Overview of the EVALITA 2018
Cross-Genre Gender Prediction (GxG) Task

Felice Dell'Orletta
ItaliaNLP Lab, ILC-CNR
Pisa, Italy
felice.dellorletta@ilc.cnr.it

Malvina Nissim
CLCG, University of Groningen
The Netherlands
m.nissim@rug.nl

Abstract

English. The Gender Cross-Genre (**GxG**) task is a shared task on author profiling (in terms of gender) on Italian texts, with a specific focus on cross-genre performance. This task has been proposed for the first time at EVALITA 2018, providing different datasets from different textual genres: Twitter, YouTube, Children writing, Journalism, Personal diaries. Results from a total of 50 different runs show that the task is difficult to learn in itself: while almost all runs beat a 50% baseline, no model reaches an accuracy above 70%. We also observe that cross-genre modelling yields a drop in performance, but not as substantial as one would expect.

Italiano. *GxG (Gender Cross-Genre) è la prima campagna di valutazione per l'identificazione del genere di un autore di testi scritti in lingua italiana e fa parte di quell'area di studio detta author profiling. In questa edizione di* ***GxG*** *particolare attenzione è stata posta nella valutazione dei sistemi in contesti di analisi cross-dominio. I domini testuali presi in esame sono stati: Twitter, YouTube, Children writing, Journalism, Personal diaries. I risultati ottenuti da un totale di 50 diverse run (prodotte da tre diversi gruppi di ricerca) mostrano che il task è complesso: mentre quasi tutti i sistemi testati superano la baseline del 50%, nessun modello raggiunge un'accuratezza superiore al 70%. Si osserva inoltre che i risultati raggiunti nel contesto di analisi cross-dominio mostrano un calo delle prestazioni non così sostanziale come ci si sarebbe potuto aspettare.*

1 Introduction

As publishing has become more and more accessible and basically cost-free, virtually anyone can get their words spread, especially online. Such ease of disseminating content doesn't necessarily go together with author identifiability. In other words: it's very simple for anyone to publicly write any text, but it isn't equally simple to always tell who the author of a text is.

In the interest of companies who want to advertise, or legal institutions, finding out at least some characteristics of an author is of crucial importance. *Author profiling* is the task of automatically discovering latent user attributes from text. Gender, which we focus on in this paper, and which is traditionally characterised as a binary feature, is one of such attributes.

Thanks to a series of tasks introduced at the PAN Labs in the last five years (pan.webis.de, and the production of a variety of gender-annotated datasets focused on social media (Verhoeven et al., 2016; Emmery et al., 2017, e.g.), gender prediction has been addressed quite substantially in NLP. State-of-the-art gender prediction on Twitter for English, the most common platform and language used for this task, is approximately 80–85% (Rangel et al., 2015; Alvarez-Carmona et al., 2015; Rangel et al., 2017; Basile et al., 2017), as obtained at the yearly PAN evaluation campaigns (pan.webis.de).

However, in the context of the 2016 PAN evaluation campaign, a cross-genre setting was introduced for gender prediction on English, Spanish, and Dutch, and best scores were recorded at an average accuracy of less than 60% (Rangel et al., 2016). This was achieved by training models on tweets, and testing them on datasets from a different source still in the social media domain, namely blogs. To further explore the cross-genre issue, (Medvedeva et al., 2017) ran additional experiments using PAN data from previous years with the model that had achieved best results at the cross-genre PAN 2016 challenge (Busger op Vollenbroek et al., 2016). The picture they obtain is mixed in terms of accuracy of cross-genre performance, eventually showing that models are not yet general enough to capture gender accurately across different datasets.

This is evidence that we have not yet found the actual dataset-independent features that do indeed capture the way females and males might write differently. To address this issue, we have designed a task specifically focused on cross-genre gender detection, and launched it within the EVALITA 2018 Campaign (Caselli et al., 2018). The rationale behind the cross-genre settings is as follows: if we can make gender prediction stable across very different genres, then we are more likely to have captured deeper gender-specific traits rather than dataset characteristics. As a by product, this task yields a variety of models for gender prediction in Italian, also shedding light on which genres favour or discourage in a way gender expression, by looking at whether they are easier or harder to model.

While Italian has featured in multi-lingual gender prediction at PAN (Rangel et al., 2015), this is the first task that addresses author profiling for Italian specifically, within and across genres.

