Simply the Best: Minimalist System Trumps Complex Models in Author Profiling

Angelo Basile[1,3][0000−0002−3312−9359], Gareth Dwyer[3][0000−0001−9024−2668], Maria Medvedeva[3][0000−0002−2972−8447], Josine Rawee[2,3][0000−0002−7603−9417], Hessel Haagsma[3][0000−0003−1514−072X], and Malvina Nissim[3][0000−0001−5289−0971]

[1] Faculty of ICT, University of Malta, Msida, Malta
angelo.basile.17@um.edu.mt
[2] Center for Mind/Brain Sciences, University of Trento, Italy
josinenelleke.rawee@studenti.unitn.it
[3] Center for Language and Cognition, University of Groningen, The Netherlands
{m.medvedeva,hessel.haagsma,m.nissim}@rug.nl, garethdwyer@gmail.com

Abstract A simple linear SVM with word and character n-gram features and minimal parameter tuning can identify the gender and the language variety (for English, Spanish, Arabic and Portuguese) of Twitter users with very high accuracy. All our attempts at improving performance by including more data, smarter features, and employing more complex architectures plainly fail. In addition, we experiment with joint and multitask modelling, but find that they are clearly outperformed by single task models. Eventually, our simplest model was submitted to the PAN 2017 shared task on author profiling, obtaining an average accuracy of 0.86 on the test set, with performance on sub-tasks ranging from 0.68 to 0.98. These were the best results achieved at the competition overall. To allow lay people to easily use and see the value of machine learning for author profiling, we also built a web application on top our models.

Keywords: author profiling · linear models · gender prediction · language variety identification · multitask learning

1 Introduction and Background

Profiling authors, that is, inferring personal characteristics from text, can reveal many things, such as their age, gender, personality traits, and/or location, even though writers might not consciously choose to put indicators of those characteristics in the text. The uses for this are obvious, for cases like targeted advertising and security, but it is also interesting from a linguistic standpoint. With the rise of social media, more and more people acquire some kind of on-line presence or persona, mostly made up of images and text. This means that these people can be considered authors, and thus that we can profile them as such.

In this contribution we explore two specific author traits, namely gender and native language variety of Twitter users, across four languages. In addition to

experimenting with a variety of features and algorithms for developing systems geared to optimal performance, we specifically investigate the benefits of modelling these two different axes *separately* or *jointly*.

Previous work has shown that the very same features could be reliable clues for classification. Indeed, for both profiling authors on Twitter as well as for discriminating between similar languages, word and character n-grams have proved to be the strongest predictors of gender as well as language varieties. For language variety discrimination, the systems that performed best at the Discriminating between Similar Languages (DSL) shared task in 2016 (on test set B, i.e. social media) used word/character n-grams, independently of the algorithm [12]. The crucial contribution of these features was also observed by [17] and [3], who participated in the 2017 DSL shared task with the two best performing systems. For author profiling, it has been shown that tf-idf weighted n-gram features, both in terms of characters and words, are very successful in capturing especially gender distinctions [25].

If different aspects, such as language variety and gender of a Twitter user, might be captured by the same features, can both tasks be modelled with the same approach? Also, if these are distinct but somewhat similar aspects, to what extent is it beneficial to model them together? We investigate such questions by building models that address the tasks separately but rely on the same set of features, and also explore the feasibility of modelling both tasks at the same time.

We built two simple SVM models based on n-gram features, using identical settings for both gender and language variety prediction (Section 3). Over such settings, we experimented with a variety of enhancements to our models which however turned out to be detrimental to performance. These include manipulating the data itself (adding more, and changing preprocessing) and using a large variety of features (Section 4.1), as well as changing strategies in modelling the problem. Specifically, we used different algorithms and paradigms, and tried to learn the two tasks jointly via Multitask Learning (Section 4.2).

We observe that simple models outperform complex models under all settings, confirming the predictive power of word and character n-grams for author profiling. The best model described in this paper (Section 3) was submitted as an official participation to the PAN 2017 shared task on author profiling (22 participants), and achieved best results overall. The system is also made available to the general public via a simple web interface.

2 Data

We use data from the 2017 shared task on author profiling [24], organised within the PAN framework [21]. Data is provided in four languages: English, Spanish, Arabic, and Portuguese, for a total of 11400 sets of tweets, each set representing a single author.[4]

[4] This is the training set released at PAN 2017. An additional test set was available for testing models during the campaign, but not anymore at the time of writing.

