classifier was used for the second layer to stay close to the setup by Custódio and Paraboni (2021). Overall, per target, the result of the experimental setup is comprised of 1200 SVMs as well stacked models, i.e., 2400 models in total for both targets. To enable a concise analysis, we opted to showcase the results for the text set for which individual inputs have a length of 500 characters. The results for the other text sets may be found in the Appendix.

3.1 Aggregate Overview for Feature-Stability

First, we will look at the aggregated results for our experiments. In Figure 1, we see the F1-scores for the classifiers trained on the dataset with a minimum character length of 500 per instance. As in previous results, the score declines markedly in the number of authors. However, especially the models trained either with different feature types as input (labeled "cumulated") or those trained similarly to the DynAA model by Custódio and Paraboni (2021) (labeled "stacked") perform consistently. Moreover, overall the results are in line with previous top-performing results on the PAN2019 dataset (Wiegmann, Stein, and Potthast, 2019). The baseline models trained on feature sets consisting of singular types have a high variance in performance. That is not surprising, when we look at individual models, each trained on one type of features. As can be seen in **??** and **??**, the performance is on the upper end for the feature types such as CHAR with an $F1_{50}^{CHAR-2}$ up to 0.88/0.82 and on the lower end for feature types such as NUM with an $F1_{50}^{NUM}$ of 0.59/0.29 for the targets gender/age. Consequently, the feature types used, as well as their combinations, not only have a great impact on the outcome, but some features do encode little or next to no information for our classification task. Hence, we exclude those from our further analysis. The same findings apply to the models trained on instances with a minimum length of 100 characters and 250 characters respectively (Figure A.1 and Figure A.3 in the Appendix). Having comparable results in terms of accuracy and f1-score to what is found in the literature for this dataset serves as a basis for our following evaluation. A high performance lends credence to the assumption that our classifier is indeed working well and extracting the relevant information from the input data. In reverse, we may therefore assume that the features used are indeed those holding the relevant information for the respective task. Thus, our approach of extracting the feature importance via the associated weights is sensible.
Figure 2 shows the results for the extended distortion calculated via Spearman's Rho. The comparison is always done between two models, varying in the amount of authors

[6] MPCDF HPC System "Raven".



(a) Results for target *gender*.

(b) Results for target *age*.

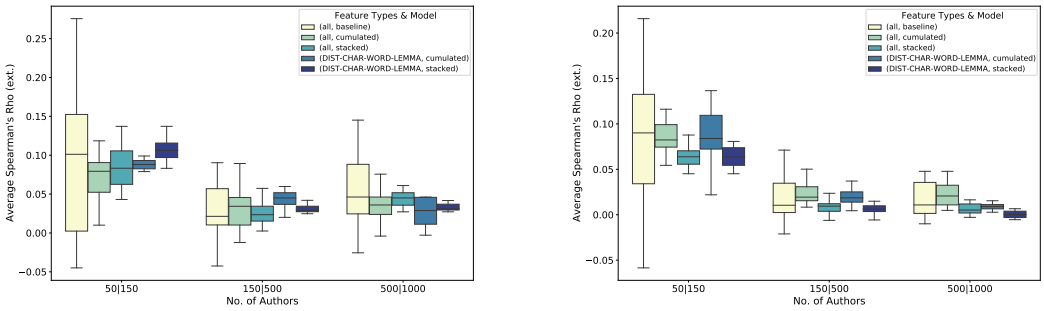
Notes: The figure shows boxplots for the F1-score of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure 1: F1-score input instance length of 500 characters.

the respective model is trained on. In the comparison, model 1 is trained on the lower number of authors (e.g., 50) whereas model 2 is trained on the next-highest number of authors (e.g., 150). Overall, we thus have 3 comparisons in the number of authors.
For target *gender*, on average, and irrespective of the model, we find only little correlation when increasing the number of authors from 50 to 150 ($0.05 < \rho < 0.20$). For the others, when increasing the number of authors, correlation goes down to a maximum of 0.05. At the same time, we show that the performance in terms of f1-score remains relatively stable. This implies that the stability in performance comes at the expense of the stability in feature importance. When increasing the number of authors, different features are therefore predictive in terms of the target. These findings do not generally imply that the previous features lose their importance. The do however mean that, when ordering all features in terms of the associated importance, the ordering changes fundamentally. That fact is reflected in Spearman's Rho. Regardless by how much we increase the number of authors, on average the correlation lies between 5% and 15% for all models. That is interesting, as the number of features additionally available when increasing the number of authors is never above 30%. Thus, it cannot be that mainly those additional features are the ones being used for predictions, as the correlation coefficient would then still be higher than what we find. Rather, it seems to be the case that the model weighs the previous and new features in such a way that the new ordering imposed differs completely from the previous one.

Similar results are found for the target *age*. However, here the decline in correlation is even more unidirectional when increasing the number of authors (increase in authors yields a decline in correlation).

The only outlier to these results is the baseline model trained on individual feature types. While on average the correlation is the same as for the models on combinations of feature types, the outliers show a very high positive correlation (up to $\rho^{NUM}_{150|50} : 0.42$,



(a) Results for target *gender*.

(b) Results for target *age*.

Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure 2: Extended Spearman correlation input instance length of 500 characters.

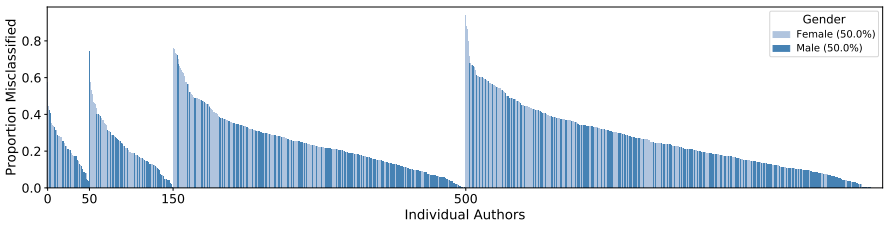
see **??**). While that may look like it stands in opposition to the other results, their low predictive power helps to explain this phenomenon. **??** shows that the F1-score for the prediction of *age* is only at $F1_{150}^{NUM}$: 0.25 with the random guess benchmark being 0.2. Thus, while the feature importance remains stable for some features, their predictive power is negligible. Thus, the importance assigned to these features may simply be random noise without any signal. As such, when looking at the average feature stability over all feature type sets, the conclusion is that the features are, on average, not stable and the distortion of importance when increasing the number of authors in a dataset is already high for a low number of authors (from 50 to 150). These findings hold regardless of the character length of the input text, as Figure A.2 and Figure A.4 in the Appendix show.

3.2 Author-Level Analysis

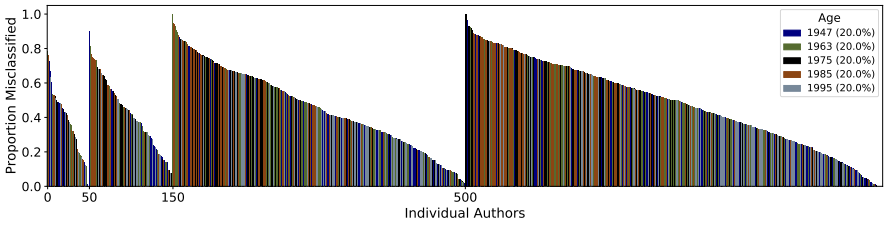
For the author-level analysis, we look at the results gained by feeding the classifier the full set of feature types in a cumulated way. That choice assures that our results are predicted making use of the full information. Furthermore, cumulating the input yields the best results overall (compare Appendix A2). In that way, the analysis is done using the best possible set. However, the findings hold for any of the feature-type approach combinations.

Figure 3a shows the error at the author level when predicting gender. We see, that overall we have very few authors for which the accuracy is lower than the random-guess accuracy (0.5), i.e, for which our prediction error is higher than 0.5. We see that the relative number of authors for which the classifier performs below the random-guess threshold stays stable, when compared to the number of authors in the set. Consequently, the relative overall-classification performance at the author level stays stable,

even when increasing the the number of authors in the set. However, there is a small but stable proportion of authors for which the classifier is *systematically unable* to predict the target correctly. When looking at the distribution of the errors across genders, we



(a) Author-level errors for target *gender*.



(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure 3: Author-level results for the full feature set with an input instance length of 500 characters.

find that a higher number of those authors for whom our classifier makes systematic errors seems to be female. On the high level, we find that the performance remains relatively stable when we increase the number of authors as depicted in Figure 4a. However, as soon as we reach the two upper-most brackets of authors, we see that the result for male authors remains relatively stable, while the outcome for female authors declines markedly from $acc_{150}^{Female} : 0.82$ to $acc_{1000}^{Female} : 0.69$. For the same sets, the mostly stable accuracy for males declines only from $acc_{150}^{Male} : 0.83$ to $acc_{1000}^{Male} : 0.80$. Looking closer, we can see that that drop for females happens when the number of authors increases from 150 to 500. The implication is thus that we add female authors who are systematically difficult to classify. Bringing this together with our observations from before, we include only those authors in Figure 4b for whom the classifier performs worse than the random-guess threshold. Here, we find some support for our previous assumption. When looking at the errors we make for female authors, we see that, starting from 500 authors onward, the number of those below the random-guess threshold jumps up.

While that seems to point to the fact that apparently female authors are difficult to classify, the true reason may be slightly more nuanced. When looking at the male authors

in Figure 4b, we see that, while obviously low in absolute numbers, the pattern is inverse for the datsets comprised of 50 and 150 authors. The underlying driver, however, does not seem to be the gender *per se*, but rather the lack of stability in regards to feature importance. Per design, we limited the amount of features for each feature type to those appearing at least in 1% of all training instances. Hence, the number of features for word-based feature types increases only *sublinearly* relative to the number of authors. Thus, the amount of author-individual fitting the classifier is able to achieve declines. Indeed, that is the very essence of reducing overfitting. However, as shown in Figure 2a, the stability of feature importance is low. Taken together, that simply means that systematic patterns within a limited number of features for authors of the same gender decline when the number of authors increases, i.e., the patterns seem to be merely correlational – they start to break down or become unstable and more complex. The underlying reason is that additional authors introduce new features, while using the old features in a different way. As the classifier is only able to estimate one weight per feature and the number of additional features is limited, the ability to represent all the necessary information declines.

(a) All authors (row-wise normalization).

(b) Below random guess accuracy (matrix-wise normalization).

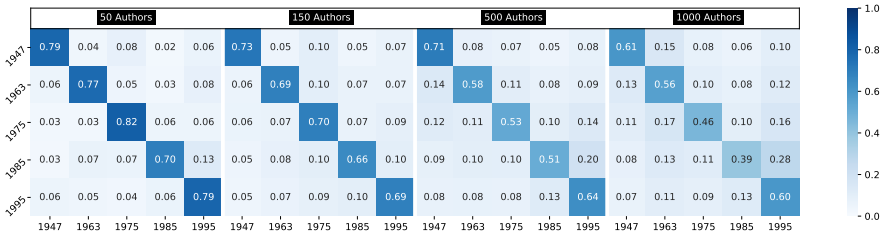
Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 500 Characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure 4: Confusion matrices for target *gender*.

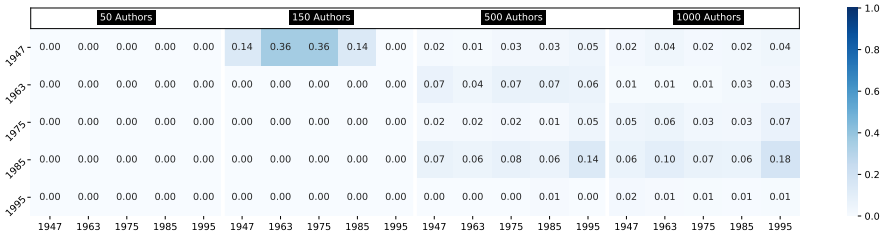
For the target *age*, we also find only comparatively few authors with an average accuracy below the random-guess threshold of 0.2 (see Figure 3b). When comparing it to the results of the target *gender*, it becomes clear that the number of those below the threshold jumps up significantly when the number of authors increases from 150

to 500. Moreover, it seems to be the case that authors of the intermediate age brackets (*1975* and *1985*) seem to be more difficult to classify. Overall, the patterns seem to be less pronounced when compared to the ones found for *gender*. Looking at the category-wise analysis presented in Figure 5a, we find that, overall there are only few pronounced patterns of confusion. As already suggested by Figure 3b, only the intermediate age brackets have a systematic pattern. Especially for the datasets consisting of 500 and 100 authors, the confusion between the true age bracket *1985* and the youngest age bracket *1995* is pronounced. Here, only 38% of the instances are classified correctly as *1985*, while 28% are confused as *1995*. While here the interpretation might be that the distinction between the youngest authors might be difficult, the results of *1975*, the intermediate category, make it more difficult. We see that the confusion with the category *1963* as well as *1995* is of similar size. It might be that, instead of predicting only *age*, the classifier picks up on a proxy in the way people express themselves. While certainly dependent on age in terms of punctuation for older authors (Flekova, Preoţiuc-Pietro, and Ungar, 2016) as well as on stability of language use in younger authors (De Jonge and Kemp, 2012), the way of expression also depends on the groups to which we belong (Chan and Fyshe, 2018). Consequently, some of those authors confused might simply be part of peer groups where the mode of expression is reflective of younger age brackets. As a consequence, they get misclassified. These relatively distinctive patterns for the oldest and youngest authors are also most likely the reason why the prediction accuracy for those is markedly high, even for the set with the highest number of authors.

When looking only at those authors below the random-guess threshold, as depicted in Figure 5b, we find that there are only two age brackets for which we have a systematic and pronounced confusion. For the set with 150 authors, this is the age bracket *1947*, which is most often confused with the two adjacent age brackets *1963* and *1975*. For the set with 1000 authors, the age bracket *1985* is mostly confused with belonging either to the youngest or the oldest age bracket. As that bracket is also overall the most confused one, it stands to reason that the variance in expression is the highest. Consequently, there is no pronounced pattern in the features on which the classifier is able to pick up.



(a) All authors (row-wise normalization).



(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 500 Characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure 5: Confusion matrices for target *age*.

4 Robustness of Feature-Importance

In order to assess how the aggregate results from the previous section hold, when looking at individual feature types as well as different classifier approaches, this section presents a more fine-grained insight. In this section, we focus on the input in terms of feature types and n-grams. We analyze these aspects with a special look on the stability of the feature importance.

4.1 Baseline

First, we now take a look at the results in feature set 1, i.e., the results gained when using each feature type individually to predict the target. **??** and **??** show the experimental results for targets *gender* and *age*, respectively. Each row shows the result for a feature type and the corresponding n-grams. The feature types themselves are sorted in an ascending order such that feature types in lower rows capture more context. The type CHAR (characters), for example, captures, in principle, less contextual information (such as topic or structural information) compared to, for example, word-based n-grams (Rocha et al., 2017). Naturally, when the n-gram window is increased, e.g., for character-based features from 2 to 4, the character n-grams also start to capture contextual information. Consequently, the n-gram combinations within the individual feature types are also sorted in an ascending fashion.

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST	2	0.6322	–	0.3481	0.0858	0.3680	0.1528	0.5403	0.2834
	2-3	0.6477	–	0.3788	0.1535	0.4365	0.1350	0.5173	0.1451
	2-3-4	0.6624	–	0.6056	0.1185	0.5462	0.0893	0.5432	0.0425
	2-3-4-5	0.6789	–	0.6402	0.0904	0.5385	0.0491	0.4858	0.0457
CHAR	2	0.8080	–	0.7580	0.2757	0.6882	0.0903	0.6719	0.0932
	2-3	0.8690	–	0.8151	0.1739	0.7638	0.0277	0.7260	0.0251
	2-3-4	0.8702	–	0.8411	0.1328	0.7746	0.0174	0.7362	0.0364
	2-3-4-5	0.8847	–	0.8094	0.0999	0.7756	0.0131	0.7411	0.0215
ASIS	2	0.8357	–	0.7995	0.2720	0.7184	0.0573	0.6916	0.0588
	2-3	0.8868	–	0.8436	0.1492	0.7750	0.0155	0.7341	0.0213
	2-3-4	0.9004	–	0.8662	0.1229	0.7885	0.0114	0.7493	0.0150
	2-3-4-5	0.9040	–	0.8158	0.1025	0.7882	0.0207	0.7488	0.0153
POS	1	0.5907	–	0.5847	0.2456	0.5775	0.3123	0.5728	0.1982
	1-2	0.6303	–	0.6403	0.0272	0.6061	-0.1515	0.5985	0.2639
	1-2-3	0.6583	–	0.6579	-0.0449	0.6174	-0.0222	0.6127	0.0582
TAG	1	0.6402	–	0.5700	-0.0086	0.5892	-0.1640	0.5771	0.0043
	1-2	0.6982	–	0.6704	0.1546	0.6265	0.0233	0.6140	0.1140
	1-2-3	0.6980	–	0.6828	0.0514	0.6230	0.0017	0.6278	0.1243
DEP	1	0.6195	–	0.5997	0.0103	0.5652	-0.0425	0.5655	0.0244
	1-2	0.6590	–	0.6277	0.1047	0.5963	-0.1028	0.5926	-0.0255
	1-2-3	0.6704	–	0.6550	-0.0349	0.6008	0.0236	0.6060	0.0737
LEMMA	1	0.7786	–	0.7679	-0.0374	0.7096	0.0221	0.6869	0.0406
	1-2	0.8007	–	0.7816	-0.0143	0.7144	0.0099	0.6925	0.0467
WORD	1	0.7588	–	0.7408	-0.0304	0.6940	0.0136	0.6782	0.0480
	1-2	0.7653	–	0.7535	-0.0003	0.6981	0.0560	0.6836	0.0304
NUM	1	0.5912	–	0.5635	0.2500	0.5229	0.0833	0.4791	0.0667

Table 2: F1-scores & stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

Looking at results for the target *gender*, we first find that the most predictive feature types are those most closely related to the words of the text, but not necessarily the structure. That can be seen from the fact that structure-capturing feature types, such as POS, TAG, and DEP, show low predictive power no matter what the number of authors within the subset is. Moreover, CHAR-2-grams already perform well ($F1_{50}^{CHAR-2} : 0.80$) on the small dataset comprised of 50 authors. However, when we increase the number of authors, the performance declines markedly ($F1_{1000}^{CHAR-5} : 0.67$), especially when compared to CHAR-(2,5)-grams ($F1_{1000}^{CHAR-5} : 0.741$), which are close to the top performance ($F1_{1000}^{ASIS-5} : 0.748$). The same pattern, although on a lower overall performance level, is visible for the text distortion features DIST capturing punctuation and other stylistic markers. For lower n-gram sizes, the performance is only negligibly above or below the random guess threshold, while for higher n-grams the performance is higher ($F1_{50}^{DIST-5} : 0.67$), but then decreases again in the number of authors. Consequently, the results show that there seems little cause to think that there are patterns in the style of authors related to gender. On the other side, CHAR-2-grams have a reliable performance ($0.67 < F1^{CHAR-2} < 0.80$); increasing the n-gram window only by 1 increases performance even more. Consequently, it can be assumed that the there seems to be a discernible pattern related to gender within the character combinations used. The underlying assumption would be that certain topics might be reflected by the use of similar words or that certain synonyms are preferred by one group over the other. We can compare this with the result for the WORD-grams. Here we see that, while we the

performance is high, it is still worse when compared to CHAR-(2,5)-grams. The latter would also capture words up to five characters long. However, if that overlap is the sole driver of performance, then WORD-grams should not be outperformed. As such, we can conclude that there is a discernible pattern related to gender in low-context CHAR-n-grams. In terms of the stability of the feature importance, the results are sobering. As in the aggregate before, the correlation tends towards zero when increasing the number of authors. Besides that, in some cases the correlation even flips signs. That implies features which were useful for predicting group A before are now either relevant for neither group or relevant for predicting group B (see, for example, **??**, $\rho_{500}^{POS-2} : -0.15$). While mostly small, all correlation coefficients are significant at the 1%-level.

When looking at age, the results shown in **??** reflect the overall findings for gender. Text distortion alone, such as punctuation reflected in the features of type DIST, does hold some, but not the majority of the information relevant to the prediction of age. That is evident from the stark decline towards random-guess accuracy, especially for low-level n-grams.

	Target	Age						
	Min. No. of Characters	500						
	No. of Authors	50		150		500	1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST	2	0.4797	–	0.2249	0.1872	0.2808	0.0083	0.2196	0.0192
	2-3	0.4809	–	0.3711	0.1142	0.2633	0.0275	0.2736	0.0478
	2-3-4	0.5294	–	0.4482	0.0268	0.2454	0.0502	0.2921	0.0099
	2-3-4-5	0.5384	–	0.4088	0.0219	0.2401	0.0371	0.2823	0.0154
CHAR	2	0.7257	–	0.6250	0.2159	0.4761	0.0197	0.4033	0.0058
	2-3	0.8024	–	0.7319	0.1388	0.5696	0.0014	0.5070	-0.0014
	2-3-4	0.8274	–	0.7579	0.1140	0.5938	0.0124	0.5346	0.0123
	2-3-4-5	0.8291	–	0.7492	0.0916	0.6114	0.0075	0.5407	0.0068
ASIS	2	0.7811	–	0.6776	0.2031	0.5142	0.0049	0.4426	-0.0024
	2-3	0.8546	–	0.7646	0.1386	0.6048	0.0027	0.5346	0.0025
	2-3-4	0.8687	–	0.7836	0.1080	0.6189	0.0032	0.5512	0.0069
	2-3-4-5	0.8542	–	0.7890	0.0925	0.6174	0.0023	0.5523	0.0011
POS	1	0.3529	–	0.2723	0.1807	0.0964	0.1046	0.1010	0.1923
	1-2	0.4794	–	0.3512	0.0762	0.2269	-0.0211	0.2283	0.0116
	1-2-3	0.5248	–	0.4124	0.0611	0.2767	0.0236	0.2491	0.0455
TAG	1	0.3934	–	0.3244	0.0886	0.1338	0.0711	0.1690	-0.0680
	1-2	0.5432	–	0.4479	0.0178	0.2755	-0.0109	0.2670	0.0441
	1-2-3	0.5878	–	0.4859	0.0825	0.3811	0.0386	0.3110	0.0474
DEP	1	0.3560	–	0.2418	0.0971	0.1000	0.0466	0.2387	0.0411
	1-2	0.5057	–	0.3820	0.0559	0.2523	0.0155	0.2358	-0.0100
	1-2-3	0.5369	–	0.4029	0.0751	0.3292	-0.0095	0.2873	-0.0020
LEMMA	1	0.6668	–	0.5763	-0.0219	0.4091	0.0128	0.3339	0.0078
	1-2	0.6920	–	0.5881	0.0010	0.4424	0.0085	0.3781	-0.0058
WORD	1	0.6282	–	0.5416	-0.0585	0.3867	0.0428	0.3439	0.0151
	1-2	0.6437	–	0.5482	-0.0112	0.4053	0.0013	0.3294	0.0118
NUM	1	0.2995	–	0.2542	0.4267	0.2386	-0.0933	0.2194	0.3033

Table 3: F1-scores & stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

When combined with CHAR, then especially higher-order n-grams (which is reflected in the feature type ASIS) hold the most information about an author's age. That seems to be in line with findings linking age to a higher adherence to linguistic rules, even in an online environment (De Jonge and Kemp, 2012; Hovy and Søgaard, 2015). However, the content of the tweets also seems to set the age categories apart, as illustrated by the fact

that TAG alone has a relatively high predictive power even for the dataset comprised of 1000 authors (F1: 0.31). The same holds true for LEMMA, implying that age groups are also set apart by the use of one set of words over another. Here again, the feature stability is low, with a $\rho \in [0, 0.05]$.

