Gender Prediction from Tweets: Improving Neural Representations with Hand-Crafted Features

Erhan Sezerer
Dept. of Computer Engineering
İzmir Institute of Technology
İzmir, Turkey
erhansezerer@iyte.edu.tr

Ozan Polatbilek
Dept. of Computer Engineering
İzmir, Turkey
ozanpolatbilek@iyte.edu.tr

Selma Tekir
Dept. of Computer Engineering
İzmir, Turkey
selmatekir@iyte.edu.tr

Abstract

Author profiling is the characterization of an author through some key attributes such as gender, age, and language. In this paper, a RNN model with Attention (RNNwA) is proposed to predict the gender of a twitter user using their tweets. Both word level and tweet level attentions are utilized to learn 'where to look'. This model[1] is improved by concatenating LSA-reduced n-gram features with the learned neural representation of a user. Both models are tested on three languages: English, Spanish, Arabic. The improved version of the proposed model (RNNwA + n-gram) achieves state-of-the-art performance on English and has competitive results on Spanish and Arabic.

1 Introduction

Author profiling is the characterization of an author through some key attributes such as gender, age, and language. It's an indispensable task especially in security, forensics, and marketing. Recently, social media has become a great data source for the potential learning approaches. Furthermore, gender prediction has been a popular profiling task.

The traditional approach to gender prediction problem is extracting a useful set of hand-crafted features and then feeding them into a standard classification algorithm. In their study, Kucukyilmaz et al. (2006) work with the style-based features of message length, stop word usage, frequency of smiley etc. and use different classifiers such as k-nearest neighbor, naive bayes, covering rules, and backpropagation to predict gender on chat messages. Similarly, Deitrick et al. (2012) select some hand-crafted features and feed them into various classifiers.

Most of the work on gender prediction rely on n-gram features (Miller et al., 2012). Daneshvar and Inkpen (2018) give Latent Semantic Analysis (LSA)-reduced forms of word and character n-grams into Support Vector Machine (SVM) and achieve state-of-the-art performance. Apart from exploiting n-gram frequencies, there are studies (Ljubešić et al., 2017), (Alowibdi et al., 2013), (van der Goot et al., 2018) to extract cross-lingual features to determine gender from tweets. Some other work (Ljubešić et al., 2017), (Sayyadiharikandeh et al., 2016) exploit user metadata besides using just tweets.

[1] https://github.com/Darg-Iztech/gender-prediction-from-tweets

Recently, neural network-based models have been proposed to solve this problem. Rather than explicitly extracting features, the aim is to develop an architecture that implicitly learns. In author profiling, both style and content-based features were proved useful (Argamon et al., 2009) and neural networks are able to capture both syntactic and semantic regularities. In general, syntactic information is drawn from the local context. On the other hand, semantic information is often captured with larger window sizes. Thus, CNNs are preferred to obtain style-based features while RNNs are the methods of choice for addressing content-based features (Goldberg, 2017). In literature, CNN (Sezerer et al., 2018) or RNN (Takahashi et al., 2018), (Kodiyan et al., 2017), (Sezerer et al., 2019a) is used on this task. Takahashi et al. (2018) obtain state-of-the-art performance among neural methods by proposing a model architecture where they process text through RNN with GRU cells. Also, the presence of an attention layer is shown to boost the performance of neural methods (Takahashi et al., 2018), (Sezerer et al., 2018).

In this work, we propose a model that relies on RNN with attention mechanism (RNNwA). A bidirectional RNN with attention mechanism both on word level and tweet level is trained with word embeddings. The final representation of the user is fed to a fully connected layer for prediction. Since combining some hand-crafted features with a learned linear layer has shown to perform well in complex tasks like Semantic Role Labeling (SRL) (Collobert and Weston, 2008), an improved version of the model (RNNwA + n-gram) is also tested with hand-crafted features. In the improved version, LSA-reduced n-gram features are concatenated with the neural representation of the user. Then the result is fed into a fully-connected layer to make prediction. Models are tested in three languages; English, Spanish, and Arabic, and the improved version achieves state-of-the-art accuracy on English, and competitive results on Spanish and Arabic corpus.