2 Task

GxG (Gender Cross-Genre) is a task on author profiling (in terms of gender) on Italian texts, with a specific focus on cross-genre performance.

Given a (collection of) text(s) from a specific genre, the gender of the author has to be predicted. The task is cast as a binary classification task, with gender represented as F (female) or M (male).

Evaluation settings were designed bearing in mind the question at the core of this task: are there indicative traits across genres that can be leveraged to model gender in a rather genre-independent way?

We hoped to provide some answers to this question by making participants train and test their models on datasets from different genres. For comparison, participants were also recommended to submit genre-specific models, i.e., tested on the very same genre they were trained on. In-genre modelling can (i) shed light on which genres might be easier to model, i.e. where gender traits are more prominent; and (ii) make it easier to quantify the loss when modelling gender across genres. Therefore, the gender prediction task must be done in two ways:

- using a model which has been trained on the same genre
- using a model which has been trained on anything but that genre.

We selected five different genres (Section 3), and asked participants to submit up to ten different models, as per the overview in Table 1. Obviously, if one participant wanted to have one single model for everything, they could submit one model for all settings. In the cross-genre setting, the only constraint is **not** using in training any single instance from the genre they are testing on. Other than that, participants were free to combine the other datasets as they wished.

Participants were also free to use external resources, provided the cross-genre settings were carefully preserved, and everything used was described in detail in their final report.

Measures As standardly done in binary classification tasks with balanced classes (see Section 3), we will evaluate performance using accuracy.

For each of the 10 models, five in the in-genre settings, and five in the cross-genre settings, we calculate the accuracy for the two classes, i.e. F and M. In order to derive two final scores, one for the in-genre and of for the cross-genre settings, we will simply average over the five accuracies obtained per genre. In-genre:

$$Acc^{in-genre} = \frac{\sum_{j=1}^{5} Acc_j^{in-genre}}{5}$$

and cross-genre:

$$Acc^{cross-genre} = \frac{\sum_{j=1}^{5} Acc_j^{cross-genre}}{5}$$

We keep the two scorings separate. For determining the official "winner", we can consider the cross-genre ranking, as more specific to this task.

Baselines For all settings, given that the datasets are balanced for gender distribution, through random assignment we will have 50% accuracy.

Table 1: Models to submit.

IN-GENRE	CROSS-GENRE
Twitter in-genre model	non-Twitter model for Twitter
YouTube in-genre model	non-YouTube model for YouTube
Children in-genre model	non-Children model for Children
Journalism in-genre model	non-Journalism model for Journalism
Diaries in-genre model	non-Diaries model for Diaries

3 Data

In order to test the portability and stability of profiling models across genres, we created datasets from five genres. We describe them below, together with the format and train/test split of the materials distributed to participants. In Figure 1 we provide a few samples to illustrate the variety of the data, and the format provided to the participations (Section 3.3).

3.1 Genres

We selected data from the following genres grounding our choice of both availability and wide variety.

Twitter Tweets were downloaded using the Twitter API and a language identification module to restrict the selection to Italian messages[1]. Names from usernames were matched with a list of unambiguous male and female names.

YouTube YouTube comments were scraped using the YouTube API and an available scraper[2]. Videos were pre-selected manually with the aim to avoid gender biases, resulting in a selection from a few general topics: travel, music, documentaries, politics. The names of the comments' authors are visible, and gender was automatically assigned via matching first names to the same list of male and female proper names used for the Twitter dataset.

Children writing This dataset is a collection of essays written by Italian L1 learners collected during the first and second year of lower secondary school called CItA (Corpus Italiano di Apprendenti L1, (Barbagli et al., 2016)). CItA contains essays written by the same students chronologically ordered and covering a two-year temporal span. The corpus contains a total of 1,352 essays written by 153 students the first year and 155 the second year. It was collected during the two school years 2012–2013 and 2013–2014.

News/journalism This dataset was created scraping two famous Italian online newspapers (*La Repubblica* and *Corriere della Sera*) and selecting only single-authored newspaper articles. Gender assignment was done manually.

Personal diaries In order to include personal writing which is more distant from social media, we collected personal diaries that are freely available as part of the Fondazione Archivio Diaristico Nazionale della Città di Pieve Santo Stefano.[3] The documents are of varying but comparable sizes, and the author's name is clearly specified in their metadata. Gender assignment was done manually.

