Exploring Combining Training Datasets for the CLIN 2019 Shared Task on Cross-genre Gender Detection in Dutch

Gerlof Bouma
Department of Swedish / Språkbanken
University of Gothenburg
gerlof.bouma@gu.se

Abstract

We present our entries to the Shared Task on Cross-genre Gender Detection in Dutch at CLIN 2019. We start from a simple logistic regression model with commonly used features, and consider two ways of combining training data from different sources.

Our in-genre models do reasonably well, but the cross-genre models are a lot worse. Post-task experiments show no clear systematic advantage of one way of combining training data sources over the other, but do suggest accuracy can be gained from a better way of setting model hyperparameters.

1 Introduction

Detection of binary author gender can be done from text alone with impressive effectiveness. For instance, Van der Goot et al. (2018) report an accuracy of 80% on Dutch tweets for their system, and the top systems for four other languages reported in Rangel et al. (2017) also perform in the low 80% accuracy range. These results concern systems that were trained on *and* applied to Twitter data, with multiple documents (i.e., tweets) per author. It is to be expected that performance suffers when such systems are applied to another genre than they were initially trained on. The *CLIN 2019 Shared Task on Cross-genre Gender Detection in Dutch* therefore invites authors to investigate "gender prediction within and across different genres in Dutch."[1] This paper reports on our participation this shared task.

The shared task consists of an in-genre setting and a cross-genre setting. In the former, models are trained on and applied to data from the same genre/data source. In the latter, models are trained on data from one or more sources, and applied to data from a genre that is assumed to be completely unknown during training. The genres supplied in the shared task were News, Twitter and YouTube.

Having access to training data from *multiple* sources raises the question of whether we can use that fact to construct models that generalize better and therefore perform better in a cross-genre setting. Our contribution to the shared task is a small investigation of the effect of how the multiple sources are combined. Building upon a basic logistic regression model with features taken from existing research on author profiling in general and gender identification in particular, we compare two ways of combining training data sources.

Section 2 gives a formal description of the used model and introduces an alternative objective that combines training datasets in a principled way. Section 3 describes the used features and gives some implementation details. The results for our systems in the shared task are presented and briefly discussed in Section 4. These results prompt a set of post-task experiments, whose outcome and implications are reflected upon in Section 5

2 Description of the models

The core of our approach is a logistic regression model. To fit a logistic regression model with L2 regularization, we need to find an intercept b_0 and feature weights B that minimize the sum of a) the normalized negative log-likelihood of the model given the data, and b) the squared magnitude of those weights. To be precise, we minimize

$$-\frac{1}{|D|}\log\mathcal{L}(b_0, B|D) + \frac{\alpha}{2}\sum_{i=1}^{|B|} b_i^2, \qquad (1)$$

where D is the training data, and α a hyperparameter that lets us set the strength of the regularization factor. (See, e.g., Hastie et al., 2009; Malouf, 2010

[1] www.let.rug.nl/clin29/shared_task.php

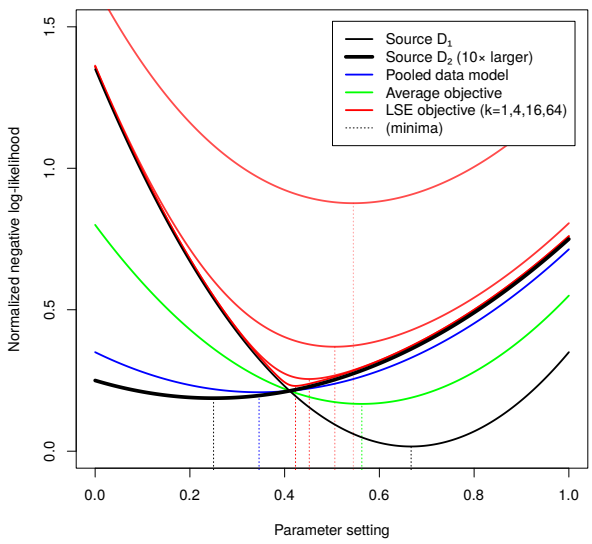

Figure 1: Hypothetical normalized negative log-likelihood curves for two source datasets D_1 and D_2, in a model with 1 parameter, and three ways of combining them. The optimal parameter setting for the two separate source datasets and their combinations are indicated by vertical dotted lines.

for proper introductions.) For the in-genre setting, we simply apply this formulation of the objective.

