Pooled LSTM for Dutch cross-genre gender classification

Matej Martinc[1] and Senja Pollak[1,2]
matej.martinc@ijs.si, senja.pollak@ijs.si

[1] Jožef Stefan Institute,Jamova 39, 1000 Ljubljana, Slovenia
[2] Usher Institute, Medical school, University of Edinburgh, UK

Abstract. We present the results of cross-genre and in-genre gender classification performed on the data sets of Dutch tweets, YouTube comments and news prepared for the CLIN 2019 shared task. We propose a recurrent neural network architecture for gender classification, in which the input word and part-of-speech sequences are fed to the LSTM layer, which is followed by average and max pooling layers. The best cross-genre accuracy of 55.2% was achieved by the model trained on YouTube comments and tweets, and tested on the balanced news corpus, while the best in-genre accuracy of 61.33% was achieved on YouTube comments. Overall, the proposed approach ranked 2nd in the global cross-genre ranking and 6th in the global in-genre ranking of CLIN 2019 shared task.

1 Introduction

Author profiling (AP) is a well-established subfield of natural language processing with a thriving community gathering data, organizing shared tasks and publishing about this topic. AP entails the prediction of an author's profile - i.e. demographic and/or psychological characteristics of the author - based on the text that he/she has written. The single most prominent author profiling task is gender classification, other tasks include the prediction of age, personality, region of origin and mental health of an author.

Gender prediction became a mainstream research topic with the influential work by Koppel et al. (2002). Based on the experiments on a subset of the British National Corpus, they found that women have a more relational writing style (e.g., using more pronouns) and men have a more informational writing style (e.g., using more determiners). Later gender prediction research remained focused on English, but in the last few years, more languages have received attention in the context of author profiling (Rangel et al., 2015, 2016), with the publication of the TwiSty corpus containing gender information on Twitter authors for six languages (Verhoeven et al., 2016) as a highlight so far.

A recent study by van der Goot et al. (2018) calls the cross-genre transferability of machine learning approaches to gender prediction into question by noticing that most of these approaches has typically focused on lexical and specialized social network features, which boosted the performance of the approaches, but on the other hand also made the approaches highly genre and topic dependent. To solve this problem, a fairly new development in the field of AP is the search for data set independent features and approaches, capable of capturing the most generic differences between male and female writing, which transfer well across different genres and languages (Dell Orletta and Nissim, 2018). This is also the main focus of the present research, in which we primarily deal with the development and testing of the system for Dutch cross-genre gender classification. In contrast to the majority of the best performing systems in the field of AP, which use hand-crafted features and traditional classifiers such as Support vector machines (SVM) and Logistic regression (Rangel et al., 2017), we opted for the neural classifier and automated feature engineering.

This paper is structured as follows. The findings from the related work are presented in Section 2. The data sets and the methodology used are presented in Section 3. Results are presented in Section 4, while in Section 5 we conclude the paper and present plans for the future work.

2 Related work

The lively AP community is centered around a series of scientific events and shared tasks on digital text forensic, such as PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[1] and VarDial (Varieties and Dialects)[2] (Zampieri et al., 2014). While VarDial is more focused on the identification of language varieties and dialects, most past PAN AP shared tasks were centered around gender classification.

The first PAN event took place in 2011 and the first AP shared task was organized in 2013 (Rangel et al., 2013). From the beginning, the PAN shared task was multilingual (Rangel et al., 2013, 2014, 2015, 2016, 2017, 2018) and two of the past competitions also had a cross-genre setting (Rangel et al., 2014, 2016). Another shared task dedicated to cross-genre gender classification on Italian documents was the EVALITA 2018 cross-genre gender prediction (GxG) task (Dell Orletta and Nissim, 2018).

The most popular approach to gender classification usually relies on bag-of-words features and SVM classifiers. For instance, winners of the PAN 2017 competition (Basile et al., 2017) used an SVM based system with very simple features (just word unigrams, bigrams and character three- to five-grams).

Some quite successful attempts of tackling the gender classification with neural networks have also been reported. A system consisting of a recurrent neural network (RNN) layer, a convolutional neural network (CNN) layer, and an attention mechanism proposed by Miura et al. (2017) ranked fourth in the PAN

[1] http://pan.webis.de/
[2] http://corporavm.uni-koeln.de/vardial/sharedtask.html

2017 shared task. In the PAN 2018 multimodal gender classification task (Rangel et al., 2018), where the task was to predict the gender of the Twitter user from their tweets and published images, deep learning approaches were prevailing and the overall winners used RNN for texts and CNN for images (Takahashi et al., 2018).

