TAG-it @ EVALITA2020: Overview of the Topic, Age, and Gender Prediction Task for Italian

Andrea Cimino
ItaliaNLP Lab, ILC-CNR
Pisa, Italy
andrea.cimino@ilc.cnr.it

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

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

Abstract

The Topic, Age, and Gender (TAG-it) prediction task in Italian was organised in the context of EVALITA 2020, using forum posts as textual evidence for profiling their authors. The task was articulated in two separate subtasks: one where all three dimensions (topic, gender, age) were to be predicted at once; the other where training and test sets were drawn from different forum topics and gender or age had to be predicted separately. Teams tackled the problems both with classical machine learning methods as well as neural models. Using the training-data to fine-tuning a BERT-based monolingual model for Italian proved eventually as the most successful strategy in both subtasks. We observe that topic and gender are easier to predict than age. The higher results for gender obtained in this shared task with respect to a comparable challenge at EVALITA 2018 might be due to the larger evidence per author provided at this edition, as well as to the availability of pre-trained large models for fine-tuning, which have shown improvement on very many NLP tasks.

1 Introduction

Author profiling is the task of automatically discovering latent user attributes from text, among which gender, age, and personality (Rao et al., 2010; Burger et al., 2011; Schwartz et al., 2013; Bamman et al., 2014; Flekova et al., 2016; Basile et al., 2017).

Past work in Natural Language Processing has contributed to advancing this task especially through the creation of resources, also in languages other than English (Verhoeven et al., 2016; Rangel et al., 2017, e.g.,), for training supervised models. Across the years, especially thanks to the organisation of shared tasks in the context of the PAN Labs, it has become evident that models that exploit lexical information, mostly in the form of word and character n-grams, make successful predictions (Rangel et al., 2017; Basile et al., 2018; Daelemans et al., 2019).

However, cross-genre experiments (Rangel et al., 2016; Busger op Vollenbroek et al., 2016; Medvedeva et al., 2017; Dell'Orletta and Nissim, 2018) have revealed that most successful approaches, exactly because they are based on lexical clues, tend to model *what* rather than *how* people write, capturing topic instead of style. As a consequence, they lack portability to new genres and more in general just new datasets.

The present work aims at shedding some more light in this direction, and at the same time increase resources and visibility for author profiling in Italian. We propose a shared task in the context of EVALITA 2020 (Basile et al., 2020) that can be broadly conceived as stemming from a previous challenge on profiling in Italian, i.e., GxG, a cross-genre gender prediction task. The new task is TAG-it (Topic, Age, and Gender prediction in Italian). With TAG-it, we introduce three main modifications with respect to GxG. One is that age is added to gender in the author profiling task. Another one is that, in one of the tasks, we conflate author and text profiling, requiring systems to simultaneously predict author traits and topic. Lastly, we restrict the task to in-genre modelling,

pan.webis.de

but we explicitly control for topic through two specific subtasks.

2 Task

TAG-it (Topic, Age and Gender prediction for Italian) is a profiling task for Italian. This can be broadly seen as a follow-up of the GxG (Dell'Orletta and Nissim, 2018) task organised in the context of EVALITA 2018 (Caselli et al., 2018), though with some differences.

GxG was concerned with gender prediction only, and had two distinctive traits: (i) models were trained and tested cross-genre, and (ii) evidence per author was for some genres (Twitter and YouTube) extremely limited (one tweet or one comment). The combination of these two aspects yielded scores that were comparatively lower than those observed in other campaigns, and for other languages. A core reason for the cross-genre setting was to remove as much as possible genre-specific traits, but also topic-related features. The two would basically coincide in most n-gram-based models, which are standard for this task.

In TAG-it, the task is revised addressing these two aspects, for a better disentanglement of the dimensions. First, only a single genre is considered (forum posts). Second, longer texts are used, which should provide better evidence than single tweets, and are more coherent than just the concatenation of more tweets. Third, "topic control" is introduced in order to assess the impact on performance of the interaction of topic and author's traits, in a more direct way than in GxG (where it was done indirectly via cross-genre prediction).

Data was collected accordingly, including information regarding topic and two profiling dimensions: gender and age. The interesting aspect of this is that we mix text profiling and author profiling, with tasks and analysis that treat their modelling both at once as well as separately. In practice, we devise and propose two tasks.