Thus, we can conclude that, for singular feature sets, the model is able to extract information from the features, especially those with higher context, as evident from the increase in predictive performance when the n-gram range is increased. However, the relevant information is not stable in the number of authors, which means that additional authors introduce a wider variation, that needs to be separated differently than the smaller range. As the number of characters is limited overall (and thus the number of features in the lower n-gram range), that automatically implies that the content and therefore the relevant features change. That seems to lead to an overall change in the way individual features are predictive. Thus, the rank correlation is low.

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST_CHAR	2	0.8230	–	0.7868	0.0911	0.7058	0.0598	0.6769	0.0998
	2-3	0.8802	–	0.8392	0.0903	0.7696	0.0331	0.7311	0.0178
	2-3-4	0.8712	–	0.8494	0.0990	0.7852	0.0445	0.7392	0.0454
	2-3-4-5	0.8951	–	0.8583	0.0829	0.7650	0.0455	0.7429	0.0461
DIST_CHAR_ASIS	2	0.8882	–	0.8623	0.0815	0.7441	0.0306	0.7494	0.0118
	2-3	0.8850	–	0.8725	0.0931	0.8036	0.0462	0.7602	0.0295
	2-3-4	0.8764	–	0.8733	0.0789	0.7641	0.0380	0.7675	0.0026
	2-3-4-5	0.8942	–	0.8729	0.0829	0.7626	0.0459	0.7673	0.0281
DIST_CHAR_ASIS_LEMMA	1	0.8925	–	0.8697	0.0935	0.7765	0.0201	0.7706	0.0152
	1-2	0.8779	–	0.8744	0.0829	0.7740	0.0428	0.7703	0.0084
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8797	–	0.8712	0.0632	0.8034	0.0388	0.7688	-0.0029
	1-2	0.8943	–	0.8702	0.0526	0.7702	0.0268	0.7723	0.0094

Table 4: F1-scores & stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

4.2 Cumulated

The previous analysis has shown that some feature types, such as POS, TAG, DEP, and NUM do hold little relevant information. We therefore constructed a subset of feature types which excludes them. That subset includes the types ASIS, CHAR, LEMMA, and WORD. Compared to the previous analysis, we now give the model the possibility to include additional information, i.e., information stemming from different feature types, in the model. As shown by the results in **??** and **??**, the additional information yields to overall increase in performance. How much additional information leads to an improvement differs by target. For *age*, we find that some additional contextual information increases the outcome. However, when the contextual information becomes larger, e.g., by including LEMMA and WORD, the result does not improve anymore. The result is consistent across a different number of authors. Consequently, the information for *age* seems to be less reliant on contextual information and content. Already single-word content and context as captured by CHAR-(2,5) and ASIS-(2,5), is enough for a high

prediction score. When we compare the outcome for *gender* with the the results in **??**, we see that using a cumulated input improves the results overall. It is especially important to note that, when faced with a high number of individual authors, increasing the context by using additional feature types such as LEMMA or WORD in addition to high-level n-grams increases performance. When taken together, our findings show that context and underlying data structure is an important driver behind the predictions of a model, as shown by the fact that the relevant features in terms of predictiveness change. At the same time, we show that the weight placed on individual features (and thus individual inputs reflecting certain contexts) is not stable. That is evident by the

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST_CHAR	2	0.7271	–	0.6349	0.1366	0.5123	0.0244	0.4434	0.0048
	2-3	0.8214	–	0.7331	0.1364	0.5898	0.0286	0.5177	0.0117
	2-3-4	0.8343	–	0.7627	0.1078	0.6114	0.0183	0.5390	0.0110
	2-3-4-5	0.8250	–	0.7595	0.0934	0.6140	0.0179	0.5397	0.0084
DIST_CHAR_ASIS	2	0.8518	–	0.7727	0.0901	0.6150	0.0189	0.5515	0.0071
	2-3	0.8726	–	0.7954	0.0815	0.6629	0.0157	0.5687	0.0102
	2-3-4	0.8753	–	0.8014	0.0848	0.6610	0.0124	0.5553	0.0053
	2-3-4-5	0.8671	–	0.8006	0.0831	0.6376	0.0198	0.5731	0.0067
DIST_CHAR_ASIS_LEMMA	1	0.8608	–	0.7966	0.0753	0.6478	0.0143	0.5717	0.0077
	1-2	0.8615	–	0.7963	0.0736	0.6629	0.0197	0.5544	0.0028
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8515	–	0.7883	0.0680	0.6558	0.0044	0.5736	0.0056
	1-2	0.8293	–	0.7727	0.0682	0.6652	0.0116	0.5758	0.0095

Table 5: F1-scores & stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

correlation scores across different author sets. The scores are $\rho^{NUM}_{150|50} : 0.42$ at the highest for target gender and $\rho^{DEP-2}_{150|50} : -0.1$ at their lowest. Consequently, while the predictive accuracy is high, the model seems to rely on correlational patterns which are not only not invariant, but also quite unstable when the dataset is changed only slightly.[7]

5 Discussion

Overall, we find that the classifier makes systematic errors at the author level. For the target *gender*, one group of authors (female) seems to be difficult to classify in general. The underlying drivers seem to be that, for the shown dataset, the group as such has a very heterogeneous pattern in the features. In other words, the second group (male) seems to be simpler to classify. That, however, may also be driven by the fact that the context in which they are active is more homogeneous than for female authors. For *age*, we find slightly more stable results, as most age brackets exhibit clear patterns that make the individual age brackets distinguishable. In terms of stability, however, the results are more mixed. While the features mainly driving the prediction are not context-reliant *per-se*, increasing the available context does increase performance markedly. This is es-

[7] All analyses –stacked as well as cumulated– were also done on the full number of feature types, as well as the different numbers of authors and the different input instance lengths. The results may be found in the Appendix A2.

pecially evident from the fact that a wide n-gram window for character features yields the greatest relative increase in performance, outperforming even those models for which additional context information is made available by including additional feature types known to capture context. That in itself is not directly surprising and not necessarily cause for any raised eyebrows. However, it is important to note that, for a low number of authors, about 10% of the prediction performance stems from an increase in context (e.g. for features CHAR and ASIS when predicting gender, see **??**). When looking at the dataset with the largest number of authors, additional information is also what makes the model perform slightly better than the random-guess threshold and pushes it into the performance ranges found within the literature for comparable data (Wiegmann, Stein, and Potthast, 2019). Moreover, building a model on top of a composite of feature types (or stacking it on top; see **??** and **??**) is when we see additional performance increases, especially for longer input texts. Thus, giving additional context on top of non-context feature types yields a better decision boundary for the classifier.

These results at first seem like technical details. However, in practice they show that the context the model is trained on and in (as simulated by varying the number of authors) largely carries over into its predictive performance. That means models trained within one context may not simply be used in another one. That is intuitive. What we show here, however, is that even by staying within one group of individuals (creators) and within one domain (Twitter), an increase in the number of possible targets changes the relevant features, and it also significantly changes the information as well as the context encoded within . That becomes evident from the fact that the stability in the relevance of features simply does not exist. For social sciences, these findings are relevant on two fronts. First, the models using the features presented here are indeed well-suited to find a pattern connecting their use to the prediction target. However, that pattern is unstable, changing with the number of authors or features available. Consequently, it hints at the fact that these patterns are merely correlations exploited by the model. Such correlations are difficult to rely upon, as their patterns – as shown by increasing the number of authors – may change at any time. Thus, this calls for a careful assessment of the validity when employing pre-trained models within the field, especially when the prediction outcome is used as input for further models or for further analysis. In other words, a change in behavior by individuals – either over time or by choice – will render the learned context irrelevant. Thus, the environment during training must be carefully compared to the one in which the model is used. In general, the findings thus paint a bleak picture for the social sciences. Our results show that there are authors of certain groups for which one has to expect above-average errors and systematic patterns of misclassification. Taken together with the apparent lack of stability in the predictiveness of features when the dataset changes slightly means that, even when these patterns are assessed during training, the researcher has little chance to assess how they will affect a the result during the time of use. The differences between training data and test data might be difficult to pinpoint. Thus, for cases where it is not clear by how much training context and use-context differ, a social scientist should be very careful in simply adopting pre-trained models as the size of the introduced error is unknown.

Another finding is of a normative nature and tied to the wider debate of transparency and proportionality, and thus affects in particular the field of law. As law enforcement is faced with the problem of combing through a large amount of online content, searching and assessing such content by hand is untenable. Thus, already today algorithms are employed by law enforcement. However, especially in such environments, it must be clear how much of the findings by an algorithm is relies merely on correlations and especially how stable these correlations are in different environments. That does not even include the fact that there might be some groups of individuals for whom the classifier makes systematic mistakes. Only then do law enforcement, the defendant, and also the courts have the possibility to assess the validity of an result *before* acting upon it. After all, how valid is a result identifying traits of a suspect when the features are context-reliant to such a high degree that changing the use of some emojis or some words would alter the result completely? How robust is a result, when the features driving the result change with the number of authors an individual is compared to? What is even more problematic in real-world terms is that the instances used for training and those used for during actual application are separated from each other by time. Thus, a real culprit could evade being identified simply because the context changes, while innocents could be systematically misidentified as culprits. Thus, as the stability in feature importance is already lacking for the relatively small changes introduced here, we should ask ourselves what requirements an algorithm should fulfill before it is being used within the law enforcement context. Optimally, we would ask for causal relationships between input and output. However, that might not be possible. The second-best would then be to have transparency for the model and a some-what stable relationship between input and output. The former assures that users, i.e., the state, as well as affected individuals are able to assess the inner workings of a model. That would enable an individual to judge whether the prediction pertaining to them might be part of a systematic error.

A reasonably stable relationship between input and output, i.e., a stable feature importance, guarantees that while there might be systematic errors, the affected groups stay at least constant, although the dateset for the predictions may vary slightly compared to the training dataset, e.g., by number of authors or point-in-time. The alternative would be, of course, to specify a "half-life" before a model has to be re-trained and re-assessed.

As the findings of this study point towards such an unstable relationship, we argue that the features used in tasks related to authorship profiling and authorship attribution need much more research. Moreover, models should be assessed with a measure for defining the boundaries of their stability. The result of that measure has to be affixed to the model so users may be able to infer its usability. Otherwise, establishing a scientifically valid link – going beyond merely showing that the model yields good correlational predictions on some datasets – might be impossible.

References

Azarbonyad, Hosein et al. (2015). "Time-aware authorship attribution for short text streams". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 727–730.

Bail, Christopher Andrew (2016). "Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media". In: *Proceedings of the National Academy of Sciences* 113.42, pp. 11823–11828.

Barberá, Pablo and Gonzalo Rivero (2015). "Understanding the political representativeness of Twitter users". In: *Social Science Computer Review* 33.6, pp. 712–729.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian (2017). "Gender stereotypes about intellectual ability emerge early and influence children's interests". In: *Science* 355.6323, pp. 389–391.

Boenninghoff, Benedikt et al. (2019). "Explainable authorship verification in social media via attention-based similarity learning". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 36–45.

Booth, Alison L and Patrick Nolen (2012). "Gender differences in risk behaviour: does nurture matter?" In: *The Economic Journal* 122.558, F56–F78.

Brocardo, Marcelo Luiz, Issa Traore, and Isaac Woungang (2015). "Authorship verification of e-mail and tweet messages applied for continuous authentication". In: *Journal of Computer and System Sciences* 81.8, pp. 1429–1440.

Burley, Timothy et al. (2020). "NLP Workflows for Computational Social Science: Understanding Triggers of State-Led Mass Killings". In: *Practice and Experience in Advanced Research Computing*, pp. 152–159.

Chan, Sophia and Alona Fyshe (2018). "Social and Emotional Correlates of Capitalization on Twitter". In: *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 10–15.

Charness, Gary and Uri Gneezy (2012). "Strong evidence for gender differences in risk taking". In: *Journal of Economic Behavior & Organization* 83.1, pp. 50–58.

Chaski, Carole E (2001). "Empirical evaluations of language-based author identification techniques". In: *Forensic Linguistics* 8, pp. 1–65.

— (2012). "Best practices and admissibility of forensic author identification". In: *JL & Pol'y* 21, p. 333.

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson (2014). "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data". In: *Journal of communication* 64.2, pp. 317–332.

Costa-jussà, Marta R (2019). "An analysis of gender bias studies in natural language processing". In: *Nature Machine Intelligence* 1.11, pp. 495–496.

Custódio, José Eleandro and Ivandré Paraboni (2021). "Stacked authorship attribution of digital texts". In: *Expert Systems with Applications* 176, p. 114866.

De Jonge, Sarah and Nenagh Kemp (2012). "Text-message abbreviations and language skills in high school and university students". In: *Journal of Research in Reading* 35.1, pp. 49–68.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, pp. 1–15.

Djuric, Nemanja et al. (2015). "Hate speech detection with comment embeddings". In: *Proceedings of the 24th international conference on world wide web*. ACM, pp. 29–30.

Flekova, Lucie, Daniel Preoţiuc-Pietro, and Lyle Ungar (2016). "Exploring stylistic variation with age and income on twitter". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 313–319.

Forstall, Christopher and Walter Scheirer (2010). "Features from frequency: Authorship and stylistic analysis using repetitive sound". In: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*. Vol. 1. 2.

Gneezy, Uri and Aldo Rustichini (2004). "Gender and competition at a young age". In: *American Economic Review* 94.2, pp. 377–381.

Halvani, Oren, Christian Winter, and Anika Pflug (2016). "Authorship verification for different languages, genres and topics". In: *Digital Investigation* 16, S33–S43.

Hovy, Dirk 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)*, pp. 483–488.

Huang, Wenjing, Rui Su, and Mizuho Iwaihara (2020). "Contribution of improved character embedding and latent posting styles to authorship attribution of short texts". In: *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, pp. 261–269.

InternetLiveStats (2019). *Twitter Usage Statistics*. Last Accessed: 20-08-2019.

Kešelj, Vlado et al. (2003). "N-gram-based author profiles for authorship attribution". In: *Proceedings of the conference pacific association for computational linguistics, PACLING*. Vol. 3. sn, pp. 255–264.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon (2011). "Authorship attribution in the wild". In: *Language Resources and Evaluation* 45.1, pp. 83–94.

Lahey, Benjamin B et al. (2000). "Age and gender differences in oppositional behavior and conduct problems: a cross-sectional household study of middle childhood and adolescence." In: *Journal of abnormal psychology* 109.3, p. 488.

Laub, Zachary (2019). *Hate Speech on Social Media: Global Comparisons*. Last Accessed: 20-08-2019.

Narayanan, Arvind et al. (2012). "On the feasibility of internet-scale author identification". In: *2012 IEEE Symposium on Security and Privacy*. IEEE, pp. 300–314.

Neal, Tempestt et al. (2017). "Surveying Stylometry Techniques and Applications". In: *ACM Computing Surveys* 50.6, pp. 1–36.

PAN (2019). *Celebrity Profiling*. Last Accessed: 20-08-2019.

Pavlick, Ellie et al. (2016). "The Gun Violence Database: A new task and data set for NLP". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1018–1024.

Peng, Fuchun et al. (2003). "Language independent authorship attribution using character level language models". In: *Proceedings of the tenth conference on European*

chapter of the Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 267–274.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why should i trust you?: Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp. 1135–1144.

Rocha, Anderson et al. (2017). "Authorship Attribution for Social Media Forensics". In: *IEEE Transactions on Information Forensics and Security* 12.1, pp. 5–33.

Sage, Manuel et al. (2020). "Investigating the Influence of Selected Linguistic Features on Authorship Attribution using German News Articles." In: *SwissText/KONVENS*.

Samek, Wojciech et al. (2019). *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

Sanchez-Perez, Miguel A et al. (2017). "Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 145–151.

Schmidt, Jan-Hinrik (2014). "Twitter and the rise of personal publics". In: *Twitter and society*, pp. 3–14.

Spitters, Martijn et al. (2016). "Authorship Analysis on Dark Marketplace Forums". In: *Proceedings - 2015 European Intelligence and Security Informatics Conference, EISIC 2015*, pp. 1–8.

Stamatatos, Efstathios (2009). "A survey of modern authorship attribution methods". In: *Journal of the American Society for information Science and Technology* 60.3, pp. 538–556.

— (2013). "On the robustness of authorship attribution based on character n-gram features". In: *Journal of Law and Policy* 21.2, pp. 421–439.

— (2017). "Authorship attribution using text distortion". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1138–1149.

Stamatatos, Efstathios et al. (2015). "Overview of the pan/clef 2015 evaluation lab". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 518–538.

Stuttgarter Nachrichten (2017). *Bitte, Facebook, Hilf uns*. URL: `https://www.stuttgarter-nachrichten.de/inhalt.polizei-als-bittsteller-bei-straftaten-bitte-facebook-hilf-uns.a3a743a3-d41e-479c-aba7-d5bc4da25286.html` (visited on 07/22/2021).

Sutter, Matthias and Daniela Glätzle-Rützler (2014). "Gender differences in the willingness to compete emerge early in life and persist". In: *Management Science* 61.10, pp. 2339–2354.

Theophilo, Antonio, Luis A. M. Pereira, and Anderson Rocha (2019). "A Needle in a Haystack? Harnessing Onomatopoeia and User-specific Stylometrics for Authorship Attribution of Micro-messages". In: pp. 2692–2696.

Varol, Onur et al. (2017). "Online human-bot interactions: Detection, estimation, and characterization". In: *Eleventh international AAAI conference on web and social media*.

Verhoeven, Ben, Walter Daelemans, and Barbara Plank (2016). "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1632–1637.

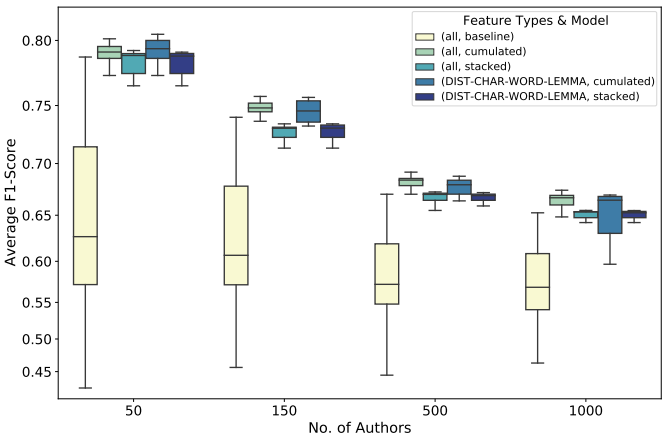
Wiegmann, Matti, Benno Stein, and Martin Potthast (2019). "Overview of the Celebrity Profiling Task at PAN 2019." In: *CLEF (Working Notes)*.

Zufall, Frederike et al. (2020). "Operationalizing the legal concept of'Incitement to Hatred'as an NLP task". In: *arXiv preprint arXiv:2004.03422*.

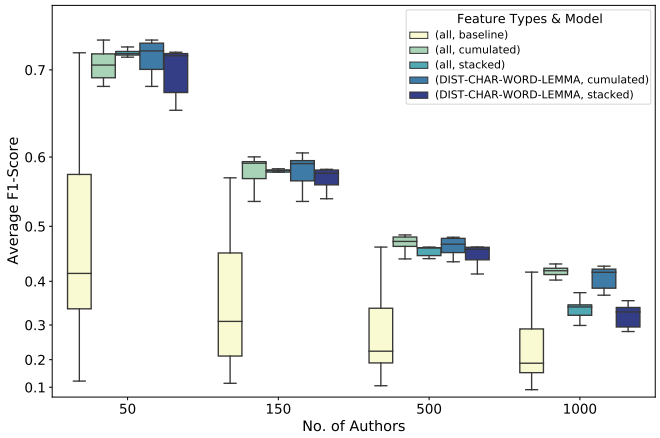
Appendix

A1 Figures

A1.1 Aggregate Overview



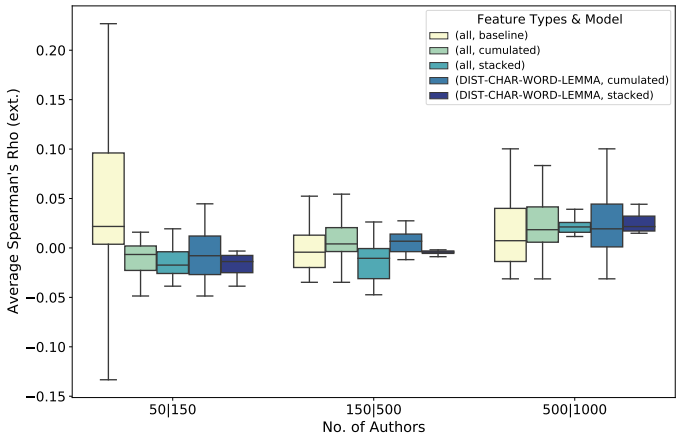
(a) Results for target *gender*



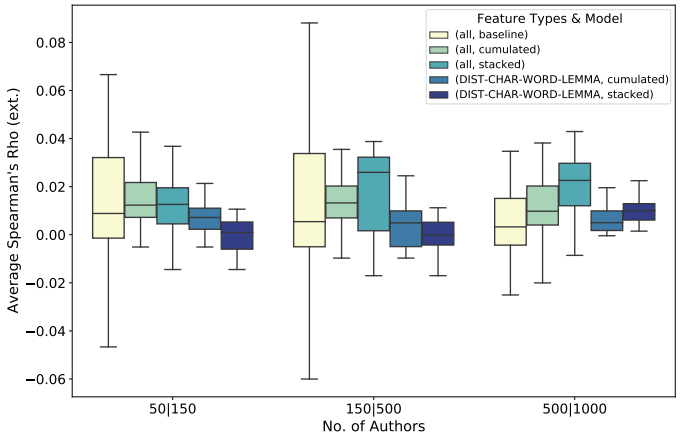
(b) Results for target *age*

Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure A.1: F1-Score for all feature type-sets for an input instance length of 100 characters.



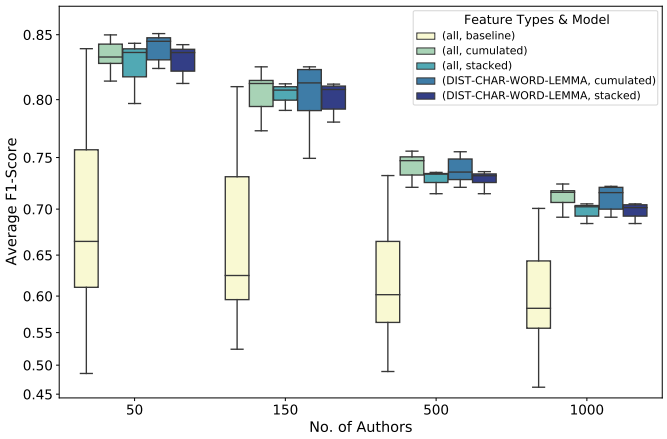
(a) Results for target *gender*



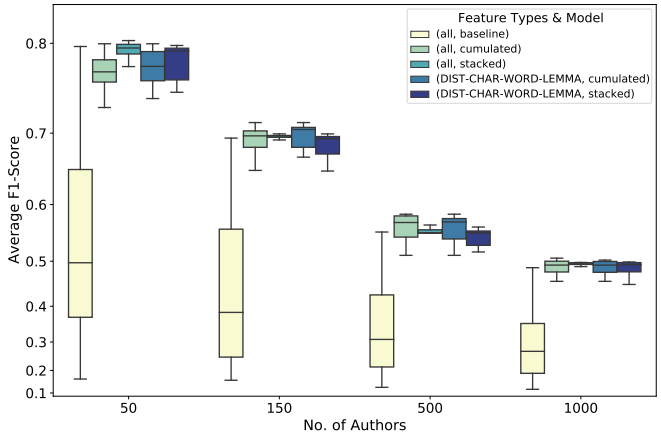
(b) Results for target *age*

Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure A.2: Extended Spearman correlations for all feature type-sets for an input instance length of 100 characters.