Gender is provided as a binary label (male/female), whereas language variety differs per language, from 2 varieties for Portuguese (Brazil and Portugal) to 7 varieties for Spanish (Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela). For each variety in each language the dataset consists of 1,000 authors, with 100 tweets each. This means that there is more data overall for the languages with the most varieties. Of these 1,000 authors per variety, 500 are male and 500 are female. The test set of each *gender* × *variety* subset contains 200 authors and the training set 300.

In order to better understand the data and gain some insights that could help the feature engineering process, we used two visualisations, one for each task. For the variety label we trained a decision tree classifier using word unigrams for English. Although the performance is poor (accuracy score of 0.63) it allowed us to see which feature values where the most distinctive (i.e. the first splits of the decision tree). We find that the most important indicators of language variety are simply geographical names: "NZ", "Dublin", "Australia", etc.

For gender, we used the Scattertext tool [9] to compare the most frequent words used by males and females in the English dataset. This revealed several interesting things about the gendered use of language. The words used often by males and very seldom by females are sport-related, and include words such as "league", and "chelsea". Conversely, tokens used frequently by females and infrequently by males include several emojis, e.g. "😫", "😊", as well as words like "kitten", "mom", "sister" and "chocolate". This kind of stereotypical usage was also observed by [26].

We also found distinctive words include both time-specific ones, like "gilmore" and "imacelebrityau", and general words from everyday life, which are less likely to be subject to time-specific trends, like "player", and "chocolate". Although time-specific words are highly useful as features within this experimental setup, where training and evaluation data are from the same time periods, the usage of such features might hamper the predictive capability on unseen future data.

3 Basic Models

Previous work suggests that character and word n-grams as features of an SVM system are excellent at capturing both gender and language variety [24,12]. Using the scikit-learn LinearSVM implementation [20], we built a simple SVM system that uses character 3- to 5-grams and word 1- to 2-grams. We employ tf-idf weighting with sublinear term frequency scaling, where instead of the standard term frequency we use: $1 + \log(tf)$.

To optimise parameters, we ran an extensive grid search over both tasks and all languages on a 64-core machine with 1 TB RAM (see Table 1 for the list of values over which the grid search was performed). The full search took about a day to complete. In particular, using min_df=2 (i.e. excluding all terms that are used by only one author) seems to have a strong positive effect and greatly reduces the feature size as there are many words that appear only once. Having different optimal parameters for different languages provided only a slight

performance boost for each language. We decided that this increase was too small to be significant, so we used the same parameter values for all languages and both tasks. Similarly, after experimenting with different tokenisation techniques for the different languages, we decided to use the default scikit-learn tokeniser for all languages as average results did not improve. Table 2 shows the results of this base system. All reported results are on the PAN 2017 training data using five-fold cross-validation, unless otherwise specified.

Table 1. The list of parameter values included in the grid search. The optimal values that we use then in our system are in **bold**.

Name	**Values**	**Description**
lowercase	**True**, False	Lowercase all words
max_df	1, **100**, None	Exclude terms appearing in more than n% documents
min_df	1, **2**, 3	Exclude terms appearing in fewer than n documents
use_idf	**True**, False	Use Inverse Document Frequency weighting
sublinear_tf	**True**, False	Replace term frequency (tf) with $1 + log(tf)$
C	0.1, 0.5, **1**, 1.5, 5	Penalty parameter for the SVM

Table 2. Accuracy of the base system using 5-fold cross-validation on the PAN 2017 training set. "Joint": single models predict gender and language variety, and the joint accuracy is inferred afterwards for global evaluation. "Merged": one models predicts the merged labels directly.

Language	**Gender**	**Variety**	**Joint**	**Merged**
Arabic	0.800	0.831	0.683	0.630
English	0.823	0.898	0.742	0.645
Spanish	0.832	0.962	0.803	0.686
Portuguese	0.845	0.981	0.828	0.792

Aside from evaluating the performance of the classifier on the two separate tasks, we also evaluated its global performance over the correct prediction of both labels at the same time. For example, for a female American user, predicting female British would lead to a correct gender prediction, a wrong prediction for variety, and therefore also a wrong prediction of the profile as a whole. Results for this joint evaluation are shown in Table 2, under the "Joint" column.