Task 1: Predict all dimensions at once Given a collection of texts (forum posts) the gender and the age of the author must be predicted, together with the topic the posts are about. The task is cast as a multi-label classification task, with gender represented as F (female) or M (male), age as five different age bins, as it has been done in past profiling tasks involving age (Rangel et al., 2015, e.g.,), and topic as 14 class values.

Task 2: Predict age and gender with topic control For posts coming from a small selection of topics not represented in the training data, systems have to predict either gender (Task 2a) or age (Task 2b).

For both tasks, participants were also free to use external resources as they wish, provided the cross-topic settings would be preserved, and that everything used would be described in detail.

3 Data

3.1 Collection

In order to generate the data for the tasks, we exploited a corpus collected by Maslennikova et al. (2019). This corpus consists of 2.5 million posts scraped from the ForumFree platform. The posts are written by 7.023 different users in 162 different forums. Information about the authors' gender and age is available.

In order to have enough data for the topic classification task, we decided to aggregate data from several forums into a single topic. For example, data from the forums *500x* and *a1audiclub* where manually classified into the *AUTO-MOTO* topic, while the forums *bellicapelli* and *farmacieonlinesicure* in the *MEDICINE-AESTHETICS* topic. At the end of the aggregation process, we obtained 31 different topics. The selection of the topics that we use in TAG-it is shown in Table 1.

For age classification, we bin age into 5 age groups: (0,19), (20, 29), (30, 39), (40, 49) and (50-100). In addition, we performed a final selection of users in order to have sufficient evidence per author. More precisely, we selected only the users that wrote at least 500 tokens across their posts. The first 500 tokens of their posts were used as textual data while the other posts from the same users were discarded. At the end of this process, we obtained posts belonging to 2,458 unique users. Table 1 reports some corpus statistics, already arranged according to the experimental splits that we used in the different tasks (see Section 3.2).

3.2 Training and test sets

The data obtained from the process described in the previous subsection was used to generate the training and test data. The training data is the same for Task 1 and Task 2. It contains a variety of topics, and we aimed at a good label distribution for both gender and age, though the forum

https://www.forumfree.it/?wiki=About

Topic	M	F	0-19	20-29	30-39	40-49	50-100
Training data for all tasks							
ANIME	133	114	77	112	33	19	6
MEDICINE-AESTHETICS	16	13	0	2	13	9	5
AUTO-MOTO	221	5	5	41	42	67	71
SPORTS	285	15	19	102	74	62	43
SMOKE	79	0	0	9	25	22	23
METAL-DETECTING	77	1	5	11	15	28	19
CELEBRITIES	23	26	1	25	8	7	8
ENTERTAINMENT	28	4	5	16	8	0	3
TECHNOLOGY	5	1	3	1	0	1	1
NATURE	24	12	7	9	9	4	7
BIKES	25	2	2	2	3	7	13
Test data for Task 1							
ANIME	46	51	27	43	13	8	6
MEDICINE-AESTHETICS	7	9	1	4	6	3	2
AUTO-MOTO	73	3	1	13	21	18	23
SPORTS	92	11	7	37	23	18	18
SMOKE	29	1	0	6	9	8	7
METAL-DETECTING	25	1	0	2	6	8	10
CELEBRITIES	7	15	0	8	5	2	7
ENTERTAINMENT	9	0	1	6	2	0	0
TECHNOLOGY	9	0	1	5	3	0	0
NATURE	7	4	1	3	6	1	0
BIKES	11	1	0	4	1	3	4
Test data for Task 2a							
GAMES	274	24	47	128	41	44	38
ROLE-GAMES	70	44	29	61	10	4	10
Test data for Task 2b							
CLOCKS	386	1	3	41	83	168	92
GAMES	274	24	47	128	41	44	38
ROLE-GAMES	70	44	29	61	10	4	10

Table 1: Number of unique users (shown by gender and age) for each topic in the training and test sets of both tasks.

data is overall rather unbalanced for these two dimensions. In the selection of test data, we had to differentiate between the two task since for Task 1 test topics should correspond to those in training, while they should differ for Task 2.

For Task 1, each topic was split into 70% for training and 30% for test. For Task 2, we picked posts from topics not present in the training data, and more specifically used the forums CLOCKS, GAMES, and ROLE-GAMES for Task 2a, and only GAMES and ROLE-GAMES for Task 2b in order to ensure more balanced data. Table 2 shows the size of the datasets in terms of tokens.

The data was distributed as simil-XML. The format can be seen in Figure 1. The test data was released blind to the participants who were given a week to return their prediction to the organisers.