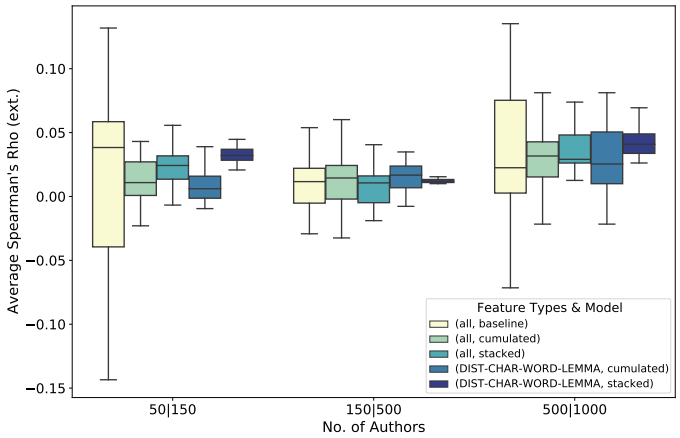


(a) Results for target *gender*

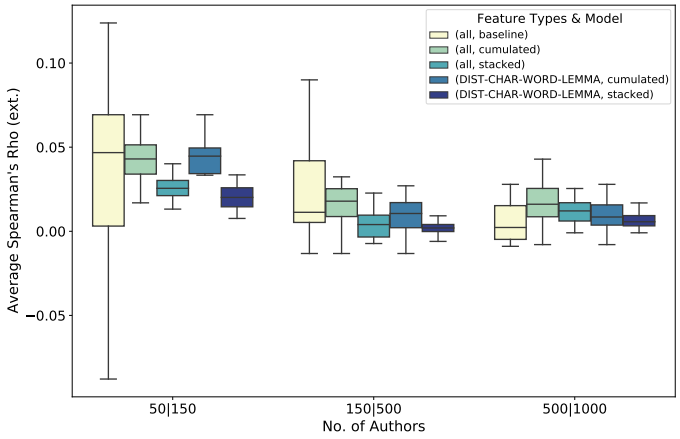


(b) Results for target *age*

Figure A.3: F1-Score for all feature type-sets for an input instance length of 250 characters



(a) Results for target *gender*

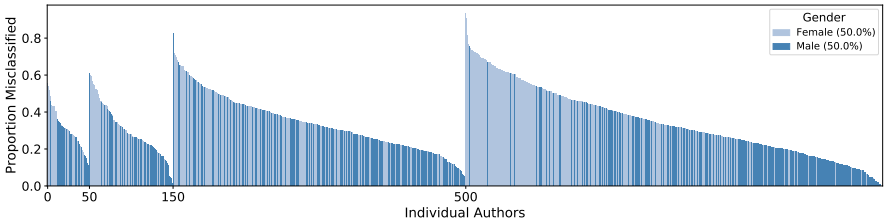


(b) Results for target *age*

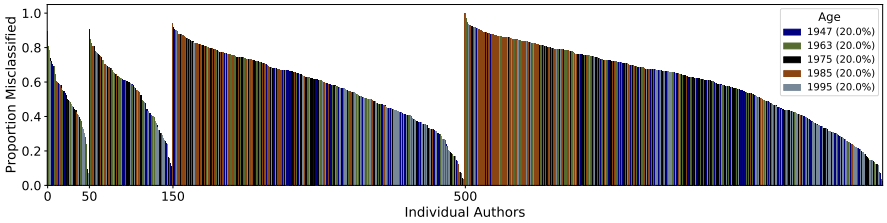
Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure A.4: Extended Spearman correlation for all feature type sets for an input instance length of 250 characters.

A1.2 Author-Level Analysis



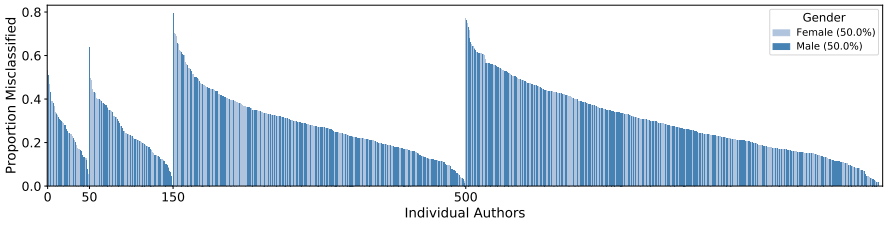
(a) Author-level errors for target *gender*.



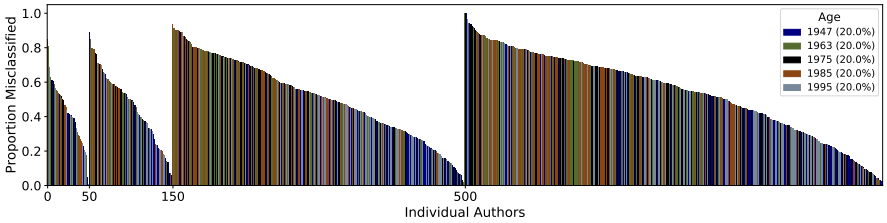
(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure A.5: Author-Level Results for the Full feature set with an input instance length of 100 characters.



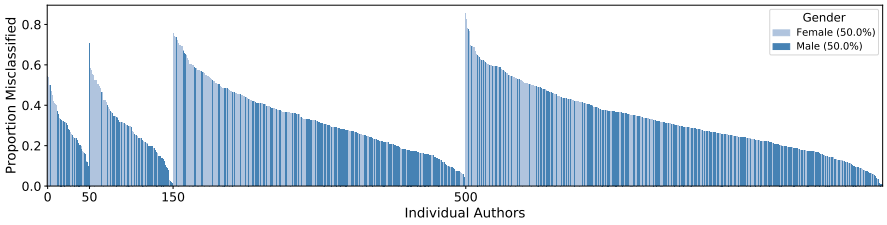
(a) Author-level errors for target *gender*.



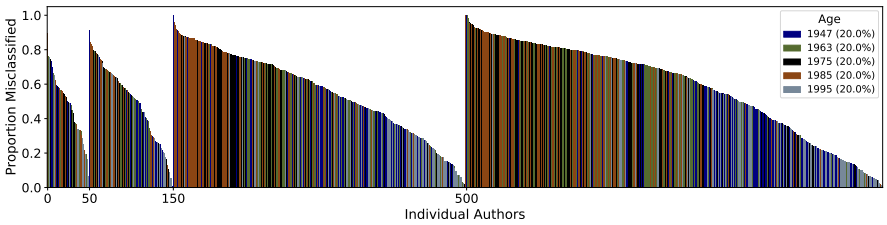
(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure A.6: Author-Level Results for the Full feature set in an input instance length of 100 characters.



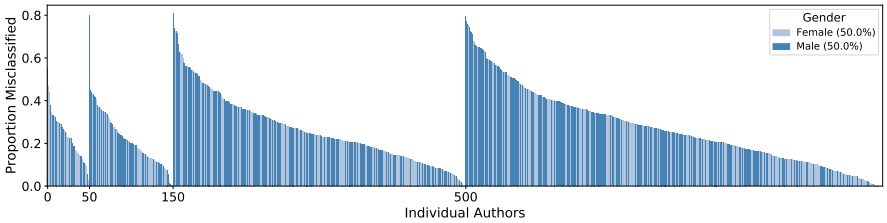
(a) Author-level errors for target *gender*.



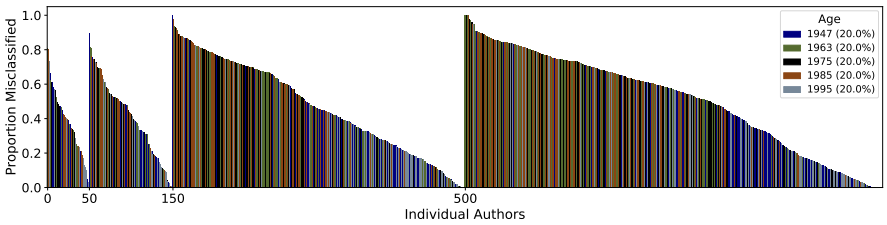
(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure A.7: Author-Level Results for the full feature set with an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.



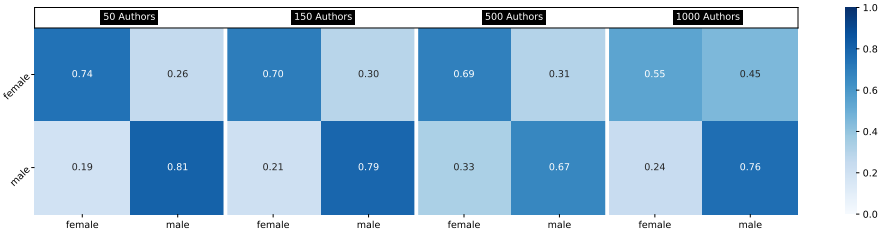
(a) Author-level errors for target *gender*.



(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure A.8: Author-Level Results for the full feature set in an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.



(a) All authors (row-wise normalization).

(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

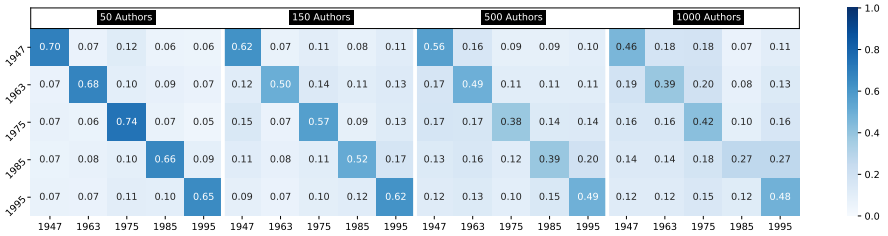
Figure A.9: Confusion matrices for target *gender* with an input instance length 100 characters - all feature types.

(a) All authors (row-wise normalization).

(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure A.10: Confusion matrices for target *gender* with an input instance length of 250 characters - all feature types.

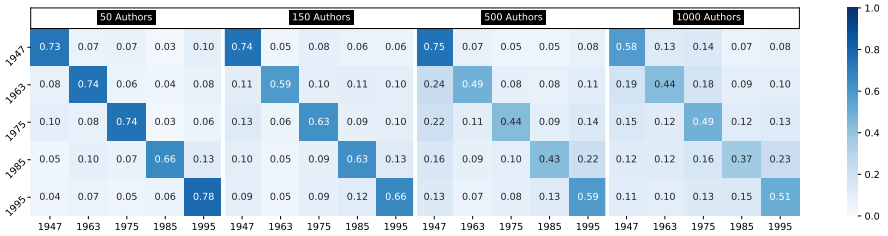


(a) All authors (row-wise normalization).

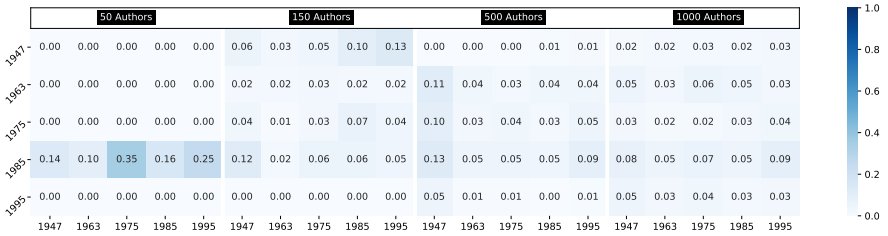
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure A.11: Confusion matrices for target *age* with an input instance length 100 characters - all feature types.



(a) All authors (row-wise normalization).



(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure A.12: Confusion matrices for target *age* with an input instance length of 250 characters - all feature types.

(a) All authors (row-wise normalization).

(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

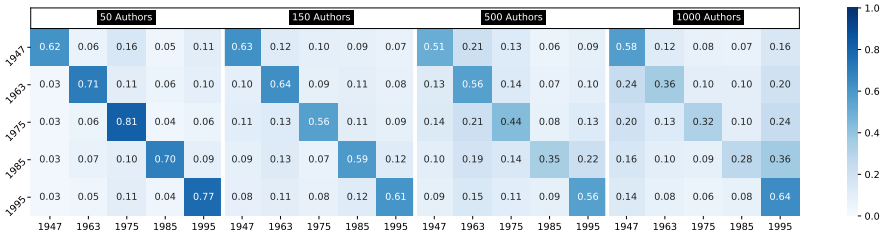
Figure A.13: Confusion matrices for target *gender* with an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.

(a) All authors (row-wise normalization).

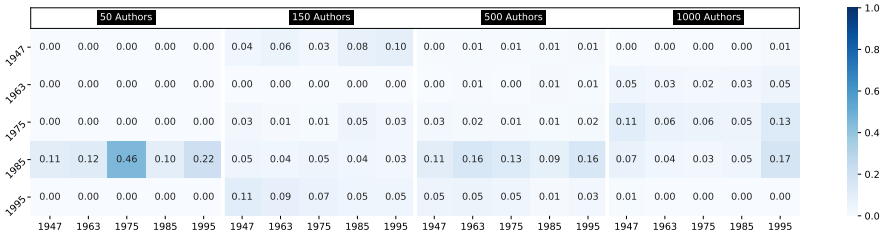
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure A.14: Confusion matrices for target *gender* with an input instance length of 250 characters - ASIS-CHAR-LEMMA-WORD.



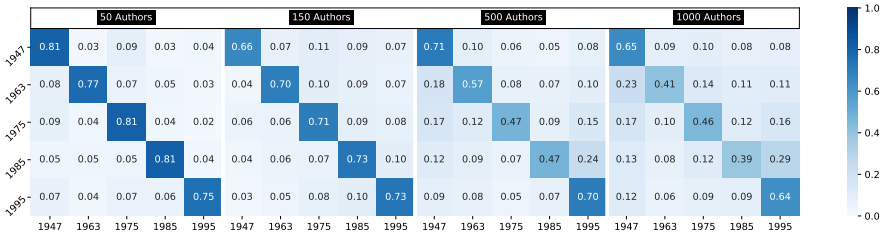
(a) All authors (row-wise normalization).



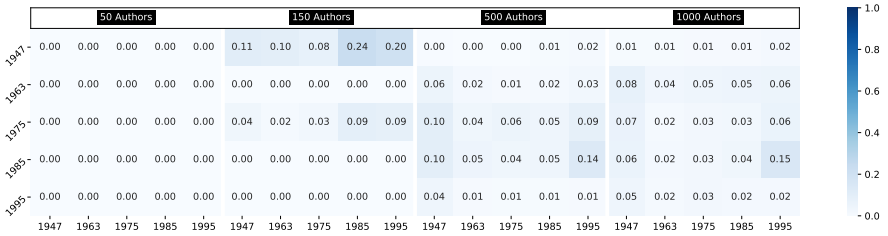
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure A.15: Confusion matrices for target *age* with an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.



(a) All authors (row-wise normalization).



(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure A.16: Confusion matrices for target *age* with an input instance length of 250 characters - ASIS-CHAR-LEMMA-WORD.

A2 Tables

A2.1 Data Set Statistics

No. of Characters	Target No. of Authors	avg_instance age	avg_instance gender	avg_tweet age	avg_tweet gender	avg_tweet_per_instance age	avg_tweet_per_instance gender
100	50	159.94	162.93	109.21	111.71	1.46	1.46
	150	160.56	162.01	107.39	112.88	1.50	1.44
	500	160.83	161.64	109.30	111.84	1.47	1.45
	1000	160.04	160.92	109.26	111.80	1.46	1.44
250	50	313.46	315.85	109.39	112.14	2.87	2.82
	150	313.21	315.31	107.38	112.75	2.92	2.80
	500	313.21	314.61	109.06	111.62	2.87	2.82
	1000	313.33	314.29	109.23	111.81	2.87	2.81
500	50	565.52	568.76	108.88	112.09	5.19	5.07
	150	565.67	568.60	107.37	112.89	5.27	5.04
	500	566.35	567.42	109.17	111.80	5.19	5.08
	1000	566.01	567.42	109.16	111.92	5.19	5.07

Table A.1: Statistics of the Dataset

No. of Characters	Target No. of Authors	avg_instance age	avg_instance gender	avg_tweet age	avg_tweet gender	avg_tweet_per_instance age	avg_tweet_per_instance gender
100	50	161.16	163.39	110.30	112.38	1.46	1.45
	150	160.11	162.66	107.11	112.99	1.49	1.44
	500	160.73	161.51	109.04	111.86	1.47	1.44
	1000	160.10	161.06	109.31	111.83	1.46	1.44
250	50	312.62	316.00	108.98	112.14	2.87	2.82
	150	313.13	316.26	107.63	113.00	2.91	2.80
	500	313.25	314.76	109.22	111.64	2.87	2.82
	1000	313.29	314.28	109.07	111.87	2.87	2.81
500	50	565.07	568.36	109.31	110.78	5.17	5.13
	150	566.11	568.92	107.51	112.82	5.27	5.04
	500	565.73	567.60	109.05	111.69	5.19	5.08
	1000	566.23	567.34	109.21	111.85	5.18	5.07

Table A.2: Statistics of the Dataset

No. of Characters	Target No. of Authors	avg_instance age	avg_instance gender	avg_tweet age	avg_tweet gender	avg_tweet_per_instance age	avg_tweet_per_instance gender
100	50	159.93	162.32	109.42	111.75	1.46	1.45
	150	161.05	162.54	107.75	112.95	1.49	1.44
	500	160.74	161.81	109.27	111.92	1.47	1.45
	1000	160.09	161.03	109.28	111.96	1.47	1.44
250	50	313.19	315.01	108.30	111.55	2.89	2.82
	150	313.55	315.26	107.24	112.85	2.92	2.79
	500	313.39	314.64	108.98	111.86	2.88	2.81
	1000	313.23	314.55	109.23	111.85	2.87	2.81
500	50	566.58	569.37	109.40	112.97	5.18	5.04
	150	566.12	568.69	107.69	112.78	5.26	5.04
	500	566.17	567.78	109.24	111.45	5.18	5.09
	1000	566.14	567.36	109.12	111.69	5.19	5.08

Table A.3: Statistics of the Dataset

A2.2 Baseline

A2.2.1 Minimum of Characters: 100

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 100 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST	2	0.5886	0.5711	0.5455	0.4567	0.5639	0.5594	0.5387	0.4881
	2-3	0.6071	0.5727	0.5992	0.5881	0.5358	0.4757	0.5549	0.5379
	2-3-4	0.6354	0.6251	0.5693	0.5099	0.5521	0.5267	0.5496	0.5167
	2-3-4-5	0.6451	0.6383	0.6076	0.5975	0.5714	0.5708	0.5554	0.5312
CHAR	2	0.7080	0.7080	0.6725	0.6724	0.6152	0.6152	0.6053	0.6047
	2-3	0.7555	0.7555	0.7127	0.7127	0.6560	0.6560	0.6389	0.6380
	2-3-4	0.7656	0.7656	0.7210	0.7210	0.6624	0.6624	0.6437	0.6423
	2-3-4-5	0.7650	0.7649	0.7201	0.7199	0.6617	0.6617	0.6434	0.6420
ASIS	2	0.7362	0.7360	0.6989	0.6982	0.6331	0.6331	0.6212	0.6212
	2-3	0.7683	0.7643	0.7351	0.7350	0.6662	0.6661	0.6475	0.6462
	2-3-4	0.7799	0.7787	0.7403	0.7403	0.6711	0.6710	0.6526	0.6525
	2-3-4-5	0.7883	0.7878	0.7403	0.7402	0.6706	0.6704	0.6523	0.6523
POS	1	0.5671	0.5450	0.5513	0.5313	0.5446	0.5425	0.5463	0.5451
	1-2	0.5950	0.5867	0.5912	0.5912	0.5622	0.5621	0.5611	0.5584
	1-2-3	0.6086	0.6028	0.6005	0.6003	0.5711	0.5711	0.5697	0.5682
TAG	1	0.5895	0.5714	0.5745	0.5740	0.5500	0.5496	0.5514	0.5498
	1-2	0.6293	0.6291	0.6117	0.6112	0.5744	0.5743	0.5709	0.5700
	1-2-3	0.6372	0.6343	0.6213	0.6213	0.5820	0.5820	0.5787	0.5779
DEP	1	0.5752	0.5612	0.5676	0.5672	0.5404	0.5389	0.5423	0.5413
	1-2	0.5951	0.5926	0.5864	0.5738	0.5569	0.5528	0.5585	0.5553
	1-2-3	0.6070	0.5987	0.6036	0.6025	0.5670	0.5670	0.5645	0.5615
LEMMA	1	0.6719	0.6653	0.6351	0.6273	0.5994	0.5968	0.5901	0.5843
	1-2	0.6958	0.6957	0.6465	0.6463	0.6019	0.5985	0.5930	0.5860
WORD	1	0.6371	0.6258	0.6271	0.6270	0.5865	0.5834	0.5817	0.5760
	1-2	0.6490	0.6485	0.6295	0.6295	0.5905	0.5901	0.5847	0.5809
NUM	1	0.5573	0.5551	0.5437	0.4778	0.5327	0.5300	0.5273	0.4637

Table A.4: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1683	0.2888	-0.0348	-0.0259	-0.0312	-0.0785
	2-3	0.0446	0.0485	0.0274	0.0075	0.1002	0.0619
	2-3-4	0.0159	0.0471	0.0272	0.0942	0.0405	-0.0265
	2-3-4-5	-0.0486	-0.0061	-0.0090	-0.0183	0.0560	0.0661
CHAR	2	0.0011	0.0221	0.0639	0.0061	-0.0048	-0.0542
	2-3	0.0047	-0.0198	-0.0146	-0.0051	0.0206	-0.0095
	2-3-4	-0.0036	-0.0135	0.0033	-0.0017	0.0115	0.0061
	2-3-4-5	0.0150	0.0178	0.0008	-0.0004	0.0053	0.0064
ASIS	2	0.0423	0.0470	0.0020	0.0311	-0.0260	0.0004
	2-3	0.0282	-0.0056	0.0137	0.0174	0.0192	0.0201
	2-3-4	0.0117	0.0059	-0.0096	-0.0011	-0.0056	0.0001
	2-3-4-5	0.0034	0.0093	-0.0170	0.0070	-0.0004	-0.0125
POS	1	0.2719	0.2719	-0.0719	-0.0719	-0.2421	-0.2421
	1-2	0.2268	0.2242	0.0524	0.0709	0.1519	0.1775
	1-2-3	0.1066	0.0676	-0.1519	-0.1720	0.0127	0.0380
TAG	1	-0.0293	-0.0293	0.0516	0.0516	-0.0968	-0.0968
	1-2	0.1570	0.1739	-0.1257	-0.1234	0.0425	0.0401
	1-2-3	-0.0161	-0.0200	0.0004	0.0193	0.0080	0.0402
DEP	1	0.2392	0.2392	-0.0207	-0.0207	0.0386	0.0386
	1-2	0.1057	0.1265	0.0345	0.0221	-0.0154	0.0064
	1-2-3	0.0077	0.0151	-0.0730	-0.0300	0.0747	0.0976
LEMMA	1	0.0673	-0.0261	0.0038	0.0636	-0.0132	0.0167
	1-2	0.0293	0.0076	-0.0149	0.0084	-0.0139	-0.0066
WORD	1	0.0276	0.0771	0.0105	0.0330	0.0067	0.0167
	1-2	0.0128	-0.0301	-0.0266	0.0056	0.0721	0.0288
NUM	1	-0.1333	-0.1333	-0.0167	-0.0167	-0.4833	-0.4833