In the cross-genre setting, we are able to train on a combination of two or more datasets from different genres. Ideally, we would be able to leverage this fact to find models that generalize better and therefore fare better when applied to a new genre. Handling data from different sources is well-studied in the domain adaptation literature. However, there one often has access to training data from both source and target domains (like in Daumé's *frustratingly easy* method, Daume III, 2007, or equivalently, multilevel regression, Finkel and Manning, 2009), or to source training data and distributional information about the target domain (e.g., work on targeting different kinds of distributional shifts). In this shared task, however, we have two or more sources, but are supposed to assume no knowledge of the target genre. Therefore, such domain adaptation methods do not apply directly.

A direct way to combine training data from multiple sources is simply to pool the data. We can then proceed to train according to Equation 1, as we would with a single dataset. When we are dealing with source datasets of unequal size, we can also consider normalizing the respective negative log-likelihoods to the sizes of the datasets, so that the larger dataset does not dominate the model. Just summing/averaging normalized negative log-likelihoods for each source could still lead to one source dominating, if that source is much easier to model. We therefore combine the negative log-likelihoods by taking their maximum. The log-sum-exp function gives us the kind of smooth maximum we need in order to be able to use standard optimization algorithms. In the formulation we use, log-sum-exp also takes a scaling parameter:

$$\mathrm{lse}(x, y; k) = \frac{1}{k} \log(e^{kx} + e^{ky}). \qquad (2)$$

With positive k, $\mathrm{lse}(x, y; k)$ is always greater than $\max(x, y)$. A higher k makes lse less smooth but closer to max.

The objective for two datasets would thus be

$$\mathrm{lse}\left(-\frac{1}{|D_1|}\log \mathcal{L}(b_0, B|D_1), -\frac{1}{|D_2|}\log \mathcal{L}(b_0, B|D_2); k\right) + \frac{\alpha}{2}\sum_{i=1}^{|B|} b_i^2, \qquad (3)$$

where α and k are hyperparameters. This is trivially extended to more datasets.

Figure 1 illustrates the three ways of combining two source datasets graphically. The hope is that by combining sources in a balanced way, we find models that generalize better to new genres. This is on the premise that we do not know anything about the target genre. If we had information that the new genre is more like one of the source genres, we might be better off building an 'unfair' model.

3 Features, data preparation, and implementation

A linear model using surface form-based n-gram features has been shown to be very effective in (in-domain) gender identification (Basile et al., 2018), and we will follow this method here, too, albeit in simplified form. The results presented in the cited paper suggest the lion's share of accuracy is contributed by simple unigram features, and Bamman et al. (2014) present an investigation of which (classes of) lexical unigrams differentiate male from female authors on Twitter. We therefore only use unigram token occurrences in our model. Van der Goot et al. (2018) show the effectiveness of 'bleached' lexical features when doing cross-lingual gender detection. Inspired by their

	$\log_{10}\alpha$	X-val acc	Eval acc	Rank
In-genre 1				
News	1	.6386	.639	4
Twitter	1	.6327	.6316	5
YouTube	0	.6183	.6294	3
–Average		.6299	.6333	4/13
In-genre 2				
News	2	.6495	.620	6
Twitter	1	.6269	.6311	6
YouTube	2	.6194	.6233	5
–Average		.6319	.6248	5/13

Table 1: Results for the in-genre models

approach, we include word lengths as features. Finally, character n-grams are a common ingredient in author profiling. Zechner (2017), on authorship attribution, shows that even character unigram frequencies carry identifying information. These therefore constitute our final feature subset. Keeping the feature set small and simple allows us to focus on the effects of model combination.

