ABI Neural Ensemble Model for Gender Prediction

Adapt Bar-Ilan Submission for the CLIN29 Shared Task on Gender Prediction

Eva Vanmassenhove[1], Amit Moryossef[2], Alberto Poncelas[1], Andy Way[1], and
Dimitar Shterionov[1]

[1] ADAPT, School of Computing, Dublin City University, Dublin, Ireland
`firstname.lastname@adaptcentre.ie`
[2] Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
`firstnamelastname@gmail.com`

Abstract. We present our system for the CLIN29 shared task on cross-genre gender detection for Dutch. We experimented with a multitude of neural models (CNN, RNN, LSTM, etc.), more "traditional" models (SVM, RF, LogReg, etc.), different feature sets as well as data pre-processing. The final results suggested that using tokenized, non-lowercased data works best for most of the neural models, while a combination of word clusters, character trigrams and word lists showed to be most beneficial for the majority of the more "traditional" (that is, non-neural) models, beating features used in previous tasks such as n-grams, character n-grams, part-of-speech tags and combinations thereof. In contradiction with the results described in previous comparable shared tasks, our neural models performed better than our best traditional approaches with our best feature set-up. Our final model consisted of a weighted ensemble model combining the top 25 models. Our final model won both the in-domain gender prediction task and the cross-genre challenge, achieving an average accuracy of 64.93% on the in-domain gender prediction task, and 56.26% on cross-genre gender prediction.

1   Introduction

In recent years, author profiling (AP) has gained a lot of interest. AP can be described as the task of predicting or identifying demographics (such as gender and age) of an author based on their writing. It has applications of growing importance in different fields such as security, marketing, etc. Of all the demographics, there has been a particular interest in gender profiling which has

been part of the shared tasks organized by PAN for the past six consecutive years (2013–2018). Tasks differed from in-domain gender prediction achieving relatively high scores to multi-modal gender prediction and cross-genre gender prediction, initially focusing on English but gradually including other languages such as Dutch, Spanish, and others.

In this paper, we describe our best system submitted for the CLIN29 2019 shared task on gender prediction for Dutch with a particular focus on out-of-domain data, along with all other systems we experimented with. The plethora of trained models and empirical evaluations we conducted for this task is one of our main contributions; the other is the ensemble system which wins both the in-domain and out-of-domain subtasks.

The rest of the paper is structured as follows: Section 2 describes the related work, Section 3 the experimental set-up. Results are discussed in Section 4. Finally, the conclusions and ideas for future work are presented in Section 5.

2   Related Work

Over the last years, there has been a significant body of research dedicated to AP, often including gender prediction. The main focus, however, has been on predicting gender using in-domain data for training and testing. The various approaches presented in the PAN shared tasks used preprocessing, lowercasing, stop-word filtering combined with specific linguistic features including character, word and POS n-grams, punctuation features, topic modeling and features specific to the given domain (e.g. hashtags or links for Twitter data). The yearly PAN evaluation campaigns (pan.webis.de) have led to the development of state-of-the-art (SOTA) in-domain gender prediction models on Twitter data for English achieving accuracies in between $80\% - 85\%$ (Alvarez-Carmona et al., 2015; Rangel et al., 2015; Basile et al., 2017; Rangel et al., 2017).

The PAN 2016 differed from previous gender prediction tasks as it was the first shared task focusing on cross-genre gender prediction. Twitter data was provided for training while the test data was another 'unknown' type of social media text. It should be noted, however, that although the test data differed from the training data, all the data still belonged to the broader 'social media' domain. The best scores recorded for gender prediction were 62%, 73% and 76% for Dutch, Spanish and English respectively (Rangel et al., 2016). An additional analysis of the cross-genre results by Medvedeva et al. (2017) revealed that the portability of the cross-genre models is only successful when the subdomains are close enough. The PAN-RUS Profiling at FIRE17 focused on predicting gender across different domains (Twitter, Facebook, essays and reviews) obtaining accuracies between $65\% - 93\%$ (Litvinova et al., 2017) depending on the domain. Similarly, in order to capture more domain-independent and thus deeper gender-specific features, the EVALITA 2018 Campaign (Caselli et al., 2018) organized a cross-genre prediction task across five domains (Children Writings, Twitter, YouTube, News, and Personal Diaries) with accuracies ranging between 51% (YouTube) and 64% (Children Writings).

An important difference between the two previous tasks on real cross-genre gender prediction and this year's CLIN shared tasks is that, unlike Russian and Italian, gender agreement with the first person is very rare in Dutch[1]. In Russian and Italian, verbs, adjectives and nouns (can) reflect the gender of the speaker, which facilitates gender prediction.

3 Experimental Setup

In this section, we describe in more detail: the datasets provided (3.1) and the different models and features used (3.2).