3.2 Train and test sets

For each genre we have a portion of training and a portion of test data. The distribution of gender labels was controlled for in each dataset (50/50). Additionally, we aimed at providing sets of comparable sizes in terms of tokens so as to avoid including training size as a relevant factor. This was intended for test, too, so as to have the same amount of evaluation samples, but due to limited availability we eventually used a smaller test set for the Diary genre.

Table 2 shows the size of the training and the test sets in terms of tokens and authors. The datasets are composed by texts written by multiple users, with possibly multiple documents per user. It is also possible that in the Twitter and YouTube datasets, different texts by the same user ended up both in training and in test. For what concerns the Children writing dataset, training and test contain texts written by same 155 children. Differently from the Children training set, the Children test set is composed by texts written on the same

[1] https://developer.twitter.com/

[2] `https://github.com/philbot9/youtube-comment-scraper`

[3] `http://archiviodiari.org/index.php/iniziative-e-progetti/brani-di-dirai.html.`

Twitter

```
<doc id="778" genre="twitter" gender="M">
@edmond644 @ilsussidiario Sarebbe vero se li avessimo eletti ma,
non avendolo fatto, "altri" se li meritano.
</doc>
```

Children

```
<doc id="1" genre="children" gender="M">
Questa estate mi sono divertito molto perché mio padre ha preso
casa nella località del Circeo. La casa era a due piani, al piano
terra c’era un giardino dove il mio gatto sela spassava.   C’era
molta ombra nel giardino e io mi ci addormivo sempre. Il mare era
poco lontano da casa e ci andavamo ogni giorno e giocavamo a fare
i subacquei. Siamo andati a mangiare la pizza  fuori ed era molto
buona.
</doc>
```

YouTube

```
<doc id="8493" genre="youtube" gender="F">
alla fine esce sempre il tuo lato gattaro! sei forte! bellissimo
video come sempre!
</doc>
```

Journalism

```
<doc id="118" genre="journalism" gender="F">
Elogio alla longevità, l’intervista bresciana a Rita Levi
Montalcini
Trent’anni fa il Nobel a Rita Levi Montalcini. Ecco l’ultima
intervista bresciana a cura di Luisa Monini: «I giovani credano
nei valori, i miei collaboratori sono tutte donne»
Tra le numerose interviste che Rita Levi Montalcini ha avuto la
bontà di concedermi, mi piace ricordare l’ultima, quella dei suoi
100 anni. Eravamo nello studio della sua Fondazione e lei era
particolarmente serena, disponibile. Elegante come sempre. [...]
</doc>
```

Diaries

```
<doc id="107" genre="diary" gender="F">
23.9.80
Sergio, volutamente stai coinvolgendo Alessandro in questa nostra
situazione, invece di tenerlo fuori: sai quanto è sensibile,
quanto è fragile, quanto è difficile anche - né puoi ignorare
che non solo lui in particolare ma nessun ragazzino di 14 anni
è in grado di subire o di affrontare o di sostenere una prova così
dolorosa.
Lo stai distruggendo, impedendogli di riflettere da solo,
martellando di parole (o scritti addirittura, come quella tua
dichiarazione) per sentirti meno solo o per annullare la sua
volontà e imporgli la tua, come volevi fare con me: ma non ti
rendi conto che non è amore il tuo, [...]
</doc>
```

Figure 1: Sample instances of all five genres from the training sets, as distributed to participants. Children, Diaries, and Journalism samples are cut due to space constraints.

Table 2: Size of datasets and label distribution.

	TRAINING			TEST		
Genre	**F**	**M**	**Tokens**	**F**	**M**	**Tokens**
Children	100	100	65,986	100	100	48,913
Diaries	100	100	82,989	37	37	22,209
Journalism	100	100	113,437	100	100	112,036
Twitter	3000	3000	101,534	3000	3000	129,846
Youtube	2200	2200	90,639	2200	2200	61,008

topic at the same time, at the end of the two school years. For News and Diaries we made sure no author was included in both training and test. We did not balance the number of users per genre, nor the number of documents per user, assuming these as rather natural conditions.