Table A.5: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 100 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST	2	0.3648	0.3249	0.2771	0.2136	0.2645	0.2207	0.2114	0.1008
	2-3	0.3689	0.3502	0.3350	0.3266	0.2162	0.1061	0.2082	0.0899
	2-3-4	0.4086	0.3943	0.3516	0.3444	0.2712	0.2120	0.2197	0.1133
	2-3-4-5	0.4309	0.4286	0.3414	0.3188	0.3018	0.2560	0.2789	0.2226
CHAR	2	0.5608	0.5608	0.4483	0.4443	0.3500	0.3291	0.2982	0.2860
	2-3	0.6430	0.6446	0.5326	0.5312	0.4334	0.4298	0.3819	0.3715
	2-3-4	0.6637	0.6642	0.5473	0.5421	0.4485	0.4477	0.4163	0.4118
	2-3-4-5	0.6649	0.6651	0.5529	0.5486	0.4474	0.4439	0.4124	0.4020
ASIS	2	0.6257	0.6254	0.4874	0.4846	0.3811	0.3769	0.3169	0.3009
	2-3	0.7053	0.7051	0.5598	0.5574	0.4483	0.4453	0.3982	0.3974
	2-3-4	0.7173	0.7172	0.5747	0.5714	0.4639	0.4618	0.4147	0.3992
	2-3-4-5	0.6978	0.6987	0.5748	0.5724	0.4673	0.4648	0.4229	0.4183
POS	1	0.2464	0.1722	0.2145	0.1158	0.2197	0.1573	0.2041	0.1218
	1-2	0.3575	0.3446	0.2257	0.1213	0.2553	0.2194	0.2176	0.1414
	1-2-3	0.4003	0.3939	0.2650	0.2222	0.2674	0.2358	0.2331	0.1823
TAG	1	0.2639	0.1977	0.2155	0.1222	0.2349	0.1735	0.2121	0.1724
	1-2	0.4195	0.4177	0.2970	0.2637	0.2743	0.2396	0.2299	0.1810
	1-2-3	0.4483	0.4456	0.3425	0.3341	0.2540	0.2068	0.2437	0.1885
DEP	1	0.3002	0.2681	0.2408	0.2056	0.2166	0.1777	0.2151	0.1850
	1-2	0.3638	0.3521	0.2843	0.2659	0.2329	0.1688	0.2283	0.1882
	1-2-3	0.3957	0.3920	0.3082	0.2939	0.2669	0.2493	0.2419	0.2160
LEMMA	1	0.4560	0.4543	0.3269	0.3092	0.2559	0.2191	0.2504	0.2193
	1-2	0.4872	0.4864	0.3484	0.3368	0.2939	0.2835	0.2367	0.1861
WORD	1	0.4193	0.4139	0.3010	0.2767	0.2441	0.2052	0.2244	0.1605
	1-2	0.4274	0.4249	0.3169	0.3090	0.2569	0.2326	0.2386	0.1985
NUM	1	0.2694	0.2420	0.2263	0.1991	0.2260	0.1932	0.2111	0.2053

Table A.6: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0348	-0.0149	0.0099	0.0315	-0.0141	0.0174
	2-3	-0.0051	-0.0070	-0.0063	-0.0108	0.0126	0.0458
	2-3-4	-0.0033	0.0366	0.0245	-0.0002	0.0020	0.0207
	2-3-4-5	0.0213	0.0143	-0.0097	-0.0042	-0.0200	-0.0321
CHAR	2	0.0052	0.0143	0.0144	-0.0001	0.0159	0.0310
	2-3	0.0018	-0.0006	-0.0138	0.0003	-0.0044	0.0119
	2-3-4	0.0129	-0.0036	0.0038	-0.0071	0.0016	0.0145
	2-3-4-5	0.0123	0.0102	0.0071	0.0085	0.0051	0.0107
ASIS	2	0.0063	0.0106	-0.0010	-0.0015	0.0112	-0.0048
	2-3	0.0066	0.0100	–	-0.0011	-0.0002	0.0086
	2-3-4	-0.0014	0.0033	0.0075	0.0063	0.0072	0.0010
	2-3-4-5	0.0099	0.0034	0.0036	0.0036	0.0101	0.0097
POS	1	-0.0312	-0.0312	0.0375	0.0375	0.1063	0.1063
	1-2	0.0078	0.0058	0.0881	0.0757	0.1056	0.0972
	1-2-3	0.0443	0.0367	0.0603	0.0675	0.1082	0.1032
TAG	1	0.1134	0.1134	-0.0600	-0.0600	-0.0528	-0.0528
	1-2	0.0647	0.0702	0.0694	0.0775	0.0347	0.0494
	1-2-3	0.0339	0.0345	0.0370	0.0329	0.0003	0.0022
DEP	1	0.0666	0.0860	-0.0819	-0.0819	-0.0113	-0.0113
	1-2	0.0529	-0.0187	0.0834	0.0891	0.0540	0.0571
	1-2-3	-0.0014	-0.0147	0.0361	0.0407	0.0603	0.0692
LEMMA	1	0.0114	0.0114	-0.0066	0.0111	-0.0040	-0.0182
	1-2	-0.0089	0.0103	0.0012	0.0139	-0.0251	-0.0304
WORD	1	0.0266	0.0294	0.0270	0.0137	0.0046	-0.0065
	1-2	-0.0109	0.0096	0.0032	-0.0119	-0.0010	0.0006
NUM	1	-0.0467	-0.0467	-0.2500	-0.2500	-0.1267	-0.1267

Table A.7: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

A2.2.2 Minimum of Characters: 250

	Target	Gender							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.6002	0.5633	0.6026	0.6026	0.5649	0.5558	0.5543	0.5311
	2-3	0.6261	0.6071	0.5801	0.5326	0.5771	0.5764	0.5541	0.5267
	2-3-4	0.6586	0.6551	0.6147	0.5988	0.5624	0.5471	0.5583	0.5314
	2-3-4-5	0.6711	0.6692	0.6203	0.6062	0.5733	0.5666	0.5675	0.5537
CHAR	2	0.7495	0.7480	0.7247	0.7242	0.6605	0.6603	0.6300	0.6218
	2-3	0.8031	0.8020	0.7781	0.7781	0.7125	0.7122	0.6793	0.6774
	2-3-4	0.8087	0.8077	0.7905	0.7903	0.7161	0.7137	0.6895	0.6880
	2-3-4-5	0.8178	0.8178	0.7804	0.7797	0.7210	0.7201	0.6897	0.6886
ASIS	2	0.7844	0.7841	0.7561	0.7556	0.6812	0.6812	0.6587	0.6575
	2-3	0.8256	0.8243	0.7964	0.7958	0.7249	0.7240	0.6905	0.6893
	2-3-4	0.8397	0.8397	0.8076	0.8068	0.7327	0.7320	0.7004	0.6996
	2-3-4-5	0.8383	0.8379	0.8106	0.8105	0.7335	0.7332	0.7013	0.7007
POS	1	0.5967	0.5923	0.5683	0.5643	0.5639	0.5639	0.5613	0.5609
	1-2	0.6294	0.6226	0.6046	0.6020	0.5863	0.5840	0.5729	0.5623
	1-2-3	0.6499	0.6486	0.6162	0.6081	0.6003	0.5997	0.5876	0.5823
TAG	1	0.6276	0.6267	0.5856	0.5853	0.5700	0.5648	0.5643	0.5643
	1-2	0.6619	0.6619	0.6324	0.6306	0.6049	0.6039	0.5879	0.5829
	1-2-3	0.6770	0.6769	0.6562	0.6546	0.6141	0.6139	0.6029	0.6020
DEP	1	0.5978	0.5962	0.5812	0.5811	0.5586	0.5557	0.5572	0.5572
	1-2	0.6192	0.6125	0.6050	0.5998	0.5802	0.5800	0.5709	0.5642
	1-2-3	0.6309	0.6297	0.6249	0.6211	0.5904	0.5875	0.5857	0.5849
LEMMA	1	0.7198	0.7179	0.7046	0.7038	0.6504	0.6479	0.6349	0.6324
	1-2	0.7388	0.7385	0.7137	0.7136	0.6554	0.6554	0.6408	0.6385
WORD	1	0.6966	0.6961	0.6869	0.6869	0.6354	0.6352	0.6294	0.6292
	1-2	0.7011	0.7000	0.6914	0.6912	0.6409	0.6396	0.6313	0.6298
NUM	1	0.5760	0.5527	0.5416	0.5414	0.5478	0.5145	0.5431	0.5431

Table A.8: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Gender					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0390	0.1945	0.0539	0.0938	0.2805	0.2587
	2-3	0.1318	0.1969	0.0012	0.1452	0.1234	0.1113
	2-3-4	0.0671	0.1703	0.0232	0.0109	0.1353	0.0908
	2-3-4-5	0.0076	0.0844	0.0169	0.0175	0.0812	0.0675
CHAR	2	0.1271	-0.0684	0.0223	0.0369	0.0445	0.0490
	2-3	0.0599	-0.0195	0.0212	0.0012	0.0459	0.0043
	2-3-4	0.0587	-0.0164	0.0135	0.0124	0.0036	0.0127
	2-3-4-5	0.0421	0.0059	0.0037	0.0049	0.0188	0.0229
ASIS	2	0.1042	0.0513	0.0164	0.0274	0.0377	0.0374
	2-3	0.0581	0.0135	0.0202	0.0295	0.0128	-0.0127
	2-3-4	0.0456	0.0216	0.0098	0.0092	0.0194	0.0276
	2-3-4-5	0.0399	0.0271	0.0061	0.0113	0.0221	0.0118
POS	1	-0.0754	-0.0754	0.3930	0.3930	0.4070	0.4070
	1-2	-0.2452	-0.1970	-0.0245	-0.0245	0.2085	0.2085
	1-2-3	-0.0732	-0.1022	-0.0618	-0.0595	-0.0122	-0.0147
TAG	1	-0.0482	0.0273	0.2040	0.2040	0.0438	0.0438
	1-2	0.0165	0.0799	-0.0851	-0.1044	-0.0175	-0.0084
	1-2-3	0.0278	—	-0.0074	0.0033	0.0023	-0.0284
DEP	1	-0.1435	-0.2558	-0.0789	-0.0943	0.2496	0.2496
	1-2	0.0537	-0.0038	-0.0292	-0.0330	-0.0715	-0.0513
	1-2-3	0.0377	-0.0120	-0.1215	-0.0453	0.0575	0.0265
LEMMA	1	-0.0620	-0.0509	0.0051	-0.0123	-0.0430	0.0474
	1-2	-0.0081	0.0508	0.0033	0.0062	-0.0042	0.0166
WORD	1	-0.0846	-0.0162	0.0407	0.0240	0.0229	0.0253
	1-2	-0.0132	0.0485	0.0153	0.0136	0.0064	-0.0079
NUM	1	0.6167	0.6167	0.1333	0.1333	-0.2333	-0.2333

Table A.9: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Age							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.3910	0.3670	0.2839	0.2198	0.2414	0.1601	0.2796	0.2179
	2-3	0.4489	0.4159	0.3855	0.3728	0.2272	0.1273	0.3201	0.3105
	2-3-4	0.4944	0.4925	0.4022	0.3925	0.3426	0.3169	0.3057	0.2629
	2-3-4-5	0.4819	0.4712	0.4192	0.4129	0.3373	0.3023	0.3209	0.2863
CHAR	2	0.6413	0.6425	0.5527	0.5512	0.4262	0.4188	0.3499	0.3432
	2-3	0.7199	0.7226	0.6369	0.6363	0.5148	0.5104	0.4514	0.4513
	2-3-4	0.7341	0.7347	0.6701	0.6691	0.5333	0.5311	0.4779	0.4728
	2-3-4-5	0.7422	0.7438	0.6661	0.6656	0.5377	0.5391	0.4825	0.4777
ASIS	2	0.6750	0.6803	0.5870	0.5838	0.4627	0.4534	0.4018	0.3890
	2-3	0.7761	0.7761	0.6804	0.6799	0.5450	0.5430	0.4757	0.4728
	2-3-4	0.7803	0.7813	0.6948	0.6938	0.5562	0.5544	0.4863	0.4742
	2-3-4-5	0.7962	0.7968	0.6916	0.6913	0.5535	0.5516	0.4938	0.4870
POS	1	0.3297	0.2573	0.2708	0.1954	0.2242	0.1283	0.2097	0.1190
	1-2	0.3974	0.3730	0.2731	0.1993	0.2838	0.2495	0.2642	0.2162
	1-2-3	0.4580	0.4464	0.3689	0.3583	0.3194	0.3139	0.2485	0.1735
TAG	1	0.3453	0.3081	0.2867	0.2571	0.2193	0.1308	0.2345	0.1638
	1-2	0.5071	0.5035	0.3718	0.3570	0.2935	0.2649	0.2798	0.2444
	1-2-3	0.5225	0.5192	0.3947	0.3762	0.3386	0.3307	0.3024	0.2927
DEP	1	0.3448	0.3216	0.2452	0.1760	0.2208	0.1383	0.2104	0.1335
	1-2	0.4338	0.4237	0.3042	0.2536	0.2771	0.2313	0.2341	0.1972
	1-2-3	0.4690	0.4677	0.3723	0.3681	0.3002	0.2653	0.2440	0.1781
LEMMA	1	0.5379	0.5390	0.4516	0.4475	0.3518	0.3446	0.3077	0.2764
	1-2	0.5433	0.5426	0.4860	0.4843	0.3675	0.3609	0.3102	0.3018
WORD	1	0.4998	0.5014	0.4085	0.4068	0.3011	0.2568	0.2697	0.2543
	1-2	0.5093	0.5080	0.4236	0.4178	0.3410	0.3344	0.3027	0.2893
NUM	1	0.2595	0.2423	0.2634	0.2383	0.2661	0.2353	0.2106	0.1918

Table A.10: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Age					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0416	0.1495	0.0072	0.0051	0.0153	0.0489
	2-3	-0.0223	0.0252	-0.0132	0.0114	0.0086	-0.0058
	2-3-4	0.0028	0.0019	0.0258	0.0180	0.0279	0.0153
	2-3-4-5	0.0040	0.0345	0.0270	0.0137	-0.0079	-0.0130
CHAR	2	0.0865	0.0131	-0.0061	0.0224	-0.0068	0.0106
	2-3	0.0681	0.0186	0.0013	–	-0.0044	-0.0007
	2-3-4	0.0614	0.0082	0.0056	0.0025	0.0104	0.0068
	2-3-4-5	0.0544	0.0085	0.0076	0.0093	0.0013	-0.0022
ASIS	2	0.0697	-0.0046	0.0127	0.0164	-0.0052	-0.0028
	2-3	0.0731	0.0092	-0.0003	0.0086	0.0048	-0.0017
	2-3-4	0.0588	0.0016	0.0051	0.0058	0.0031	0.0043
	2-3-4-5	0.0546	0.0129	0.0092	0.0109	0.0028	0.0018
POS	1	0.3351	0.3351	0.3225	0.3225	-0.0642	-0.0642
	1-2	0.0164	0.0153	0.0506	0.0512	0.0775	0.0510
	1-2-3	0.0483	0.0510	0.0544	0.0511	-0.0014	0.0060
TAG	1	-0.0185	0.0673	0.0453	0.0453	0.0256	0.0256
	1-2	0.0452	0.0696	0.0474	0.0469	0.0464	0.0452
	1-2-3	0.0400	0.0834	0.0099	0.0097	0.0017	0.0006
DEP	1	0.1238	0.0699	0.0320	0.0320	-0.0798	-0.0798
	1-2	0.1172	0.0457	-0.0097	-0.0097	0.0150	0.0150
	1-2-3	0.0376	0.0561	0.0511	0.0444	0.0211	0.0173
LEMMA	1	-0.0878	0.0197	0.0073	0.0009	-0.0049	0.0194
	1-2	-0.0661	0.0356	0.0217	-0.0039	-0.0011	0.0040
WORD	1	-0.1184	0.0144	-0.0030	0.0417	-0.0089	-0.0003
	1-2	-0.0686	0.0209	0.0142	0.0203	-0.0035	0.0018
NUM	1	0.2100	0.2100	0.0900	0.0900	0.2667	0.2667

Table A.11: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

A2.2.3 Minimum of Characters: 500

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.6367	0.6322	0.5043	0.3481	0.5095	0.3680	0.5582	0.5403
	2-3	0.6531	0.6477	0.5202	0.3788	0.5278	0.4365	0.5518	0.5173
	2-3-4	0.6672	0.6624	0.6266	0.6056	0.5609	0.5462	0.5636	0.5432
	2-3-4-5	0.6820	0.6789	0.6404	0.6402	0.5584	0.5385	0.5452	0.4858
CHAR	2	0.8080	0.8080	0.7590	0.7580	0.6901	0.6882	0.6720	0.6719
	2-3	0.8694	0.8690	0.8164	0.8151	0.7641	0.7638	0.7265	0.7260
	2-3-4	0.8713	0.8702	0.8413	0.8411	0.7751	0.7746	0.7369	0.7362
	2-3-4-5	0.8848	0.8847	0.8122	0.8094	0.7759	0.7756	0.7412	0.7411
ASIS	2	0.8366	0.8357	0.7996	0.7995	0.7186	0.7184	0.6925	0.6916
	2-3	0.8871	0.8868	0.8444	0.8436	0.7753	0.7750	0.7354	0.7341
	2-3-4	0.9006	0.9004	0.8664	0.8662	0.7886	0.7885	0.7494	0.7493
	2-3-4-5	0.9042	0.9040	0.8197	0.8158	0.7882	0.7882	0.7493	0.7488
POS	1	0.6032	0.5907	0.5864	0.5847	0.5776	0.5775	0.5730	0.5728
	1-2	0.6413	0.6303	0.6404	0.6403	0.6062	0.6061	0.5988	0.5985
	1-2-3	0.6597	0.6583	0.6587	0.6579	0.6189	0.6174	0.6128	0.6127
TAG	1	0.6410	0.6402	0.5930	0.5700	0.5893	0.5892	0.5772	0.5771
	1-2	0.6990	0.6982	0.6710	0.6704	0.6274	0.6265	0.6140	0.6140
	1-2-3	0.6987	0.6980	0.6830	0.6828	0.6299	0.6230	0.6279	0.6278
DEP	1	0.6219	0.6195	0.6017	0.5997	0.5691	0.5652	0.5678	0.5655
	1-2	0.6590	0.6590	0.6322	0.6277	0.5976	0.5963	0.5935	0.5926
	1-2-3	0.6705	0.6704	0.6566	0.6550	0.6053	0.6008	0.6066	0.6060
LEMMA	1	0.7788	0.7786	0.7680	0.7679	0.7096	0.7096	0.6876	0.6869
	1-2	0.8008	0.8007	0.7816	0.7816	0.7145	0.7144	0.6935	0.6925
WORD	1	0.7591	0.7588	0.7434	0.7408	0.6942	0.6940	0.6785	0.6782
	1-2	0.7653	0.7653	0.7535	0.7535	0.6983	0.6981	0.6837	0.6836
NUM	1	0.5950	0.5912	0.5635	0.5635	0.5542	0.5229	0.5332	0.4791

Table A.12: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0858	0.1370	0.1528	0.2654	0.2834	0.2834
	2-3	0.1535	0.1134	0.1350	0.1049	0.1451	0.1875
	2-3-4	0.1185	-0.0050	0.0893	0.0610	0.0425	0.0354
	2-3-4-5	0.0904	0.0550	0.0491	0.0644	0.0457	0.0703
CHAR	2	0.2757	0.0476	0.0903	0.1012	0.0932	0.0932
	2-3	0.1739	0.0056	0.0277	0.0341	0.0251	0.0317
	2-3-4	0.1328	0.0002	0.0174	0.0285	0.0364	0.0302
	2-3-4-5	0.0999	-0.0035	0.0131	0.0221	0.0215	0.0244
ASIS	2	0.2720	0.0435	0.0573	0.0488	0.0588	0.0656
	2-3	0.1492	0.0154	0.0155	0.0369	0.0213	0.0390
	2-3-4	0.1229	-0.0056	0.0114	0.0142	0.0150	0.0139
	2-3-4-5	0.1025	0.0215	0.0207	0.0084	0.0153	0.0196
POS	1	0.2456	0.2456	0.3123	0.3123	0.1982	0.1982
	1-2	0.0272	0.0314	-0.1515	-0.1286	0.2639	0.2440
	1-2-3	-0.0449	-0.0707	-0.0222	-0.0120	0.0582	0.0586
TAG	1	-0.0086	-0.0086	-0.1640	-0.1460	0.0043	0.0043
	1-2	0.1546	0.1327	0.0233	-0.0280	0.1140	0.1091
	1-2-3	0.0514	0.0577	0.0017	-0.0141	0.1243	0.1212
DEP	1	0.0103	-0.1889	-0.0425	-0.0425	0.0244	0.0244
	1-2	0.1047	0.0867	-0.1028	-0.1028	-0.0255	-0.0422
	1-2-3	-0.0349	-0.0498	0.0236	0.0366	0.0737	0.0692
LEMMA	1	-0.0374	-0.0415	0.0221	0.0459	0.0406	-0.0087
	1-2	-0.0143	0.0461	0.0099	-0.0155	0.0467	-0.0293
WORD	1	-0.0304	0.0282	0.0136	-0.0115	0.0480	0.0523
	1-2	-0.0003	0.0252	0.0560	0.0179	0.0304	-0.0080
NUM	1	0.2500	0.2500	0.0833	0.0833	0.0667	0.0667