We follow common praxis in authorship attribution (see e.g., Smith and Aldridge, 2011), in restricting the set of features to just the most frequently occurring types. The cut-off points where chosen on the basis of non-systematic trial-and-error investigation of in-genre classification. Feature frequencies are estimated from the training data, by (macro-)averaging frequency distributions from different training data sources. Also following results from authorship attribution, we use z-scores of frequencies as feature values. All in all, the feature vector for a document is made up of z-scores for the 2500 most frequent words, z-scores for the 50 most frequent characters and z-scores for the 10 most frequent word lengths.

Texts were tokenized using Cutter (Graën et al., 2018), with some provisions to treat ascii emoticons '`:P`', repeated punctuation marks '`???`' and sequences of unicode emoji '🐻🐶🐼' as single words. Other punctuation was included as any other other 'word'. All text was lower-cased before constructing the feature vectors.

Fitting the logistic regression models was done with L-BFGS using the facilities supplied by SciPy[2] and Autograd.[3]

[2] `scipy.optimize.minimize`, see `scipy.org`
[3] See `github.com/HIPS/autograd`

4 Entries to the shared task

For the in-genre models, we entered two groups: full models according to the specifications above ('in-genre 1'), and models that only uses the 2500 lexical features ('in-genre 2'). We set the regularization hyperparamater α by 5-fold cross-validation.

The results are in Table 1. Cross-validation gives fair estimates of the task evaluation results (except for one overestimate). In the task evaluation, the full models do better than the reduced. We speculate that the character and word length features, being less sparse, make the models more robust. Compared to the other entries to the shared task, both kinds of model perform reasonably well, with accuracies consistently in the top half. It should be noted that all entries to the shared task perform well below the 80% mentioned in the introduction. This is probably related to the training data set sizes and the fact that the task requires prediction on the basis of just one document.

For the cross-genre models, we also entered two groups of models: one group combining the datasets using the lse-based formulation of the objective ('cross-genre 1') and one pooling the data ('cross-genre 2'). The scaling hyperparameter k for lse was kept constant at 20, no attempts where made to optimize it, and α was set by 5-fold cross-validation. The hyperparameter setting for the model with the highest macro-averaged accuracy between data sources was chosen.

The results are in Table 2. The cross-validated accuracies are all lower than in the in-genre case: Apparently the models suffer more from having to deal with two different sources than they benefit from having larger training datasets. Comparing accuracies of cross-genre 1 (lse) to cross-genre 2 (pooling), we can observe that the pooled data model tends to cater better for the larger source dataset (viz., Twitter or YouTube), although this is only very pronounced in the model combing News and Twitter training data. The average cross-validation accuracies of the two combination methods are very similar, except in this last case, where lse has a slight advantage. The task evaluation accuracies are also very similar between the two model types. As is to be expected from the shift in genre, the cross-validation accuracy here is a very poor indication of evaluation accuracy: the latter is on average almost 8 percentage points lower. Compared to the other entries in the shared task,

			X-val acc					
		$\log_{10} \alpha$	News	Twitter	YouTube	Avg	Eval acc	Eval rank
Cross-genre 1	News	0		.6202	.6047	.6125	.510	10
	Twitter	0	.5993		.6127	.6060	.5428	7
	YouTube	0	.6183	.6175		.6084	.5252	7
	— Average					.6090	.5260	11 of 13
Cross-genre 2	News	0		.6225	.6029	.6127	.508	11
	Twitter	0	.5971		.6157	.6064	.5494	5
	YouTube	2	.5354	.6298		.5826	.5236	8
	— Average					.6006	.5270	10 of 13

Table 2: Results for the cross-genre models

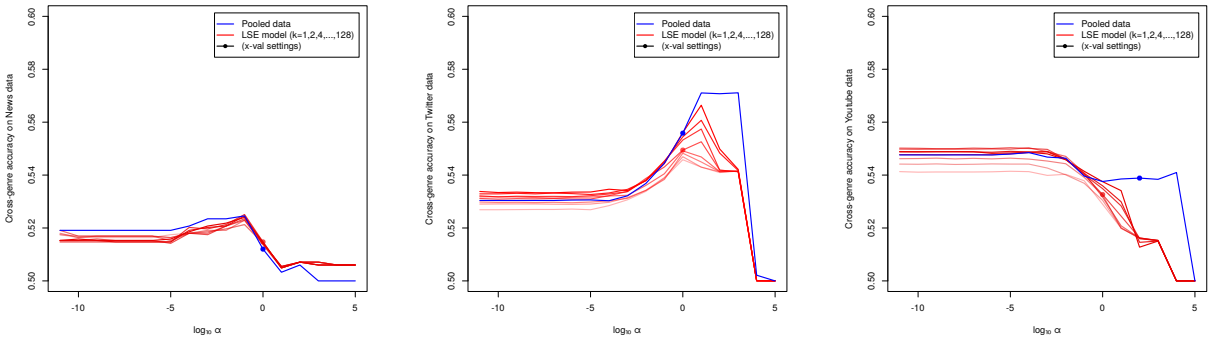

Figure 2: Cross-genre gender detection accuracy per hyperparameter setting: evaluation on News data (left), Twitter data (middle), and YouTube data (right).

we now do a lot worse, performing in the bottom segment. The poor results on the News genre – the least similar genre – are to blame for this, as we are in the middle bracket for the other two genres.

5 Post-task experiments

As mentioned, the cross-genre models do not fare as well as the in-genre ones in the shared task. In addition, the differences between lse and pooling is also small. To see to what extent these results depend on our choice of hyperparamaters, we trained models on two genres at different levels of α and k, and evaluated on a third genre. We only used the shared task's training data for these experiments. The results are in Figure 2. Note that the results of the shared task evaluation are not in here, since we did not use the task evaluation data. There does not seem to be a clear, systematic difference between accuracies for the two ways of combining data. However, we can see that the hyperparameter settings from cross-validation are suboptimal for both methods in all three datasets. In addition, choosing the right hyperparameter setting for k can make a real difference in performance, although overall it seems that a higher k is preferable.

6 Conclusions

We have presented our efforts in the Cross-genre Gender Detection shared task, where we aimed to compare two ways of combining data sources: simply pooling the data vs optimizing an objective that combines the respective negative log-likelihoods with log-sum-exp. The two methods perform similarly, and we have not seen evidence of a real advantage of using the more involved method. However, a set of post-task experiments does show that there is performance to be gained from a better way of picking the hyperparamaters in both methods.

In this work, we have not focussed on the feature set definition nor studied the effectiveness of different kinds of features in any depth. In theory, these issues are orthogonal to what we presented in our report. We thus reserve the investigation of model combination methods in the context of known state of the art feature sets for future work.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2018. Simply the best: Minimalist system trumps complex models in author profiling. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 143–156, Cham. Springer International Publishing.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, Boulder, Colorado. 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)*, pages 383–389. Association for Computational Linguistics.

Johannes Graën, Mara Bertamini, and Martin Volk. 2018. Cutter – a universal multilingual tokenizer. In *Proceedings of the 3rd Swiss Text Analytics Conference - SwissText 2018*, pages 75–81, Winterthur.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer-Verlag, New York.

Rob Malouf. 2010. Maximum entropy models. In Alex Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*, pages 133–155. Wiley Blackwell.

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. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*.

Peter Smith and W. Aldridge. 2011. Improving authorship attribution: Optimizing burrows' delta method. *Journal of Quantitative Linguistics*, 18(1):63–88.

Niklas Zechner. 2017. *A Novel Approach to Text Classification*. Ph.D. thesis, Umeå University.