ItaliaNLP @ TAG-IT: UmBERTo for Author Profiling at TAG-it 2020

Daniela Occhipinti***, Andrea Tesei*****, Maria Iacono*****, Carlo Aliprandi*** **and Lorenzo De Mattei**$^{\diamond\,\dagger\,\star}$
* Aptus.AI / Pisa, Italy
$^{\diamond}$ Dipartimento di Informatica, Università di Pisa / Pisa, Italy
† Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab / Pisa, Italy
{daniela,andrea,maria,carlo}@aptus.ai
lorenzo.demattei@di.unipi.it

Abstract

In this paper we describe the systems we used to participate in the task TAG-it of EVALITA 2020. The first system we developed uses linear Support Vector Machine as learning algorithm. The other two systems are based on the pretrained Italian Language Model UmBERTo: one of them has been developed following the Multi-Task Learning approach, while the other following the Single-Task Learning approach. These systems have been evaluated on TAG-it official test sets and ranked first in all the TAG-it subtasks, demonstrating the validity of the approaches we followed.

1 Introduction

Author Profiling (AP) is a known Natural Language Processing task consisting in the extraction or the prediction of information about the authors of some disputed documents. Such information can include the age and the gender of the authors. The AP problem is assuming more and more importance in several fields, such as security, forensics, marketing and sales, and so on. For example, in forensics, detecting the age and the gender of the author of a given document can be very helpful for determining whether a person should be considered as a suspect or not; from the marketing and sales' perspective, companies can understand what kind of people may or not like their products on the basis of the analysis performed on people's reviews or blog and social network posts (Rangel et al., 2015).

In the context of EVALITA 2020 (Basile et al., 2020), the periodic evaluation campaign of Natural Language Processing and speech tools for the Italian language, the task TAG-it (Cimino et al., 2020) is proposed. TAG-it is an AP task in which the goal is to provide a system capable of predicting the gender and the age of the authors of several blog posts and their topics. This task can be considered as a follow-up of the EVALITA 2018's GxG task (Dell'Orletta and Nissim, 2018) in which the goal was the prediction of the author's gender for Twitter posts, YouTube comments, Children Essays, Diaries and News; in GXG models were trained and tested *cross-genre*. These two aspects led to scores lower than ones observed in other campaigns and languages. In order to address this problem and get better performances, in TAG-it only blogs' genre is considered and longer texts are used, since they provide more evidence than tweets and Youtube comments, which are shorter than blog posts. Moreover, with respect to GxG, TAG-it adds the topic control with the aim of evaluating the interaction of topic and lexically rich models on performances in a more direct way than in GxG, in which this was indirectly done via cross-genre prediction. TAG-it is divided in two subtasks: the goal of the first one (Subtask 1) is to classify gender, age and topic at once, while the goal of the second one is to predict age (Subtask 2a) and gender (Subtask 2b) separately and with topic control.

De Mattei and Cimino (2018) and Cimino et al. (2018) demonstrated the validity of Multi-Task Learning approach to establish the state of the art for several Italian NLP task, in the context of GxG, Cimino et al. (2018) developed the best system for this task based on Bidirectional LSTM (Bi-LSTMs) trained using a Multi-Task Learning approach. For TAG-it we replicated the same approach: we developed a baseline system based on SVM, and two neural systems, the first one exploiting a Single-Task Learning approach, the second one a Multi-Task Learning approach. Instead of the Bi-LSTM model used by Cimino et al. (2018) for TAG-it we exploited a deeper neural pretrained language model: BERT (Devlin et al., 2019).

2 Description of the Systems

We implemented and tested three different systems. Our early experiments were led on a training set and a test set obtained by shuffling and splitting (80% training - 20% test) the training set provided by the organisers in order to analyse the classifiers' performances on a labeled dataset. At the end of our experiments, we trained our best classifiers on the whole training set and run them on the TAG-it test sets provided by the organisers.

For our experiments and runs, as a preprocessing phase, we filtered out all posts less than 20 characters in length and labeled each post of the dataset with the corresponding author's id, gender, age and topic. In Table 1 we report the distributions of the classes of the TAG-it dataset.

