CapetownMilanoTirana for GxG at Evalita2018. Simple n-gram based models perform well for gender prediction. Sometimes. (Short Paper)

Angelo Basile
Symanto Research
angelo.basile@symanto.net

Gareth Dwyer
CoGrammar
garethdwyer@gmail.com

Chiara Rubagotti
Independent Researcher
chiara.rubagotti@gmail.com

Abstract

In this paper we describe our participation in the Evalita 2018 GxG cross-genre/domain gender prediction shared task for Italian. Building on previous results obtained on in-genre gender prediction, we try to assess the robustness of a linear model using n-grams in a cross-genre setting. We show that performance drops significantly when the training and testing genres differ. Furthermore, we experiment with abstract features in trying to capture genre-independent features. We achieve an average F1-score of 0.55 on the official in-genre test set — being thus ranked first out of five submissions — and 0.51 on the cross-genre test set.

In questo articolo presentiamo il nostro contributo per lo shared task GxG di Evalita 2018 per l'analisi predittiva del genere di un autore su corpora di dominî diversi. Usiamo un modello basato su una macchina a vettori supporto con kernel lineare che usa gli n-grammi come feature. Il modello, che ha ottenuto in passato risultati eccellenti nella predizione di genere quando allenato e valutato all'interno di un unico dominio (Twitter), crolla significativamente nella performance in questo lavoro, anche quando usato in combinazione con una serie di feature astratte. Sul test set ufficiale il nostro modello raggiunge una F1-score pari a 0.55 (permettendoci di piazzarci in cima alla classifica) nel contesto in-genre e 0.51 nel contesto cross-genre.

1 Introduction

Gender prediction is the task of profiling authors to infer their gender based on their writing. This task has been carried out so far with success within one-genre/single-domain data sets, reaching state-of-the-art accuracy of 85% on English tweets (Basile et al., 2017). Cross-domain gender classification, on the other hand, has proven to be more difficult, with state-of-the-art accuracy halting at 60% (Medvedeva et al., 2017).

The theoretical assumption behind gender prediction from text is language variation: the same meaning can be expressed in different forms and this variation can be explained in terms of social variables, such as social status, personality, age and, indeed, gender (Labov, 2006; Verhoeven et al., 2016; Johannsen et al., 2015).

Proof of significant variation in language use between men and women has been found at the morpho-syntactic level (Johannsen et al., 2015) and indeed syntactic features have been used effectively for gender attribution (Sarawgi et al., 2011). Using syntax for attribution tasks has the benefit of modelling the problem in a space which is more resilient to topic and genre effects. However, we do not experiment here with deep syntactic features, but instead we try to leverage surface and frequency-based features.

In this paper we use a model built, trained, and tested for gender prediction on a single domain (i.e. Twitter). Instead of experimenting with new techniques, our aim is to sound the existing model's resilience in the different context of a cross-genre train and test setting (**RQ1**). Since we expect our model to fail on this task, we set out to design an experiment using a set of abstract features that has recently been applied for cross-lingual gender prediction (van der Goot et al., 2018): this way we want to investigate if surface features can be used to mitigate topic and genre effects (**RQ2**).

We organise this work as follows: in Section 2 we present an overview of the data set released by the organisers; in Section 3 we describe the exper-

imental set up, the model and the features used; in Section 4 we give an overview of the results and finally we conclude our work in Section 5.

The contributions of this work for the GxG task for EVALITA 2018 are the following:

- we test if a gender prediction model achieving state-of-the-art performances when trained and tested on the same domain can also achieve good results when tested in a cross-domain setting
- we experiment with abstract features in order to factor out domain-dependent effects
- we release all the code for further reproducibility at `https://github.com/anbasile/gxg-partecipation`

Task description The GxG task is a document classification task. Given a document belonging to a given genre, we have to predict the gender of the author. Thus the task is a binary classification task. The task is composed by two sub-tasks: in-genre prediction and cross-genre prediction. In the first case, the training set and the test set belong to the same genre. In the cross-genre sub-task, on the other hand, models will be trained on four genres and then tested on the single genre which they have not been exposed to during training.

2 Data

We use only the data released by the task organisers: that is, texts from five different genres. The genres are as follows.

- YouTube comments
- Tweets
- Children's writing
- News
- Personal diaries

From the data description given by the organisers we know that one author can possibly have authored multiple documents. We provide an overview of the data set in Table 1.

Data set	F	M	Tokens
Children	100	100	65986
Diaries	100	100	82989
Journalism	100	100	113437
Twitter	3000	3000	101534
Youtube	2200	2200	90639

Table 1: Overview of number of instances and corpus size per genre and label distribution.

3 Experiments

In this section we describe the feature extraction process and the model that we built. We develop one single model and train it in ten different configurations, as required by the task assignment. We train and test on five different domains, in-domain and across domains.

3.1 Pre-processing

We decide not to pre-process the data in any way, since we have no linguistic (nor non-linguistic) reasons for doing so. As a tokenization strategy for the lexicalized models we simply split on all white space tokens. For building the abstract feature representation we use `spaCy's` Italian tokenizer (Honnibal and Johnson, 2015).

3.2 Model and Features

We build a sparse linear model for approaching this task.