3.3 Format

The data was distributed as simil-XML. The format can be seen in Figure 1. Although we distributed one file per genre, we still included the genre information in the XML so as to ease the combination of the different files.

4 Participants and Results

Following a call for interest, 15 teams registered for the task and thus obtained the training data. Eventually, three teams submitted their predictions, for a total of 50 runs. Three different runs were allowed per task, and one team experimented with three different models submitting three different predictions for each of the 10 subtasks. A summary of participants is provided in Table 3, while Tables 4 and 5 report the final results on the test sets of the EVALITA 2018 GxG Task.

CapetownMilanoTirana proposed a classifier based on Support Vector Machine (SVM) as learning algorithm. They tested different n-gram features extracted at the word level as well as at the character level. In addition, they experimented feature abstraction transforming each word into a list of symbols and computing the length of the obtained word and its frequency (Basile et al., 2018).

UniOr tested several binary classifiers based on different learning algorithms. For the official run, they used their two best systems based on Logistic Regression (LR) and Random Forest (RF) depending on the dataset analyzed. As features, they exploited linguistic parameters extracted using stylometric analysis, such as the vocabulary richness, use of the first or third person, etc.[4]

ItaliaNLP tested three different classification models: one based on linear SVM, and two based on Bi-directional Long Short Term Memory (Bi-LSTM). The two deep neural network architectures use 2-layers of Bi-LSTM. The first Bi-LSTM layer encodes each sentence as a token sequence, the second layer encodes the sentence sequence. These two architectures differ in the learning approaches they use: Single-Task Learning (STL) and Multi-Task Learning (MTL) (Cimino et al., 2018).

5 Analysis and Discussion

In this section we provide both a discussion of the approaches and an analysis of the results.

5.1 Approaches

Participants experimented with more classical machine learning approaches as well as with neural networks. Results show that while neural models achieve globally more accurate results, feature engineered SVMs are as competitive. This holds both in the in-genre and in the cross-genre settings.

All models suffer to some extent from the shift to cross-genre, though the CapetownMilanoTirana-SVM system appears to be the most robust. This might be due more to the choice of (abstract) features, rather than the learning algorithm itself. This system also employs *bleaching* (a technique to fade out lexicon in favour of more abstract token representation) in this GxG cross-genre setting, after it had shown promise in a cross-lingual profiling task, where it was firstly introduced (van der Goot et al., 2018). However, from their cross-validation

[4] The participation of this team was not followed by a system description paper.

Table 3: Participants to the EVALITA 2018 GxG Task with number of runs.

Team Name	**Research Group**	**# Runs**
CapetownMilanoTirana	Symanto Research, CoGrammar, freelance researcher	10
UniOr	Università Orientale di Napoli	10
ItaliaNLP Lab	ItaliaNLP Lab, ILC-CNR, Pisa	30

Table 4: Results in terms of Accuracy of the EVALITA 2018 GxG In-Domain Task.

Team Name-Model	**CH**	**DI**	**JO**	**TW**	**YT**	TOT
CapetownMilanoTirana-SVM	0.615	0.635	0.480	0.545	0.547	0.564
UniOr-LR-RF	0.550	0.550	**0.585**	0.49	0.500	0.535
ItaliaNLP Lab-SVM	0.550	0.649	0.555	0.567	**0.555**	0.575
ItaliaNLP Lab-STL	0.545	0.541	0.500	**0.595**	0.512	0.538
ItaliaNLP Lab-MTL	**0.640**	**0.676**	0.470	0.561	0.546	0.578
avg-Accuracy	0.580	0.610	0.518	0.552	0.532	0.558

Table 5: Results in terms of Accuracy of the EVALITA 2018 GxG Cross-Domain Task.

Team Name-Model	**CH**	**DI**	**JO**	**TW**	**YT**	TOT
CapetownMilanoTirana-SVM	0.535	**0.635**	**0.515**	0.555	0.503	0.549
UniOr-LR-RF	0.525	0.550	0.415	0.500	0.500	0.498
ItaliaNLP Lab-SVM	0.540	0.514	0.505	0.586	**0.513**	0.532
ItaliaNLP Lab-STL	**0.640**	0.554	0.495	**0.609**	0.510	0.562
ItaliaNLP Lab-MTL	0.535	0.595	0.510	0.500	0.500	0.528
avg-Accuracy	0.555	0.570	0.488	0.550	0.505	0.534

results on training data, where they also perform an evaluation of feature contribution, it seems that bleaching in this context does not yield the expected benefits (Basile et al., 2018).