There are many datasets created for this task (Pardo et al., 2018), (Sezerer et al., 2019b). In this work, we have used the dataset and benchmarks provided by the PAN 2018 shared task on author profiling (Pardo et al., 2018). As the dataset contains a constant number of 100 tweets per user, accuracy tests are performed both on user and tweet level (tweet-level predictions are made by removing the user-level attention). Tweet-level accuracy tests show interesting results during hyperparameter optimization. When the tweet-level predictions are averaged to produce user-level predictions, it is seen that the hyperparameters that gave the best results in terms of tweet-level accuracy, performs worse in user-level accuracy. The better user-level models, with different hyperparameters, that gave the highest user-level accuracy are observed to slightly overfit on tweet-level. It leads us to believe that the overfitting in the tweet-level predictions in best user-level models acts similar to an attention mechanism by over-emphasizing some distinctive tweets and ignoring the rest.

2 Model architecture

In author profiling, both style-based and content-based features must be addressed (Argamon et al., 2009). An appropriate baseline for this task is a CNN-based model that is able to capture style-based information (Sezerer et al., 2018). The proposed RNN-based model relies on extracting content-based features. In addition, in order to improve its accuracy, the proposed model is combined with some hand-crafted features. For all of the models, Adam optimizer (Kingma and Ba, 2014) is used with cross-entropy loss along with the L2 regularization to prevent from overfitting.

2.1 Baseline CNN model

CNN model (denoted CNNwA on results) is based on Sezerer et al. (2018) where each character in the tweet is represented with a character embedding of size 25, which is trained along the neural network. All characters are lower-cased. Non-alphabetical characters such as punctuation are kept with a view to capturing some information on the profile of the user since they are heavily used in twitter as emoticons.

Filters of size 3×3, 6×6 and 9×9 are used for each language, and the number of filters is determined by performing grid search on validation set. Among the tested range (50-125 with intervals of 25), the number of filters that gives the best accuracy is 100 (per each filter), for all languages.