Table A.13: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.4831	0.4797	0.2924	0.2249	0.3224	0.2808	0.2819	0.2196
	2-3	0.4969	0.4809	0.3987	0.3711	0.3089	0.2633	0.3029	0.2736
	2-3-4	0.5338	0.5294	0.4561	0.4482	0.2905	0.2454	0.3208	0.2921
	2-3-4-5	0.5427	0.5384	0.4385	0.4088	0.3021	0.2401	0.3217	0.2823
CHAR	2	0.7240	0.7257	0.6254	0.6250	0.4829	0.4761	0.4143	0.4033
	2-3	0.8023	0.8024	0.7320	0.7319	0.5762	0.5696	0.5104	0.5070
	2-3-4	0.8268	0.8274	0.7575	0.7579	0.6009	0.5938	0.5368	0.5346
	2-3-4-5	0.8290	0.8291	0.7494	0.7492	0.6136	0.6114	0.5439	0.5407
ASIS	2	0.7801	0.7811	0.6779	0.6776	0.5188	0.5142	0.4515	0.4426
	2-3	0.8540	0.8546	0.7651	0.7646	0.6086	0.6048	0.5371	0.5346
	2-3-4	0.8682	0.8687	0.7843	0.7836	0.6253	0.6189	0.5559	0.5512
	2-3-4-5	0.8544	0.8542	0.7893	0.7890	0.6236	0.6174	0.5572	0.5523
POS	1	0.3740	0.3529	0.2990	0.2723	0.2140	0.0964	0.2056	0.1010
	1-2	0.4875	0.4794	0.3689	0.3512	0.2864	0.2269	0.2678	0.2283
	1-2-3	0.5321	0.5248	0.4249	0.4124	0.3220	0.2767	0.2844	0.2491
TAG	1	0.4083	0.3934	0.3316	0.3244	0.2301	0.1338	0.2291	0.1690
	1-2	0.5476	0.5432	0.4492	0.4479	0.3318	0.2755	0.2998	0.2670
	1-2-3	0.5877	0.5878	0.4856	0.4859	0.3875	0.3811	0.3402	0.3110
DEP	1	0.3704	0.3560	0.2994	0.2418	0.2089	0.1000	0.2504	0.2387
	1-2	0.5089	0.5057	0.3869	0.3820	0.3140	0.2523	0.2762	0.2358
	1-2-3	0.5370	0.5369	0.4142	0.4029	0.3468	0.3292	0.3108	0.2873
LEMMA	1	0.6679	0.6668	0.5773	0.5763	0.4239	0.4091	0.3651	0.3339
	1-2	0.6923	0.6920	0.5924	0.5881	0.4493	0.4424	0.3916	0.3781
WORD	1	0.6282	0.6282	0.5433	0.5416	0.3993	0.3867	0.3523	0.3439
	1-2	0.6438	0.6437	0.5524	0.5482	0.4168	0.4053	0.3551	0.3294
NUM	1	0.3099	0.2995	0.2714	0.2542	0.2807	0.2386	0.2440	0.2194

Table A.14: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Age					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1872	0.0727	0.0083	0.0109	0.0192	0.0501
	2-3	0.1142	0.0359	0.0275	0.0294	0.0478	0.0012
	2-3-4	0.0268	0.0157	0.0502	0.0416	0.0099	0.0147
	2-3-4-5	0.0219	0.0064	0.0371	0.0096	0.0154	0.0220
CHAR	2	0.2159	0.0025	0.0197	0.0100	0.0058	-0.0009
	2-3	0.1388	0.0013	0.0014	0.0022	-0.0014	0.0042
	2-3-4	0.1140	0.0123	0.0124	0.0118	0.0123	0.0141
	2-3-4-5	0.0916	0.0113	0.0075	0.0069	0.0068	0.0055
ASIS	2	0.2031	0.0121	0.0049	0.0141	-0.0024	0.0076
	2-3	0.1386	0.0059	0.0027	-0.0003	0.0025	0.0066
	2-3-4	0.1080	0.0070	0.0032	0.0008	0.0069	0.0038
	2-3-4-5	0.0925	0.0066	0.0023	0.0045	0.0011	0.0056
POS	1	0.1807	0.1807	0.1046	0.1046	0.1923	0.1923
	1-2	0.0762	0.0317	-0.0211	-0.0158	0.0116	0.0116
	1-2-3	0.0611	0.0675	0.0236	0.0195	0.0455	0.0432
TAG	1	0.0886	0.0509	0.0711	0.0010	-0.0680	-0.0357
	1-2	0.0178	-0.0089	-0.0109	-0.0351	0.0441	0.0250
	1-2-3	0.0825	0.0528	0.0386	0.0268	0.0474	0.0551
DEP	1	0.0971	0.1030	0.0466	0.0466	0.0411	0.0411
	1-2	0.0559	0.0636	0.0155	0.0068	-0.0100	-0.0088
	1-2-3	0.0751	0.0724	-0.0095	-0.0003	-0.0020	0.0079
LEMMA	1	-0.0219	0.0181	0.0128	-0.0047	0.0078	0.0045
	1-2	0.0010	-0.0082	0.0085	0.0027	-0.0058	0.0056
WORD	1	-0.0585	0.0064	0.0428	0.0232	0.0151	0.0098
	1-2	-0.0112	-0.0047	0.0013	0.0081	0.0118	0.0092
NUM	1	0.4267	0.4267	-0.0933	-0.0933	0.3033	0.3033

Table A.15: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

A2.3 DIST, CHAR, ASIS, WORD, and LEMMA

A2.3.1 Minimum of Characters: 100 & Cumulated

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7343	0.7338	0.6967	0.6947	0.6284	0.6276	0.6089	0.5966
	2-3	0.7746	0.7738	0.7331	0.7330	0.6647	0.6645	0.6387	0.6301
	2-3-4	0.7882	0.7881	0.7379	0.7370	0.6712	0.6711	0.6412	0.6312
	2-3-4-5	0.7836	0.7829	0.7354	0.7342	0.6710	0.6710	0.6409	0.6305
DIST_CHAR_ASIS	2	0.7905	0.7904	0.7453	0.7453	0.6775	0.6775	0.6452	0.6340
	2-3	0.7973	0.7972	0.7537	0.7534	0.6835	0.6833	0.6645	0.6645
	2-3-4	0.7913	0.7908	0.7478	0.7459	0.6849	0.6848	0.6660	0.6660
	2-3-4-5	0.7980	0.7979	0.7461	0.7437	0.6844	0.6844	0.6658	0.6658
DIST_CHAR_ASIS_LEMMA	1	0.8014	0.8013	0.7550	0.7549	0.6868	0.6861	0.6688	0.6688
	1-2	0.7999	0.7998	0.7541	0.7534	0.6882	0.6882	0.6689	0.6689
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8010	0.8007	0.7550	0.7548	0.6828	0.6801	0.6700	0.6694
	1-2	0.8045	0.8044	0.7567	0.7566	0.6832	0.6803	0.6707	0.6702

Table A.16: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1683	0.2888	-0.0348	-0.0259	-0.0312	-0.0785
	2-3	0.0446	0.0485	0.0274	0.0075	0.1002	0.0619
	2-3-4	0.0159	0.0471	0.0272	0.0942	0.0405	-0.0265
	2-3-4-5	-0.0486	-0.0061	-0.0090	-0.0183	0.0560	0.0661
DIST_CHAR	2	-0.0481	-0.0285	-0.0020	-0.0628	0.0686	0.0497
	2-3	-0.0355	0.0028	0.0120	0.0464	0.0665	0.0553
	2-3-4	-0.0022	0.0259	0.0037	0.0052	0.0198	0.0238
	2-3-4-5	-0.0103	0.0157	0.0228	0.0138	0.0165	0.0021
DIST_CHAR_ASIS	2	-0.0215	0.0133	-0.0312	0.0060	0.0065	0.0232
	2-3	-0.0055	-0.0018	0.0044	0.0263	-0.0050	0.0227
	2-3-4	-0.0265	-0.0274	0.0009	-0.0053	0.0245	0.0168
	2-3-4-5	-0.0284	-0.0212	-0.0118	-0.0198	-0.0143	-0.0140
DIST_CHAR_ASIS_LEMMA	1	-0.0142	-0.0172	0.0200	0.0075	0.0022	0.0267
	1-2	0.0151	-0.0069	0.0094	-0.0110	0.0216	0.0211
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0075	0.0216	0.0092	-0.0109	-0.0022	0.0149
	1-2	0.0111	0.0057	0.0093	-0.0034	0.0189	0.0100

Table A.17: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.5983	0.5994	0.4887	0.4861	0.3739	0.3555	0.3088	0.2673
	2-3	0.6907	0.6908	0.5595	0.5583	0.4424	0.4382	0.3888	0.3706
	2-3-4	0.7031	0.7037	0.5749	0.5714	0.4523	0.4435	0.4023	0.3817
	2-3-4-5	0.6804	0.6827	0.5378	0.5386	0.4624	0.4589	0.4017	0.3870
DIST_CHAR_ASIS	2	0.7161	0.7161	0.5828	0.5810	0.4677	0.4662	0.4223	0.4062
	2-3	0.7275	0.7276	0.5944	0.5926	0.4759	0.4726	0.4336	0.4241
	2-3-4	0.7270	0.7267	0.6014	0.6002	0.4744	0.4672	0.4309	0.4160
	2-3-4-5	0.7292	0.7295	0.5952	0.5939	0.4788	0.4738	0.4320	0.4203
DIST_CHAR_ASIS_LEMMA	1	0.7056	0.7042	0.5915	0.5902	0.4850	0.4799	0.4372	0.4295
	1-2	0.7213	0.7220	0.5950	0.5942	0.4838	0.4801	0.4377	0.4299
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7261	0.7265	0.5994	0.5989	0.4845	0.4818	0.4329	0.4201
	1-2	0.7218	0.7221	0.6055	0.6053	0.4842	0.4813	0.4356	0.4243

Table A.18: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0348	-0.0149	0.0099	0.0315	-0.0141	0.0174
	2-3	-0.0051	-0.0070	-0.0063	-0.0108	0.0126	0.0458
	2-3-4	-0.0033	0.0366	0.0245	-0.0002	0.0020	0.0207
	2-3-4-5	0.0213	0.0143	-0.0097	-0.0042	-0.0200	-0.0321
DIST_CHAR	2	0.0013	-0.0092	0.0077	0.0500	0.0047	0.0328
	2-3	0.0064	0.0009	-0.0084	-0.0053	0.0196	0.0099
	2-3-4	-0.0009	0.0049	0.0101	0.0123	-0.0004	-0.0037
	2-3-4-5	0.0157	0.0139	0.0107	-0.0099	0.0016	0.0008
DIST_CHAR_ASIS	2	0.0108	0.0058	-0.0019	0.0102	0.0090	0.0187
	2-3	0.0089	0.0029	0.0068	0.0036	0.0053	0.0122
	2-3-4	0.0119	0.0045	0.0045	0.0047	0.0087	0.0154
	2-3-4-5	0.0074	0.0087	-0.0049	-0.0032	0.0018	0.0034
DIST_CHAR_ASIS_LEMMA	1	0.0026	0.0128	0.0053	0.0030	0.0128	0.0147
	1-2	0.0072	0.0095	-0.0051	-0.0010	0.0127	0.0073
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0039	-0.0016	0.0130	0.0050	0.0054	0.0082
	1-2	0.0071	0.0105	-0.0009	-0.0024	0.0037	0.0097

Table A.19: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

A2.3.2 Minimum of Characters: 250 & Cumulated

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender 250 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7844	0.7841	0.7495	0.7493	0.6736	0.6736	0.6514	0.6511
	2-3	0.8249	0.8247	0.7771	0.7738	0.7219	0.7218	0.6916	0.6916
	2-3-4	0.8310	0.8309	0.7884	0.7858	0.7306	0.7305	0.7001	0.7001
	2-3-4-5	0.8316	0.8315	0.7943	0.7927	0.7283	0.7279	0.6995	0.6995
DIST_CHAR_ASIS	2	0.8434	0.8434	0.8019	0.8002	0.7419	0.7418	0.7084	0.7084
	2-3	0.8423	0.8418	0.8120	0.8104	0.7475	0.7473	0.7166	0.7165
	2-3-4	0.8499	0.8499	0.8260	0.8260	0.7497	0.7495	0.7176	0.7172
	2-3-4-5	0.8479	0.8479	0.8243	0.8243	0.7485	0.7484	0.7173	0.7168
DIST_CHAR_ASIS_LEMMA	1	0.8477	0.8477	0.8238	0.8237	0.7350	0.7298	0.7221	0.7221
	1-2	0.8508	0.8508	0.8174	0.8165	0.7362	0.7311	0.7228	0.7228
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8470	0.8470	0.8238	0.8237	0.7552	0.7552	0.7220	0.7218
	1-2	0.8472	0.8471	0.8250	0.8250	0.7553	0.7553	0.7222	0.7218

Table A.20: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender 250 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0390	0.1945	0.0539	0.0938	0.2805	0.2587
	2-3	0.1318	0.1969	0.0012	0.1452	0.1234	0.1113
	2-3-4	0.0671	0.1703	0.0232	0.0109	0.1353	0.0908
	2-3-4-5	0.0076	0.0844	0.0169	0.0175	0.0812	0.0675
DIST_CHAR	2	0.0017	0.0251	0.0348	0.0854	0.0363	0.0600
	2-3	-0.0013	0.0141	-0.0077	0.0357	0.0245	0.0117
	2-3-4	0.0015	0.0345	0.0257	0.0152	0.0389	0.0345
	2-3-4-5	-0.0060	0.0096	0.0168	0.0283	0.0403	0.0370
DIST_CHAR_ASIS	2	0.0181	0.0458	0.0196	-0.0173	-0.0216	-0.0147
	2-3	-0.0017	0.0099	0.0257	0.0034	0.0041	-0.0010
	2-3-4	0.0151	-0.0199	0.0072	-0.0106	0.0263	0.0332
	2-3-4-5	0.0060	0.0030	0.0127	-0.0121	0.0074	-0.0042
DIST_CHAR_ASIS_LEMMA	1	0.0061	0.0041	0.0156	0.0043	0.0136	0.0103
	1-2	0.0062	-0.0082	0.0058	-0.0069	0.0152	0.0150
DIST_CHAR_ASIS_LEMMA_WORD	1	-0.0095	0.0059	0.0053	0.0094	0.0108	0.0203
	1-2	-0.0079	-0.0089	0.0166	0.0134	0.0077	0.0049

Table A.21: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Age / 250 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.6772	0.6770	0.5772	0.5772	0.4543	0.4439	0.3928	0.3686
	2-3	0.7627	0.7629	0.6684	0.6689	0.5239	0.5115	0.4643	0.4581
	2-3-4	0.7534	0.7555	0.6707	0.6732	0.5471	0.5440	0.4860	0.4795
	2-3-4-5	0.7380	0.7413	0.6847	0.6848	0.5454	0.5361	0.4927	0.4875
DIST_CHAR_ASIS	2	0.7749	0.7751	0.6993	0.6996	0.5607	0.5567	0.4827	0.4712
	2-3	0.7918	0.7920	0.7083	0.7067	0.5742	0.5705	0.5002	0.4942
	2-3-4	0.7981	0.7996	0.7140	0.7131	0.5773	0.5736	0.4982	0.4902
	2-3-4-5	0.7928	0.7941	0.7059	0.7071	0.5862	0.5845	0.5049	0.5020
DIST_CHAR_ASIS_LEMMA	1	0.7737	0.7743	0.7069	0.7054	0.5857	0.5829	0.5066	0.5014
	1-2	0.7867	0.7863	0.7075	0.7079	0.5782	0.5728	0.5060	0.4989
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7774	0.7778	0.7030	0.7037	0.5803	0.5757	0.5056	0.4972
	1-2	0.7911	0.7916	0.7070	0.7074	0.5862	0.5820	0.5082	0.5022

Table A.22: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Age / 250 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0416	0.1495	0.0072	0.0051	0.0153	0.0489
	2-3	-0.0223	0.0252	-0.0132	0.0114	0.0086	-0.0058
	2-3-4	0.0028	0.0019	0.0258	0.0180	0.0279	0.0153
	2-3-4-5	0.0040	0.0345	0.0270	0.0137	-0.0079	-0.0130
DIST_CHAR	2	0.0620	0.0267	0.0122	0.0015	0.0181	0.0090
	2-3	0.0686	0.0548	0.0189	0.0225	0.0152	0.0079
	2-3-4	0.0693	0.0148	0.0045	0.0058	0.0072	0.0068
	2-3-4-5	0.0496	0.0082	0.0093	0.0005	-0.0054	-0.0113
DIST_CHAR_ASIS	2	0.0390	0.0173	0.0187	0.0134	0.0050	0.0077
	2-3	0.0346	0.0165	0.0025	0.0088	0.0027	0.0066
	2-3-4	0.0333	0.0131	0.0004	0.0048	0.0168	0.0067
	2-3-4-5	0.0450	0.0078	-0.0004	0.0097	0.0206	-0.0040
DIST_CHAR_ASIS_LEMMA	1	0.0474	0.0163	0.0010	0.0054	0.0083	0.0038
	1-2	0.0445	0.0153	0.0165	0.0143	0.0040	-0.0025
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0495	0.0140	0.0119	0.0137	0.0022	0.0153
	1-2	0.0449	0.0180	0.0132	0.0067	0.0151	0.0067

Table A.23: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

A2.3.3 Minimum of Characters: 500 & Cumulated

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 500 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.8231	0.8230	0.7868	0.7868	0.7062	0.7058	0.6795	0.6769
	2-3	0.8802	0.8802	0.8396	0.8392	0.7698	0.7696	0.7313	0.7311
	2-3-4	0.8727	0.8712	0.8498	0.8494	0.7852	0.7852	0.7407	0.7392
	2-3-4-5	0.8953	0.8951	0.8583	0.8583	0.7679	0.7650	0.7433	0.7429
DIST_CHAR_ASIS	2	0.8887	0.8882	0.8623	0.8623	0.7521	0.7441	0.7511	0.7494
	2-3	0.8861	0.8850	0.8725	0.8725	0.8036	0.8036	0.7612	0.7602
	2-3-4	0.8779	0.8764	0.8734	0.8733	0.7700	0.7641	0.7676	0.7675
	2-3-4-5	0.8947	0.8942	0.8730	0.8729	0.7688	0.7626	0.7674	0.7673
DIST_CHAR_ASIS_LEMMA	1	0.8930	0.8925	0.8697	0.8697	0.7805	0.7765	0.7707	0.7706
	1-2	0.8782	0.8779	0.8744	0.8744	0.7785	0.7740	0.7708	0.7703
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8799	0.8797	0.8712	0.8712	0.8037	0.8034	0.7698	0.7688
	1-2	0.8943	0.8943	0.8704	0.8702	0.7754	0.7702	0.7727	0.7723

Table A.24: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 500 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0858	0.1370	0.1528	0.2654	0.2834	0.2834
	2-3	0.1535	0.1134	0.1350	0.1049	0.1451	0.1875
	2-3-4	0.1185	-0.0050	0.0893	0.0610	0.0425	0.0354
	2-3-4-5	0.0904	0.0550	0.0491	0.0644	0.0457	0.0703
DIST_CHAR	2	0.0911	0.0265	0.0598	0.0466	0.0998	0.0889
	2-3	0.0903	-0.0365	0.0331	0.0278	0.0178	0.0419
	2-3-4	0.0990	0.0314	0.0445	0.0261	0.0454	0.0362
	2-3-4-5	0.0829	-0.0121	0.0455	0.0365	0.0461	0.0232
DIST_CHAR_ASIS	2	0.0815	-0.0217	0.0306	0.0077	0.0118	0.0141
	2-3	0.0931	0.0012	0.0462	0.0171	0.0295	0.0138
	2-3-4	0.0789	-0.0137	0.0380	0.0166	0.0026	0.0088
	2-3-4-5	0.0829	0.0014	0.0459	0.0050	0.0281	0.0243
DIST_CHAR_ASIS_LEMMA	1	0.0935	0.0170	0.0201	0.0087	0.0152	0.0141
	1-2	0.0829	0.0101	0.0428	0.0172	0.0084	0.0070
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0632	0.0040	0.0388	0.0146	-0.0029	0.0213
	1-2	0.0526	0.0023	0.0268	0.0127	0.0094	0.0274

Table A.25: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7262	0.7271	0.6350	0.6349	0.5133	0.5123	0.4465	0.4434
	2-3	0.8192	0.8214	0.7342	0.7331	0.5938	0.5898	0.5199	0.5177
	2-3-4	0.8321	0.8343	0.7623	0.7627	0.6151	0.6114	0.5437	0.5390
	2-3-4-5	0.8232	0.8250	0.7590	0.7595	0.6186	0.6140	0.5451	0.5397
DIST_CHAR_ASIS	2	0.8508	0.8518	0.7725	0.7727	0.6203	0.6150	0.5566	0.5515
	2-3	0.8727	0.8726	0.7954	0.7954	0.6640	0.6629	0.5724	0.5687
	2-3-4	0.8753	0.8753	0.8017	0.8014	0.6618	0.6610	0.5643	0.5553
	2-3-4-5	0.8673	0.8671	0.8007	0.8006	0.6421	0.6376	0.5764	0.5731
DIST_CHAR_ASIS_LEMMA	1	0.8611	0.8608	0.7971	0.7966	0.6518	0.6478	0.5773	0.5717
	1-2	0.8620	0.8615	0.7961	0.7963	0.6655	0.6629	0.5643	0.5544
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8517	0.8515	0.7885	0.7883	0.6585	0.6558	0.5771	0.5736
	1-2	0.8290	0.8293	0.7736	0.7727	0.6653	0.6652	0.5798	0.5758

Table A.26: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.1872	0.0727	0.0083	0.0109	0.0192	0.0501
	2-3	0.1142	0.0359	0.0275	0.0294	0.0478	0.0012
	2-3-4	0.0268	0.0157	0.0502	0.0416	0.0099	0.0147
	2-3-4-5	0.0219	0.0064	0.0371	0.0096	0.0154	0.0220
DIST_CHAR	2	0.1366	0.0099	0.0244	0.0326	0.0048	0.0061
	2-3	0.1364	0.0224	0.0286	0.0257	0.0117	0.0144
	2-3-4	0.1078	0.0159	0.0183	0.0196	0.0110	-0.0038
	2-3-4-5	0.0934	0.0227	0.0179	0.0167	0.0084	0.0023
DIST_CHAR_ASIS	2	0.0901	-0.0056	0.0189	0.0107	0.0071	0.0017
	2-3	0.0815	0.0144	0.0157	0.0055	0.0102	0.0100
	2-3-4	0.0848	0.0158	0.0124	0.0120	0.0053	0.0009
	2-3-4-5	0.0831	0.0126	0.0198	0.0020	0.0067	0.0095
DIST_CHAR_ASIS_LEMMA	1	0.0753	0.0081	0.0143	0.0131	0.0077	0.0040
	1-2	0.0736	0.0087	0.0197	0.0150	0.0028	0.0084
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0680	0.0152	0.0044	0.0052	0.0056	0.0053
	1-2	0.0682	0.0210	0.0116	0.0111	0.0095	0.0098

Table A.27: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

A2.3.4 Minimum of Characters: 100 & Stacked

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7223	0.7220	0.6879	0.6864	0.6232	0.6230	0.6142	0.6136
	2-3	0.7658	0.7658	0.7159	0.7139	0.6594	0.6594	0.6426	0.6424
	2-3-4	0.7776	0.7776	0.7259	0.7253	0.6655	0.6655	0.6478	0.6475
	2-3-4-5	0.7732	0.7731	0.7249	0.7243	0.6638	0.6637	0.6477	0.6474
DIST_CHAR_ASIS	2	0.7761	0.7760	0.7217	0.7200	0.6657	0.6656	0.6487	0.6484
	2-3	0.7881	0.7880	0.7314	0.7300	0.6694	0.6693	0.6518	0.6516
	2-3-4	0.7914	0.7914	0.7361	0.7348	0.6723	0.6723	0.6543	0.6540
	2-3-4-5	0.7890	0.7889	0.7338	0.7324	0.6708	0.6707	0.6540	0.6539
DIST_CHAR_ASIS_LEMMA	1	0.7898	0.7897	0.7342	0.7329	0.6707	0.6707	0.6537	0.6535
	1-2	0.7913	0.7912	0.7349	0.7336	0.6702	0.6700	0.6537	0.6535
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7909	0.7909	0.7360	0.7347	0.6726	0.6726	0.6550	0.6548
	1-2	0.7906	0.7903	0.7357	0.7345	0.6722	0.6722	0.6549	0.6546