With an eye to the performance on the whole profile, we also trained our system to predict both language variety and the gender of each user simultaneously, instead of predicting each task separately, by simply merging the two labels. As expected, since the task is harder, the performance goes down when compared to a model trained independently on each task. In other words: the

derived joint prediction is better than the joint prediction learnt directly from the merged labels (see Table 2).

4 Variations

4.1 Data and Features

As potential improvements over the base models, we experimented with more training data, and by including more features.

Adding Previous PAN Data We extended the training dataset by adding data and gender labels from the PAN 16 Author Profiling shared task [25], based on the expectation that having a larger amount of training data might yield better performance. To confirm this, we attempted to train on English data from PAN 17 and predict gender labels for the English data from PAN 16, as well as vice versa. Training on the PAN 16 data resulted in an accuracy score of 0.754 for the PAN 17 task, and training on PAN 17 gave an accuracy score of 0.700 for PAN 16, both scores significantly lower than cross-validated results.

One possible explanation for this is that our unigram model captures aspects that are tied specifically to the PAN 17 dataset, because it contains topics that may not be present in datasets that were collected in a different time period. This is in line with previous findings, as [16] show that simple author profiling models tend to generalise poorly to datasets from different genres or time periods.

Using the Twitter 14k dataset We attempted to classify the English tweets by Gender using only the data collected by [1]. This dataset consists of aggregated word counts by gender for about 14,000 Twitter users and 9 million Tweets. We used this data to calculate whether each word in our dataset was a 'male' word (used more by males), or a 'female' word, and classified users as male or female based on a majority count of the words they used. Using this method we achieved 0.712 accuracy for the English gender data, showing that this simple method can provide a reasonable baseline to the gender task.

PoS tags We added PS-tags to the English tweets using the *spaCy*[5] tagger, and experimented with a model that included both regular unigrams and part-of-speech information. The results of both models are shown in Table 3. Compared to the model using only unigrams performance dropped slightly for gender and a bit more for variety. It is not clear whether the missed increase in performance is due to the fact that the PoS tagger does not perform well on Twitter data (the PoS tagger is not Twitter specific) or to the fact that our classifier does not perform well with PoS tags.

[5] https://spacy.io/

Table 3. Accuracy scores on gender and variety classification of only an only unigram model with and without part-of-speech tags on the PAN 17 English training data using 5-fold cross-validation

	Gender	Variety
Unigrams	0.826	0.895
Unigrams + Part-of-Speech	0.818	0.853

Emojis (😀😁😂😃😄😅😆😇😈😉😊😋)

In April 2015, SwiftKey did an extensive report[6] on emoji use by country. They discovered that emoji use varies across languages and across language varieties. For example, they found that Australians use double the average amount of alcohol-themed emoji and use more junk food and holiday emoji than anywhere else in the world.

We tried to leverage these findings but the results were disappointing. We used a list of emojis[7] as a vocabulary for the td/idf vectorizer. Encouraged by the data in the SwiftKey report, we tried first to use emojis as the only vocabulary for predicting gender. The results on the Spanish training set using 5-fold cross validation are surprisingly high (0.67 accuracy) and clearly higher than a random baseline, but fall clearly short of the score of the simple unigram model (0.79 accuracy). Adding emojis as extra features to the unigram model did not yield any improvement.

Excluding Specific Word Patterns We looked at accuracy scores for the English gender and variety data more closely. We tried different representations of the tweet texts, to see what kind of words were most predictive of variety and gender. Specifically, we look at using only words that start with an uppercase letter, only words that start with a lowercase letter, only Twitter handles (words that start with an "@") and all the text excluding the handles. Results are presented in Table 4.

It is interesting that the accuracies are so high although we are using only a basic unigram model, without looking at the character n-grams that we include in our final model. Representing each text only by the Twitter handles used in that text results in 0.77 accuracy for variety, probably because users tend to interact with other users who are in the same geographic area. However, excluding handles from the texts barely decreases performance for the variety task, showing that while the handles can be discriminative, they are not necessary for this task. It is also interesting to note that for this dataset, looking only at words beginning with an uppercase character results in nearly the same score for the gender task as we get when using all of the available text, while using only lowercase words decreases performance. The opposite is true for the variety task,

[6] https://blog.swiftkey.com/americans-love-skulls-brazilians-love-cats-swiftkey-emoji-meanings-report/

[7] http://www.unicode.org/emoji/charts/full-emoji-list.html

where using lowercase-only words results in as good performance as using all the text, but using only uppercase words decreases accuracy by over 10 percent.