The use of external resources was globally little explored, with the exception of generic word embeddings (ItaliaNLP Lab). While such embeddings do not seem to have contributed much to performance, specialised lexica or embeddings could be something to be investigated in the future.

From a learning settings' perspective, teams chose quite straightforward strategies. In-genre, all models were trained using data from the target genre only, with the exception of the ItaliaNLP Lab-MTL model, where the adopted multitask strategy could use the knowledge from other genres, even in the in-genre settings. This seems confirmed comparing the ItaliaNLP Lab-MTL's results with the twin model ItaliaNLP Lab-STL (the same architecture with a single-task setting).

In the cross-genre scenario, all systems have used as training all of the available datasets apart from the dataset from the target genre. It appears that no team tried to exploit functions of (dis)similarity among genres in order to select subportions of data for training when testing on a given genre. The only different model in the way the training data from the other genres is used is the ItaliaNLP Lab-MTL, but its performance on the cross-genre setting indicates that this approach is not robust for this task.

5.2 Results

The in-domain results (Table 4) are useful to identify which genres are overall easier to model in terms of the author's gender, and provide an overview of gender detection in Italian.

As could be expected, Diaries are the easiest genre to model. This might result from the fact that the texts are longer, and are characterised by a more personal and subjective writing style. For example, the collected diaries present an extensive use of the first and second singular person verbs and a higher distribution of possessive adjectives.

Due to availability, the Diaries test set is smaller than the others, providing thus fewer instances for evaluation (see Table 2) and possibly weaker reliability of results. However, from the analysis of results reported by (Basile et al., 2018), we see that even in the cross-validated training set (100 samples), accuracy on Diaries is the highest out of the five genres.

We also see that Children writings carry better signal towards gender detection than social media. This might be due to the fact that the Children test set is composed by documents characterised by a common prompt (in the original collection settings, this was meant to provide evidence of how students perceive the different writing instructions received in the considered school years). This feature makes the children texts a reflexive textual typology, typically characterised by a more subjective writing style as we observed also for Diaries.

Both Twitter and YouTube score above a 50% baseline, but are clearly harder to model. This could be due to the short texts, which in some cases offer very little evidence to go by. Compared to previous results reported for gender detection on Twitter in Italian, as obtained at the 2015 PAN Lab challenge on author profiling (Rangel et al., 2015), and on the TwiSty dataset (Verhoeven et al., 2016), scores at GxG are substantially lower (PAN 2015's best performance on Italian: .8611, TwiSty's reported F-measure: .7329). The reason for this could be the fact that in the PAN Lab and Twisty datasets authors are represented by a collection of tweets, while we do not control for this aspect at all, under the assumption that if gender traits emerge, these should not depend on having large evidence from a single author. It could therefore be that the PAN's and TwiSty's results are inflated by partially modelling authors rather than gender. Another interesting observation regarding Twitter is that when cross-validating the training set, (Basile et al., 2018) report accuracy in the 70s, while their in-genre results on the test data are just above 50% (Table 4).

The most difficult genre is Journalism. While texts can be as long as diaries, results suggest that the requested jargon and writing conventions typical of this genre overshadow the gender signal. Moreover, while we have selected only articles written by a single author, there is always the chance that revisions are made by an editor before the piece is published. This is the only genre where some models fail to beat the 50% baseline. However, it is interesting to note that the highest score in-genre is achieved by UniOr, which uses a selection of stylometric features (Koppel et al., 2002; Schler et al., 2006, e.g.), which have long been thought to capture unconscious behaviour better than just lexical choice (on which most of the other models are based, as they mainly use n-grams). The highest Journalism score in the cross-settings is achieved by CapetownMilanoTirana.

Cross-genre, we observe that results are on average lower, but only by 2.5 percentage points (55.8 vs 53.4), which is less than one would expect. Some models clearly drop more heavily from in-genre to cross-genre (ItaliaNLP Lab-MTL: .578 vs .528 average accuracy, ItaliaNLP Lab-SVM: .575 vs .532, UniOr: .535 vs .498). However, others appear more stable in both settings (CapetownMilanoTirana-SVM: .564 vs .549), or even better at the cross- rather than in-genre prediction (ItaliaNLP Lab-STL: .538 vs .562).