3.1 Datasets

The datasets initially provided belonged to three different domains: Twitter (TW), YouTube (YT) and News (N). For the 3 in-domain scenarios we used 90% of the data for training and 10% for validation. For the 3 out-of-domain scenarios we used all the out-of-domain data available. For example, for the out-of-domain YouTube prediction, we used all the given News and Twitter data and validated on the YouTube data provided. For some of our models (i.e., our second set of submissions), we also used external CSI (Verhoeven and Daelemans, 2014) and TwiSty (Verhoeven et al., 2016) data. We made sure there was no overlap between the data provided and the external data we added. An overview of the different scenarios and the amount of training/validation examples can be found in Table 1.

IN-DOMAIN	# Train	# Valid
N (90-10)	1 648	184
TW (90-10)	18 000	2 000
YT (90-10)	13 269	1 475

OUT-DOMAIN	# Train	# Valid
N+TW |YT	21 832	14 744
N+YT |TW	16 576	20 000
TW+YT |N	34 744	1 832

EXTERNAL	# Train
CSI	3 113
TWisty	68 907

Table 1. Number of training and validation examples per scenario.

---

[1] Exceptions would be certain sentences where the noun agrees in gender with the subject, e.g., 'Ik ben een acteur' (masc.) vs 'Ik ben een actrice' (fem) [English: 'I am an actor/actress']

3.2   Systems Descriptions

Our winning approach consisted of an ensemble model of our strongest models. First, we will first describe all the neural and traditional models we experimented with as well as the effects of different feature sets. After, the ensembling of the best models will be described in more detail.

Neural Models We experimented with the following neural networks: SpaCy *TextCategorizer* models (SpaCy) (Honnibal and Montani, 2017), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Long Short-Term Memory with Attention (LSTMa), Region-based Convolutional Neural Networks (RCNN), Recurrent Neural Network (RNN) and Self Attention (SA). All models were trained with and without frozen `fasttext` embeddings [2] (Joulin et al., 2016) using a publicly available pytorch implementation.[3]

All neural network models we trained using the following parameters: **batch size** = 32, *hidden state size* = 256 and *embedding length* = 300; The *learning optimizer* was Adam (Kingma and Ba, 2014) and we used a *learning rate* = $2e - 5$. For the SpaCy models, the default parameters were used.[4]

Traditional Models From previous shared tasks it resulted that more traditional statistical models, in general, still outperform neural models. The traditional approaches we experimented with include: Statistical Language Models (LMs), Support Vector Machines (SVMs), K-Nearest Neighbour Classifier (KNN), Logistic Regression Classifier (LogReg), Random Forest Classifier (RF) and Bernouilli Naive Bayes (NB) using the `scikit-learn` toolkit (Pedregosa et al., 2011).

We ought to note some differences between our LM-based models and the other approaches. In particular, for the LM-based models, we: (i) split the data into two subsets – a female and a male subset – which we (ii) use separately to train two KenLM (Heafield et al., 2013) models, one for each of the two subsets. At training time we build 3-, 4-, 5- and 6- gram KenLM models with pruning of singleton 2-grams and above; each paragraph was also segmented into single sentences. For each sentence in a test paragraph, we generated two scores – one for the female and one for the male models. The scores for a paragraph are the averages over all sentence scores in that paragraph. To classify the test paragraph we identify whether the female or the male KenLM model leads to higher score. This approach contrasts with the rest, where female and male data is used jointly to train a model that predicts (probabilistically) one of two labels.

Features As previous shared tasks showed the usefulness of features such as word *n*-grams, character *n*-grams, Part-Of-Speech (POS) tags, and article counts, we compared how the 'traditional' classifiers performed with such features. We experimented with 4 additional features:

---

[2] As the fasttext embeddings are extracted from Wikipedia, yet a different domain, the usage of these embeddings does not interfere with the objective of the cross-genre prediction task.

[3] `https://github.com/prakashpandey9/Text-Classification-Pytorch`

[4] See: `https://spacy.io/usage/training` for details.

clusters: Inspired by the cross-genre task and the necessity for generalization we added a cluster feature. To do so, we clustered words together based on their `fasttext` embeddings and subsequently replaced unique words by their cluster number. Clusters containing only singletons or more than 500 words were removed.

The three following features were inspired by the findings of Keune (2012) whose thesis investigated the difference between male and female speech.

words used more by men: Keune (2012) concluded from her research on the Corpus Gesproken Nederlands (CGN)[5] that there were certain words linked to a specific gender. Similar observations were made in Vanmassenhove et al. (2018), showing that certain words are more frequent in male/female speech. Words such as: 'feitelijk', 'voornamelijk', 'degelijk', 'oorspronkelijk', 'tamelijk', 'onmiddellijk', 'je', 'd'r', 'ja', 'nee' and 'neen' are used more by men than by women. Therefore, we decided to count the number of occurrences of these words.

words used more by females: Similarly, some other concrete words are used significantly more by women: 'ik', 'hij', 'dadelijk', 'vriendelijk', 'lelijk', 'vrolijk', 'eindelijk' and 'verschrikkelijk'.

diminutives count: We also counted the diminutives appearing in every data example as diminutives are used more by women than by men (Keune, 2012).