Table 4. Accuracy scores on gender and variety prediction using 5-fold cross-validation with the base system on the English training data, with and without the exclusion of specific groups of words.

	Gender	**Variety**
All text	0.816	0.876
Handles only	0.661	0.769
Exclude handles	0.814	0.869
Uppercase only	0.802	0.767
Lowercase only	0.776	0.877

Place Names and Twitter Handles We tried enriching the data to improve the unigram model. For each of the language varieties, we obtained 100 geographical location names, representing the cities with the most inhabitants. When this location was mentioned in the tweet, the language variety the location was part of was added to the tweet.

We attempted to use Twitter handles in a similar manner. The 100 most-followed Twitter users per language variety were found and the language variety was added to the text when one of its popular Twitter users was mentioned.

Unfortunately, these methods did not improve our model's performance. We suspect that the information is already captured by the word n-grams, so encoding this information explicitly does not improve performance.

GronUP Combinations We have tried the partial setup of last year's winning system, GronUP [5], with the distinction that we had to classify language variety instead of age groups. We have excluded the features that are language-dependent (i.e. PoS-tagging and misspelling/typos), and experimented with various feature combinations of the rest while keeping word and character n-grams the same. We achieved average accuracy scores ranging from 0.810 to 0.830, which is clearly lower than our simple final model, which achieved an average accuracy score of 0.872 using 5 fold cross validation of the training data.

4.2 Different Approaches

As an alternative to adding data and features, we tried to improve the performance of our base system employing different algorithms and modelling strategies.

FastText We experimented with Facebook's FastText system, which is an out-of-the-box supervised learning classifier [8]. We used only the data for the English gender task, trying both tweet-level and author-level classification. We pre-processed all text with the NLTK Tweet Tokenizer and used the classification-example script provided with the FastText code base.

Training on 3,000 authors and testing on 600 authors gave an accuracy score of 0.64, compared to average English gender performance of SVM of 0.823. Changing the FastText parameters such as number of epochs, word n-grams, and learning rate showed no improvement. We achieved an accuracy of 0.79 when we attempted to classify on a per-tweet basis (300,000 tweets for training and 85,071 for test), but this is an easier task as some authors are split over the training and test sets. There are various ways to summarise per-tweet predictions into author-predictions, but we did not experiment further as the SVM system clearly worked better for the amount of data we have.

Multi-task learning Multi-task learning (MTL, [6]) has proven successful in a variety of NLP tasks [7,4,10,15], including author profiling [2]. Usually, one main task is learned while one or more auxiliary tasks are learned at the same time in order to provide some additional signal, and reduce overfitting.

We used MTL to investigate whether learning the two tasks at the same time would be beneficial. Practically, we used the *DyNet* framework [19] to build a neural network that learns both tasks simultaneously, defining gender as the main task, and language variety as the auxiliary task. The reason for this choice is the observation that, while language variety is predicted by the SVM with high accuracy, performance on gender is lower, suggesting that it could benefit from an additional signal.

We compute two different losses, one per task, and back-propagate their sum to train the model. Our network structure consists of an embedding layer, two Bi-LSTM layers, and two multi-layer perceptrons on top, one for each task. The hidden layers are shared. We trained the network for 20 iterations, using a constant learning rate. For these experiments, the only pre-processing that we applied consisted of lower-casing all the words.

The final accuracy of the MTL model is 48.3%, thus below the baseline. Due to resource constraints, we did not perform proper tuning of the hyper-parameters of the network, which can be a reason for the low performance. For the moment, to better understand and contextualise these results, we trained the same network two more times, one per task, thus treating them separately again, and using a single loss. The rationale behind this is to verify whether it is the architecture itself that is not learning the two problems well, or whether the poor performance derives mainly by treating them jointly.

Accuracy is 70.2% for the language variety model, and 51.4% for the gender model. The models are slightly better than chance at predicting gender, while for variety there seems to be some signal that could potentially be amplified with more training data and/or hyper-parameter tuning. With the current settings, results are still far below what we achieved with our n-gram based SVM. As

such, it is likely that the low MTL performance is due to the chosen network architecture, and not necessarily due to the joint learning of the two tasks.

