Author Profiling with Bidirectional RNNs using Attention with GRUs

Notebook for PAN at CLEF 2017

Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak

Zurich University of Applied Sciences
kodiydon@students.zhaw.ch, hardeflo@students.zhaw.ch,
neut@zhaw.ch, ciel@zhaw.ch

Abstract This paper describes our approach for the Author Profiling Shared Task at PAN 2017. The goal was to classify the gender and language variety of a Twitter user solely by their tweets. Author Profiling can be applied in various fields like marketing, security and forensics. Twitter already uses similar techniques to deliver personalized advertisement for their users. PAN 2017 provided a corpus for this purpose in the languages: English, Spanish, Portuguese and Arabic. To solve the problem we used a deep learning approach, which has shown recent success in Natural Language Processing. Our submitted model consists of a bidirectional Recurrent Neural Network implemented with a Gated Recurrent Unit (GRU) combined with an Attention Mechanism. We achieved an average accuracy over all languages of 75,31% in gender classification and 85,22% in language variety classification.

1 Introduction

Social media has become an important platform for communication and exchange of information. In contrast to classical letters and emails, the language on social media is much more personal. This raises the question whether the text style and content allows to draw conclusions about demographics traits of its author, such as age, gender, or language variety. Such insights can be used in various applications, such as forensics, security, or marketing. For instance, on the basis of such profiles it would be possible to determine which users could be interested in a new product or campaign, how urgent a complaint is, or if a profile in an online forum might be a fake profile.

The Author Profiling Shared Task of the PAN shared task aims to answer these question by extracting information about authors based on their linguistic style of writing [14,13]. The goal of the 2017 shared task at PAN is to detect the author's gender and dialect from his/her Twitter texts. Both training and test data is provided in four different languages: English, Spanish, Portuguese and Arabic.

We have implemented a solution that is based on a bidirectional recurrent neural network (bi-RNN) using gated recurrent units (GRUs) in combination with an attention mechanism.

The paper is structured as follows. In Section 3, we give a short overview of related work. Then, in Section 4, we describe our model, and Section 5 compares the different attempts and their results on test data. Conclusions are drawn in the last section.

2 PAN

PAN is a series of different digital text forensics tasks. It organizes shared task evaluations. Shared Tasks are computer science events of a specific problem of interest. This paper is the result of our participation at the Author Profiling Shared Task of 2017. Author Profiling includes gender and language variety predictions of an author of a given Twitter document. To solve this problems, training and test datasets are available [16].

PAN 2017 Training Data. PAN 2017 Training Data consists of Twitter profiles in four different languages: English, Spanish, Portuguese and Arabic. The corpus was annotated with gender and language variety information about the authors.

For each of the language varieties, there are 600 Twitter profiles. In each language there are the same number of male and female profiles. The dataset includes exactly 100 tweets for each author.

Language Variety. Language Variety is defined as a specific variation of an author's native language. For instance, one has to identify whether an English author has a language variation from Australia, Canada, Great Britain, Ireland, New Zealand or the United States.

Table 1. Distribution of data for language variety in the PAN 2017 training corpus

Native Language	Author Profiles	Language Variations
English	3600	Australia, Canada, Great Britain, Ireland, New Zealand, United States
Spanish	4200	Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela
Portuguese	1200	Brazil, Portugal
Arabic	2400	Egypt, Gulf, Levantine, Maghrebi

TIRA. TIRA is an evaluation-as-a-service platform [12]. The submission for the PAN shared task was done with this tool. The submitted models were self-evaluated on a virtual machine which was hosted by the organizers. The test data was only available on this virtual machine and was not visible to the participants.

Evaluation. The performance measure of the submissions at PAN 2017 is done with accuracy. The individual accuracy for gender and variety identification was calculated for each language as follows:

$$\text{accuracy} = \frac{\text{correct predicted}}{\text{total}}. \tag{1}$$

The *joint* accuracy is calculated when both gender and variety are properly predicted together. The final ranking is calculated with the averaged accuracy over all four languages.

3 Related Work

In this chapter we provide an overview of the most relevant works for the Author Profiling Task with neural networks.

Neural Networks. Neural networks have achieved great results in natural language processing in the past few years. In many tasks like machine translation [10] and sentiment analysis [7], neural networks have proven to be very successful. The two state-of-the-art neural networks used today are recurrent neural networks (RNN) and convolutional neural networks (CNN). The main challenge in most NLP tasks is to simplify the input sequence and keep the most important information. Research on neural machine translation (NMT) already focuses heavily on this challenge. For that reason we applied techniques from NMT to the Author Profiling Task.

RNNs and CNNs. The recent success of RNNs are achieved through long short-term memory networks (LSTM) and gated recurrent unit networks (GRU) [6]. With their capabilities of long-term dependencies, LSTMs and GRUs have achieved state-of-the-art results in various NLP tasks. The work of Bahdanau et al. [1] proposed an attention mechanism to simplify a sequence. In combination with a bidirectional RNN (bi-RNN) learns this approach to automatically weigh the most relevant information of the input sequence. This leads to substantial improvements in machine translation and other fields like automatic summarization [4]. The latest research of Gehring et al. [10] has shown that CNNs are capable of achieving state-of-the-art results in NMT. Those results were achieved by applying the attention mechanism to CNNs. CNNs are computationally less expensive compared to LSTMs and GRUs, which makes them preferable for large datasets.

4 Methodology

In this chapter we describe the technical solution. Main focus is on the system architecture of the neural networks.

4.1 Preprocessing

Every single tweet was preprocessed by converting them to lower-case. We replaced URLs and usernames with a standardized token. We converted hashtags to regular words and used the TweetTokenizer from NLTK [2] to tokenize the tweets. We use a vocabulary to map tokens with an token-ID. The IDs point to a vector representation of the token, which is used later. After the preprocessing step we receive a list of tweets of each author and each tweet is a list of token-IDs.

4.2 Embeddings

Each token in a tweet is represented by pretrained word embeddings [8]. For English and Spanish we used embeddings created with word2vec [11]. For both languages a

corpus of 200 million unlabelled tweets were used. The skip-gram algorithm was used for training with window-size 5, sample size of 1e-05, minimum frequency of 15 and 200 dimensions.

For Portuguese and Arabic we used pretrained embeddings from [3], which were trained on Wikipedia corpus[1]. They have an output dimension of 300.

4.3 Architecture

In this section we describe our model, which consists of a bi-RNN with GRUs followed by an attention mechanism.

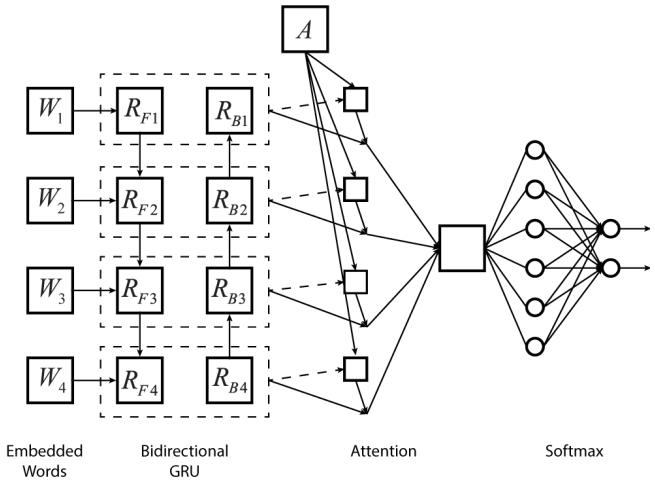

Figure 1. Representation of the bi-GRU+Attention model. We used $n = 4$ and $u = 3$ for visualization purposes.

Embedding Layer. The embedding layer is used to map the token-IDs with their vector representation. The token-ID is used to lookup the word-vector in the embeddings. Those vectors get concatenated and are passed to the next layer. This results in an output matrix $S \in \mathbb{R}^{d \times n}$, where d stands for the dimension of the word vector and n for the size of the input. To determine n, we took the tweet with the biggest amount of tokens from our training dataset and rounded the number up to the next 10. This resulted in a maximum input size of $n = 60$. Shorter inputs were padded with zeros to match that size. To reduce the effect of unknown and padded words we used masking [5]. This way our model only uses known words and skips zero-values.

GRU Layer. This layer consists of two GRUs with u number of units. We used a GRU for each direction, which resulted two matrices $R_F \in \mathbb{R}^{u \times n}$ and $R_B \in \mathbb{R}^{u \times n}$. Finally both matrices were concatenated and resulted a matrix $R \in \mathbb{R}^{2u \times n}$. For our model we used $u = 50$.

[1] http://wikipedia.org

Attention Layer. This layer is used to weight the most important parts of the GRU encoded input and deliver a more simplified matrix of the input. The output-matrix R of the previous layer, the weight-matrix $W_a \in \mathbb{R}^{2u \times 2u}$ and the bias $b \in \mathbb{R}^{2u}$ is used to calculate a hidden state h_t:

$$h_t = \tanh(W_a R + b). \tag{2}$$

The hidden state h_t and the weight-vector $W_u \in \mathbb{R}^{2u}$ used to calculate the final attention a for each word by

$$a = \text{softmax}(h_t W_u). \tag{3}$$

The attention-vector a is then multiplied with R and the result summed together. This results a summarized representation of the sentence as a vector $s_a \in \mathbb{R}^{2u}$.

Softmax Layer. As the final layer we used a fully connected layer with softmax as the activation function. The number of output nodes were depending on the number of classification possibilities. For gender prediction were 2 nodes required, for language variety predictions were between 2 and 7 nodes required, depending on the language.

Dropout. Dropout drops individual nodes during training with a probability of p and is therefore used to reduce overfitting [15]. We used dropout on our softmax layer with $p = 0.2$.

Optimization. Our model is trained using the AdaDelta optimizer [17]. We used $\epsilon = 10^{-5}$ and default values for the other hyper-parameters.

Author Prediction. Our model is trained to classify single tweets. To get the classification of an author, his tweets are classified separately. The outputs of our model, which is the output of the softmax layer, is then summed together and the class with the highest value is the final prediction. For example, if we want to predict the gender of an user u who has three tweets t_1, t_2, t_3, we first classify the tweets separately. This could result following predictions: $t_1 = [0.4, 0.6], t_2 = [0.3, 0.7], t_3 = [1.0, 0.0]$. The first number of each output indicates the probability that the tweet is written by a female and the second number indicates the probability that the author is a male. The outputs of the tweets t_1, t_2, t_3 are summed together and results $[1.7, 1.3]$. In this example, user u would be predicted as a female.

4.4 Training

To train our models for submission we used 90% of the training data and the remaining 10% were used as validation set. The validation set was used to select a model checkpoint during training. For more details in model checkpoints, see Section 5.1.

5 Evaluation

We distinguish between the evaluation during development, and the benchmark measured on actual test data on TIRA. The results during the development phase were achieved on the provided training corpus with cross validation.

Cross Validation. Our models were trained with 10-fold cross validation. We used cross validation to calculate a representative score for the model. The data in each fold was used as follows: 80% training data, 10% validation data and 10% test data. The evaluation on the test data does not influence the training and is only used to evaluate the model. We used a validation set in combination with model checkpoints to prevent overfitting. Model checkpoints will be explained in the following section.

F1 Score. During the training phase we used F1 score to find the best model. The F1 score considers both *precision* and *recall* to compute the score. We used the F1 score, because it penalizes one-sided predictions of a model. The abbreviations tp, fp, fn indicate in the following calculations *true positives*, *false positives* and *false negatives*. Precision is the ratio between correct predicted (tp) to all classified data of this class ($tp + fp$):

$$\text{precision} = \frac{tp}{tp + fp}. \tag{4}$$

Recall is the ratio between correct classified data (tp) to the number of total data in the corresponding class ($tp + fn$):

$$\text{recall} = \frac{tp}{tp + fn}. \tag{5}$$

The harmonic mean of this two scores is called F1 score. The F1 score is calculated as follows:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{6}$$

5.1 Model Checkpoints

The accuracy and F1 score of the model were measured during training. The scores were evaluated on a validation and a test dataset. If the model achieved a higher F1 score on the validation data than a previous one, the model (and its weights) was saved. An example of the measured scores is shown in Figure 2.

The goal is to select the best weights for a model during the training phase. Figure 2 shows that our model performs very similar on validation and test data. That means by choosing the best weights on the validation set, the chances are high that the model performs equally on the test set. This makes our model very stable and predictable.

5.2 Analysis of the Attention

While working with attention mechanism we developed a tool to represent how the different words in a tweet are weighted. This tool helped us to understand which words are more important for our model. An example on language variety is shown in Figure 3, where multiple tweets of British and American authors are compared.

In Figure 3 the attention of the words are highlighted. As we can see some typical American English and British English words are marked. For example, in the first tweet

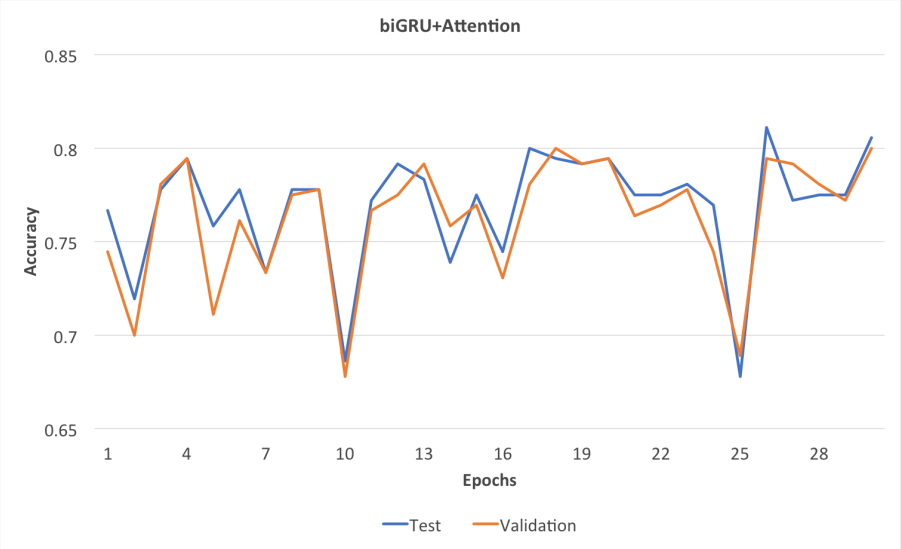

Figure 2. Accuracy graphs of the bi-GRU+Attention model during training. Visualized comparison between validation (orange) and test (blue) accuracy scores on author level. On the X axis is the number of epochs represented and on the Y axis the corresponding accuracy value.

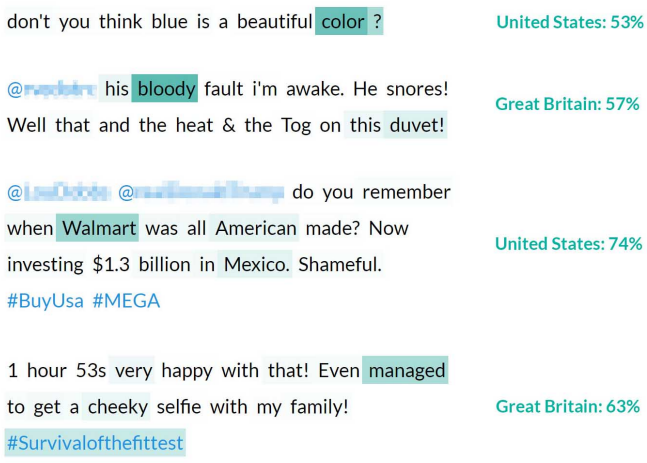

Figure 3. Visualized attention weights comparison between British and American Twitter users. The left side visualizes the attention of each word in a tweet. The darker the background color, the stronger those words are weighted. On the right side the final prediction and its probability is shown. In these examples are all predictions correct.

is the word "color" and in the third tweet "Walmart" marked as very important, which are common words in American English. In the second and fourth tweet are the words "bloody" and "cheeky" marked as significant for British English, which are common words in British English.

5.3 Cross Validation Results

During our preparation for the PAN shared task several models were tested and compared. Our baseline was a CNN model [7] which already participated in PAN 2016. The model has a 2-layer CNN architecture with a fully-connected softmax layer at the end.

The experiments have shown that the bi-GRU+Attention model has the best performance on both classification tasks (gender, variety). The measured scores of both models are shown in Table 2 and Table 3.

Table 2. Evaluation results of classifying gender on PAN 2017 training datasets using cross validation.

	Gender				
Model	**English**	**Spanish**	**Portuguese**	**Arabic**	**Average**
bi-GRU+Attention	79,03%	72,57%	79,50%	71,58%	75,67%
CNN	73,24%	72,93%	79,83%	70,88%	74,22%

Table 3. Evaluation results of classifying language variety on PAN 2017 training datasets using cross validation.

	Language Variety				
Model	**English**	**Spanish**	**Portuguese**	**Arabic**	**Average**
bi-GRU+Attention	79,03%	92,05%	98,76%	78,71%	87,11%
CNN	70,90%	89,67%	98,75%	78,38%	84,22%

5.4 PAN 2017 Results

We trained two distinct models for each language: one for gender and one for variety. These models were uploaded to the virtual machine and were evaluated on the actual test dataset. In Table 4 the results obtained on the PAN 2017 Author Profiling test dataset are shown.

The highest score on gender prediction was achieved in English. Portuguese gender prediction follows with 0.075% less accuracy. The gender predictions in Spanish and Arabic are lower than the others. We assume that this issue is related to the worse vocabulary usage: For both languages Spanish and Arabic, the vocabulary coverage is below 80%, in contrast to around 90% coverage of the vocabularies in English and Portuguese.

In general, good scores are achieved for variety prediction. Outstanding is the variety accuracy of 91,43% for the Spanish language, which consists of seven language variations. The score dropped only in English and Arabic below 80%. The lowest score

Table 4. Evaluation results in terms of accuracy for the bi-GRU+Attention model on PAN 2017 Author Profiling test dataset.

Language	Joint	Gender Accuracy	Number of Language Variations	Language Variety Accuracy
English	62,63%	78,88%	6	79,08%
Portuguese	73,00%	78,13%	2	93,50%
Spanish	66,46%	72,17%	7	91,43%
Arabic	56,88%	71,50%	4	76,88%

76,88% is achieved for variety prediction on Arabic, due to low vocabulary coverage. The exact vocabulary coverage of the used embeddings is shown in Table 5.

Table 5. Vocabulary usage of the embeddings on PAN 2017 training dataset in comparison to the achieved gender prediction accuracies. The languages are sorted by gender accuracy score.

Language	Gender Accuracy	Vocabulary Coverage
English	78,88%	90,85%
Portuguese	78,13%	88,33%
Spanish	72,17%	79,68%
Arabic	71,50%	77,66%

The results in Table 5 seems to imply that the accuracy for gender prediction correlates with vocabulary coverage.

6 Conclusion

In this paper, we presented deep learning models to predict gender and language variety of Twitter profiles. We described a bidirectional RNN with GRU and an attention mechanism. We compared the average accuracy of our models over all languages with a previously developed CNN model. The RNN exceeds the CNN in gender prediction by 1,45% and in variety prediction by 2,69% on average over four languages on PAN 2017 training data.

For future work, we would like to see if a combination of several high-quality solutions for Author Profiling with a random forest could even outperform each of the subsystems. This has been done successfully for sentiment analysis [9], and it would be interesting to see if it works for Author Profiling as well.

7 References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. CoRR abs/1409.0473 (2014)

2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. CoRR abs/1607.04606 (2016)
4. Cho, K., Courville, A., Bengio, Y.: Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks. IEEE Transactions on Multimedia 17(11), 1875–1886 (Nov 2015)
5. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
6. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014)
7. Deriu, J., Cieliebak, M.: Sentiment Analysis using Convolutional Neural Networks with Multi-Task Training and Distant Supervision on Italian Tweets. In: Evaluation of NLP and Speech Tools for Italian (EVALITA) (2016)
8. Deriu, J., Lucchi, A., Luca, V.D., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., Jaggi, M.: Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1045–1052 (2017)
9. Dürr, O., Uzdilli, F., Cieliebak, M.: JOINT_FORCES: Unite Competing Sentiment Classifiers with Random Forest. SemEval 2014-Proceedings of the 8th International Workshop on Semantic Evaluation pp. 366–369 (2014)
10. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional Sequence to Sequence Learning. ArXiv e-prints (May 2017)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. CoRR abs/1310.4546 (2013)
12. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
13. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017)
14. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15(1), 1929–1958 (Jan 2014)
16. Tan, L., Zampieri, M., Ljubešic, N., Tiedemann, J.: Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In: Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC). pp. 11–15. Reykjavik, Iceland (2014)
17. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method. CoRR abs/1212.5701 (2012)