	Train	**Test1**	**Test2a**	**Test2b**
M	15070	315	344	730
F	3113	96	68	69
0-19	2232	39	76	79
20-29	5412	131	189	230
30-39	3569	95	51	134
40-49	3577	69	48	216
50-100	3393	77	48	140
ANIME	3925	97	0	0
AUTO-MOTO	3648	76	0	0
BIKES	468	12	0	0
CELEBRITIES	1063	22	0	0
ENTERTAINMENT	534	9	0	0
MEDICINE-AESTHETICS	370	16	0	0
METAL-DETECTING	1471	26	0	0
NATURE	481	11	0	0
SMOKE	1574	30	0	0
SPORTS	4593	103	0	0
TECHNOLOGY	56	9	0	0
GAMES	0	0	298	298
ROLE-GAMES	0	0	114	114
CLOCKS	0	0	0	387

Table 1: TAG-it datasets distributions

As a first step, our systems make their predictions by classifying the three dimensions post by post. Then they use a voting mechanism according to which the gender, the age and the topic of an author are represented by the most frequent values assigned by the classifiers to his/her posts.

The first system we implemented uses linear Support Vector Machine as learning algorithm and we used different features for predicting the core dimensions of the dataset, the second system is based on a Single-Task Learning BERT model and the third system is based on a Multi-Task Learning BERT model. In particular, we used UmBERTo[1], an Italian pretrained Language Model developed by Musixmatch.

In the following subsections we will describe these systems in detail.

2.1 Support Vector Machine Classifiers

As regards the system based on three linear SVM statistical models, we used the scikit-learn[2] Python library and we conducted several experiments by testing different configurations for feature extraction. In all the experiments we used the TF-IDF vectorizer, but we changed the tokenizer and the *n*-grams context window. In particular we tested five different kinds of features: character *n*-grams, word *n*-grams, lemma *n*-grams, Part-Of-Speech *n*-grams and bleached tokens. As regards the bleached tokens features, they were extracted after performing a bleach tokenization consisting in fading out lexicon in favour of an abstract token representation (van der Goot et al., 2018). The word *n*-grams, lemma *n*-grams and Part-Of-Speech *n*-grams features were extracted by using the linguistic pipeline for the Italian language provided by spaCy[3]. For the multi-class classification we applied the One-Vs-Rest method (Rennie and Rifkin, 2001). In Table 2 we report the performances in terms of micro-average f-score of the SVM models tested in our experiments.

These results led us to choose the best SVM classifiers for the official runs on the provided test set; analysing them, we can state that the best SVM classifiers tested in our experiments are the following:

- Topic Detection: One-Vs-Rest Linear SVM using features extracted through a TF-IDF Vectorizer considering character *n*-grams;
- Age Detection: One-Vs-Rest Linear SVM using features extracted through a TF-IDF Vectorizer considering lemma *n*-grams;
- Gender Detection: Linear SVM using features extracted through a TF-IDF Vectorizer considering word *n*-grams.

[1] `https://github.com/musixmatchresearch/umberto`

[2] `https://scikit-learn.org/stable/`

[3] `https://spacy.io`

	Gender	Age	Topic
word n-gram	**0.933**	0.3873	0.7882
char n-gram	0.9284	0.3739	**0.8333**
lemma n-gram	0.9265	**0.4189**	0.7928
pos n-gram	0.9223	0.3063	0.3873
bleached words	0.9223	0.3739	0.4775

Table 2: SVM classifiers' micro-average f1-scores on validation set

2.2 Single-Task BERT-based Classifiers

Our second system consists of three different BERT models and a classifier on top of each of them. More precisely, we used the UmBERTo language model, which was pretrained on a large Italian Corpus: OSCAR (Ortiz Suárez et al., 2020).

This language model have 12-layer, 768-hidden, 12-heads, 110M parameters. On top of the language model we added a ReLU classifier (Nair and Hinton, 2010). We applied dropout (Srivastava et al., 2014) to prevent overfitting. As loss function we used the sum of loss functions of the three classifiers. For each classifier, we used Cross Entropy as loss function.

In Table 3 we report the system's performances in terms of f1-score obtained on the validation set.

	f1-score
Gender	0.86
Age	0.35
Topic	0.66

Table 3: Single-Task Learning BERT-based system micro-average f1-scores on validation set

2.3 Multi-task BERT-based Classifier

Our last system is based on a unique UmBERTo model and three classifiers on top of it, each one responsible of predicting one of the three core dimensions of the dataset according to the Multi-Task Learning approach used in (Cimino et al., 2018). On top of the model we added three ReLU classifiers, we applied the dropout method and we used the sum of the Cross-Entropy loss functions of the three classifiers as loss function.