As features we use n-grams extracted at the word level as well as at the character level. We use 3-10 n-grams and binary TF-IDF. We feed these features to a Support Vector Machine (SVM) model with a linear kernel; we use the implementation included in `scikit-learn` (Pedregosa et al., 2011). This model in this same configuration has achieved excellent results during the PAN 2017 evaluation campaign (Potthast et al., 2017; Basile et al., 2017).

Furthermore, we experiment with feature abstraction: we follow the bleaching approach recently proposed by (van der Goot et al., 2018). First, we transform each word into a list of symbols that 1) represents the shape of the individual characters and 2) abstracts from meaning by still approximating the vowels and characters that compose the word; then, we compute the length of the word and its frequency (while taking care of padding the first one with a zero in order to

	IN					CROSS				
	DI	YT	TW	CH	JO	DI	YT	TW	CH	JO
Words (W)	0.73	0.60	0.73	0.55	0.65	0.64	0.53	0.55	0.58	0.55
Chars (C)	0.68	0.62	0.74	0.52	0.54	0.67	0.56	0.54	0.59	0.52
W+C	0.70	0.62	0.74	0.54	0.62	0.62	0.57	0.52	0.60	0.56
Bleaching	0.67	0.59	0.67	0.53	0.54	0.53	0.53	0.50	0.53	0.53

Table 2: Accuracy results on the training set per genre (DI (diaries), YT (YouTube), TW (Twitter), CH (children writing), JO (journalism)), in-genre and cross-genre. Scores are obtained via cross-validation. The cross-genre results are obtained by training on everything but the tested genre.

avoid feature collision); finally, we use a boolean label for explicitly distinguishing words from non-alphanumeric tokens (e.g. emojis). Table 3 shows an example of this feature abstraction process.

	SHAPE	FREQ	LEN	ALPHA
Questo	Cvvccv	46	06	True
è	v	650	01	True
solo	cvcv	116	04	True
un	vc	1	02	True
esempio	vcvccvv	1	07	True
.	.	60	01	False
😎	😎	1	01	False

Table 3: An illustration of the bleaching process.

(van der Goot et al., 2018) proposed this bleaching approach for successfully modelling gender across languages, by leveraging the language-independent nature of these features: here, we test if this approach is sound for mitigating the genre effect on the model.

4 Evaluation and Results

Since the data set labels are evenly distributed across the two classes, we use accuracy to evaluate our model. First, we report results obtained via a 10-fold cross-validation on the training set; then, we report results from the official test set, whose labels have been released.

4.1 Development Results

We report the development results obtained by using different text representations. Table 2 presents an overview of these results. Overall, we see that all the different feature representation formats lead to comparable results; the combination of words and characters seems to be the best combination. This is the same combination that we use for the bleached representation.

4.2 Test Results

We present official test results in Table 4. We submitted only one run. For one genre (*Diaries*) we obtained exactly the same score in both the in- and cross-genre settings: this outcome is extremely unlikely, however we ran the models several times and inspected the code for bugs and yet the results remained identical. In the cross-genre setting, we did not tune the hyper-parameters of our model considering the target genre.

The overview of the results is puzzling. First, compared to related in-domain work (Potthast et al., 2017), the overall performance is considerably lower, even taking into account domain variance. Second, the drop in performance from the in-genre to the cross-genre setting is not as high as expected. Third, even in the in-genre setting the difference in performance between genres is not trivial and it seems to be independent from training corpus's size: the two social network domains are considerably bigger in size and yet the testing scores are lower when compared to other genres.

GENRE	IN		CROSS	
	acc.	f1	acc.	f1
Diaries	0.635	0.624	0.635	0.624
YouTube	0.547	0.527	0.503	0.461
Twitter	0.545	0.542	0.555	0.540
Children	0.615	0.614	0.535	0.533
Journalism	0.480	0.460	0.515	0.381
Average	0.564	0.553	0.548	0.507

Table 4: Official test results

5 Conclusions

We presented our participation to the GxG cross-genre gender prediction task and we obtained good

results using a simple system. On top of that, we experimented with abstract features and got sub-optimal results.

Based on our experiment with a gender-prediction model which obtained state-of-the-art in-domain performance in the past, we conclude that genre plays a crucial role in gender prediction: not only genre, but the notion of *variety space*, as instructed by (Plank, 2016), should be taken into consideration for a fuller account of social variability and for building more robust systems (**RQ1**). We then attempted to improve our stock model using abstract, delexicalized features, but we failed to demonstrate any substantial improvement (**RQ2**). Furthermore, from our results it emerges that not all the examined genres pose the same challenges for gender prediction purposes: in some genres, namely journalism, the personality of the author, whether male or female, is hedged by the domain style, which favours objectivity and neutrality over self-expression and abandonment in writing. Therefore we suspect that genre-inherent style elements might make it harder for the model to carry out effective profiling of the author (be it for gender or for other social variables).

Recently, Variational Auto Encoders (VAE) (Kingma and Welling, 2013) are emerging as a good tool for properly modelling language in presence of latent variables: we plan to investigate the effectiveness of VAEs in predicting gender while modelling genre as a latent variable.

References

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, 9. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

William Labov. 2006. *The social stratification of English in New York city*. Cambridge University Press.

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.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *arXiv preprint arXiv:1608.07836*.

Martin Potthast, Francisco M. Rangel Pardo, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of pan'17 - author identification, author profiling, and author obfuscation. In *CLEF*.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86. Association for Computational Linguistics.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 383–389.

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.