We started by using the classifier with the strongest baseline on unigrams for the 3 out-of-domain validation datasets. This appeared to be the LogReg classifier with an average accuracy of 53.11%. We first evaluated all features separately, then combinations of two and three features. Using 4 features or more lead to decline in the accuracy. Interestingly, as can be observed in Table 2, by using a single feature, the *cluster* feature outperformed all other features on all three cross-genre datasets. The highest average accuracy for the traditional models (53.24%) was obtained by combining *clusters*, *words used more by men* and *char 3grams*.

		Features		
		Clusters	Char 3-grams	Unigrams
Valid	YouTube	**53.55%**	52.84%	53.28%
	Twitter	**53.00%**	52.93%	50.99%
	News	**53.11%**	52.78%	51.20%

Table 2. The performance of the best traditional model (LogReg) on the different cross-genre datasets based on the three best single features: clusters, char 3-grams and unigrams

The Winning Approach From the experiments, it resulted that, both for in-genre and cross-genre experiments, the neural models outperformed (with many configurations) the traditional approaches with different feature sets as well as the statistical language models. Therefore, we decided to continue working

---

[5] 'Corpus of spoken Dutch'

with the neural models of which the best ones were SpaCy, LSTMa, and SA. We trained 5 instances of these models, each with a different random seed. In addition, for the LSTMa and SA, 5 instances were trained with pretrained frozen word vectors and 5 without. This resulted in a total of 25 models.

Ensembling We opted for a weighted ensemble, where each model contributes its deviation from randomness, i.e. if the validation score of a model is 0.55, it will have a positive weight of 0.05 $(0.55 - 0.50)$, however, if the validation score is 0.40, the weight will be -0.10 $(0.40 - 0.50)$. Afterwards, we computed the output for every model (1 or -1) and added the prediction multiplied by the weights.

4   Results

In the following paragraphs, we present the results of our final weighted ensemble model.

The shared task allowed to submit 2 sets of results for every scenario, allowing thus for maximum of 12 different models. Our first set of results (**E-1**) are presented in Table 3. These results are all generated by the ensemble models described above. No external data was used here.

Test set	in-genre	cross-genre
Twitter	64.75%	57.89%
Youtube	62.47%	56.98%
News	66.60%	53.50%
AVG	64.61%	56.12%

Table 3. The accuracies of our first set of results on the different test sets for in-genre and cross-genre gender prediction (**E-1**).

Our second set of submissions (**E-2**) were generated by models trained with external data from CSI (Verhoeven and Daelemans, 2014) and TwiSty (Verhoeven et al., 2016). From Table 4, it results that, the scores for the in-genre data are on average 8% higher than those for the cross-genre prediction task. With in-genre prediction scores ranging between 63.49 and 66.10, the performance on all three domains are relatively similar, YouTube being the hardest genre to predict.

Test set	in-genre	cross-genre
Twitter	65.01%	55.89%
Youtube	63.49%	57.10%
News	66.30%	55.80%
AVG	64.94%	56.26%

Table 4. The accuracies of our best performing submission on the different test sets for in-genre and cross-genre gender prediction. (E-2)

With respect to the results obtained in the cross-genre setting, we observe low accuracies (55.80–57.10). However, during our experiments we observed that the models consistently got an accuracy of more than 50% on all validation data, i.e., better than random predictions, indicating that there are speaker-specific language or style features that aid to identify the gender of the speaker/writer. However, a further exploration of the outputs of the models is needed in order to draw more concrete conclusions.

5   Conclusions and Future Work

More and more research has focused on author profiling and gender prediction more particularly. The performance of models largely depends on the languages and domains involved. In this work, we explored different methods and models that achieve state-of-the-art results on various Natural Language Processing tasks and applied them on the task of gender prediction for Dutch. The plethora of empirical data we collected confirms the complexity of this task and we can not help but wonder whether we are to expect further improvements on this type of tasks for Dutch.

Acknowledgements

Bibliography

Miguel A Alvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villasenor-Pineda, and Hugo Jair-Escalante. 2015. Inaoes participation at pan15: Author profiling task. *Working Notes Papers of the CLEF*.

Angelo Basile, Gareth Dwyer, and Chiara Rubagotti. 2017. Capetownmilanotirana for gxg at evalita2018. simple n-gram based models perform well for gender prediction. sometimes.(short paper). *Children*, 100(100):65986.

Tomasso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018), Turin, Italy*, pages 211–223.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Karen Keune. 2012. *Explaining register and sociolinguistic variation in the lexicon: Corpus studies on Dutch*. Netherlands Graduate School of Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Tatiana Litvinova, Francisco M Rangel Pardo, Paolo Rosso, Pavel Seredin, and Olga Litvinova. 2017. Overview of the rusprofiling pan at fire track on cross-genre gender identification in russian. In *FIRE (Working Notes)*, pages 1–7.

Maria Medvedeva, Hessel Haagsma, and Malvina Nissim. 2017. An analysis of cross-genre and in-genre performance for author profiling in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 211–223. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016*

Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al., pages 750–784.

Francisco Manuel Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Ben Verhoeven and Walter Daelemans. 2014. Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC 2014-NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, pages 3081–3085.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al.*, pages 1–6.