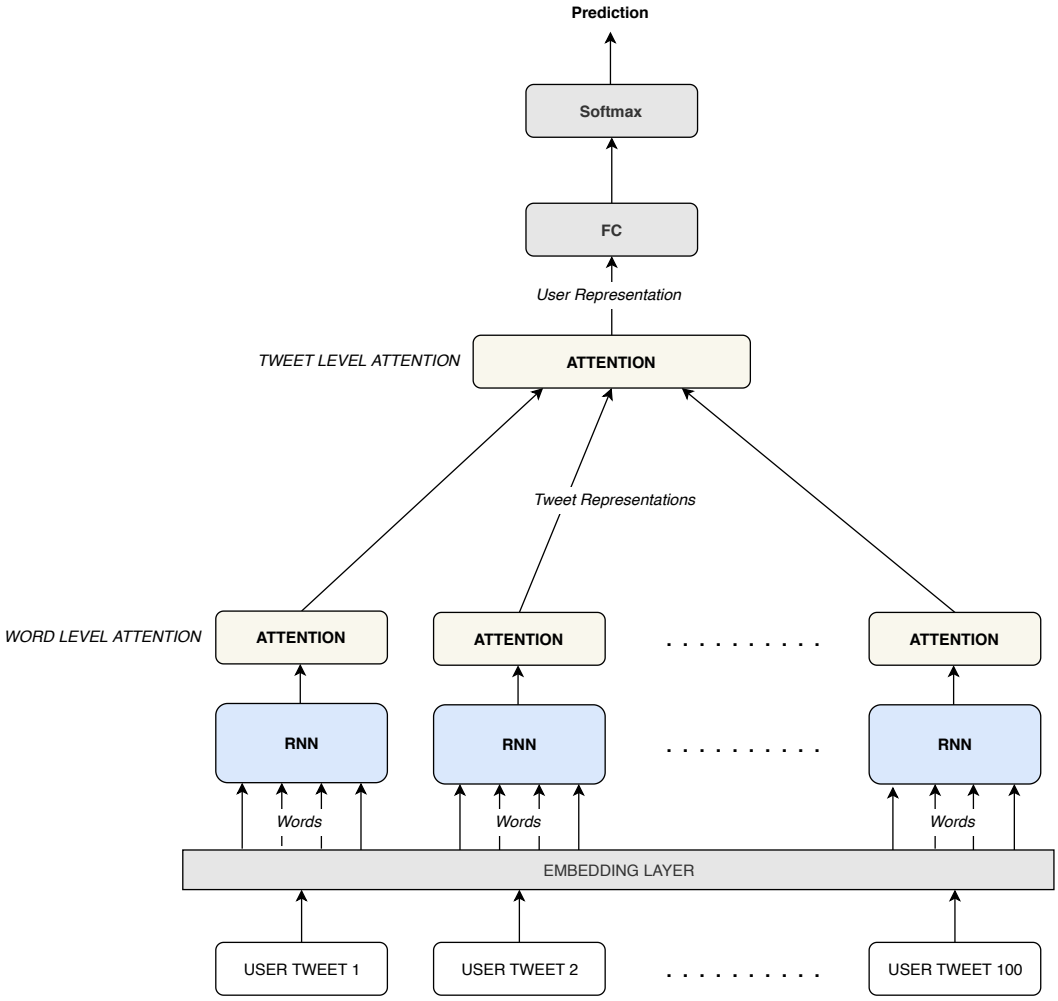

Figure 1: Proposed model.

2.2 RNN Model

Since the dataset is not big enough to train word embeddings, Glove word embeddings (Pennington et al., 2014) of size 200 are used in the proposed RNN Model (denoted RNNwA on results) due to their success at various NLP tasks and their multi-linguality: They encompass all the languages in the test set. In addition, the Glove embeddings are also trained on Twitter data which make them reflect the nature of the dataset better than other alternatives.

A bidirectional RNN with GRU (Chung et al., 2014) cells are used in this model where the number of cells is a hyperparameter. Among the tested range (50-150 with intervals of 25), best accuracy on validation set is obtained by 150 cells in English and 100 cells in Spanish and Arabic. An attention mechanism is used on word-level in addition to tweet-level to capture the important parts of each tweet as shown in Figure 1.

A feature vector for each tweet is created by feeding tweets to RNN separately. In order to discriminate tweets with respect to their information carrying capacity on its author's gender, Bahdanau attention mechanism (Bahdanau et al., 2014) is used to combine the tweets rather than concatenating them before feeding to the network or averaging their predictions later. Figure 2 shows the tweet-level attention layer in detail which is calculated by the following formulas:

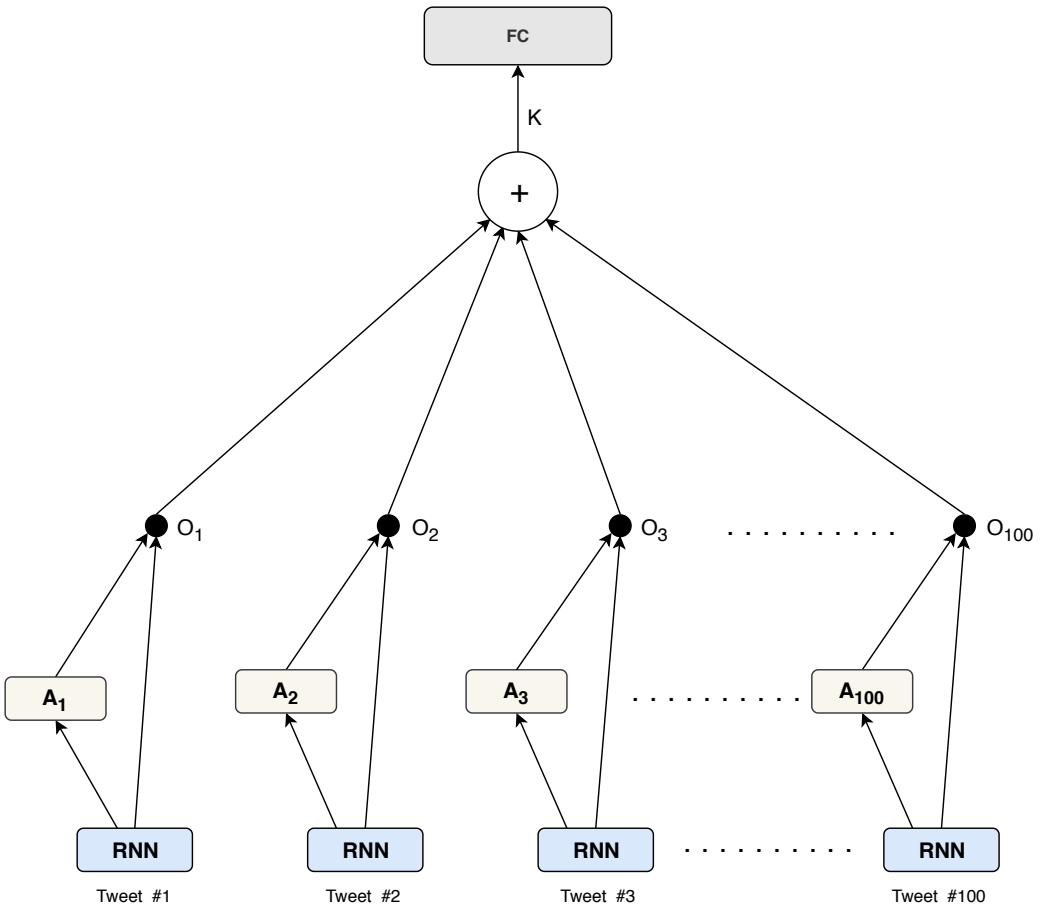

Figure 2: Tweet-level Attention Layer in Detail.

$$
\begin{aligned}
A_i &= tanh(\boldsymbol{W_\alpha} t_i + b) \\
v_i &= \frac{exp(A_i w_i)}{\sum_j exp(A_j w_j)} \\
o_i &= v_i t_i \\
K &= \sum_i o_i
\end{aligned}
$$

where $\boldsymbol{W_\alpha}$ is a learnable weight matrix that is used to multiply each output of the RNN, t_i is the feature vector of ith tweet, b is a learnable bias vector, w_i is a learnable attention weight, A_i is the attention context vector, v_i is the attention value for ith tweet, o_i is attention output vector for the corresponding tweet, K is the output vector for user. Matrix $\boldsymbol{W_\alpha}$ and vectors w_i and b are learned parameters.

Attention layer outputs a single feature vector that corresponds to a user, which is then fed to a fully-connected layer to lower the dimension to the number of classes.

There are two different attention layers on the model. One is a word level attention where it amplifies the signal coming from important words, the other one is on tweet level where it combines the signals coming from each tweet and creates the final representation of a user.

Table 1: Tweet Level Accuracy of the CNN and RNN Models without Attention.

Model	English	Spanish	Arabic
RNN	62.600	62.163	62.170
CNN	59.675	59.700	59.267

Table 2: User Level Accuracy of the Proposed Model (RNNwA) along with the Baselines.

Model	English	Spanish	Arabic
CNN	74.947	71.772	72.100
CNNwA	78.474	75.000	71.800
RNN	79.316	74.091	77.100
RNNwA	81.789	78.227	78.500

2.3 RNN with N-gram Model

For this model (denoted RNNwA + n-gram on results), n-gram features are collected with the same method described in Daneshvar and Inkpen (2018). At the beginning, word level and character level n-gram features are obtained and concatenated. Then they are normalized with tf-idf transformation. For reducing the number of features and sparsity in n-gram vectors, tuples that have frequency less than 2 are ignored. For character level n-gram N is selected as $3, 4$, and 5 and for word level n-gram, N is $1, 2$ for Spanish and Arabic; $1, 2, 3$ for English. The dimension of the vector is reduced by LSA to 300. Then the vector is concatenated with neural representation which is produced right after tweet level attention in RNNwA model. The resultant representation is fed to a fully- connected layer that produces predictions.

2.4 Dataset

Models are tested on the PAN 2018 author profiling dataset (Pardo et al., 2018), which provides tweets in three languages: English, Spanish and Arabic with training/test datasets of sizes (3000 users, 1900 users), (3000 users, 2200 users), and (1500 users, 1000 users) respectively, where each user has 100 tweets. Each training set is further partitioned randomly into training and validation sets with the ratio $(0.8, 0.2)$ respectively for hyper-parameter optimization.

3 Results

In order to measure the effectiveness of the attention mechanism, in addition to the CNN baseline model (CNNwA) and RNNwA, two new models (denoted as CNN and RNN) are created by removing the tweet level attention layer (word level attention stays the same) and generating a prediction for each tweet then just simply taking an average to give a user level prediction. Tweet level accuracies for these models are shown in Table 1.

In Table 2, user level accuracy results for the proposed model (RNNwA) along with the baseline models are given. As can be seen in the results, tweet level attention mechanism increases the score of all baseline models with the only exception of the CNNwA model in Arabic.

Also, compared to the best neural model (Takahashi et al., 2018) where max pooling is used instead of an attention mechanism on the outputs of RNN, the proposed model (RNNwA) gives better results in terms of accuracy on English and Arabic datasets, and produces similar accuracy levels on Spanish dataset (Table 3). These results show that an attention layer is able to learn "where/how to look" for features that are helpful in identifying the gender of a user.

On the other hand, the improved model (RNNwA + n-gram), where neural and hand-crafted features are concatenated, increases the accuracy of the proposed model by approximately $0,5\%$ on English and approximately 2% in Spanish and Arabic. This also supports our intuition that the performance of neural models can be improved by hand-crafted features, which is based on the study of Collobert and Weston (2008). As can be seen in Table 3, the improved model outperforms the state-of-the-art method of Daneshvar and Inkpen (2018) in English and produces competitive results in Spanish and Arabic.

Table 3: Accuracy on PAN 2018 test set.

Model	English	Spanish	Arabic
Daneshvar and Inkpen (2018)[1]	81.52	**82.00**	**80.90**
Takahashi et al. (2018)[2]	79.68	78.64	77.10
Proposed Model (RNNwA)	81.79	78.23	78.50
Improved Model (RNNwA + n-gram)	**82.31**	80.22	80.50

There is an interesting observation concerning the models without tweet level attention (RNN and CNN) in hyper-parameter optimization. During the hyperparameter optimization of the models RNN and CNN, we saved both the models that gave the best tweet-level accuracy and the models that gave the best user-level accuracy. The expectation is to see that the best setup on tweet-level also gives the best performance in user-level, but the outcome is the opposite: Best setups on tweet-level always fall behind best user-level setups. Performance differences between various setups can be seen in Figure 3 where accuracies of the best three models in terms of tweet-level and best three models in terms of user-level are shown for all languages. It can be observed that the best tweet-level setups are almost 4% worse in terms of user-level accuracy. Deeper investigation shows that the best user-level models exhibit slight overfitting on tweet-level, in training. Although overfitting normally leads to poor generalization, in this case we believe that this overfitting acts similar to an attention mechanism by over-emphasizing some important tweets and ignoring uninformative ones in the process. Even though this leads to poor tweet-level accuracy, it improves the user-level accuracy of the models as it can be seen from the Figure 3.

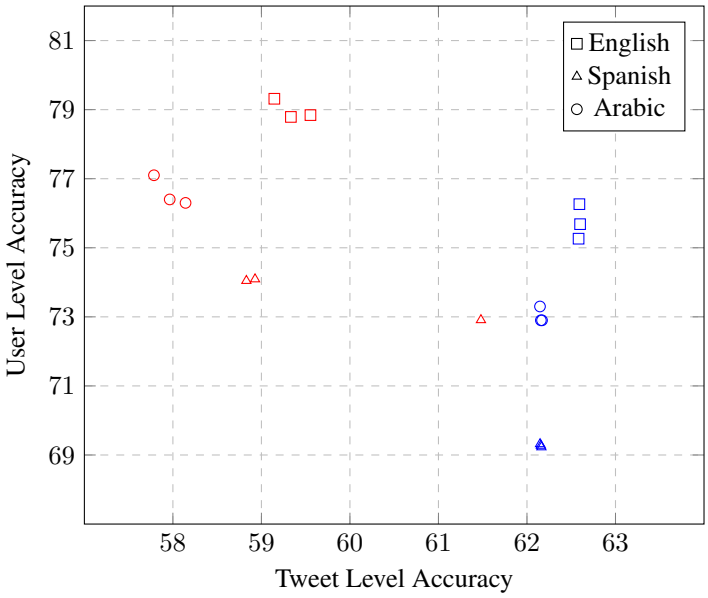

Figure 3: Comparison of Tweet-Level and User-level accuracy of RNN Model. Best three user-level models (colored in red) and best three tweet-level models (colored in blue) are selected for each language.

[1]In their paper, authors report a result of 82.21 in English but we couldn't verify their accuracy in our repetitions by using their software and the same dataset.

[2]Since their software is not provided, we directly take the accuracy values from their paper.

4 Conclusion

In this work, a neural network-based model namely RNN with attention (RNNwA) is proposed on the task of gender prediction from tweets. The proposed model is further improved by hand-crafted features which are obtained by LSA-reduced n-grams and concatenated with the neural representation from RNNwA. User representations that is the result of this model is then fed to a fully-connected layer to make prediction. This improved model achieved state-of-the-art accuracy on English and has a competitive performance on Spanish and Arabic.

We also would like to kindly remind our readers that although the model is self-learning, there might still exist a gender bias in the evaluation of the model due to the data itself. Since the model learns to predict the gender directly from tweets of the twitter users, any bias the twitter users have might be reflected in the model predictions.

Acknowledgments

We would like to thank Computer Vision Research Group from Izmir Institute of Technology for providing us the hardware for performing the tests in this research.

The Titan V used for this research was donated by the NVIDIA Corporation.

References

Alowibdi, J. S., U. A. Buy, and P. Yu (2013). Language independent gender classification on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pp. 739–743. ACM.

Argamon, S., M. Koppel, J. W. Pennebaker, and J. Schler (2009, February). Automatically profiling the author of an anonymous text. *Commun. ACM 52*(2), 119–123.

Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Chung, J., Ç. Gülçehre, K. Cho, and Y. Bengio (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*.

Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML*, Volume 307 of *ACM International Conference Proceeding Series*, pp. 160–167. ACM.

Daneshvar, S. and D. Inkpen (2018). Gender identification in twitter using n-grams and LSA: notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

Deitrick, W., Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu (2012, 01). Author gender prediction in an email stream using neural networks. *Journal of Intelligent Learning Systems and Applications 04*, 169–175.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *CoRR*.

Kodiyan, D., F. Hardegger, S. Neuhaus, and M. Cieliebak (2017). Author profiling with bidirectional rnns using attention with grus. In *CLEF*.

Kucukyilmaz, T., B. B. Cambazoglu, C. Aykanat, and F. Can (2006). Chat mining for gender prediction. In T. Yakhno and E. J. Neuhold (Eds.), *Advances in Information Systems*, pp. 274–283. Springer Berlin Heidelberg.

Ljubešić, N., D. Fišer, and T. Erjavec (2017). Language-independent gender prediction on twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 1–6. Association for Computational Linguistics.

Miller, Z., B. Dickinson, and W. Hu (2012, 01). Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science 02*, 143–148.

Pardo, F. M. R., P. Rosso, M. M. y Gómez, M. Potthast, and B. Stein (2018). Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In *CLEF*.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Sayyadiharikandeh, M., G. L. Ciampaglia, and A. Flammini (2016, November). Cross-domain gender detection in twitter. In *Proceedings of the Workshop on Computational Approaches to Social Modeling (ChASM 2016)*.

Sezerer, E., O. Polatbilek, O. Sevgili, and S. Tekir (2018). Gender prediction from tweets with convolutional neural networks. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*.

Sezerer, E., O. Polatbilek, and S. Tekir (2019a, Apr). Gender prediction from turkish tweets with neural networks. *Signal, Image and Video Processing (in print)*.

Sezerer, E., O. Polatbilek, and S. Tekir (2019b, August). A Turkish dataset for gender identification of twitter users. In *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy, pp. 203–207. Association for Computational Linguistics.

Takahashi, T., T. Tahara, K. Nagatani, Y. Miura, T. Taniguchi, and T. Ohkuma (2018). Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*.

van der Goot, R., N. Ljubešić, I. Matroos, M. Nissim, and B. 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: Short Papers)*, pp. 383–389. Association for Computational Linguistics.