5 The System in Practice

5.1 Participation in the PAN 2017 Shared Task

N-GrAM (New Groningen Author-profiling Model), our best system as described in Section 3, was submitted as official participant to the PAN 2017 evaluation campaign on author profiling. Overall, N-GrAM came first in the shared task, with a score of 0.8253 for gender 0.9184 for variety, a joint score of 0.8361 and an average score of 0.8599 (final rankings were taken from this average score) on the official PAN 2017 test set [24]. For the global scores, all languages are combined.

Table 5. Results (accuracy) on the test set for variety, gender and their joint prediction.

Task	System	Arabic	English	Portuguese	Spanish	Average	+ 2nd
Gender	N-GrAM	**0.8006**	**0.8233**	**0.8450**	**0.8321**	**0.8253**	0.0029
	LDR	0.7044	0.7220	0.7863	0.7171	0.7325	
Variety	N-GrAM	**0.8313**	0.8988	0.9813	0.9621	0.9184	0.0013
	LDR	0.8250	**0.8996**	**0.9875**	**0.9625**	**0.9187**	
Joint	N-GrAM	**0.6831**	**0.7429**	**0.8288**	**0.8036**	**0.7646**	0.0101
	LDR	0.5888	0.6357	0.7763	0.6943	0.6738	

We present finer-grained scores showing the breakdown per language in Table 5. We compare our gender and variety accuracies against the LDR-baseline [23], a low dimensionality representation especially tailored to language variety identification, provided by the organisers. The final column, *+ 2nd* shows the difference between N-GrAM and the score achieved by the second-highest ranked system (excluding the baseline).

Results are broken down per language, and are summarised as both *joint* and *average* scores. The joint score is is the percentage of texts for which both gender and variety were predicted correctly at the same time, while still running single models. The average is calculated as the mean over all languages.

N-GrAM ranked first in all cases except for the language variety task. In this case, the baseline was the top-ranked system, and ours was second by a minimal margin. Our system significantly out-performed the baseline on the joint evaluation, as the baseline scored significantly lower for the gender task than for the variety task. These scores are highly similar to the scores on the cross-validated training set that were described in Table 2.

N-GrAM compared to other systems Although we have tried a large amount of different approaches to the task, everything boiled down to a simple Linear SVM system with n-gram features and slightly adapted tf-idf parameters.

When looking at the other participating systems at the PAN 2017 shared task, it appears we were not alone [24]. Out of the top-ranked seven teams (including us), six used Logistic Regression [14,22] or SVMs [27,11,13] and only one used Neural Attention Networks [18]. Interestingly, the latter system performed much better than any other system when predicting Portuguese gender, but was beaten by linear classifiers in other subtasks.

Although many participants have experimented with various preprocessing methods and normalisation, such as removing Twitter handles, URLs and lowercasing, as well as tried to take advantage of emojis, the majority of the systems have also used n-grams as the main set of features and the difference in scores often came down to small alternation within n-gram length.

5.2 Online System Demonstration

We worked with a group of software engineers to make author profiling and author identification more accessible, even outside an academic context. Under our direction, these engineers built a web application through which anyone can easily submit text and see instant author profiling results, with no need for any technical or academic experience.

The web application encompasses author attribution as well as author profiling. The author profiling follows the PAN 2016 settings [25], attempting to predict gender and age instead of gender and variety. Nonetheless, the model used for gender identification is built on N-GrAM.

Users of the web application do not require any software except a standard web browser. On visiting the page, they see a brief description of what author profiling is. After navigating to the author profiling page, they can choose to paste text into a box, upload a plain text file, or load an example. After submitting text through any of the three options, they see some visualisations which depict the predicted gender and age of the text's author. The submission screen of the web application can be seen in Figure 1 and the full application can be used online.[8]

We believe that taking research such as ours, which is all-too-often presented only in academic papers and code repositories, and making it available to and accessible by members of the public who are not necessarily academics or programmers, is highly important. Not only does it help solve a disconnect between active areas of research and the public perception of research, but it furthermore moves towards a goal of ensuring that a gap does not develop between those who understand and can use machine learning and those who cannot. This is part of a larger conversation which is well summarised in a feature released by Microsoft, titled *Democratizing AI*.[9]

[8] https://aabeta.herokuapp.com
[9] https://news.microsoft.com/features/democratizing-ai/