4 Evaluation

System evaluation was performed using both standard (accuracy, precision, recall, and f-score), as well as ad hoc measures.

DATASET	M	F	0-19	20-29	30-39	40-49	50-100
Training for all Tasks	533,195	114,723	74,349	199,902	132,518	132,130	109,019
Test Task1	180,646	70,407	24,259	77,869	53,955	40,196	54,774
Test Task2a	225,416	43,318	47,659	135,347	29,337	27,623	28,768
Test Task2b	438,759	43,834	50,583	158,704	76,986	117,721	78,599

Table 2: Number of tokens for gender and age contained in training and test data.

Team Name	Research Group	# Runs
UOBIT	Computer Science Department, Universidad de Oriente, Santiago de Cuba, Cuba	9
UO4to	Computer Science Department, Universidad de Oriente, Santiago de Cuba, Cuba	2
ItaliaNLP	Aptus.AI, Computer Science Department, ItaliaNLP Lab (ILC-CNR), Pisa, Italy	9

Table 3: Participants to the EVALITA 2020 TAG-it Task with number of runs.

```
<user id="2" topic="BIKES" age="40-49" gender="M">

<post>
perfetto direi veramente ingegnoso
</post>

<post>
Ma come hai carpito queste notizie certe?
Hai fermato le signore ad un posto di blocco
spacciandoti per agente di polizia?
</post>

<post>
A chent'annos Alessandro.
</post>

[...]

</user>
```

Figure 1: Sample of a training instance.

For Task 1, the performance of each system was evaluated according to two different measures, which yielded two different rankings. In the first ranking we use a partial scoring scheme (Metric 1), which assigns 1/3 to each dimension correctly predicted. Therefore, if no dimension is predicted correctly, the system is scored with 0, if one dimension is predicted correctly the score is 1/3, if two dimensions are correct the score is 2/3, and if all of age, gender, and topic are correctly assigned, then the score for the given instance is 1.

In the second ranking (Metric 2), 1 point is assigned if all the dimensions are predicted correctly simultaneously, 0 otherwise. This corresponds to the number of '1' points assigned in Metric 1.

For each ranking, the final score is the sum of the points achieved by the system across all the test instances, normalized by the total number of instances in the test set.

For Task 2, the standard micro-average f-score was be used as scoring function. For carrying out further analysis, we also report macro-f.

Baselines For all tasks, we introduced two baselines. One is a data-based majority baseline, which assign the most frequent label in the training data to all test instances. The other one is an SVM-based model (*SVM baseline* hereafter), as SVMs are known to perform well in profiling tasks (Basile et al., 2018; Daelemans et al., 2019).

This classifier is implemented using scikit-learn's `LinearSVC` (Pedregosa et al., 2011) with default parameters, using as features up to 5-grams of characters and up to 3-grams of words (frequency counts).

5 Participants

Following a call for interest, 24 teams registered for the task and thus obtained the training data. Eventually, three teams submitted their predictions, for a total of 20 runs. Three different runs were allowed per task. A summary of participants is provided in Table 3.

Overall, participants experimented with more classical machine learning approaches as well as with neural networks, with some of them employing language model based neural networks models such as multilingual BERT (Devlin et al., 2019) and UmBERTo. While the UO4to team (Artigas Herold and Castro Castro, 2020) proposed a classical feature engineered ensamble approach, UOBIT (Labadie et al., 2020) and Ital-

https://github.com/musixmatchresearch/umberto iaNLP (Occhipinti et al., 2020) experimented different deep learning techniques. UOBIT proposed a novel approach based on a combination of different learning components, aimed at capturing different level of information, while ItaliaNLP experimented with both SVM and Single and Multi task learning settings using a state-of-the-art language model specifically tailored for the Italian language.

Even if allowed, the use of external resources was not explored most probably due to great performances already provided by the latest deep learning language models w.r.t featured engineered models.

The following paragraphs provide a summary of each team's approach for ease of reference.

UOBIT tested a deep learning architecture with 4 components aimed at capturing different information from documents. More precisely, they extracted information from the layers of a finedtuned multilingual version of BERT (T), used information from a LSTM trained with FastText input vectors (RNN-W), they added raw features for stylistic feature extraction (STY) and finally they extracted information from a sentence encoder (RNN-S). The information from all the four components is finally concatenated and fed into a dense layer.

