Quanti anni hai? Age Identification for Italian

Aleksandra Maslennikova[•], Paolo Labruna[•], Andrea Cimino[◇], Felice Dell'Orletta[◇]
[•] Università di Pisa
a.maslennikova@studenti.unipi.it
pielleunipi@gmail.com
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{andrea.cimino, felice.dellorletta}@ilc.cnr.it

Abstract

English. We present the first work to our knowledge on automatic age identification for Italian texts. For this work we built a dataset consisting of more than 2.400.000 posts extracted from publicly available forums and containing authorship attribution metadata, such as age and gender. We developed an age classifier and performed a set of experiments with the aim of evaluating the possibility of assigning the correct age of an user and which information is useful to tackle this task: lexical or linguistic information spanning across different levels of linguistic descriptions. The performed experiments show the importance of lexical information in age classification, but also that exists writing style that relates to the age of an user.

Italiano. *In questo articolo presentiamo il primo lavoro a nostra conoscenza sul riconoscimento automatico dell'età per la lingua italiana. Per condurre il lavoro abbiamo costruito un dataset composto da più di 2.400.000 di post estratti da forum pubblici e associati a informazioni rispetto all'età e al genere degli autori. Abbiamo sviluppato un sistema di classificazione dell'età dello scrittore di un testo e condotto una serie di esperimenti per valutare se è possibile definire l'età e attraverso quali informazioni estratte dal testo: lessicali o di descrizione linguistica a diversi livelli. I risultati ottenuti dimostrano l'importanza del lessico nella classificazione, ma anche l'esistenza di uno stile di scrittura correlato all'età.*

1 Introduction

Social media platforms such as Facebook, Twitter and public forums allow users to communicate and share their opinions and to build social relations. The proliferation of such platforms allowed the scientific community to study many communication phenomena such as the analysis of the sentiment (Pak et al., 2010) or irony (Hernández Farías et al, 2016). Another related research field is the "author profiling" one, where the features that allow to discriminate age, gender, or native language of a person are analyzed. These studies are conducted both for forensic and marketing reasons, since the classification of these characteristics allow companies to better focus their marketing campaigns. In the author profiling scenario, many are the studies conducted by the scientific community, that were generally focused on English and Spanish language. The majority of these studies were performed in PAN [1] (Rangel et al., 2016), a lab at CLEF [2] that holds each year and in which many shared tasks related to the "authorship attribution" research topic are run. In these shared tasks participants were asked to identify the gender or the age using manually annotated training data from social media platforms. Among the most successful approaches proposed by participants the ones that achieved the best results (op Vollenbroek et al., 2016), (Modaresi et al., 2016) are based on SVM classifiers exploiting a wide variety of lexical and linguistic features, such as word n–grams, part–of–speech, and syntax. Only recently deep learning based approaches were proposed and have showed very good results especially when dealing with multi–modal data, i.e. text and images posted on Twitter (Takahashi et al., 2018).

In the present work we tackle a specific author-

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

[2] http://www.clef-initiative.eu/association/steering-committee

ship attribution task: the age detection for the Italian language. To our knowledge, this is the first time that such task is performed on Italian. For this reason, we built a multi–topic corpus, developed a classifier which exploits a wide range of linguistic features, and conducted several experiments to evaluate both the newly introduced corpus and the classifier.

The main contributions of this work are: i) an automatically built corpus for the age detection task for the Italian language; ii) the development of an age detection system; iii) the study of the impact of linguistic and lexical features.

2 Dataset construction

With the aim of building an automatic dataset from the web, we needed a set of Italian texts with the age of authors publicly available. Nowadays collecting this information is a challenging task, since the majority of the available platforms, for the sake of privacy, prefer not to make the user's age public. So, first-of-all, we had to find a website with such data. We choose the ForumFree platform[3] which allows users to create their own forums without any coding skills, using an existing template. Having all the forums based on the same templates makes them perfect for automated crawling. We extracted all the posts of the users that decided to show publicly their age. We tried to collect the data from the top 200 most active forums. Not all the forums had users with all the user information filled and, in the end of the processes, we fetched messages from 162 different forums. Since our goal was to build a corpus with author profiling purposes, and such task is very difficult with very small comments, we selected only posts with a minimum length of 20 words.