Another related research we looked at was the use of part-of-speech (POS) tags in existing gender classification approaches, since we hypothesized that POS based features would be less topic and genre-dependent, and therefore appropriate for the cross-genre task at hand. Mukherjee and Liu (2010) showed that sequences of POS tags can be successfully used for gender prediction as a standalone feature or in combination with other features. POS tag sequences were also successfully used in combination with other features in the PAN 2017 AP shared task by Martinc et al. (2017), who overall ranked second in the competition and also tested their model in a cross-genre setting (Martinc and Pollak, 2018).

3 Experimental setup

This section describes the data sets, methodology and the conducted experiments.

3.1 Data sets

CLIN 2018 shared task organizers provided six data sets from three different genres. Altogether, they provided 30,000 tweets, 19,658 YouTube comments and 2,832 news, each of them split into a gender labeled train set and an unlabeled test set. All data sets are balanced in terms of number of documents written by male and female authors. A more detailed description of all the data sets in terms of document size and word length is given in Table 1.

Dataset	**Documents**	**Words**
Twitter train	20,000	380,074
Twitter test	10,000	192,306
YouTube train	14,744	280,498
YouTube test	4,914	87,038
News train	1,832	336,602
News test	1,000	401,235

Table 1. Data sets used in the experiments

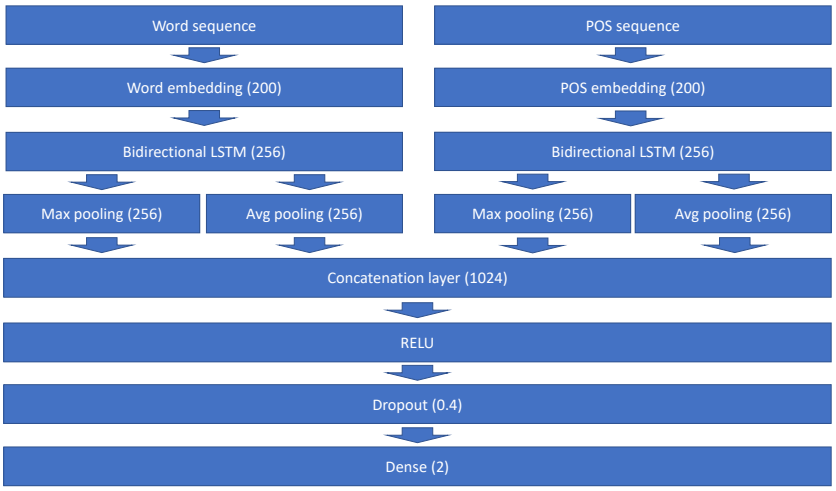

Fig. 1. Infrastructure of the proposed Pooled LSTM network

3.2 Methodology

Altogether, six classification models, three in-genre and three cross-genre, were trained and later used for prediction in our experiments. For the in-genre experiments, the train set for a specific genre was randomly split into a train set containing 90% of the documents and a validation set containing 10% of the documents. For the cross-genre experiments, we trained the Twitter cross-genre model on a concatenation of YouTube and news train sets (Twitter train set was used as a validation set during training), YouTube cross-genre model was trained on a concatenation of Twitter and news train sets (YouTube train set was used as a validation set during training) and news cross-genre model was trained on tweets and YouTube comments (news train set was used as a validation set during training).

Text preprocessing is light, we only replace hashtags in some of the data sets with *#HASHTAG* tokens, URLs with *HTTPURL* tokens and mentions with *@MENTION* tokens. We also limit the text vocabulary to 30,000 most frequent words and replace the rest with the <*unk*> token.

We decided on a neural approach to the task at hand, mostly because of the relatively large sizes of the available train and test sets (described in Section 3.1). Taking into the consideration some of the findings from the related work, we opted for the bidirectional recurrent architecture, which was successfully employed for gender prediction in the past (Miura et al., 2017; Takahashi et al., 2018). Initial experiments and previous research (Martinc et al., 2017; Martinc and Pollak, 2018) also suggested that adding POS tag information improves the performance of the model (especially in the cross-genre setting), therefore POS sequences are fed to the network together with the preprocessed texts.

	Twitter	YouTube	News	Average
Validation set in-genre	0.6245	0.6270	**0.6477**	0.6331
Validation set cross-genre	0.5473	**0.5580**	0.5573	0.5542
Official test set in-genre	0.6099	**0.6133**	0.5990	0.6074
Official test set cross-genre	0.5427	0.5507	**0.5520**	0.5485

Table 2. Results of the in-genre and cross-genre classification

Embedding vectors of size 200 are produced for input word and POS tag sequences, with the help of two randomly initialized embedding layers, and then fed to two distinct Bidirectional Long short-term memory networks (BiLSTM) with 256 neurons, which both produce a two dimensional matrix (with the time-step dimension and the feature vector dimension) representation for every token in the sequence. In order to find the words/POS tags with the highest predictive power, we use an approach similar to the one proposed by Lai et al. (2015), and employ one-dimensional max pooling and average-pooling operations (Collobert et al., 2011) on the time-step dimension to obtain two fixed-length vectors for each of the inputs.