UO4to participated to Task 1 with two different ensemble classifiers, using Random Forest, Nearest Centroid and OneVsOneClassfier learning algorithms provided by the scikit-learn library (Pedregosa et al., 2011). They used n-grams of characters using term frequency or TF-IDF depending on the used configuration.

ItaliaNLP tested three different systems. The first one is based on three different SVM models (one for each dimension to be predicted), using character n-grams, word n-grams, Part-Of-Speech n-grams and *bleached* (van der Goot et al., 2018) tokens. The second one is based on three different BERT-based classifier using UmBERTo as a pre-trained language model, modelling each task separately. Finally, they tested a multi–task learning approach to jointly learn the three tasks, again using UmBERTo as a language model.

6 Results and Analysis

Tables 4 and 6 report the final results on the test sets of the EVALITA 2020 TAG-it Task 1 and Task 2 respectively, using the official evaluation metrics. For all tasks, the ItaliaNLP system achieves the best scores. Before delving into the specifics of each task, and into a deeper analysis of the results, we want to make a general observation regarding approaches. SVMs have longed proved to be successful at profiling, and this trend emerged also at the last edition of the PAN shared task on author profiling (Daelemans et al., 2019). In our tasks, we also observe that the SVM baseline that we have trained for comparison is competitive. However, the submitted model that achieves best results is neural.

Team Name-MODEL	**Metric 1**	**Metric 2**
Majority baseline	0.445	0.083
SVM baseline	0.674	0.248
UOBIT-(RNN-W T STY)	0.686	0.250
UOBIT-(RNN-S T STY)	0.674	0.243
UOBIT-(RNN-W RNN-S T STY)	0.699	0.251
UO4to-ENSAMBLE-1	0.416	0.092
UO4to-ENSAMBLE-2	0.444	0.092
ItaliaNLP-STL-SVM	0.663	0.253
ItaliaNLP-MTL-UmBERTo	0.718	0.309
ItaliaNLP-STL-UmBERTo	**0.735**	**0.331**

Table 4: Results according to TAG-it's Metric 1 and Metric 2 for Task 1.

Task 1 The best ItaliaNLP model achieves the scores of 0.735 for Metric 1 and 0.331 for Metric 2, which accounts for correctly predicted instances according to all dimensions at once. The other systems' performance is quite a bit lower. For Metric 1 UOBIT's best system still performs above all baselines, while UO4to only above majority baseline. Also according to Metric 2, UO4to performs above majority baseline but not better than the SVM.

For a deeper understanding of the results in Task 1, we look at the separate performance on the various dimensions, including both micro-F and macro-F scores, as label distribution is not balanced (Table 5).

What clearly emerges from the table is that classification of gender and topic is much easier than classification of age. This seems to suggest that textual cues are more indicative of these dimensions than age. Gap between best submitted (neural) model and SVM is way wider for topic and gender than for age.

	Micro-F			Macro-F		
Team Name-MODEL	**Topic**	**Gender**	**Age**	**Topic**	**Gender**	**Age**
Majority baseline	0.251	0.766	0.319	0.036	0.434	0.097
SVM baseline	0.808	0.832	0.382	0.565	0.683	0.319
UOBIT-(RNN-W T STY)	0.859	0.842	0.358	0.751	0.736	0.343
UOBIT-(RNN-S T STY)	0.835	0.856	0.331	0.724	0.797	0.303
UOBIT-(RNN-W RNN-S T STY)	0.869	0.869	0.360	0.791	0.811	0.337
UO4to-ENSAMBLE-1	0.333	0.523	0.392	0.172	0.517	0.341
UO4to-ENSAMBLE-2	0.470	0.521	0.341	0.394	0.515	0.302
ItaliaNLP-STL-SVM	0.774	0.810	0.404	0.502	0.619	0.347
ItaliaNLP-MTL-UmBERTo	0.873	0.873	0.406	0.716	0.716	0.358
ItaliaNLP-STL-UmBERTo	**0.898**	**0.891**	**0.416**	**0.804**	**0.834**	**0.377**

Table 5: Results according to micro and macro F-score for TAG-it's Task 1, for each separate dimension.