In Table 4 we report the system's performances in terms of f1-score obtained on the validation set.

	f1-score
Gender	0.86
Age	0.39
Topic	0.64

Table 4: Multi-Task Learning BERT-based system f1-scores on validation set

3 Results and Evaluation

We run all our three systems on the test sets provided by the task organisers. The performances of our systems are reported in Table 5.

For the Task 1 scoring, TAG-it considers two different rankings. The first ranking is obtained using a partial scoring scheme, giving 0 points if no correct predictions are provided for the three dimensions of the dataset, 1/3 points if one out of three correct answers is given, 2/3 points if two out of three correct answers are given and 1 point if all the answers given by the system are correct. The second ranking assigns 0 points if no correct predictions are provided for the three dimensions of the dataset and 1 point if all the answers given by the system are correct. In both cases, the final score is the sum of the points achieved by the system across all the documents normalized with respect to the number of documents in the test set. For the Task 2, the micro-average f-score is used as scoring function.

	STL-SVM	MTL-BERT	STL-BERT
Task 1 metric 1	0,6626	0,7178	**0,7348**
Task 1 metric 2	0,253	0,3090	**0,3309**
Task 2a	0,8519	**0,9247**	0,9053
Task 2b	0,3742	0,3667	**0,4093**

Table 5: Systems' performances evaluation with TAG-it metrics

Analysing the scores in Table 5, we can state that the best system in the TAG-it context is the one based on BERT using the Single-Task Learning (STL-BERT) approach, obtaining the best scores in Task 1 and Task 2b (age prediction). In Task 2a, consisting in gender prediction with topic control, the best system is the Multi-Task Learning BERT-based system (MTL-BERT). Hence, the systems based on deeper neural models outperform the systems based on traditional machine learning techniques, i.e. the SVM (STL-SVM).

Task 1: In order to compare classifiers' predictions on Task 1 with regard to each dimension and to understand the correlation between labels, we plotted and analysed some distributions.

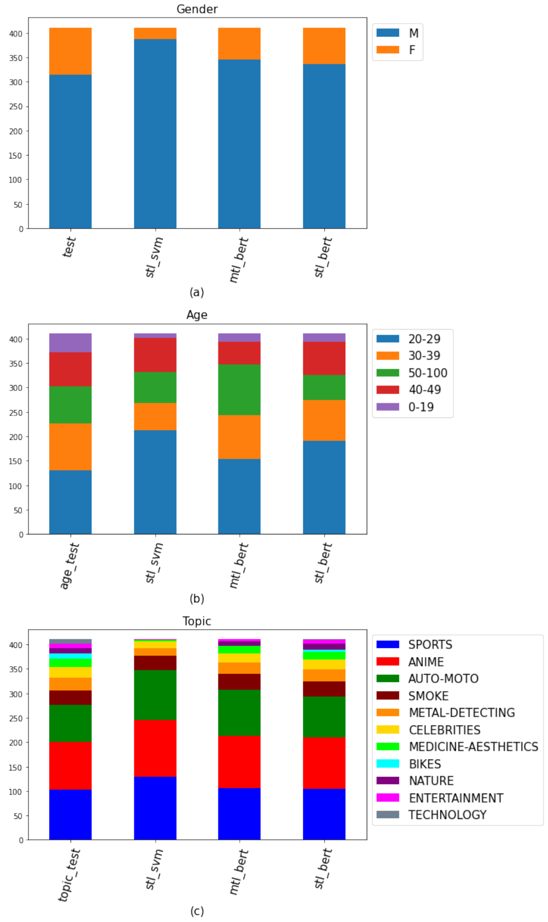

Figure 1: Task 1, Distributions of the dimensions' classes in test set and classifiers' predictions.

In Figure 1, we reported the distribution of the labels in the test set and in the classifiers' output. As regards the gender prediction (a), we can note that the STL-SVM classifier overestimates the M class, most likely because the M and F classes are very unbalanced in the training set. STL-BERT and MTL-BERT's distributions, on the contrary, are closer to the test set's one: in our setting the neural models appear less affected by the imbalance of a training set.

Observing the distributions of the Age classes in Figure 1 (b), we can observe that for all the three systems the distributions of the labels are not close to the distribution of the test set. The nearest distribution is the one of MTL-BERT's output.