From a genre perspective, the drop is more substantial for some genres, with Diaries losing the most, with large variation though across systems. For example, the model that achieves best performance on Diaries in-genre (ItaliaNLP Lab-MTL, .676) suffers a drop of almost eight percentage points on the same dataset cross-genre (.595). Conversely, CapetownMilanoTirana preserve a high performance on the Diaries testset in both in- and cross-settings (.635), yielding the highest cross-performance on this genre. Twitter shows the least variation between in- and cross-genre testing. Not only the losses for all systems are minimal, but in some cases we even observe higher scores in the cross-genre setting. ItaliaNLP Lab-STL obtains the highest score on this test set in both settings, but the performance is higher for cross- than in-genre (.609 vs .595).

Finally, some possibly interesting insights also emerge from looking at the precision and recall for the two classes (which we do not report in tables as there are too many data points to show). For example, we observe that for some genres only one gender ends up being assigned almost entirely by all classifiers. This happens in the cross-genre settings, while the same test set has rather balanced assignments of the two classes in the in-genre settings. In the case of Journalism, all systems in cross-genre modelling almost only predict female as the author's gender. At the opposite side of the spectrum we find YouTube, where almost all test instances are classified as male by almost all systems, cross-genre. In this case though we also see high recall for male in the in-genre setting, though not so dominant. While more grounded considerations are left to a deeper analysis in the future, we could speculate that some genres are globally seen by classifiers as more characteristic of one gender or the other, as learnt from a large amount of mixed-genre, gender-balanced data.

6 Conclusions

Gender detection was for the first time the focus of a dedicated task at EVALITA. The GxG task specifically focussed on comparing performance within and across genres, building on previous observations that high performing systems were likely to be modelling datasets rather than gender, as their accuracy substantially dropped when tested on different, even related, domains.

Results from 50 different runs were mostly above baseline for most prediction tasks, both in-genre and cross-genre, but not particularly high overall. Also, the drop between in-genre and cross-genre performance is noticeable, but marginal. Neural models appear to perform only slightly better than a more classic SVM which leverages character and word n-grams. The use of the recently introduced *text bleaching* strategy among the engineered features (aimed at reducing lexicon bias (van der Goot et al., 2018)), does not seem to yield the desired performance in the cross-genre settings.

In the near future, it will be interesting to compare these results to human performance, as it has been observed that for profiling human do not perform usually better than machines (Flekova et al., 2016; van der Goot et al., 2018), to which they provide complementary performance in terms of correctly predicted instances.

Acknowledgments

We would like to thank Eleonora Cocciu and Claudia Zaghi for their help with data collection and annotation. We are also grateful to the EVALITA chairs for hosting this task.

References

Miguel A Alvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villasenor-Pineda, and Hugo Jair Escalante. 2015. INAOE's participation at PAN'15: Author profiling task. In *CLEF Working Notes*.

Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. Cita: an l1 italian learners corpus to study the development of writing competence. In *In Proceedings of the 10 thConference on Language Resources and Evaluation (LREC 2016)*, pages 88–95.

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-gram: New groningen author-profiling model. In *Proceedings of the CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland (Sept. 2017).*

Angelo Basile, Gareth Dwyer, and Chiara Rubagotti. 2018. *CapetownMilanoTirana* for GxG at Evalita2018. Simple n-gram based models perform well for gender prediction. Sometimes. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

M. Busger op Vollenbroek, T. Carlotto, T. Kreutz, M. Medvedeva, C. Pool, J. Bjerva, H. Haagsma, and M. Nissim. 2016. Gron-UP: Groningen user profiling notebook for PAN at CLEF. In *CLEF 2016 Evaluation Labs and Workshop, Working Notes*.

Tommaso 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 Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55, Copenhagen, Denmark.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany, August.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17:401–412.

Maria Medvedeva, Hessel Haagsma, and Malvina Nissim. 2017. An analysis of cross-genre and in-genre performance for author profiling in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017*, pages 211–223.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *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*, pages 750–784.

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. *CLEF Working Notes*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI.

Rob van der Goot, Nikola Ljubesic, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 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 International Conference on Language Resources and Evaluation (LREC 2016)*.