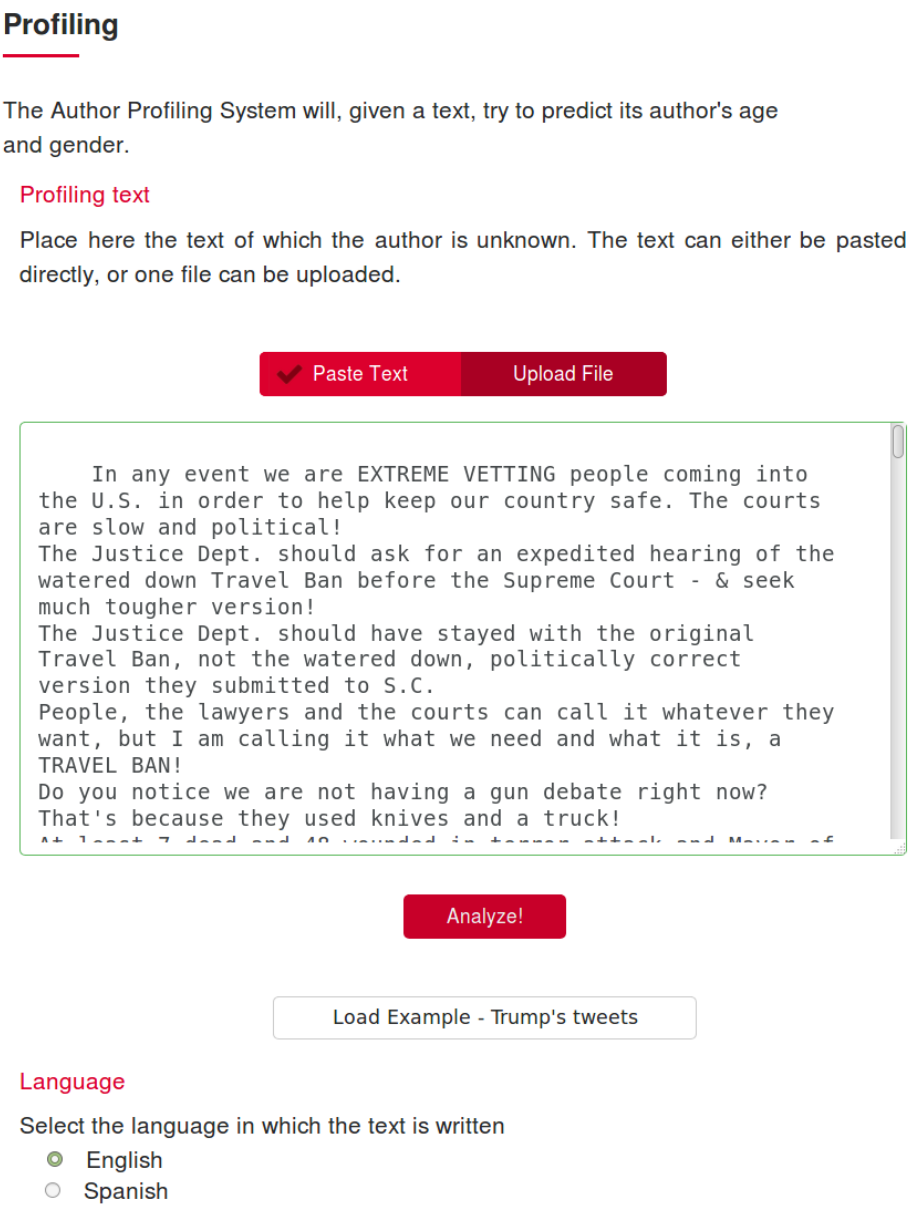

Figure 1. Example page from the author analysis web application.

6 Conclusion

For the author profiling task at hand, a seemingly simple system using word and character n-grams and an SVM classifier proves very hard to beat. Indeed, our simple system, N-GrAM, turned out to be the best-performing out of the 22 systems submitted in the PAN 2017 shared task. Using additional training data, 'smart' features, and hand-crafted resources hurt rather than helped performance. A possible lesson to take from this would be that manually crafting features serves only to hinder a machine learning algorithm's ability to find patterns in a dataset, and perhaps it is better to focus one's efforts on parameter optimisation instead of feature engineering.

Our preliminary experiments, including a setting that has proved beneficial for a variety of language processing tasks, namely multitask learning, do not however show the superiority of neural models compared to the SVM that one might have expected. Nevertheless, we believe that this is too strong a conclusion to draw from this limited study, since several factors specific to this setting need to be taken into account. We expect that while an SVM is the best choice for the given amount of training data, with more training data, and proper parameter optimisation, a neural network-based approach might achieve better results.