	Task 2a		Task 2b	
Team Name-MODEL	**Micro-F**	**Macro-F**	**Micro-F**	**Macro-F**
Majority baseline	0.835	0.455	0.288	0.089
SVM baseline	0.862	0.618	0.393	0.304
UOBIT-(RNN-W T STY)	0.852	0.692	0.278	0.272
UOBIT-(RNN-S T STY)	0.883	0.796	0.370	0.320
UOBIT-(RNN-W RNN-S T STY)	0.893	0.794	0.308	0.303
ItaliaNLP-STL-SVM	0.852	0.608	0.374	0.300
ItaliaNLP-MTL-UmBERTo	**0.925**	**0.846**	0.367	0.328
ItaliaNLP-STL-UmBERTo	0.905	0.816	**0.409**	**0.344**

Table 6: Results according to micro and macro F-score for TAG-it's Task 2a (gender) and Task 2b (age).

Task 2 As for Task 1, the best system is a neural model submitted by ItaliaNLP, both for Task 2a (gender) and Task 2b (age). All of the models perform above majority baseline, in spite of this task being potentially more complex since train and test data are drawn from different topics. As observed before, the gap between models and both baselines is higher for gender than for age. In addition to the previous observation that textual clues could be more indicative of gender than age, this lower performance could also be due to the fact that gender prediction is cast as a binary task while age is cast as a multiclass problem, turning a continuous scale into separate age bins.

In-depth Analysis Although official results are provided as micro-F score, we also report macro-F since classes are unbalanced and it is important to assess the systems' ability to discriminate well both classes. In gender prediction (Task 2a), comparing macro and micro F-scores, we observe that the gap between the two metrics is much lower for the neural models than for the SVMs (both our baseline as well as the system submitted by ItaliaNLP). This suggests that neural models are better able to detect correct cases of both classes, rather than majority class only.

We can also observe that in both tasks, results for age are not only globally lower than for gender, but also closer to one another across the submissions. We therefore zoom in on the age prediction task by comparing the confusion matrices of our SVM baseline and the best ItaliaNLP model, both in Task 1 (just the age prediction part) and in Task 2b. These are shown in Figure 2 and Figure 3 respectively.

What can be observed right away is that errors are not random, rather they are more condensed in classes closer to each other, underlining the ability of the systems. This is particularly true for the neural model (left in the Figures), where we

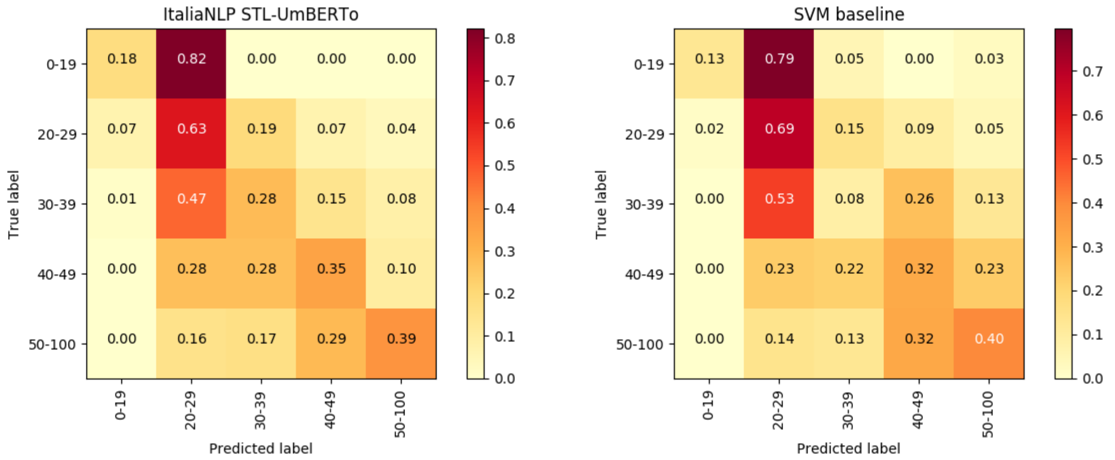

Figure 2: Normalized confusion matrices of the best ItaliaNLP system and the SVM baseline for Task 1 on the age dimension.