Another problem we faced is that users are not age-balanced in the forums: for example, anime dedicated forum have mostly users aged under 35. Another example are cars dedicated forums, where usually users are more mature with respect to anime forums. Only a couple of forums have very balanced information, which usually is the best data for training machine learning based classifiers. For this reason, we decided to group the forums by their topics, because in this scenario it is more probable to gather enough textual data for each age gap. We manually looked the content of all forums and assigned the topic for each one of them. We didn't have a preassigned settled list of possible topics. Instead, we were adding them in the process. For example, if we have an entire forum which discusses about only watches, we wouldn't assign some general "Hobby" tag, but we would create a special group "Watches" specifically for this forum.

At the and of the collection process, we collected 2.445.012 posts from 7.023 different users and 162 forums, that we divided in 30 different topic groups. All the information regarding the dataset are shown in Table 1.

[3] https://www.forumfree.it/?wiki=About

3 The Age classifier

We implemented a document age classifier that operates on morpho–syntactically tagged and dependency parsed texts. The classifier exploits widely used lexical, morpho-syntatic and syntactic features that are used to build the final statistical model. This statistical model is finally used to predict the age range of unseen documents. We used linear SVM implemented in LIBLINEAR (Rong-En et al., 2008) as machine learning algorithm. The input documents were automatically POS tagged by the Part–Of–Speech tagger described in (Cimino and Dell'Orletta, 2016) and dependency–parsed by the DeSR parser (Attardi et al., 2009).

3.1 Features

Raw and Lexical Text Features

Word n-grams, calculated as presence or absence of a word n-gram in the text.

Lemma n-grams, calculated as the frequency of each lemma n-gram in the text and normalized with respect to the number of tokens in the text.

Morpho–syntactic Features

Coarse and fine grained Part-Of-Speech n-grams, calculated as the logarithm of the frequency of each coarse/fine grained PoS n-gram in the text and normalized with respect to the number of tokens of the text.

Syntactic Features

Linear dependency types n-grams, calculated as the frequency of each dependency n-gram in the text with respect to the surface linear ordering of words and normalized with respect to the number of tokens in the text.

Hierarchical dependency types n-grams calculated as the logarithm of the frequency of each hierarchy dependency n-gram in the text and nor-