The four resulting vectors are concatenated and fed into the rectified linear unit (RELU) activation function, on the output of which we conduct a dropout operation, in which 40% of input units are dropped in order to reduce overfitting. The resulting vector is passed on to a fully connected layer (*Dense*) responsible for producing the final binary gender prediction.

We use the Python Pytorch library (Paszke et al., 2017) for the implementation of the system. For optimization, we use an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001. Each of the models is trained on the train set for one hundred epochs and tested on the validation set after every epoch. The model with the best performance on the validation set is chosen for the test set predictions. For POS tagging, a Perceptron tagger from NLTK (Bird and Loper, 2004) is used and for measuring the performance of the classifier, accuracy is used.

4 Results

Classification results are presented in Table 2. On the official test sets, the highest cross-genre accuracy (55.20%) was achieved on news. Slightly worse was the accuracy on the data set of YouTube comments (55.07%), while the accuracy on the tweet test set was almost 1% lower. When it comes to the official in-genre results, the highest accuracy was achieved on the test set of YouTube comments (61.33%) and lowest on news (59.99%).

Results on the validation sets are in all cases better than the results on the official test sets, when same genres and same types of classification on validation and test sets are compared. This suggests some overfitting, which is generally more alarming in the in-genre setting, where the training sets were smaller.

Overfitting is the worst in the news in-genre setting, where the difference in performance on the official test set and validation set is almost 5%.

When we compare these results to the results of other teams in the CLIN shared task, our approach yields good performance in the cross-genre part of the competition, where we ranked second as a team, although it should be mentioned that the first ranked team submitted two runs which both performed better than our submitted run. On the other hand, our approach yields worse results in the in-genre setting, where we ranked sixth out of eight teams with the ninth best run.

5 Conclusion

In this paper we presented the results of the CLIN 2019 cross-genre and in-genre gender classification shared task performed on the data set of Dutch tweets, YouTube comments and news. A neural network architecture, which takes word and POS sequences as input, is capable of detecting relatively good features by performing max and average pooling on the output matrix of the LSTM layer. On the official CLIN 2019 test sets, our team ranked second in the cross-genre setting and sixth in the in-genre setting.

Not surprisingly, the models trained and tested on the same genre achieve much better performance than the models with train and test sets from different genres, even though the train sets in the cross-genre setting are much larger in all the cases. The performance of our classifier is quite consistent across all genres, which is against our expectations, since we expected better performance on the news data set because of the on average much longer documents and therefore more per-instance information for the classifier.

Dutch gender classification is still a tough problem, which becomes clear, if we compare the low performances of all the approaches in the shared task with the performances usually achieved on the English data sets in PAN shared tasks. In order to narrow this gap, for the short term future work we plan to test our approach on other languages, just to get the better picture of the difficulty of cross-genre and in-genre gender classification across different languages. We will also be conducting a comprehensive error analysis, which will help us identify language- and genre-independent features that work well across different genres and languages. In the long term, we will try to improve our approach by testing numerous state-of-the-art neural architectures and employ transfer learning techniques.

Acknowledgments

The work presented in this paper has been supported by European Unions Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors acknowledge also the financial

support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The Titan Xp used for this research was donated by the NVIDIA Corporation.

Bibliography

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Felice Dell Orletta and Malvina Nissim. 2018. Overview of the evalita 2018 cross-genre gender prediction (gxg) task. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA18), Turin, Italy. CEUR. org*.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273.

Matej Martinc and Senja Pollak. 2018. Reusable workflows for gender prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. 2017. Pan 2017: Author profiling-gender and language variety prediction. *Cappellato et al.[13]*.

Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma. 2017. Author profiling with word+ character neural attention network. In *CLEF (Working Notes)*.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2017. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. Available at `https://pytorch.org/`.

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

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*. CEUR.

Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giancomo Inches. 2013. Overview of the author profiling task at pan 2013. In *CLEF 2013 Evaluation Labs and Workshop Working Notes Papers*. CEUR.

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*.

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 *CLEF 2016 Working Notes*. CEUR-WS.org.

Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2018. Text and image synergy with feature cross technique for gender identification. *Working Notes Papers of the CLEF*.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: a multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. ELRA, Portorož, Slovenia.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67.