Table A.28: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	-0.0387	-0.0005	0.0055	-0.0134	0.0438	0.0420
	2-3	-0.0308	-0.0107	-0.0109	-0.0139	0.0442	0.0409
	2-3-4	-0.0293	-0.0093	-0.0037	-0.0112	0.0369	0.0404
	2-3-4-5	-0.0168	0.0059	-0.0041	-0.0094	0.0306	0.0362
DIST_CHAR_ASIS	2	-0.0102	0.0104	-0.0034	-0.0049	0.0243	0.0322
	2-3	-0.0078	0.0036	0.0263	-0.0042	-0.0069	0.0140
	2-3-4	-0.0236	-0.0164	-0.0019	-0.0071	0.0207	0.0259
	2-3-4-5	-0.0155	-0.0105	-0.0049	-0.0039	0.0203	0.0195
DIST_CHAR_ASIS_LEMMA	1	-0.0120	-0.0057	-0.0043	0.0013	0.0177	0.0192
	1-2	-0.0053	0.0045	-0.0064	-0.0022	0.0154	0.0157
DIST_CHAR_ASIS_LEMMA_WORD	1	-0.0031	0.0093	-0.0053	0.0002	0.0149	0.0158
	1-2	-0.0069	-0.0138	-0.0088	-0.0012	0.0225	0.0174

Table A.29: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 100 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.5814	0.5806	0.4549	0.4544	0.2092	0.0892	0.2029	0.0707
	2-3	0.6560	0.6565	0.5434	0.5426	0.4313	0.4281	0.2607	0.1626
	2-3-4	0.6726	0.6728	0.5630	0.5621	0.4346	0.4147	0.3547	0.2988
	2-3-4-5	0.6769	0.6773	0.5647	0.5638	0.4531	0.4464	0.3450	0.2828
DIST_CHAR_ASIS	2	0.6955	0.6954	0.5636	0.5627	0.4507	0.4473	0.3738	0.3267
	2-3	0.7126	0.7128	0.5758	0.5750	0.4604	0.4584	0.3820	0.3378
	2-3-4	0.7164	0.7163	0.5829	0.5822	0.4703	0.4649	0.3707	0.3001
	2-3-4-5	0.7156	0.7156	0.5835	0.5828	0.4695	0.4638	0.3786	0.3584
DIST_CHAR_ASIS_LEMMA	1	0.7158	0.7158	0.5845	0.5839	0.4703	0.4644	0.3673	0.3416
	1-2	0.7178	0.7177	0.5836	0.5829	0.4704	0.4645	0.3679	0.3422
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7168	0.7167	0.5840	0.5832	0.4706	0.4645	0.3827	0.3481
	1-2	0.7173	0.7172	0.5837	0.5830	0.4706	0.4646	0.3810	0.3449

Table A.30: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 100 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	-0.0261	-0.0180	-0.0041	0.0083	0.0156	0.0209
	2-3	-0.0058	-0.0031	-0.0170	-0.0032	0.0285	0.0297
	2-3-4	-0.0065	0.0028	0.0036	0.0192	-0.0086	-0.0014
	2-3-4-5	-0.0145	-0.0003	-0.0099	0.0049	0.0225	0.0185
DIST_CHAR_ASIS	2	0.0049	-0.0039	0.0051	0.0080	0.0112	0.0043
	2-3	-0.0002	-0.0051	-0.0048	0.0081	0.0121	0.0122
	2-3-4	0.0066	0.0059	-0.0038	0.0006	0.0089	0.0184
	2-3-4-5	0.0106	0.0062	-0.0028	0.0105	0.0121	0.0134
DIST_CHAR_ASIS_LEMMA	1	0.0093	0.0081	0.0066	0.0132	0.0015	0.0014
	1-2	0.0021	0.0123	0.0027	0.0058	0.0019	0.0066
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0032	-0.0001	0.0112	0.0060	0.0075	0.0043
	1-2	-0.0002	0.0096	0.0055	0.0137	0.0080	0.0116

Table A.31: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

A2.3.5 Minimum of Characters: 250 & Stacked

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 250 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.7737	0.7735	0.7414	0.7414	0.6647	0.6642	0.6455	0.6436
	2-3	0.8138	0.8130	0.7813	0.7813	0.7156	0.7156	0.6858	0.6849
	2-3-4	0.8218	0.8217	0.7912	0.7911	0.7261	0.7261	0.6933	0.6926
	2-3-4-5	0.8230	0.8230	0.7927	0.7926	0.7266	0.7266	0.6936	0.6928
DIST_CHAR_ASIS	2	0.8287	0.8287	0.7983	0.7983	0.7282	0.7280	0.6947	0.6940
	2-3	0.8397	0.8396	0.8076	0.8076	0.7323	0.7323	0.6996	0.6990
	2-3-4	0.8392	0.8391	0.8108	0.8106	0.7353	0.7353	0.7051	0.7050
	2-3-4-5	0.8359	0.8357	0.8114	0.8113	0.7345	0.7345	0.7046	0.7045
DIST_CHAR_ASIS_LEMMA	1	0.8428	0.8427	0.8090	0.8088	0.7338	0.7338	0.7041	0.7041
	1-2	0.8381	0.8379	0.8109	0.8108	0.7345	0.7345	0.7044	0.7043
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8387	0.8385	0.8118	0.8117	0.7370	0.7369	0.7050	0.7044
	1-2	0.8426	0.8425	0.8126	0.8125	0.7365	0.7364	0.7054	0.7054

Table A.32: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 250 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0315	0.0538	0.0180	0.0214	0.0739	0.0638
	2-3	0.0250	0.0498	0.0184	0.0121	0.0694	0.0464
	2-3-4	0.0295	0.0412	0.0155	0.0154	0.0480	0.0440
	2-3-4-5	0.0248	0.0451	0.0103	0.0112	0.0500	0.0452
DIST_CHAR_ASIS	2	0.0336	0.0458	0.0110	0.0130	0.0486	0.0443
	2-3	0.0315	0.0388	0.0124	0.0391	0.0262	0.0248
	2-3-4	0.0327	0.0429	0.0126	0.0066	0.0414	0.0400
	2-3-4-5	0.0557	-0.0074	0.0112	0.0075	0.0403	0.0336
DIST_CHAR_ASIS_LEMMA	1	0.0366	-0.0120	0.0111	0.0280	0.0339	0.0347
	1-2	0.0447	-0.0151	0.0100	0.0073	0.0339	0.0312
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0380	-0.0132	0.0121	0.0084	0.0332	0.0308
	1-2	0.0207	0.0447	0.0108	0.0081	0.0305	0.0263

Table A.33: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 250 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.6667	0.6672	0.5623	0.5621	0.4216	0.4205	0.3466	0.3196
	2-3	0.7480	0.7484	0.6506	0.6498	0.5183	0.5180	0.4535	0.4513
	2-3-4	0.7588	0.7592	0.6741	0.6735	0.5337	0.5308	0.4774	0.4757
	2-3-4-5	0.7627	0.7630	0.6722	0.6718	0.5325	0.5292	0.4829	0.4793
DIST_CHAR_ASIS	2	0.7759	0.7763	0.6793	0.6788	0.5455	0.5419	0.4849	0.4810
	2-3	0.7940	0.7943	0.6919	0.6914	0.5571	0.5567	0.4930	0.4892
	2-3-4	0.7911	0.7915	0.6996	0.6991	0.5553	0.5529	0.4994	0.4962
	2-3-4-5	0.7933	0.7933	0.6966	0.6962	0.5553	0.5530	0.5000	0.4972
DIST_CHAR_ASIS_LEMMA	1	0.7947	0.7948	0.6989	0.6986	0.5562	0.5540	0.5001	0.4976
	1-2	0.7959	0.7961	0.6956	0.6949	0.5596	0.5559	0.5002	0.4977
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7962	0.7964	0.6963	0.6956	0.5641	0.5630	0.5006	0.4985
	1-2	0.7977	0.7979	0.6961	0.6954	0.5613	0.5579	0.4975	0.4967

Table A.34: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 250 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0132	-0.0179	-0.0008	0.0021	0.0057	0.0131
	2-3	0.0076	-0.0155	0.0017	0.0131	0.0161	0.0122
	2-3-4	0.0248	0.0128	-0.0031	0.0107	0.0037	0.0109
	2-3-4-5	0.0154	—	0.0055	0.0092	0.0169	0.0218
DIST_CHAR_ASIS	2	0.0218	0.0024	-0.0060	-0.0045	0.0118	0.0127
	2-3	0.0311	-0.0075	0.0108	0.0040	0.0065	0.0082
	2-3-4	0.0289	-0.0034	0.0092	0.0006	-0.0009	-0.0029
	2-3-4-5	0.0335	0.0136	0.0001	0.0041	0.0002	—
DIST_CHAR_ASIS_LEMMA	1	0.0249	0.0066	0.0030	-0.0050	0.0019	0.0093
	1-2	0.0185	-0.0032	0.0022	-0.0045	0.0050	0.0150
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0150	0.0033	0.0011	0.0009	0.0057	0.0093
	1-2	0.0094	0.0105	0.0036	-0.0011	0.0085	0.0197

Table A.35: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

A2.3.6 Minimum of Characters: 500 & Stacked

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 500 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.8192	0.8179	0.7780	0.7778	0.6948	0.6935	0.6765	0.6764
	2-3	0.8743	0.8740	0.8369	0.8369	0.7639	0.7636	0.7286	0.7285
	2-3-4	0.8874	0.8874	0.8485	0.8485	0.7772	0.7769	0.7400	0.7399
	2-3-4-5	0.8884	0.8883	0.8501	0.8500	0.7773	0.7769	0.7411	0.7410
DIST_CHAR_ASIS	2	0.8920	0.8919	0.8527	0.8527	0.7767	0.7761	0.7412	0.7411
	2-3	0.8983	0.8982	0.8619	0.8619	0.7844	0.7841	0.7467	0.7466
	2-3-4	0.8832	0.8830	0.8670	0.8670	0.7862	0.7855	0.7505	0.7504
	2-3-4-5	0.8789	0.8787	0.8664	0.8663	0.7856	0.7850	0.7515	0.7513
DIST_CHAR_ASIS_LEMMA	1	0.9015	0.9013	0.8665	0.8665	0.7852	0.7846	0.7522	0.7521
	1-2	0.8983	0.8979	0.8662	0.8661	0.7858	0.7852	0.7521	0.7520
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8992	0.8992	0.8666	0.8666	0.7863	0.7856	0.7524	0.7524
	1-2	0.8986	0.8983	0.8664	0.8663	0.7862	0.7855	0.7519	0.7519

Table A.36: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 500 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.1274	0.0535	0.0573	0.0718	0.0552	0.0749
	2-3	0.1182	0.0385	0.0420	0.0543	0.0388	0.0574
	2-3-4	0.1085	0.0315	0.0355	0.0490	0.0417	0.0531
	2-3-4-5	0.0951	0.0258	0.0311	0.0432	0.0336	0.0474
DIST_CHAR_ASIS	2	0.1148	0.0277	0.0340	0.0439	0.0364	0.0494
	2-3	0.1371	-0.0178	0.0286	0.0425	0.0308	0.0454
	2-3-4	0.1046	0.0192	0.0263	0.0358	0.0282	0.0381
	2-3-4-5	0.0987	0.0262	0.0281	0.0320	0.0272	0.0380
DIST_CHAR_ASIS_LEMMA	1	0.1079	-0.0122	0.0277	0.0331	0.0282	0.0344
	1-2	0.0976	-0.0065	0.0255	0.0252	0.0299	0.0284
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0886	-0.0027	0.0247	0.0228	0.0312	0.0300
	1-2	0.0831	-0.0038	0.0293	0.0243	0.0300	0.0239

Table A.37: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7413	0.7422	0.6347	0.6350	0.4899	0.4848	0.4263	0.4239
	2-3	0.8250	0.8252	0.7357	0.7356	0.5868	0.5840	0.5140	0.5114
	2-3-4	0.8415	0.8416	0.7623	0.7622	0.6156	0.6141	0.5433	0.5432
	2-3-4-5	0.8397	0.8397	0.7605	0.7603	0.6172	0.6161	0.5433	0.5416
DIST_CHAR_ASIS	2	0.8473	0.8472	0.7703	0.7701	0.6201	0.6187	0.5489	0.5469
	2-3	0.8651	0.8652	0.7795	0.7795	0.6280	0.6258	0.5550	0.5537
	2-3-4	0.8749	0.8748	0.7912	0.7911	0.6400	0.6396	0.5630	0.5630
	2-3-4-5	0.8700	0.8700	0.7878	0.7876	0.6372	0.6357	0.5623	0.5623
DIST_CHAR_ASIS_LEMMA	1	0.8700	0.8696	0.7843	0.7834	0.6392	0.6368	0.5623	0.5624
	1-2	0.8727	0.8723	0.7833	0.7824	0.6391	0.6366	0.5628	0.5628
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8695	0.8691	0.7899	0.7897	0.6389	0.6365	0.5633	0.5634
	1-2	0.8722	0.8725	0.7820	0.7810	0.6382	0.6359	0.5629	0.5630

Table A.38: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0733	0.0228	0.0064	0.0045	-0.0029	0.0157
	2-3	0.0645	-0.0005	0.0075	0.0223	-0.0016	0.0015
	2-3-4	0.0493	0.0004	0.0092	0.0073	0.0037	0.0017
	2-3-4-5	0.0451	0.0075	0.0236	0.0156	-0.0009	0.0045
DIST_CHAR_ASIS	2	0.0799	0.0056	0.0117	0.0167	0.0066	0.0003
	2-3	0.0807	0.0021	0.0149	0.0020	0.0013	0.0205
	2-3-4	0.0751	0.0083	-0.0057	-0.0054	0.0046	0.0045
	2-3-4-5	0.0688	-0.0053	0.0013	0.0040	0.0018	0.0051
DIST_CHAR_ASIS_LEMMA	1	0.0628	0.0107	0.0042	0.0054	0.0044	-0.0015
	1-2	0.0622	0.0067	-0.0011	0.0012	-0.0054	0.0015
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0553	0.0101	0.0061	0.0045	-0.0038	0.0071
	1-2	0.0517	0.0057	0.0058	0.0032	-0.0036	0.0008

Table A.39: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

A2.4 Full Set of feature types

A2.4.1 Minimum of Characters: 100 & Cumulated

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 100 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.7343	0.7338	0.6967	0.6947	0.6284	0.6276	0.6089	0.5966
	2-3	0.7746	0.7738	0.7331	0.7330	0.6647	0.6645	0.6387	0.6301
	2-3-4	0.7882	0.7881	0.7379	0.7370	0.6712	0.6711	0.6412	0.6312
	2-3-4-5	0.7836	0.7829	0.7354	0.7342	0.6710	0.6710	0.6409	0.6305
DIST_CHAR_ASIS	2	0.7905	0.7904	0.7453	0.7453	0.6775	0.6775	0.6452	0.6340
	2-3	0.7973	0.7972	0.7537	0.7534	0.6835	0.6833	0.6645	0.6645
	2-3-4	0.7913	0.7908	0.7478	0.7459	0.6849	0.6848	0.6660	0.6660
	2-3-4-5	0.7980	0.7979	0.7461	0.7437	0.6844	0.6844	0.6658	0.6658
DIST_CHAR_ASIS_POS	1	0.7994	0.7994	0.7519	0.7516	0.6834	0.6826	0.6660	0.6660
	1-2	0.8012	0.8012	0.7475	0.7462	0.6855	0.6855	0.6663	0.6662
	1-2-3	0.7959	0.7949	0.7500	0.7489	0.6861	0.6857	0.6681	0.6681
DIST_CHAR_ASIS_POS_TAG	1	0.7933	0.7920	0.7483	0.7469	0.6871	0.6871	0.6683	0.6682
	1-2	0.7972	0.7968	0.7484	0.7478	0.6863	0.6863	0.6694	0.6694
	1-2-3	0.7988	0.7988	0.7512	0.7511	0.6846	0.6828	0.6701	0.6698
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7954	0.7952	0.7530	0.7529	0.6861	0.6849	0.6700	0.6696
	1-2	0.7913	0.7912	0.7509	0.7507	0.6879	0.6879	0.6703	0.6701
	1-2-3	0.7939	0.7937	0.7534	0.7533	0.6877	0.6872	0.6713	0.6712
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7919	0.7918	0.7544	0.7544	0.6872	0.6858	0.6738	0.6738
	1-2	0.7929	0.7927	0.7559	0.7555	0.6876	0.6861	0.6739	0.6739
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7898	0.7895	0.7574	0.7573	0.6921	0.6921	0.6749	0.6747
	1-2	0.7883	0.7883	0.7510	0.7503	0.6913	0.6913	0.6751	0.6749

Table A.40: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 100 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.1683	0.2888	-0.0348	-0.0259	-0.0312	-0.0785
	2-3	0.0446	0.0485	0.0274	0.0075	0.1002	0.0619
	2-3-4	0.0159	0.0471	0.0272	0.0942	0.0405	-0.0265
	2-3-4-5	-0.0486	-0.0061	-0.0090	-0.0183	0.0560	0.0661
DIST_CHAR	2	-0.0481	-0.0285	-0.0020	-0.0628	0.0686	0.0497
	2-3	-0.0355	0.0028	0.0120	0.0464	0.0665	0.0553
	2-3-4	-0.0022	0.0259	0.0037	0.0052	0.0198	0.0238
	2-3-4-5	-0.0103	0.0157	0.0228	0.0138	0.0165	0.0021
DIST_CHAR_ASIS	2	-0.0215	0.0133	-0.0312	0.0060	0.0065	0.0232
	2-3	-0.0055	-0.0018	0.0044	0.0263	-0.0050	0.0227
	2-3-4	-0.0265	-0.0274	0.0009	-0.0053	0.0245	0.0168
	2-3-4-5	-0.0284	-0.0212	-0.0118	-0.0198	-0.0143	-0.0140
DIST_CHAR_ASIS_POS	1	0.0138	0.0230	-0.0061	0.0064	0.0482	0.0538
	1-2	-0.0100	-0.0181	0.0198	0.0108	0.0833	0.0697
	1-2-3	-0.0009	0.0076	0.0354	0.0258	0.0346	0.0240
DIST_CHAR_ASIS_POS_TAG	1	-0.0063	0.0042	0.0069	-0.0115	0.0310	0.0311
	1-2	-0.0366	-0.0332	-0.0115	-0.0066	0.0082	0.0118
	1-2-3	-0.0141	-0.0133	0.0296	0.0455	0.0383	0.0351
DIST_CHAR_ASIS_POS_TAG_DEP	1	-0.0154	-0.0017	0.0028	-0.0029	0.0159	0.0247
	1-2	-0.0077	0.0165	-0.0028	0.0099	0.0089	0.0170
	1-2-3	0.0008	0.0040	0.0022	0.0015	-0.0248	-0.0046
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0058	0.0071	0.0076	0.0101	0.0037	-0.0090
	1-2	-0.0004	0.0174	0.0129	-0.0012	0.0129	0.0178
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0031	0.0001	0.0124	0.0225	0.0171	0.0133
	1-2	0.0016	-0.0074	0.0278	0.0162	0.0380	0.0168

Table A.41: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 100 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.5983	0.5994	0.4887	0.4861	0.3739	0.3555	0.3088	0.2673
	2-3	0.6907	0.6908	0.5595	0.5583	0.4424	0.4382	0.3888	0.3706
	2-3-4	0.7031	0.7037	0.5749	0.5714	0.4523	0.4435	0.4023	0.3817
	2-3-4-5	0.6804	0.6827	0.5378	0.5386	0.4624	0.4589	0.4017	0.3870
DIST_CHAR_ASIS	2	0.7161	0.7161	0.5828	0.5810	0.4677	0.4662	0.4223	0.4062
	2-3	0.7275	0.7276	0.5944	0.5926	0.4759	0.4726	0.4336	0.4241
	2-3-4	0.7270	0.7267	0.6014	0.6002	0.4744	0.4672	0.4309	0.4160
	2-3-4-5	0.7292	0.7295	0.5952	0.5939	0.4788	0.4738	0.4320	0.4203
DIST_CHAR_ASIS_POS	1	0.7229	0.7231	0.6008	0.5990	0.4822	0.4804	0.4324	0.4223
	1-2	0.7283	0.7283	0.5954	0.5942	0.4743	0.4679	0.4354	0.4266
	1-2-3	0.7053	0.7041	0.5854	0.5851	0.4850	0.4815	0.4392	0.4313
DIST_CHAR_ASIS_POS_TAG	1	0.7159	0.7162	0.5971	0.5952	0.4783	0.4756	0.4294	0.4164
	1-2	0.7004	0.6989	0.5946	0.5937	0.4823	0.4776	0.4364	0.4297
	1-2-3	0.6975	0.6957	0.5950	0.5931	0.4875	0.4843	0.4380	0.4277
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7124	0.7132	0.5923	0.5914	0.4856	0.4817	0.4367	0.4257
	1-2	0.7081	0.7076	0.5920	0.5910	0.4843	0.4809	0.4338	0.4204
	1-2-3	0.7075	0.7080	0.5952	0.5940	0.4865	0.4838	0.4382	0.4341
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.6861	0.6845	0.5950	0.5939	0.4885	0.4842	0.4362	0.4247
	1-2	0.7079	0.7082	0.5969	0.5956	0.4884	0.4844	0.4371	0.4269
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7058	0.7055	0.5969	0.5957	0.4890	0.4857	0.4349	0.4218
	1-2	0.7046	0.7043	0.5893	0.5879	0.4887	0.4856	0.4357	0.4228

Table A.42: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 100 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0348	-0.0149	0.0099	0.0315	-0.0141	0.0174
	2-3	-0.0051	-0.0070	-0.0063	-0.0108	0.0126	0.0458
	2-3-4	-0.0033	0.0366	0.0245	-0.0002	0.0020	0.0207
	2-3-4-5	0.0213	0.0143	-0.0097	-0.0042	-0.0200	-0.0321
DIST_CHAR	2	0.0013	-0.0092	0.0077	0.0500	0.0047	0.0328
	2-3	0.0064	0.0009	-0.0084	-0.0053	0.0196	0.0099
	2-3-4	-0.0009	0.0049	0.0101	0.0123	-0.0004	-0.0037
	2-3-4-5	0.0157	0.0139	0.0107	-0.0099	0.0016	0.0008
DIST_CHAR_ASIS	2	0.0108	0.0058	-0.0019	0.0102	0.0090	0.0187
	2-3	0.0089	0.0029	0.0068	0.0036	0.0053	0.0122
	2-3-4	0.0119	0.0045	0.0045	0.0047	0.0087	0.0154
	2-3-4-5	0.0074	0.0087	-0.0049	-0.0032	0.0018	0.0034
DIST_CHAR_ASIS_POS	1	0.0141	0.0144	0.0166	0.0247	0.0187	0.0186
	1-2	0.0264	0.0325	0.0420	0.0360	0.0075	0.0030
	1-2-3	0.0127	0.0213	0.0150	0.0151	0.0125	0.0151
DIST_CHAR_ASIS_POS_TAG	1	0.0108	0.0118	0.0413	0.0368	0.0266	0.0279
	1-2	0.0067	0.0070	0.0243	0.0161	0.0117	0.0148
	1-2-3	0.0105	0.0177	0.0271	0.0232	0.0076	0.0183
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0139	0.0094	0.0129	0.0161	0.0093	0.0124
	1-2	0.0209	0.0192	0.0183	0.0169	0.0102	0.0107
	1-2-3	0.0229	0.0251	0.0072	0.0043	-0.0013	0.0013
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0245	0.0257	0.0189	0.0198	0.0154	0.0243
	1-2	0.0081	0.0135	0.0135	0.0140	0.0276	0.0304
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	-0.0017	-0.0044	0.0071	0.0104	0.0224	0.0262
	1-2	0.0203	0.0132	0.0189	0.0189	0.0241	0.0298