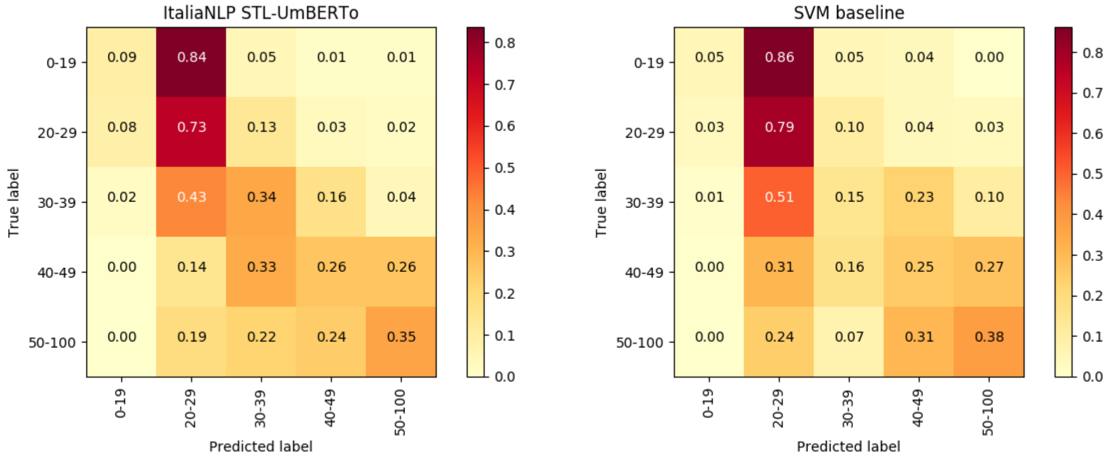

Figure 3: Normalized confusion matrices of the best ItalianNLP system and the SVM baseline for Task 2b.

can see the most confounded classes are the closest ones, thus generating a more uniform darker cluster along the diagonal.

Comparison to GxG As mentioned, TAG-it could be seen as a continuation of the GxG task at EVALITA 2018. In the latter, teams were asked to predict gender within and across five different genres. In TAG-it, in terms of profiling, we add age, which we cannot obviously compare to performances in GxG, and we use one genre only (forum posts), but implement a cross-topic setting.

We observe that results at TAG-it for gender prediction are higher than in GxG both within and cross-domain. We believe these are ascribable mainly to two relevant differences between the two tasks: (i) in this editions authors were represented by multiple texts, while in GxG, for some domains, evidence per author was minimal, and (ii) texts in TAG-it are probably less noisy, at least in comparison to some of the GxG genres (e.g., tweets and YouTube comments). Lastly, methods evolve fast, and since GxG was run in 2018, the use of Transformer-based models was not as spread as today. It would thus be interesting to assess the impact of fine-tuning large pre-trained models (as it's done in the best model at TAG-it) to gain further improvements in gender prediction.

One aspect that seems relevant in this respect is the appropriateness of the pre-trained model. Both ItaliaNLP and UOBIT used fine-tuned pre-trained models. However, while the latter used multilingual BERT as base, the former used the monolingual UmBERTo, obtaining higher results. This suggests, as it has been recently shown for a variety of tasks (Nozza et al., 2020), that monolingual models are a better choice for language-specific downstream tasks.

References

Maria Fernanda Artigas Herold and Daniel Castro Castro. 2020. TAG-it 2020: Ensemble of Machine Learning Methods. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

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. 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.*

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 *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 143–156. Springer.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. 2016. GronUP: Groningen user profiling notebook for PAN at CLEF. In *CLEF 2016 Evaluation Labs and Workshop: Working Notes Papers.*

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview on 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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Walter Daelemans, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Michael Tschuggnall, et al. 2019. Overview of pan 2019: Bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 402–416. Springer.

Felice Dell'Orletta and Malvina Nissim. 2018. Overview of the EVALITA 2018 cross-genre gender prediction (GxG) task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

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. Association for Computational Linguistics.

Roberto Labadie, Daniel Castro Castro, and Reynier Ortega Bueno. 2020. UOBIT@TAG-it: Exploring a multi-faceted representation for profiling age, topic and gender in Italian texts. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Aleksandra Maslennikova, Paolo Labruna, Andrea Cimino, and Felice Dell'Orletta. 2019. Quanti anni hai? age identification for italian. In *Proceedings of 6th Italian Conference on Computational Linguistics (CLiC-it), 13-15 November, 2019, Bari, Italy.*

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. Proceedings of the 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017*, pages 211–223.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making sense of language-specific BERT models. arXiv:2003.02912.

Daniela Occhipinti, Andrea Tesei, Maria Iacono, Carlo Aliprandi, and Lorenzo De Mattei. 2020. ItaliaNLP @ TAG-IT: UmBERTo for Author Profiling at TAG-it 2020. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR Workshop Proceedings*.

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*.

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*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, 8(9):e73791.

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, pages 383–389.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).