Topic		≤20	21-30	31-40	41-50	51-60	≥61
Cars	Users	36	158	187	209	158	45
	Posts	6056	50281	46746	62002	48939	15867
Bicycles	Users	10	11	12	35	25	1
	Posts	2056	2284	5532	13418	16959	6
Smoking	Users	3	52	78	69	46	18
	Posts	7	21399	41470	38149	17981	4742
Anime/Manga	Users	392	438	142	62	16	6
	Posts	60367	99165	39939	29086	3873	228
Role playing	Users	115	104	14	8	6	7
	Posts	22953	40652	3893	3945	534	2060
Gaming	Users	235	358	113	131	48	7
	Posts	54584	81535	20379	20055	4560	1323
Spirituality	Users	11	25	21	13	11	2
	Posts	336	1427	1342	1095	1517	965
Aesthetic medicine	Users	7	36	27	29	17	1
	Posts	1345	6135	11767	8208	3384	1
Sport	Users	215	338	192	136	52	24
	Posts	82495	310220	158382	103027	34627	16084
Culinary	Users	0	1	4	10	4	4
	Posts	0	52	10130	2414	747	438
Pets	Users	10	21	11	4	2	3
	Posts	4307	13222	7357	2592	5383	10353
Celebrities	Users	21	76	26	24	17	4
	Posts	548	21114	5820	6150	3139	1248
Politics	Users	0	2	4	10	6	0
	Posts	0	330	2801	3548	576	0
Different topics	Users	52	45	34	43	34	15
	Posts	9453	12000	21667	16316	4759	24418
Fishing	Users	11	57	79	62	30	5
	Posts	3040	14805	24306	17131	13155	8356
Institution community	Users	6	6	0	2	5	1
	Posts	13	12	0	18	11130	4364
Rail transport modelling	Users	0	6	7	5	5	1
	Posts	0	3597	2289	999	2470	751
Culture	Users	4	10	4	7	4	0
	Posts	1855	560	653	1174	219	0
Tourism	Users	0	2	2	4	1	2
	Posts	0	16	10	1378	2	14
Sexuality	Users	11	31	18	10	2	1
	Posts	185	2540	8201	1421	7	1179
Metal Detecting	Users	25	34	78	121	55	11
	Posts	7750	9830	19299	31288	16547	3529
Music	Users	12	25	15	0	0	0
	Posts	8731	15720	5276	0	0	0
Parenting	Users	1	4	1	1	0	0
	Posts	719	2250	626	420	0	0
Technologies	Users	37	47	12	4	8	5
	Posts	185	266	431	26	19	23
Nature	Users	5	9	10	6	6	2
	Posts	998	1304	3653	2171	292	10
Religion	Users	0	5	6	1	0	0
	Posts	0	2618	4125	896	0	0
Films	Users	25	26	10	5	1	2
	Posts	9476	6135	503	43	4	2477
Psychology	Users	12	14	2	0	1	2
	Posts	291	912	44	0	1	11
Gambling	Users	0	3	3	10	11	7
	Posts	0	458	134	364	715	274
Watches	Users	29	153	317	302	109	32
	Posts	5158	52623	114074	101869	50243	18085

Table 1: Distribution of number of users and posts per age gap in different topics in the corpus

malized with respect to the number of tokens in the text. In addition to the dependency relationship, the feature takes into account whether a node is a left or a right child with respect to its parent.

4 Experiments

In order to test the corpus and the classifier, we performed a set of experiments. The experiments were devised in order to test real-word scenarios where 1) we were interested to classify a set of posts written by a single user rather then a single post; 2) we always classified unseen users, i.e. no training data was available for such users. For these reasons, we merged all the posts of a single user in the original corpus in a single document. We then considered only the users that wrote a minimum of 200 tokens and limited the final merged document to a 'soft' limit of 1000 tokens for each user. When the soft limit was exceeded, we included the whole post that exceeded the soft limit. The described procedure allows training and test splits to never contain the same user. For the age detection tasks, similarly as in (Rangel et al., 2016), we considered age-splits as the classification classes. More precisely, we took into account two different age group splits: the first one, which we will refer with the name *5–class*, in which we split the documents in 5 different age groups: 20-29, 30-39, 40-49, 50-59, 60-69. The second age group split, which we will refer with the name *2–class*, is composed by the following age group splits: $\leq$29, $\geq$50-69 (excluding all the documents written by users that did not belong to these age groups). We conducted two different kind of experiments. In the first experiment (*in–domain*), we evaluated the performance of the classifier on in-domain texts, more precisely we selected three different topics starting from the main corpus and on each of the topics we trained the classifier on the 80% of the data, and evaluated the performance of the classifier on the remaining 20%. For this experiment we choose the the following domains: Sports, Watches and Cars. In the second experiment (*out–domain*) we trained the classifier on the all the 3 topics used for the *in–domain* experiments and evaluated the performance of the classifier on other 3 different topics (Smoking, Celebrities, Metal Detecting).

In addition, we devised 3 different machine learning models based on 3 different sets of features. The first one (*Lexicon*), which uses only word and lemmas features, the second one (*Syntax*), which uses only the morpho–syntactic and syntactic features. Finally, the last model (*All*), which uses both the lexical, morpho–syntactic and syntactic features. We considered as baseline model a classifier which predicts always the most frequent class.

4.1 Results

Tables 2 and 3 report the results achieved by the classifier for the in–domain and out–domain experiments respectively. For what concerns all the experiments, we can notice that the results achieved by our classifier are higher than the baseline results, showing that there are features that are able to discriminate among the considered classes. The in–domain results show that the lexical features are the ones that have the most discriminative power with respect to the syntax ones. The f-score achieved by the *lexicon* model is 3-4 times better than the baseline in the 5–class setting, and 2 times better in average in the 2–class setting. The *syntax* model shown very good results but, as expected, lower than the results achieved by the lexicon model. This is an important result since it shows that syntax and morpho–syntax are relevant characteristics in each age-group, both in the 5–class and 2–class settings. Surprisingly, the *All* model didn't show in any experiment an increase in classification performance. The classification patterns revealed in the in–domain experiments are similarly shown also in the out–domain experiments. The results achieved in this setting as expected are lower than results achieved in the in–domain settings. The 5–class experiments show a drop in performance achieved by the considered learning models of 8-10% f–score points in average w.r.t. to the in–domain experiments. When we move to the 2–class experiments, no significant drop in performance is noticed. This shows that in case of domain shifting, the machine learning models are still able to well discriminate between young and aged people.

Figures 1 and 2 report the confusion matrices of the in–domain and out–domain experiments using the 5-class age-groups. More precisely, the in–domain confusion matrix is obtained by training the *All* model on all the three training in–domain topics and testing the model on the respective testset (f–score: 0.47). Similarly, the out-domain confusion matrix is obtained by training

	5-class				2-class			
Topic	Baseline	Lexicon	Syntax	All	Baseline	Lexicon	Syntax	All
Sport	0.27	0.45	0.42	0.48	0.74	0.74	0.75	0.75
Watches	0.19	0.43	0.35	0.42	0.44	0.85	0.75	0.83
Cars	0.12	0.54	0.34	0.45	0.47	0.87	0.77	0.84

Table 2: Results achieved in the in–domain experiments in terms of f–score

	5-class				2-class			
Topic	Baseline	Lexicon	Syntax	All	Baseline	Lexicon	Syntax	All
Smoking	0.14	0.30	0.25	0.32	0.42	0.79	0.68	0.79
Celebrities	0.33	0.45	0.39	0.47	0.62	0.83	0.73	0.81
Metal Detecting	0.21	0.36	0.27	0.34	0.52	0.80	0.66	0.78

Table 3: Results achieved in the out–domain experiments in terms of f–score

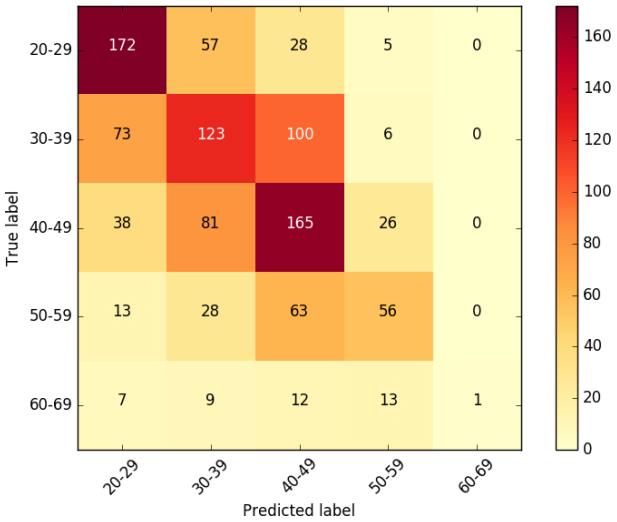

Figure 1: Confusion matrix calculated on the documents belonging to the in-domain topics