Table A.43: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

A2.4.2 Minimum of Characters: 250 & Cumulated

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 250 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.7844	0.7841	0.7495	0.7493	0.6736	0.6736	0.6514	0.6511
	2-3	0.8249	0.8247	0.7771	0.7738	0.7219	0.7218	0.6916	0.6916
	2-3-4	0.8310	0.8309	0.7884	0.7858	0.7306	0.7305	0.7001	0.7001
	2-3-4-5	0.8316	0.8315	0.7943	0.7927	0.7283	0.7279	0.6995	0.6995
DIST_CHAR_ASIS	2	0.8434	0.8434	0.8019	0.8002	0.7419	0.7418	0.7084	0.7084
	2-3	0.8423	0.8418	0.8120	0.8104	0.7475	0.7473	0.7166	0.7165
	2-3-4	0.8499	0.8499	0.8260	0.8260	0.7497	0.7495	0.7176	0.7172
	2-3-4-5	0.8479	0.8479	0.8243	0.8243	0.7485	0.7484	0.7173	0.7168
DIST_CHAR_ASIS_POS	1	0.8483	0.8483	0.8155	0.8145	0.7474	0.7470	0.7171	0.7171
	1-2	0.8439	0.8436	0.8170	0.8163	0.7486	0.7485	0.7184	0.7183
	1-2-3	0.8430	0.8430	0.8138	0.8128	0.7505	0.7505	0.7193	0.7193
DIST_CHAR_ASIS_POS_TAG	1	0.8441	0.8440	0.8159	0.8152	0.7472	0.7459	0.7176	0.7174
	1-2	0.8390	0.8389	0.8169	0.8168	0.7509	0.7509	0.7185	0.7184
	1-2-3	0.8377	0.8377	0.8134	0.8132	0.7521	0.7520	0.7185	0.7180
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8345	0.8342	0.8131	0.8126	0.7505	0.7505	0.7210	0.7210
	1-2	0.8345	0.8345	0.8152	0.8151	0.7513	0.7513	0.7169	0.7152
	1-2-3	0.8327	0.8327	0.8142	0.8141	0.7518	0.7516	0.7205	0.7203
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8298	0.8291	0.8173	0.8171	0.7429	0.7399	0.7178	0.7150
	1-2	0.8332	0.8328	0.8162	0.8161	0.7386	0.7340	0.7206	0.7186
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8294	0.8286	0.8133	0.8133	0.7559	0.7555	0.7232	0.7226
	1-2	0.8289	0.8286	0.8064	0.8051	0.7562	0.7558	0.7252	0.7251

Table A.44: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 250 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0390	0.1945	0.0539	0.0938	0.2805	0.2587
	2-3	0.1318	0.1969	0.0012	0.1452	0.1234	0.1113
	2-3-4	0.0671	0.1703	0.0232	0.0109	0.1353	0.0908
	2-3-4-5	0.0076	0.0844	0.0169	0.0175	0.0812	0.0675
DIST_CHAR	2	0.0017	0.0251	0.0348	0.0854	0.0363	0.0600
	2-3	-0.0013	0.0141	-0.0077	0.0357	0.0245	0.0117
	2-3-4	0.0015	0.0345	0.0257	0.0152	0.0389	0.0345
	2-3-4-5	-0.0060	0.0096	0.0168	0.0283	0.0403	0.0370
DIST_CHAR_ASIS	2	0.0181	0.0458	0.0196	-0.0173	-0.0216	-0.0147
	2-3	-0.0017	0.0099	0.0257	0.0034	0.0041	-0.0010
	2-3-4	0.0151	-0.0199	0.0072	-0.0106	0.0263	0.0332
	2-3-4-5	0.0060	0.0030	0.0127	-0.0121	0.0074	-0.0042
DIST_CHAR_ASIS_POS	1	0.0431	0.0190	-0.0015	-0.0213	0.0100	0.0060
	1-2	0.0327	0.0124	-0.0106	-0.0153	0.0417	0.0422
	1-2-3	0.0142	0.0324	0.0102	-0.0236	0.0124	0.0116
DIST_CHAR_ASIS_POS_TAG	1	-0.0230	-0.0049	0.0602	0.0363	0.0459	0.0400
	1-2	-0.0051	-0.0015	-0.0061	-0.0297	0.0271	0.0223
	1-2-3	0.0022	0.0010	-0.0143	-0.0176	0.0368	0.0398
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0084	0.0033	-0.0038	-0.0138	0.0160	0.0227
	1-2	0.0184	0.0234	0.0126	0.0124	0.0381	0.0358
	1-2-3	0.0184	0.0299	-0.0325	-0.0191	0.0387	0.0331
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0384	0.0402	-0.0095	-0.0214	0.0229	0.0133
	1-2	0.0124	0.0285	0.0307	0.0042	0.0130	0.0092
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0093	0.0051	0.0216	0.0340	0.0195	0.0214
	1-2	0.0252	0.0271	0.0238	0.0170	-0.0156	-0.0159

Table A.45: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 250 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.6772	0.6770	0.5772	0.5772	0.4543	0.4439	0.3928	0.3686
	2-3	0.7627	0.7629	0.6684	0.6689	0.5239	0.5115	0.4643	0.4581
	2-3-4	0.7534	0.7555	0.6707	0.6732	0.5471	0.5440	0.4860	0.4795
	2-3-4-5	0.7380	0.7413	0.6847	0.6848	0.5454	0.5361	0.4927	0.4875
DIST_CHAR_ASIS	2	0.7749	0.7751	0.6993	0.6996	0.5607	0.5567	0.4827	0.4712
	2-3	0.7918	0.7920	0.7083	0.7067	0.5742	0.5705	0.5002	0.4942
	2-3-4	0.7981	0.7996	0.7140	0.7131	0.5773	0.5736	0.4982	0.4902
	2-3-4-5	0.7928	0.7941	0.7059	0.7071	0.5862	0.5845	0.5049	0.5020
DIST_CHAR_ASIS_POS	1	0.7622	0.7611	0.6947	0.6963	0.5817	0.5785	0.4990	0.4899
	1-2	0.7842	0.7846	0.7106	0.7104	0.5852	0.5838	0.4983	0.4921
	1-2-3	0.7857	0.7854	0.7052	0.7056	0.5856	0.5842	0.4992	0.4968
DIST_CHAR_ASIS_POS_TAG	1	0.7847	0.7851	0.7071	0.7068	0.5749	0.5713	0.4929	0.4799
	1-2	0.7752	0.7764	0.7012	0.7016	0.5699	0.5649	0.5002	0.4960
	1-2-3	0.7825	0.7829	0.7008	0.7012	0.5694	0.5635	0.5064	0.4999
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7744	0.7747	0.6959	0.6970	0.5797	0.5749	0.4959	0.4950
	1-2	0.7713	0.7721	0.6947	0.6960	0.5842	0.5829	0.4969	0.4925
	1-2-3	0.7823	0.7825	0.6981	0.6980	0.5832	0.5815	0.5061	0.5003
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7666	0.7668	0.6957	0.6956	0.5766	0.5709	0.5078	0.5031
	1-2	0.7642	0.7647	0.7027	0.7019	0.5735	0.5673	0.5108	0.5060
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7652	0.7674	0.6895	0.6883	0.5845	0.5821	0.5070	0.5042
	1-2	0.7686	0.7692	0.6882	0.6875	0.5860	0.5842	0.5078	0.5019

Table A.46: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 250 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0416	0.1495	0.0072	0.0051	0.0153	0.0489
	2-3	-0.0223	0.0252	-0.0132	0.0114	0.0086	-0.0058
	2-3-4	0.0028	0.0019	0.0258	0.0180	0.0279	0.0153
	2-3-4-5	0.0040	0.0345	0.0270	0.0137	-0.0079	-0.0130
DIST_CHAR	2	0.0620	0.0267	0.0122	0.0015	0.0181	0.0090
	2-3	0.0686	0.0548	0.0189	0.0225	0.0152	0.0079
	2-3-4	0.0693	0.0148	0.0045	0.0058	0.0072	0.0068
	2-3-4-5	0.0496	0.0082	0.0093	0.0005	-0.0054	-0.0113
DIST_CHAR_ASIS	2	0.0390	0.0173	0.0187	0.0134	0.0050	0.0077
	2-3	0.0346	0.0165	0.0025	0.0088	0.0027	0.0066
	2-3-4	0.0333	0.0131	0.0004	0.0048	0.0168	0.0067
	2-3-4-5	0.0450	0.0078	-0.0004	0.0097	0.0206	-0.0040
DIST_CHAR_ASIS_POS	1	0.0487	0.0185	0.0226	0.0315	0.0268	0.0170
	1-2	0.0371	0.0085	0.0278	0.0285	0.0101	0.0091
	1-2-3	0.0359	0.0068	0.0253	0.0272	0.0106	0.0085
DIST_CHAR_ASIS_POS_TAG	1	0.0467	0.0368	0.0249	0.0209	0.0238	0.0231
	1-2	0.0567	0.0221	0.0192	0.0230	0.0132	0.0162
	1-2-3	0.0444	0.0179	0.0295	0.0261	0.0359	0.0319
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0169	-0.0045	0.0155	0.0171	0.0337	0.0358
	1-2	0.0212	0.0112	0.0235	0.0135	0.0254	0.0214
	1-2-3	0.0463	0.0107	0.0254	0.0261	0.0086	0.0097
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0495	0.0303	0.0171	0.0123	0.0171	0.0151
	1-2	0.0392	0.0134	0.0085	0.0048	0.0258	0.0210
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0342	0.0096	0.0089	0.0097	0.0285	0.0263
	1-2	0.0310	0.0073	0.0136	0.0129	0.0133	0.0160

Table A.47: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

A2.4.3 Minimum of Characters: 500 & Cumulated

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 500 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.8231	0.8230	0.7868	0.7868	0.7062	0.7058	0.6795	0.6769
	2-3	0.8802	0.8802	0.8396	0.8392	0.7698	0.7696	0.7313	0.7311
	2-3-4	0.8727	0.8712	0.8498	0.8494	0.7852	0.7852	0.7407	0.7392
	2-3-4-5	0.8953	0.8951	0.8583	0.8583	0.7679	0.7650	0.7433	0.7429
DIST_CHAR_ASIS	2	0.8887	0.8882	0.8623	0.8623	0.7521	0.7441	0.7511	0.7494
	2-3	0.8861	0.8850	0.8725	0.8725	0.8036	0.8036	0.7612	0.7602
	2-3-4	0.8779	0.8764	0.8734	0.8733	0.7700	0.7641	0.7676	0.7675
	2-3-4-5	0.8947	0.8942	0.8730	0.8729	0.7688	0.7626	0.7674	0.7673
DIST_CHAR_ASIS_POS	1	0.9048	0.9045	0.8717	0.8717	0.8043	0.8042	0.7660	0.7656
	1-2	0.8976	0.8975	0.8670	0.8668	0.8040	0.8040	0.7678	0.7677
	1-2-3	0.8937	0.8936	0.8701	0.8700	0.7985	0.7976	0.7660	0.7650
DIST_CHAR_ASIS_POS_TAG	1	0.8996	0.8996	0.8696	0.8696	0.7980	0.7971	0.7687	0.7687
	1-2	0.8881	0.8881	0.8678	0.8677	0.7755	0.7704	0.7695	0.7694
	1-2-3	0.8792	0.8786	0.8640	0.8639	0.8058	0.8057	0.7692	0.7691
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8740	0.8729	0.8676	0.8676	0.8056	0.8055	0.7698	0.7698
	1-2	0.8615	0.8596	0.8634	0.8633	0.8037	0.8031	0.7698	0.7697
	1-2-3	0.8677	0.8664	0.8615	0.8615	0.7804	0.7761	0.7710	0.7709
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8835	0.8835	0.8628	0.8628	0.8050	0.8049	0.7754	0.7754
	1-2	0.8786	0.8786	0.8594	0.8591	0.7841	0.7801	0.7707	0.7695
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8704	0.8698	0.8639	0.8636	0.8057	0.8055	0.7759	0.7757
	1-2	0.8819	0.8814	0.8644	0.8644	0.8062	0.8060	0.7769	0.7767

Table A.48: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 500 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.0858	0.1370	0.1528	0.2654	0.2834	0.2834
	2-3	0.1535	0.1134	0.1350	0.1049	0.1451	0.1875
	2-3-4	0.1185	-0.0050	0.0893	0.0610	0.0425	0.0354
	2-3-4-5	0.0904	0.0550	0.0491	0.0644	0.0457	0.0703
DIST_CHAR	2	0.0911	0.0265	0.0598	0.0466	0.0998	0.0889
	2-3	0.0903	-0.0365	0.0331	0.0278	0.0178	0.0419
	2-3-4	0.0990	0.0314	0.0445	0.0261	0.0454	0.0362
	2-3-4-5	0.0829	-0.0121	0.0455	0.0365	0.0461	0.0232
DIST_CHAR_ASIS	2	0.0815	-0.0217	0.0306	0.0077	0.0118	0.0141
	2-3	0.0931	0.0012	0.0462	0.0171	0.0295	0.0138
	2-3-4	0.0789	-0.0137	0.0380	0.0166	0.0026	0.0088
	2-3-4-5	0.0829	0.0014	0.0459	0.0050	0.0281	0.0243
DIST_CHAR_ASIS_POS	1	0.1044	0.0228	0.0343	-0.0083	0.0389	0.0358
	1-2	0.0661	0.0052	-0.0052	-0.0298	0.0173	0.0290
	1-2-3	0.0603	0.0113	0.0389	0.0158	0.0362	0.0372
DIST_CHAR_ASIS_POS_TAG	1	0.0473	-0.0021	0.0340	0.0171	0.0295	0.0218
	1-2	0.0578	-0.0141	0.0384	0.0072	-0.0040	-0.0065
	1-2-3	0.0267	-0.0247	0.0100	-0.0060	0.0340	0.0434
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0737	0.0128	0.0054	-0.0020	0.0309	0.0268
	1-2	0.0797	0.0354	-0.0123	-0.0162	0.0159	0.0262
	1-2-3	0.0488	0.0061	0.0163	-0.0007	0.0464	0.0227
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0535	-0.0062	0.0102	-0.0180	0.0378	0.0332
	1-2	0.0436	0.0064	0.0343	0.0316	0.0089	0.0107
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0380	-0.0058	0.0038	-0.0200	0.0358	0.0473
	1-2	0.0617	0.0180	0.0094	0.0061	0.0258	0.0262

Table A.49: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7262	0.7271	0.6350	0.6349	0.5133	0.5123	0.4465	0.4434
	2-3	0.8192	0.8214	0.7342	0.7331	0.5938	0.5898	0.5199	0.5177
	2-3-4	0.8321	0.8343	0.7623	0.7627	0.6151	0.6114	0.5437	0.5390
	2-3-4-5	0.8232	0.8250	0.7590	0.7595	0.6186	0.6140	0.5451	0.5397
DIST_CHAR_ASIS	2	0.8508	0.8518	0.7725	0.7727	0.6203	0.6150	0.5566	0.5515
	2-3	0.8727	0.8726	0.7954	0.7954	0.6640	0.6629	0.5724	0.5687
	2-3-4	0.8753	0.8753	0.8017	0.8014	0.6618	0.6610	0.5643	0.5553
	2-3-4-5	0.8673	0.8671	0.8007	0.8006	0.6421	0.6376	0.5764	0.5731
DIST_CHAR_ASIS_POS	1	0.8633	0.8632	0.7846	0.7846	0.6552	0.6541	0.5766	0.5736
	1-2	0.8736	0.8733	0.7838	0.7841	0.6565	0.6545	0.5797	0.5772
	1-2-3	0.8451	0.8466	0.7900	0.7893	0.6411	0.6362	0.5611	0.5498
DIST_CHAR_ASIS_POS_TAG	1	0.8526	0.8531	0.7741	0.7742	0.6620	0.6614	0.5714	0.5656
	1-2	0.8575	0.8579	0.7851	0.7850	0.6645	0.6640	0.5764	0.5727
	1-2-3	0.8362	0.8383	0.7776	0.7770	0.6530	0.6500	0.5609	0.5498
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8388	0.8391	0.7795	0.7792	0.6606	0.6606	0.5829	0.5806
	1-2	0.8482	0.8487	0.7723	0.7717	0.6534	0.6517	0.5753	0.5719
	1-2-3	0.8388	0.8399	0.7694	0.7683	0.6360	0.6308	0.5651	0.5569
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8495	0.8486	0.7737	0.7735	0.6416	0.6376	0.5705	0.5625
	1-2	0.8531	0.8526	0.7795	0.7790	0.6531	0.6509	0.5803	0.5756
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8353	0.8368	0.7718	0.7715	0.6654	0.6651	0.5751	0.5708
	1-2	0.8575	0.8575	0.7755	0.7751	0.6598	0.6591	0.5676	0.5600

Table A.50: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST	2	0.1872	0.0727	0.0083	0.0109	0.0192	0.0501
	2-3	0.1142	0.0359	0.0275	0.0294	0.0478	0.0012
	2-3-4	0.0268	0.0157	0.0502	0.0416	0.0099	0.0147
	2-3-4-5	0.0219	0.0064	0.0371	0.0096	0.0154	0.0220
DIST_CHAR	2	0.1366	0.0099	0.0244	0.0326	0.0048	0.0061
	2-3	0.1364	0.0224	0.0286	0.0257	0.0117	0.0144
	2-3-4	0.1078	0.0159	0.0183	0.0196	0.0110	-0.0038
	2-3-4-5	0.0934	0.0227	0.0179	0.0167	0.0084	0.0023
DIST_CHAR_ASIS	2	0.0901	-0.0056	0.0189	0.0107	0.0071	0.0017
	2-3	0.0815	0.0144	0.0157	0.0055	0.0102	0.0100
	2-3-4	0.0848	0.0158	0.0124	0.0120	0.0053	0.0009
	2-3-4-5	0.0831	0.0126	0.0198	0.0020	0.0067	0.0095
DIST_CHAR_ASIS_POS	1	0.0836	0.0140	0.0336	0.0287	0.0294	0.0322
	1-2	0.0750	0.0146	0.0199	0.0174	0.0331	0.0223
	1-2-3	0.0830	0.0122	0.0178	0.0230	0.0221	0.0137
DIST_CHAR_ASIS_POS_TAG	1	0.0816	0.0207	0.0313	0.0318	0.0364	0.0321
	1-2	0.0806	0.0242	0.0142	0.0162	0.0363	0.0337
	1-2-3	0.0773	0.0271	0.0191	0.0185	0.0177	0.0214
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0728	0.0171	0.0158	0.0135	0.0362	0.0267
	1-2	0.0760	0.0222	0.0086	0.0084	0.0380	0.0439
	1-2-3	0.0780	0.0198	0.0306	0.0297	0.0268	0.0312
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0704	0.0164	0.0138	0.0153	0.0159	0.0129
	1-2	0.0574	0.0144	0.0129	0.0065	0.0227	0.0188
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0544	0.0200	0.0196	0.0187	0.0297	0.0340
	1-2	0.0652	0.0201	0.0148	0.0152	0.0180	0.0197

Table A.51: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

A2.4.4 Minimum of Characters: 100 & Stacked

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 100 / 50 / Accuracy	F1-score	150 / Accuracy	F1-score	500 / Accuracy	F1-score	1000 / Accuracy	F1-score
DIST_CHAR	2	0.7223	0.7220	0.6879	0.6864	0.6232	0.6230	0.6142	0.6136
	2-3	0.7658	0.7658	0.7159	0.7139	0.6594	0.6594	0.6426	0.6424
	2-3-4	0.7776	0.7776	0.7259	0.7253	0.6655	0.6655	0.6478	0.6475
	2-3-4-5	0.7732	0.7731	0.7249	0.7243	0.6638	0.6637	0.6477	0.6474
DIST_CHAR_ASIS	2	0.7761	0.7760	0.7217	0.7200	0.6657	0.6656	0.6487	0.6484
	2-3	0.7881	0.7880	0.7314	0.7300	0.6694	0.6693	0.6518	0.6516
	2-3-4	0.7914	0.7914	0.7361	0.7348	0.6723	0.6723	0.6543	0.6540
	2-3-4-5	0.7890	0.7889	0.7338	0.7324	0.6708	0.6707	0.6540	0.6539
DIST_CHAR_ASIS_POS	1	0.7893	0.7893	0.7334	0.7318	0.6709	0.6708	0.6535	0.6532
	1-2	0.7889	0.7888	0.7340	0.7325	0.6708	0.6707	0.6533	0.6530
	1-2-3	0.7903	0.7903	0.7335	0.7320	0.6711	0.6711	0.6537	0.6534
DIST_CHAR_ASIS_POS_TAG	1	0.7889	0.7888	0.7340	0.7324	0.6721	0.6721	0.6546	0.6542
	1-2	0.7898	0.7897	0.7329	0.7314	0.6709	0.6709	0.6537	0.6534
	1-2-3	0.7900	0.7900	0.7329	0.7313	0.6717	0.6716	0.6538	0.6535
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7890	0.7889	0.7335	0.7319	0.6724	0.6724	0.6541	0.6539
	1-2	0.7892	0.7891	0.7295	0.7273	0.6720	0.6720	0.6537	0.6534
	1-2-3	0.7911	0.7910	0.7324	0.7306	0.6720	0.6720	0.6537	0.6534
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7907	0.7906	0.7326	0.7310	0.6718	0.6718	0.6539	0.6536
	1-2	0.7912	0.7911	0.7328	0.7311	0.6718	0.6718	0.6537	0.6534
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7927	0.7926	0.7340	0.7324	0.6733	0.6733	0.6552	0.6549
	1-2	0.7924	0.7923	0.7336	0.7320	0.6724	0.6723	0.6550	0.6548