Looking at the Topic classes distributions in Figure 1 (c), we can observe, once again, that the SVM-based system's one is the less close to the test set in that it has the tendency to overestimate the SPORT, ANIME and AUTO-MOTO classes and it does not recognise the BIKES and TECHNOLOGY classes as they are underrepresented in the training set (respectively the 2.574% and the 0.308% of training set). For the same reason, it has difficulties in recognising the classes ENTERTAINMENT, MEDICINE-AESTHETICS and NATURE (which are respectively the 2.937%, 2.035% and 2.645% of the training set). The two BERT-based systems, on the contrary, are less affected by this imbalance of the training set and their predictions reflect more the reality of the test set, even though, as STL-SVM, also MTL-BERT cannot recognise the BIKES and TECHNOLOGY classes.

In Figure 2 we report the distribution of the Age classes with respect to the Topic classes. Figure 2 (b) shows that in the STL-SVM's output the 0-19 age class is only related to the ANIME topic, the age 20-29 is related more or less with all the detected topics, the 30-39 class is mostly related to SMOKE and MEDICINE-AESTHETICS, the 40-49 class to the METAL-DETECTING, AUTO-MOTO and SMOKE topics and the 50-100 class mostly to AUTO-MOTO, SPORTS and CELEBRITIES. This distribution is quite far from the test set one and it seems that the relation between the class 0-19 and the topics is overestimated. In Figure 2 (c), which refers to MTL-BERT, we can note that authors classified as having age 20-29 are predicted to talk mostly about ANIME, CELEBRITIES, NATURE and SPORTS and are less related to ENTERTAINMENT, MEDICINE-AESTHETICS and NATURE topics than in STL-SVM's output; the relation between the 30-39 class and ENTERTAINMENT and MEDICINE-AESTHETICS categories on one hand, and 50-100 and AUTO-MOTO, MEDICINE-AESTHETICS, METAL-DETECTING, NATURE and SMOKE on the other is stronger than in STL-SVM's results. Also this distribution, though, is quite far from the test set's one, even if ages seem to be more distributed than in STL-SVM's output. As shown in Figure 2 (d), in STL-BERT's distribution, the age 0-19 seems mostly related to TECHNOLOGY and ANIME. The class BIKES, which has not been recognised by the other systems, is related to the classes 30-39, 40-49 and, mostly, 50-100. As regards the 20-29 class, its relations are quite similar to the ones found in the STL-SVM's

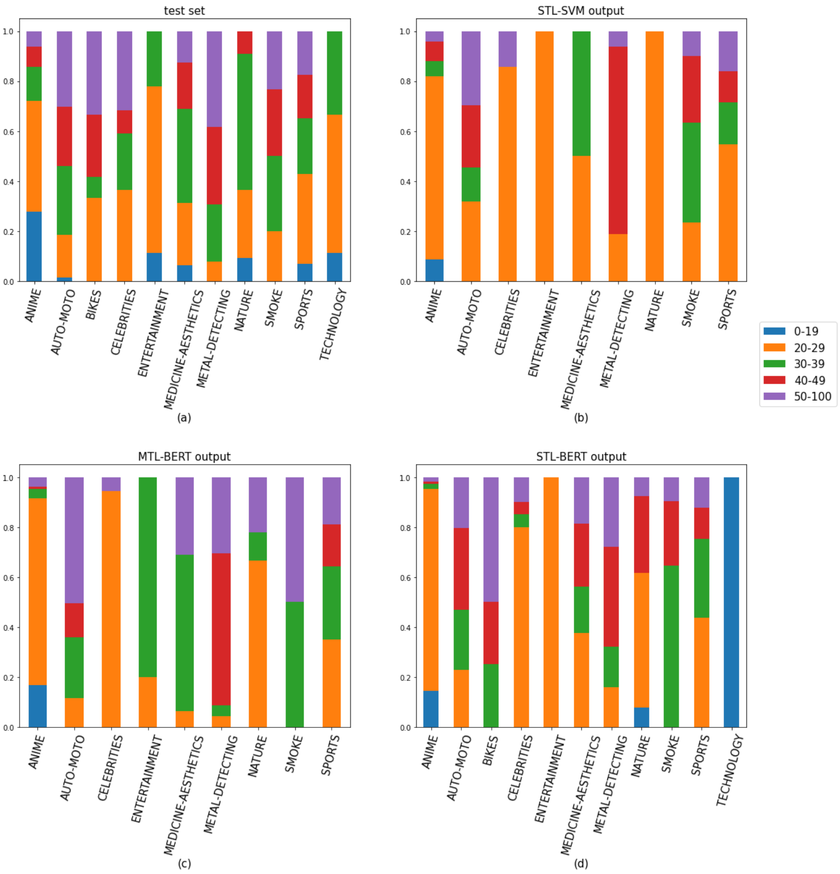

Figure 2: Task 1, Distributions of the Topic and Age dimensions in test set and classifiers' predictions.

results, except for the class NATURE, which is related also to the ages 0-19, 40-49 and 50-100. Also this distribution is quite far from the test's one. All the three distributions differ considerably from the test set because systems do not perform well enough in age prediction.