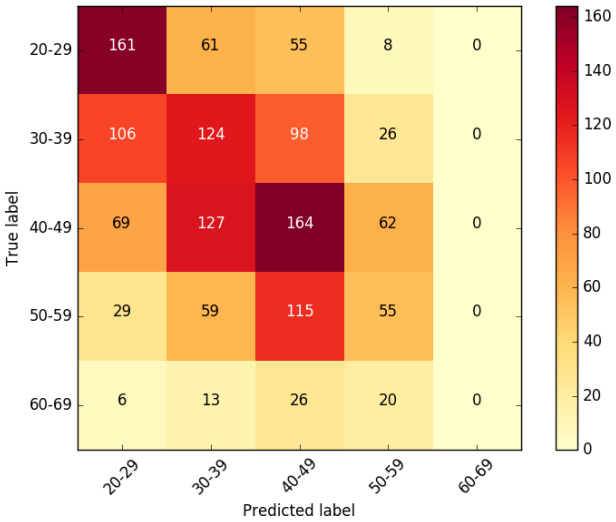

Figure 2: Confusion matrix calculated on the documents belonging to the out-domain topics

the *All* model on all the in-domain topics (including the test-sets), and testing the model on the out-domain documents of the selected 3 topics. As it can be seen, the errors both on the in–domain and out–domain experiments show very good performances of the classifier, i.e., in case of errors, usually it makes a mistake of a range of $\pm$ 10 years. Such results show also that the automatically built corpus is a very useful resource for the age classification task. Finally, it is interesting to notice that the most correct predicted classes are the ranges 20-29 and 40-49, both in the in–domain and out–domain settings, while the worst predicted class in both experiments is the 60-69 age range, most probably because is the most underrepresented class in the training set.

5 Conclusions

We presented the first automatically built corpus for the age detection task for the Italian language. By exploiting the publicly available information on the FreeForum platform, we built a corpus consisting of more than 2.400.000 posts and 7.000 different users containing the user's age information. The first experiments performed through a machine learning based classifier that uses a wide range of linguistic features showed promising results in two different range classification tasks both in the in–domain and out–domain settings. The conducted experiments show that lexicon plays a fundamental role in the age classification task both in in–domain and out–domain scenarios. Lastly, the experiments shown that the corpus, even though if automatically generated, is suitable for real–world applications. We plan to release the full corpus as soon as privacy and legal issues will be fully investigated.

Acknowledgments

This work was partially supported by the 2-year project ARTILS, Augmented RealTime Learning for Secure workspace, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian. 2009. *Accurate dependency parsing with a stacked multilayer perceptron.* In Proceedings of the 2nd Workshop of Evalita 2009. December, Reggio Emilia, Italy.

Andrea Cimino and Felice Dell'Orletta. 2016. *Building the state-of-the-art in POS tagging of italian tweets.* In Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA), December 5-7.

Delia Irazú Hernández Farías, Viviana Patti and Paolo Rosso 2016. *Irony Detection in Twitter: The Role of Affective Content.* In ACM Transactions on Internet Technology (TOIT), Volume 15, number 3.

Pashutan Modaresi, Matthias Liebeck and Stefan Conrad. 2016. *Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016.* In Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.

Francisco Manuel Rangel Pardo, 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 of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.

Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, and Hessel Haagsma and Malvina Nissim. 2016. *Gronup: Groningen user profiling* In Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.

Alexander Pak and Patrick Paroubek. 2010. *Twitter as a corpus for sentiment analysis and opinion mining.* In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta

Fan Rong-En, Chang Kai-Wei, Hsieh Cho-Jui, Wang Xiang-Rui and Lin Chih-Jen. 2008. *LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:1871–1874.

Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi and Tomoko Ohkuma. 2016. *Text and image synergy with feature cross technique for gender identification.* In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, 10 - 14 September, 2018.