Table A.52: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

Feature types	Target / Min. No. of Characters / No. of Authors / Score / N-gram ranges	Gender / 100 / 150 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 / Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	-0.0387	-0.0005	0.0055	-0.0134	0.0438	0.0420
	2-3	-0.0308	-0.0107	-0.0109	-0.0139	0.0442	0.0409
	2-3-4	-0.0293	-0.0093	-0.0037	-0.0112	0.0369	0.0404
	2-3-4-5	-0.0168	0.0059	-0.0041	-0.0094	0.0306	0.0362
DIST_CHAR_ASIS	2	-0.0102	0.0104	-0.0034	-0.0049	0.0243	0.0322
	2-3	-0.0078	0.0036	0.0263	-0.0042	-0.0069	0.0140
	2-3-4	-0.0236	-0.0164	-0.0019	-0.0071	0.0207	0.0259
	2-3-4-5	-0.0155	-0.0105	-0.0049	-0.0039	0.0203	0.0195
DIST_CHAR_ASIS_POS	1	-0.0262	-0.0216	-0.0101	-0.0092	0.0001	-0.0007
	1-2	-0.0185	-0.0130	0.0032	0.0068	0.0391	0.0421
	1-2-3	-0.0077	0.0013	-0.0343	-0.0375	0.0188	0.0232
DIST_CHAR_ASIS_POS_TAG	1	-0.0131	-0.0047	-0.0303	-0.0333	0.0116	0.0157
	1-2	-0.0179	-0.0111	-0.0474	-0.0499	0.0216	0.0252
	1-2-3	-0.0209	-0.0095	-0.0277	-0.0273	0.0166	0.0257
DIST_CHAR_ASIS_POS_TAG_DEP	1	-0.0184	-0.0077	-0.0274	-0.0269	0.0178	0.0263
	1-2	-0.0330	-0.0194	-0.0215	-0.0223	0.0134	0.0237
	1-2-3	-0.0284	-0.0184	-0.0342	-0.0277	0.0249	0.0359
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	-0.0257	-0.0151	-0.0325	-0.0235	0.0232	0.0351
	1-2	0.0076	0.0090	-0.0325	-0.0245	0.0215	0.0322
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0084	0.0118	-0.0307	-0.0221	0.0209	0.0316
	1-2	0.0082	0.0058	-0.0320	-0.0221	0.0256	0.0320

Table A.53: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

	Target	Age						
	Min. No. of Characters	100						
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.5814	0.5806	0.4549	0.4544	0.2092	0.0892	0.2029	0.0707
	2-3	0.6560	0.6565	0.5434	0.5426	0.4313	0.4281	0.2607	0.1626
	2-3-4	0.6726	0.6728	0.5630	0.5621	0.4346	0.4147	0.3547	0.2988
	2-3-4-5	0.6769	0.6773	0.5647	0.5638	0.4531	0.4464	0.3450	0.2828
DIST_CHAR_ASIS	2	0.6955	0.6954	0.5636	0.5627	0.4507	0.4473	0.3738	0.3267
	2-3	0.7126	0.7128	0.5758	0.5750	0.4604	0.4584	0.3820	0.3378
	2-3-4	0.7164	0.7163	0.5829	0.5822	0.4703	0.4649	0.3707	0.3001
	2-3-4-5	0.7156	0.7156	0.5835	0.5828	0.4695	0.4638	0.3786	0.3584
DIST_CHAR_ASIS_POS	1	0.7158	0.7159	0.5835	0.5828	0.4693	0.4635	0.3673	0.3417
	1-2	0.7158	0.7159	0.5826	0.5819	0.4693	0.4634	0.3673	0.3417
	1-2-3	0.7163	0.7164	0.5831	0.5825	0.4693	0.4635	0.3679	0.3423
DIST_CHAR_ASIS_POS_TAG	1	0.7162	0.7163	0.5832	0.5825	0.4696	0.4635	0.3831	0.3488
	1-2	0.7159	0.7161	0.5832	0.5825	0.4695	0.4634	0.3833	0.3491
	1-2-3	0.7183	0.7184	0.5835	0.5828	0.4695	0.4635	0.3831	0.3489
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7183	0.7184	0.5832	0.5825	0.4569	0.4440	0.3812	0.3464
	1-2	0.7185	0.7186	0.5833	0.5826	0.4697	0.4634	0.3812	0.3465
	1-2-3	0.7187	0.7189	0.5833	0.5826	0.4697	0.4634	0.3812	0.3464
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7179	0.7184	0.5842	0.5835	0.4703	0.4638	0.3793	0.3440
	1-2	0.7191	0.7192	0.5826	0.5820	0.4704	0.4639	0.3797	0.3443
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7197	0.7198	0.5851	0.5843	0.4706	0.4639	0.4036	0.3761
	1-2	0.7196	0.7197	0.5853	0.5846	0.4705	0.4639	0.4019	0.3724

Table A.54: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	-0.0261	-0.0180	-0.0041	0.0083	0.0156	0.0209
	2-3	-0.0058	-0.0031	-0.0170	-0.0032	0.0285	0.0297
	2-3-4	-0.0065	0.0028	0.0036	0.0192	-0.0086	-0.0014
	2-3-4-5	-0.0145	-0.0003	-0.0099	0.0049	0.0225	0.0185
DIST_CHAR_ASIS	2	0.0049	-0.0039	0.0051	0.0080	0.0112	0.0043
	2-3	-0.0002	-0.0051	-0.0048	0.0081	0.0121	0.0122
	2-3-4	0.0066	0.0059	-0.0038	0.0006	0.0089	0.0184
	2-3-4-5	0.0106	0.0062	-0.0028	0.0105	0.0121	0.0134
DIST_CHAR_ASIS_POS	1	0.0128	0.0108	0.0211	0.0260	0.0288	0.0307
	1-2	0.0068	0.0038	0.0031	0.0109	0.0196	0.0211
	1-2-3	0.0035	-0.0061	0.0233	0.0247	0.0061	0.0055
DIST_CHAR_ASIS_POS_TAG	1	0.0132	0.0031	0.0259	0.0330	0.0098	0.0031
	1-2	0.0141	0.0051	0.0348	0.0395	0.0161	0.0098
	1-2-3	0.0125	-0.0036	0.0332	0.0297	0.0227	0.0170
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0149	0.0027	0.0260	0.0241	0.0154	0.0133
	1-2	0.0281	0.0163	0.0388	0.0381	0.0344	0.0309
	1-2-3	0.0178	0.0115	0.0328	0.0333	0.0331	0.0304
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0309	0.0213	0.0342	0.0276	0.0324	0.0308
	1-2	0.0239	0.0197	0.0286	0.0290	0.0267	0.0261
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0181	0.0120	0.0313	0.0288	0.0259	0.0256
	1-2	0.0074	0.0104	0.0267	0.0247	0.0264	0.0268

Table A.55: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

A2.4.5 Minimum of Characters: 250 & Stacked

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Gender 250 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7737	0.7735	0.7414	0.7414	0.6647	0.6642	0.6455	0.6436
	2-3	0.8138	0.8130	0.7813	0.7813	0.7156	0.7156	0.6858	0.6849
	2-3-4	0.8218	0.8217	0.7912	0.7911	0.7261	0.7261	0.6933	0.6926
	2-3-4-5	0.8230	0.8230	0.7927	0.7926	0.7266	0.7266	0.6936	0.6928
DIST_CHAR_ASIS	2	0.8287	0.8287	0.7983	0.7983	0.7282	0.7280	0.6947	0.6940
	2-3	0.8397	0.8396	0.8076	0.8076	0.7323	0.7323	0.6996	0.6990
	2-3-4	0.8392	0.8391	0.8108	0.8106	0.7353	0.7353	0.7051	0.7050
	2-3-4-5	0.8359	0.8357	0.8114	0.8113	0.7345	0.7345	0.7046	0.7045
DIST_CHAR_ASIS_POS	1	0.8218	0.8200	0.8128	0.8127	0.7345	0.7345	0.7041	0.7040
	1-2	0.8401	0.8400	0.8095	0.8092	0.7348	0.7348	0.7042	0.7040
	1-2-3	0.8419	0.8419	0.8105	0.8104	0.7342	0.7342	0.7034	0.7027
DIST_CHAR_ASIS_POS_TAG	1	0.8401	0.8399	0.8127	0.8126	0.7357	0.7357	0.7040	0.7034
	1-2	0.8437	0.8437	0.8094	0.8090	0.7347	0.7347	0.7038	0.7037
	1-2-3	0.8105	0.8085	0.8104	0.8103	0.7348	0.7348	0.7034	0.7028
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8417	0.8417	0.8085	0.8082	0.7362	0.7362	0.7031	0.7026
	1-2	0.8376	0.8375	0.8082	0.8079	0.7350	0.7350	0.7038	0.7037
	1-2-3	0.8376	0.8374	0.8105	0.8104	0.7351	0.7351	0.7026	0.7021
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8387	0.8386	0.8072	0.8069	0.7330	0.7330	0.7027	0.7022
	1-2	0.8397	0.8396	0.8105	0.8104	0.7351	0.7351	0.7027	0.7021
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8365	0.8363	0.8067	0.8064	0.7363	0.7359	0.7042	0.7036
	1-2	0.8383	0.8381	0.8054	0.8050	0.7364	0.7361	0.7054	0.7054

Table A.56: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Gender 250 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0315	0.0538	0.0180	0.0214	0.0739	0.0638
	2-3	0.0250	0.0498	0.0184	0.0121	0.0694	0.0464
	2-3-4	0.0295	0.0412	0.0155	0.0154	0.0480	0.0440
	2-3-4-5	0.0248	0.0451	0.0103	0.0112	0.0500	0.0452
DIST_CHAR_ASIS	2	0.0336	0.0458	0.0110	0.0130	0.0486	0.0443
	2-3	0.0315	0.0388	0.0124	0.0391	0.0262	0.0248
	2-3-4	0.0327	0.0429	0.0126	0.0066	0.0414	0.0400
	2-3-4-5	0.0557	-0.0074	0.0112	0.0075	0.0403	0.0336
DIST_CHAR_ASIS_POS	1	0.0237	0.0340	0.0405	0.0371	0.0685	0.0623
	1-2	0.0525	0.0020	0.0072	0.0242	0.0643	0.0586
	1-2-3	0.0106	0.0140	-0.0034	-0.0059	0.0298	0.0240
DIST_CHAR_ASIS_POS_TAG	1	0.0069	0.0149	0.0095	0.0072	0.0283	0.0229
	1-2	0.0019	-0.0313	0.0135	0.0283	0.0243	0.0202
	1-2-3	0.0152	0.0117	-0.0018	-0.0021	0.0263	0.0163
DIST_CHAR_ASIS_POS_TAG_DEP	1	-0.0049	-0.0251	0.0210	0.0162	0.0381	0.0286
	1-2	-0.0067	-0.0289	0.0067	0.0067	0.0194	0.0124
	1-2-3	0.0186	0.0083	-0.0189	-0.0083	0.0308	0.0178
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0143	0.0056	-0.0178	-0.0085	0.0274	0.0191
	1-2	0.0173	0.0120	-0.0170	-0.0071	0.0277	0.0177
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0133	0.0108	-0.0146	-0.0058	0.0275	0.0180
	1-2	0.0136	0.0149	-0.0143	-0.0054	0.0260	0.0156

Table A.57: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

Feature types	Target / N-gram ranges	Age							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
DIST_CHAR	2	0.6667	0.6672	0.5623	0.5621	0.4216	0.4205	0.3466	0.3196
	2-3	0.7480	0.7484	0.6506	0.6498	0.5183	0.5180	0.4535	0.4513
	2-3-4	0.7588	0.7592	0.6741	0.6735	0.5337	0.5308	0.4774	0.4757
	2-3-4-5	0.7627	0.7630	0.6722	0.6718	0.5325	0.5292	0.4829	0.4793
DIST_CHAR_ASIS	2	0.7759	0.7763	0.6793	0.6788	0.5455	0.5419	0.4849	0.4810
	2-3	0.7940	0.7943	0.6919	0.6914	0.5571	0.5567	0.4930	0.4892
	2-3-4	0.7911	0.7915	0.6996	0.6991	0.5553	0.5529	0.4994	0.4962
	2-3-4-5	0.7933	0.7933	0.6966	0.6962	0.5553	0.5530	0.5000	0.4972
DIST_CHAR_ASIS_POS	1	0.7869	0.7874	0.6967	0.6962	0.5553	0.5530	0.5000	0.4975
	1-2	0.7930	0.7930	0.6969	0.6965	0.5553	0.5529	0.5001	0.4976
	1-2-3	0.7986	0.7990	0.6975	0.6971	0.5554	0.5530	0.5000	0.4975
DIST_CHAR_ASIS_POS_TAG	1	0.7984	0.7989	0.6974	0.6970	0.5547	0.5522	0.4973	0.4965
	1-2	0.7986	0.7988	0.6965	0.6963	0.5547	0.5523	0.4973	0.4965
	1-2-3	0.7984	0.7990	0.6981	0.6974	0.5550	0.5525	0.4975	0.4966
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7979	0.7985	0.6976	0.6969	0.5625	0.5625	0.4974	0.4967
	1-2	0.7986	0.7992	0.6978	0.6970	0.5626	0.5626	0.4975	0.4968
	1-2-3	0.7981	0.7985	0.6986	0.6979	0.5628	0.5628	0.4976	0.4969
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7893	0.7899	0.6971	0.6964	0.5560	0.5537	0.4974	0.4968
	1-2	0.7925	0.7930	0.6966	0.6959	0.5554	0.5531	0.4972	0.4967
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7962	0.7964	0.6967	0.6960	0.5604	0.5568	0.4972	0.4967
	1-2	0.8003	0.8005	0.6973	0.6966	0.5565	0.5544	0.4974	0.4969

Table A.58: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

Feature types	Target / N-gram ranges	Age					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0132	-0.0179	-0.0008	0.0021	0.0057	0.0131
	2-3	0.0076	-0.0155	0.0017	0.0131	0.0161	0.0122
	2-3-4	0.0248	0.0128	-0.0031	0.0107	0.0037	0.0109
	2-3-4-5	0.0154	—	0.0055	0.0092	0.0169	0.0218
DIST_CHAR_ASIS	2	0.0218	0.0024	-0.0060	-0.0045	0.0118	0.0127
	2-3	0.0311	-0.0075	0.0108	0.0040	0.0065	0.0082
	2-3-4	0.0289	-0.0034	0.0092	0.0006	-0.0009	-0.0029
	2-3-4-5	0.0335	0.0136	0.0001	0.0041	0.0002	—
DIST_CHAR_ASIS_POS	1	0.0386	0.0154	0.0174	0.0131	0.0050	0.0073
	1-2	0.0268	0.0082	0.0128	0.0042	0.0088	0.0037
	1-2-3	0.0241	0.0055	-0.0044	-0.0052	0.0085	0.0093
DIST_CHAR_ASIS_POS_TAG	1	0.0239	0.0045	-0.0050	-0.0100	0.0033	0.0047
	1-2	0.0252	0.0051	0.0043	-0.0025	0.0144	0.0131
	1-2-3	0.0344	0.0300	-0.0073	-0.0073	0.0085	0.0086
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0329	0.0234	-0.0059	-0.0065	0.0145	0.0159
	1-2	0.0298	0.0162	-0.0067	-0.0084	0.0237	0.0231
	1-2-3	0.0259	0.0234	-0.0017	0.0031	0.0254	0.0271
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0212	0.0223	0.0092	0.0033	0.0190	0.0194
	1-2	0.0212	0.0179	0.0037	0.0047	0.0215	0.0199
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0212	0.0178	0.0053	-0.0020	0.0219	0.0186
	1-2	0.0144	0.0180	0.0050	0.0030	0.0124	0.0166

Table A.59: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

A2.4.6 Minimum of Characters: 500 & Stacked

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.8192	0.8179	0.7780	0.7778	0.6948	0.6935	0.6765	0.6764
	2-3	0.8743	0.8740	0.8369	0.8369	0.7639	0.7636	0.7286	0.7285
	2-3-4	0.8874	0.8874	0.8485	0.8485	0.7772	0.7769	0.7400	0.7399
	2-3-4-5	0.8884	0.8883	0.8501	0.8500	0.7773	0.7769	0.7411	0.7410
DIST_CHAR_ASIS	2	0.8920	0.8919	0.8527	0.8527	0.7767	0.7761	0.7412	0.7411
	2-3	0.8983	0.8982	0.8619	0.8619	0.7844	0.7841	0.7467	0.7466
	2-3-4	0.8832	0.8830	0.8670	0.8670	0.7862	0.7855	0.7505	0.7504
	2-3-4-5	0.8789	0.8787	0.8664	0.8663	0.7856	0.7850	0.7515	0.7513
DIST_CHAR_ASIS_POS	1	0.8756	0.8754	0.8653	0.8652	0.7858	0.7852	0.7526	0.7525
	1-2	0.8986	0.8982	0.8673	0.8672	0.7860	0.7854	0.7516	0.7515
	1-2-3	0.8819	0.8817	0.8663	0.8662	0.7860	0.7854	0.7513	0.7512
DIST_CHAR_ASIS_POS_TAG	1	0.8815	0.8814	0.8666	0.8665	0.7851	0.7845	0.7522	0.7522
	1-2	0.8940	0.8940	0.8659	0.8659	0.7861	0.7855	0.7513	0.7511
	1-2-3	0.8924	0.8923	0.8664	0.8663	0.7864	0.7859	0.7514	0.7512
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8897	0.8897	0.8647	0.8646	0.7878	0.7873	0.7525	0.7525
	1-2	0.8871	0.8871	0.8664	0.8663	0.7875	0.7869	0.7515	0.7514
	1-2-3	0.8933	0.8933	0.8665	0.8664	0.7864	0.7859	0.7511	0.7509
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8858	0.8854	0.8674	0.8673	0.7865	0.7859	0.7511	0.7510
	1-2	0.8989	0.8987	0.8672	0.8671	0.7862	0.7856	0.7509	0.7508
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8828	0.8822	0.8673	0.8672	0.7872	0.7866	0.7523	0.7523
	1-2	0.8966	0.8963	0.8684	0.8684	0.7872	0.7866	0.7514	0.7514

Table A.60: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.1274	0.0535	0.0573	0.0718	0.0552	0.0749
	2-3	0.1182	0.0385	0.0420	0.0543	0.0388	0.0574
	2-3-4	0.1085	0.0315	0.0355	0.0490	0.0417	0.0531
	2-3-4-5	0.0951	0.0258	0.0311	0.0432	0.0336	0.0474
DIST_CHAR_ASIS	2	0.1148	0.0277	0.0340	0.0439	0.0364	0.0494
	2-3	0.1371	-0.0178	0.0286	0.0425	0.0308	0.0454
	2-3-4	0.1046	0.0192	0.0263	0.0358	0.0282	0.0381
	2-3-4-5	0.0987	0.0262	0.0281	0.0320	0.0272	0.0380
DIST_CHAR_ASIS_POS	1	0.1100	0.0431	0.0500	0.0536	0.0403	0.0503
	1-2	0.0886	0.0270	0.0025	0.0091	0.0610	0.0674
	1-2-3	0.0702	0.0068	0.0181	0.0232	0.0334	0.0421
DIST_CHAR_ASIS_POS_TAG	1	0.0644	0.0050	0.0067	0.0122	0.0316	0.0398
	1-2	0.0779	0.0190	0.0207	0.0189	0.0429	0.0500
	1-2-3	0.0695	0.0170	0.0159	0.0142	0.0485	0.0553
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0664	0.0061	0.0203	0.0188	0.0472	0.0537
	1-2	0.0733	0.0239	0.0040	0.0025	0.0411	0.0456
	1-2-3	0.0548	0.0065	0.0150	0.0154	0.0521	0.0573
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0522	-0.0209	0.0153	0.0168	0.0516	0.0543
	1-2	0.0481	0.0099	0.0146	0.0128	0.0516	0.0498
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0449	-0.0142	0.0145	0.0117	0.0515	0.0499
	1-2	0.0431	-0.0145	0.0179	0.0132	0.0499	0.0451

Table A.61: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 50 Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7413	0.7422	0.6347	0.6350	0.4899	0.4848	0.4263	0.4239
	2-3	0.8250	0.8252	0.7357	0.7356	0.5868	0.5840	0.5140	0.5114
	2-3-4	0.8415	0.8416	0.7623	0.7622	0.6156	0.6141	0.5433	0.5432
	2-3-4-5	0.8397	0.8397	0.7605	0.7603	0.6172	0.6161	0.5433	0.5416
DIST_CHAR_ASIS	2	0.8473	0.8472	0.7703	0.7701	0.6201	0.6187	0.5489	0.5469
	2-3	0.8651	0.8652	0.7795	0.7795	0.6280	0.6258	0.5550	0.5537
	2-3-4	0.8749	0.8748	0.7912	0.7911	0.6400	0.6396	0.5630	0.5630
	2-3-4-5	0.8700	0.8700	0.7878	0.7876	0.6372	0.6357	0.5623	0.5623
DIST_CHAR_ASIS_POS	1	0.8713	0.8709	0.7878	0.7877	0.6375	0.6360	0.5621	0.5621
	1-2	0.8718	0.8715	0.7881	0.7879	0.6375	0.6360	0.5621	0.5621
	1-2-3	0.8749	0.8746	0.7870	0.7868	0.6372	0.6357	0.5620	0.5621
DIST_CHAR_ASIS_POS_TAG	1	0.8767	0.8767	0.7869	0.7867	0.6395	0.6385	0.5621	0.5621
	1-2	0.8793	0.8793	0.7843	0.7833	0.6395	0.6385	0.5636	0.5636
	1-2-3	0.8820	0.8821	0.7910	0.7908	0.6417	0.6415	0.5628	0.5629
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8802	0.8803	0.7905	0.7904	0.6419	0.6418	0.5626	0.5627
	1-2	0.8736	0.8735	0.7911	0.7909	0.6376	0.6351	0.5626	0.5627
	1-2-3	0.8793	0.8794	0.7912	0.7911	0.6418	0.6417	0.5631	0.5632
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8740	0.8741	0.7911	0.7909	0.6387	0.6362	0.5632	0.5633
	1-2	0.8776	0.8773	0.7914	0.7912	0.6387	0.6361	0.5633	0.5634
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8758	0.8756	0.7946	0.7945	0.6399	0.6385	0.5634	0.5635
	1-2	0.8762	0.8761	0.7911	0.7908	0.6389	0.6374	0.5634	0.5635

Table A.62: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

Feature types	Target Min. No. of Characters No. of Authors Score N-gram ranges	Age 500 150 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0733	0.0228	0.0064	0.0045	-0.0029	0.0157
	2-3	0.0645	-0.0005	0.0075	0.0223	-0.0016	0.0015
	2-3-4	0.0493	0.0004	0.0092	0.0073	0.0037	0.0017
	2-3-4-5	0.0451	0.0075	0.0236	0.0156	-0.0009	0.0045
DIST_CHAR_ASIS	2	0.0799	0.0056	0.0117	0.0167	0.0066	0.0003
	2-3	0.0807	0.0021	0.0149	0.0020	0.0013	0.0205
	2-3-4	0.0751	0.0083	-0.0057	-0.0054	0.0046	0.0045
	2-3-4-5	0.0688	-0.0053	0.0013	0.0040	0.0018	0.0051
DIST_CHAR_ASIS_POS	1	0.0877	0.0359	0.0129	0.0079	0.0084	0.0085
	1-2	0.0772	0.0195	0.0121	0.0085	0.0058	0.0069
	1-2-3	0.0694	0.0280	-0.0024	0.0009	0.0064	0.0082
DIST_CHAR_ASIS_POS_TAG	1	0.0568	0.0114	-0.0062	—	0.0164	0.0198
	1-2	0.0600	0.0070	0.0046	-0.0013	0.0009	0.0146
	1-2-3	0.0652	0.0193	0.0145	0.0120	0.0032	0.0120
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0593	0.0090	0.0097	0.0081	0.0046	0.0094
	1-2	0.0666	0.0221	0.0106	0.0098	0.0118	0.0109
	1-2-3	0.0577	0.0178	0.0123	0.0135	0.0120	0.0110
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0462	0.0048	0.0117	0.0081	0.0019	0.0065
	1-2	0.0491	0.0121	0.0052	-0.0006	0.0025	0.0052
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0519	0.0136	0.0101	0.0044	0.0128	0.0159
	1-2	0.0459	0.0123	0.0077	0.0107	0.0098	0.0004

Table A.63: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set