Regarding the frustrating lack of benefit from more advanced features than n-grams, a possible explanation comes from a closer inspection of the data. Both the decision tree model and the Scattertext visualisation give us an insight in the most discriminating features in the dataset. In the case of language variety, we see that place names can be informative features, and could therefore be used as a proxy for geographical location, which in turn serves as a proxy for language variety. Adding place names explicitly to our model did not yield performance improvements, which we take to indicate that this information is already captured by n-gram features.

In the case of gender, many useful features are ones that are highly specific to the Twitter platform (*#iconnecthearts*), time (*cruz*), and topics (*pbsnewshour*) in this dataset, which have been shown not to carry over well to other datasets [16], but provide high accuracy in this case. Conversely, features designed to capture gender in a more general sense do not yield any benefit over the more specific features, although they would likely be useful for a robust, cross-dataset system and should definitely be further investigated.

Acknowledgements

We are grateful to the organisers of PAN 2017 for making the data available. We also would like to thank Barbara Plank for her advice on the MTL architecture and the anonymous reviewers for providing valuable insights.

References

1. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender identity and lexical variation in social media. Journal of Sociolinguistics **18**(2), 135–160 (2014)
2. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. vol. 1, pp. 152–162 (2017)
3. Bestgen, Y.: Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In: Proceedings of the VarDial Workshop (2017)
4. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: Joint learning of words and meaning representations for open-text semantic parsing. In: AISTATS. vol. 351, pp. 423–424 (2012)
5. Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: GronUP: Groningen User Profiling. In: Working Notes of CLEF. pp. 846–857. CEUR Workshop Proceedings, CEUR-WS.org (2016)
6. Caruana, R.: Multitask learning. Machine Learning **28**, 41–75 (1997)
7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)
8. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
9. Kessler, J.S.: Scattertext: a browser-based tool for visualizing how corpora differ. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations. Association for Computational Linguistics, Vancouver, Canada (2017)
10. Liu, X., Gao, J., He, X., Deng, L., Duh, K., Wang, Y.Y.: Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In: Proc. NAACL (2015)
11. López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Solorio, T.: Social-media users can be profiled by their similarity with other users. In: Working Notes of CLEF. CEUR Workshop Proceedings, CEUR-WS.org (2017)
12. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). pp. 1–14. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016)
13. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. In: Working Notes of CLEF. CEUR Workshop Proceedings, CEUR-WS.org (2017)
14. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling – gender and language variety prediction. In: Working Notes of CLEF. CEUR Workshop Proceedings, CEUR-WS.org (2017)
15. Martínez Alonso, H., Plank, B.: When is multitask learning effective? semantic sequence prediction under varying data conditions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 44–53. Association for Computational Linguistics, Valencia, Spain (April 2017), http://www.aclweb.org/anthology/E17-1005

16. Medvedeva, M., Haagsma, H., Nissim, M.: An analysis of cross-genre and in-genre performance for author profiling in social media. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 211–223. Springer International Publishing (2017)
17. Medvedeva, M., Kroon, M., Plank, B.: When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 156–163. Association for Computational Linguistics (2017)
18. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T.: Author profiling with word+character neural attention network. In: Working Notes of CLEF. CEUR Workshop Proceedings, CEUR-WS.org (2017)
19. Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., Duh, K., Faruqui, M., Gan, C., Garrette, D., Ji, Y., Kong, L., Kuncoro, A., Kumar, G., Malaviya, C., Michel, P., Oda, Y., Richardson, M., Saphra, N., Swayamdipta, S., Yin, P.: Dynet: The dynamic neural network toolkit. arXiv preprint arXiv:1701.03980 (2017)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
21. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin, Heidelberg, New York (2017)
22. Poulston, A., Waseem, Z., Stevenson, M.: Using TF-IDF n-gram and word embedding cluster ensembles for author profiling. In: Working Notes of CLEF. CEUR Workshop Proceedings, CEUR-WS.org (2017)
23. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. arXiv preprint arXiv:1705.10754 (2017)
24. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)
25. Rangel, F., Rosso, P., Verhoeven, B., Potthast, W.D.M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In: Working Notes of CLEF. pp. 750–784 (2016)
26. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73791 (2013)
27. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D.: Gender and language variety identification with MicroTC. In: Working Notes of CLEF. CEUR Workshop Proceedings, CEUR-WS.org (2017)