The distributions of the topics with respect to gender in the test set and the predictions are reported in Figure 3. As shown in the figure, all the three systems results relate the F class mostly to the ANIME topic, as it is also in the test set. In the STL-SVM's output, though, this relation seems to be overestimated. Moreover, in STL-SVM the F class, besides ANIME, is only related to a much lesser extent to SMOKE. The relation between M and SMOKE seems to be overestimated too with respect to the test set. As regards the F class in MLT-BERT and STL-BERT outputs, topics are more distributed than in STL-SVM, but the nearest to the test set's one is STL-BERT: MLT-BERT, in fact, seems to overestimate the relation between F and BIKES and ENTERTAINMENT and to underestimate the relation between F and MEDICINE-AESTHETIC and SPORTS. For what concerns the M class in MLT-BERT and STL-BERT distributions, we can state once again that the distribution which is closer to the test set one is given by STL-BERT: STL-SVM, MLT-BERT overestimates the relation between M and SMOKE and NATURE.

Task 2:

The results reported in Table 5 show that for Task 2a (gender prediction with topic control) the best classifier is MLT-BERT. In this subtask, BERT-based systems outperform in a significant way the system based on SVM.

As regards the Task 2b, consisting in the age prediction, the best metrics belong to the STL-BERT. In the age prediction the gap between all the systems' metrics is not very high. In this case, in which only the age dimension must be predicted, the best classifier is the one using a Single-Task Learning approach.

4 Conclusions

In this paper we reported the performances and the results of the systems we used to participate to the TAG-it task of EVALITA 2020. We com-

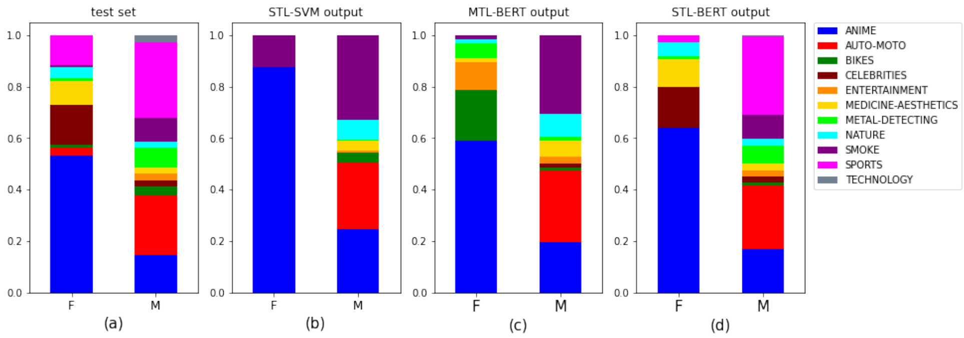

Figure 3: Task 1, Distributions of the Topic and Gender dimensions in test set and classifiers' predictions.

pared our systems' performances and noted that in the case in which the goal is to predict topic, age and gender dimensions at once, and in the case in which only the age must be predicted, the best classifier is the one developed using a Single-Task Learning approach and based on transformers. In the case in which the goal is the gender prediction only a Multi-task Learning approach combined with transformers have slightly better performances. These results prove that the proposed systems based on transformers, are more effective than traditional machine learning techniques in topic, age and gender classification achieving the state of the art for TAG-it shared task. Using deep pretrained language models on this task Multi-Task Learning does not provide any relevant boost of performances.

References

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.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural networks at EVALITA 2018. In *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.

Andrea Cimino, Dell'Orletta Felice, and Nissim Malvina. 2020. Tag-it@evalita2020: Overview of the topic, age, and gender prediction task 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.

Lorenzo De Mattei and Andrea Cimino. 2018. Multi-task learning in deep neural network for sentiment polarity and irony classification. In *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence*, November.

Felice Dell'Orletta and Malvina Nissim. 2018. Overview of the evalita 2018 cross-genre gender prediction (gxg) task.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July. Association for Computational Linguistics.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015.

Jason D. M. Rennie and Ryan Rifkin. 2001. Improving multiclass text classification with the support vector machine.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Rob van der Goot, Nikola Ljubesic, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction.