Overview of the Cross-Genre Gender Prediction Shared Task on Dutch at CLIN29

Hessel Haagsma[1][0000−0003−1514−072X], Tim Kreutz[2][0000−0001−9633−1995], Masha Medvedeva[1][0000−0002−2972−8447], Walter Daelemans[2][0000−0002−9832−7890], and Malvina Nissim[1][0000−0001−5289−0971]

[1] University of Groningen, Groningen, Netherlands
{hessel.haagsma, m.medvedeva, m.nissim}@rug.nl
[2] University of Antwerp, Antwerpen, Belgium
{tim.kreutz, walter.daelemans}@uantwerpen.be

Abstract. This overview presents the results of the cross-genre gender prediction task (GxG) organized at CLIN29. Teams were tasked with training a system to predict the gender of authors of tweets, YouTube comments and news articles. In the cross-genre setting, systems were trained on two genres, and tested on the other to assess domain adaptivity of the solutions. Eight teams participated in the shared task. Performance was generally better in the in-genre setting. In the cross-genre settings, performance on news articles declined the most compared to other target genres.

1 Introduction

In this paper we give an overview of the GxG shared task at CLIN29. The next section explains the motivation and setup for the task. Section 3 gives an overview of the genres and the data that was used. We then briefly outline the participating systems and their solutions to the posed task in Section 4. Section 5 will list the final results and highlights a few notable findings.

2 Task and Settings

Despite considerable progress and increasing research activity in author profiling from text, as witnessed for example by the PAN competition[1], the problem is far from solved. One obstacle is the absence of solutions to cross-genre profiling: when models trained on one genre are applied to another, accuracy usually decreases dramatically. Typically, gender profiling for languages like English is

[1] https://pan.webis.de

in the 80-85% range, but drops to the 60% range in a cross-genre setting [2]. Clearly, features that work well for one genre may not work at all for other genres. To investigate the cross-genre profiling task in more depth for the case of gender, a shared task was organized in association with CLIN 29, the 29th conference on computational linguistics in the Netherlands. For comparability, we chose a set-up similar to the cross-genre gender detection shared task at Evalita 2018 [1].

The task is cast as a binary classification problem. Given a text, a system has to predict whether its author was male or a female. The models are trained in two settings: within the same genre and in a cross-genre setting. Teams were allowed to submit up to two runs per setting, for a potential total of 12 runs per team, six in-genre and six cross-genre.

3 Data

Data was collected for three different genres. Two genres (Twitter and YouTube) consisted of short user posts as documents. The third genre consisted of online news paper articles from ten Flemish and three Dutch news outlets. The labels were balanced in each of the genres and care was taken to provide equivalent numbers of tokens in the training portion despite different document lengths. See Table 1 for an overview of the data.

To determine an author's gender, Twitter and YouTube user profiles were cross-checked with lists of known Dutch male and female names. For the news collection, only articles that were written by a single author were considered. We then looked up their full name to determine whether the author was male or female. The final collection of news articles was written by 767 different authors, 437 male and 330 female.

Table 1. Data Overview

Genre	Train		Test	
	Documents	Tokens	Documents	Tokens
Twitter	20,000	372,361	10,000	187,893
YouTube	14,744	300,691	4,914	93,113
News	2,444	485,103	1,000	452,448

4 Participating Systems

A total of eight teams participated in the shared task. Table 2 summarizes the participants, with their affiliations, the number of submitted runs, and the letter code used in the tables reporting the results.

Table 2. List of Participants

Code	Affiliation(s)	#Runs
A	Jožef Stefan Institute, Ljubljana, Slovenia & Usher Institute, Medical school, University of Edinburgh, UK	6
B	Department of Information Science University of Groningen, The Netherlands	6
C	National Research University Higher School of Economics, Moscow, Russia	12
D	Anonymized	6
E	ADAPT, School of Computing, Dublin City University, Dublin, Ireland & Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel	12
F	Department of Information Science University of Groningen, The Netherlands	12
G	Department of Swedish / Språkbanken, University of Gothenburg, Sweden	12
H	Fraunhofer IAIS Sankt Augustin, Germany (Fraunhofer Center for Machine Learning)	12

System A uses a neural approach, more specifically a BiLSTM, trained on both word and part-of-speech n-gram features.

System B (Rob's Angels) applied a basic SVM approach with many different features and preprocessing steps. They also experimented with trying different training genres in the cross-genre setting.

System C made use of lexical features like lemmata, syntactic features like dependency relations and more abstract character-level features based on a text bleaching approach. These features were tested separately and combined in a logistic regression classifier; lexical features proved to be most effective overall.

System D did not submit a system description paper and as such we cannot report their system setup.

System E experimented with word clusters based on word embeddings. Additional features used were word unigrams and character trigrams. The eventual (winning) setup used an ensemble of different neural models whose output was weighted by their validation score

System F, (wUGs) tried a basic character n-gram SVM and a combination of SVM and logistic regression using co-training and pre-processing by normalization.

System G used a basic logistic regression model using token and character unigram features, and word lengths as features. The paper focuses on two different ways to combine the different training genres: using an LSE-based formulation of the objective and pooling the data.

System H investigated a bidirectional LSTM on word sequences and topic modeling features, and a random forest classifier using topic modeling features and function word patterns.

5 Results

Table 3 shows all results for the in-genre settings. The top runs performed best when training and testing on news. The difference in performance between the genres may be caused by the number of tokens provided for each genre. More data was available for news in comparison to Twitter and YouTube.

As expected, the performance in the cross-genre settings (Table 4) was lower for all systems. The influence of data size seems less apparent in this setting, as there is no clear difference between the three genres. However, the scores on news data have clearly suffered the most in the cross-genre setting. This may be because the Twitter and YouTube data, both being social media texts, share more commonalities. In the setting where a system is trained on the social media genres and tested on journalism, it may pick up artifacts that affect the decision process.

The outcomes of the submitted runs can be directly compared to the cross-genre gender prediction shared task at Evalita 2018 [1]. This task used Twitter, YouTube, children's writing, journalism and personal diaries as genres. Tweets and YouTube comments were collected and annotated using the same methods as used here. The journalism genre mirrors the news genre as it consists of single-author newspaper articles with gender being manually annotated.

In the cross-genre setting, results at Evalita were comparable. The best performing system at Evalita did better when training on other genres and testing on Twitter (.609 accuracy) but worse on YouTube (.510) and journalism (.495).

Because of the fluctuating scores of teams in both tasks, we cannot conclude that cross-genre profiling was more successful in either. However, the best performing team in GxG CLIN29 had consistent performance on all genres, achieving the best score for YouTube and news and the second best score for Twitter.

References

1. Dell'Orletta, F., Nissim, M.: Overview of the evalita 2018 cross-genre gender prediction (gxg) task. In: EVALITA@ CLiC-it (2018)
2. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: 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. pp. 750–784 (2016)

Table 3. Accuracy for all submitted runs, in-genre settings.

Team/Run	Twitter	YouTube	News	AVG
E-2	0.6501	0.6349	0.663	0.6493
F-1	0.6367	0.6156	0.689	0.6471
E-1	0.6475	0.6247	0.666	0.6461
G-1	0.6316	0.6294	0.639	0.6333
C-1	0.6235	0.6331	0.637	0.6312
G-2	0.6311	0.6233	0.620	0.6248
C-2	0.6115	0.6231	0.619	0.6179
B-1	0.6482	0.6091	0.594	0.6171
A-1	0.6099	0.6133	0.599	0.6074
F-2	0.6241	0.5849	0.583	0.5973
H-1	0.5945	0.5566	0.503	0.5514
H-2	0.5915	0.5511	0.494	0.5455
D-1	0.4848	0.5254	0.502	0.5041
AVG	0.6142	0.6019	0.601	0.6056

Table 4. Accuracy for all submitted runs, cross-genre settings.

Team/Run	Twitter	YouTube	News	AVG
E-2	0.5589	0.5710	0.558	0.5626
E-1	0.5789	0.5698	0.535	0.5612
A-1	0.5427	0.5507	0.552	0.5485
B-1	0.5549	0.5594	0.528	0.5474
C-1	0.5567	0.5413	0.534	0.5440
C-2	0.5467	0.5220	0.554	0.5409
H-1	0.5425	0.5227	0.548	0.5377
F-2	0.5376	0.5212	0.553	0.5373
F-1	0.5406	0.5360	0.526	0.5342
G-2	0.5494	0.5236	0.508	0.5270
G-1	0.5428	0.5252	0.510	0.5260
H-2	0.5177	0.5094	0.504	0.5104
D-1	0.4946	0.4969	0.501	0.4975
AVG	0.5434	0.5345	0.532	0.5365