Duluth at PAN-2013: The Author Profiling Task

Saptarshi Sengupta
Department of Computer Science
University of Minnesota
Duluth, MN 55812, USA
sengu059@d.umn.edu

Abstract

In this paper, we present two neural models developed for the Author Profiling Task at PAN 2013. The task required systems to identify the age and gender of the author of a given document. Both of our models are feed-forward, but they have different configurations and are trained using two types of features viz. TFIDF and GloVe embeddings. The TFIDF based model achieved an accuracy of 0.34, whereas the latter obtained 0.31.

1 Introduction

Imagine writing an award-winning novel only to be told by everyone that they do not believe that you have authored it, or being accredited with having written a horrible invective against someone dear! What tragic situations to be in! Such issues are prevalent in real life and author identification/profiling aims to tackle them.

Traditional author identification involves determining who, among a set of authors, wrote a certain document. Here, however, we are dealing with author profiling which requires categorizing authors based on gender, age, writing style, etc. Thus, profiling encompasses a broader identification task in which we search for common traits among a class of authors rather than isolating the nuances of a single author.

Social media is one area where author profiling finds much use, and for a good reason. There have been several cases of predators, and people masquerading as others to warrant the use of author profiling, for safeguarding innocent users against such threats. Employing such techniques can help us better understand the linguistic styles of such authors and thereby hone in on them quickly.

2 Task Description

The objective of the author profiling task of PAN-2013 (Rangel et al., 2013a) was to identify the age-range ($13-17/23-27/33-47$) and gender (male/female) of the author of a certain document. The task was carried out for two languages viz. English (EN) and Spanish (ES).

The dataset (training and test) was sourced from three locations; blog posts from open-source online repositories, samples from conversations involving sexual predators, and sex-related conversations between adults. Including male and female authors, the training data consisted of 236K EN and 76K ES documents.

The dataset for evaluation was provided in two stages, the first being for an early bird trial run (EN-21K and ES-68K), and the second was for the final evaluation run (EN-25K and ES-8K).

For ranking a system, the accuracy of joint identification of age and gender, averaged over both languages, was considered.

3 Related Work

At the time of proposing the author profiling task, the problem was still in its nascent stage. The only prior work seems to have been conducted by Argamon et al. (2009) who trained a Bayesian Multinomial Regression (BMR) classifier using a combination of style and content-based features for detecting an author's gender, age, native language, and neuroticism [1] level (a prolonged tendency to be more moody or anxious than the average person).

All of the submitted models took a supervised learning approach, seeing as the supplied data was labeled. Cruz et al. (2013) took an ensemble approach in which they trained a linear Support Vector Machine (SVM) using n-grams (with n = 1,2,3)

[1] https://www.medicalnewstoday.com/articles/246608

and Part-of-Speech (POS) features followed by training it on style-based features such as accents and punctuation marks. The feature sets from both models were then combined and passed on to the JRip classifier. This approach led them to achieve 0.39 overall accuracy for ES and 0.31 for EN.

Alemán et al. (2013) proposed two models, one for each language. Both models trained a Random Forest classifier on different sets of features. For EN, the feature vector was made up of frequencies of emoticons, contractions, misspelled words, etc. A more sophisticated approach was taken for ES. After splitting the training set by gender and age-range, graphs were generated for the documents in each class according to a star topology. The objective of building the graph was to discover unigrams *exclusively* used by one gender for the age-range and not the other. For EN, they achieved an overall accuracy of 0.33 while for ES it was 0.16.

4 Ethical Considerations

While pondering the ethics of author profiling, a few observations were made. The dataset which was curated for the task exploited some known stereotypes for both sexes (Rangel et al., 2013a). For example, it was assumed that men would be more likely to talk about *beer* and *football* as opposed to women who were more likely to engage in conversations about *clothes* and *shopping*. However, it should be mentioned that the curators of the dataset were in no way trying to reinforce these ideas. On the contrary, they were trying to develop a diverse corpus to propose a more realistic task.

When it comes to women authors, it is sometimes seen that they publish articles under gender-neutral or male names. Famous cases of this include *J.K. Rowling* and *E.L. James*. They deliberately chose to do so, so as not to get prejudged by their audience. Speaking in an interview with CNN [2], Rowling said,

> My publisher, who published *Harry Potter*, they said to me, we think this is a book that will appeal to boys and girls. And I said, oh, great. And they said, so could we use your initials?

Thus, our models need to be cognizant of this fact and explore more subtle features so as not to incorrectly recognize a female-authored article as being written by a man.

A more precarious extension of author profiling would be criminal profiling wherein a mistaken identity could lead to devastating effects. Such was the case with *Raymond Jennings* who after being wrongly accused of the murder of *Michelle O'Keefe* by ex-FBI profiler *Mark Safarik's* testimony, was sentenced to 11 years in prison [3]. Fortunately, albeit quite late, Jennings was exonerated and let go. This just goes to show how a small lapse in judgment could very well destroy a person's life. Therefore, taking heed from these cases, our models should be fairly certain when delivering their predictions in practice.

Finally, an important consideration when conducting such research is user anonymity. For instance, we might envisage situations in which a user, playing the role of a *whistleblower*, would certainly not like to get deanonymized. As can be inferred from Townsend and Wallace (2016), special care must be taken when tackling such real-world author profiling/identification issues so as not to jeopardize someone's safety.

5 Methodology

We developed two feed-forward neural networks (FFNN) for the task [4] trained using embedding and TFIDF features. To evaluate their performance, we provide results from a majority classifier, which we used as a baseline.

5.1 Preprocessing

The dataset for the task (Rangel et al., 2013b) consisted of XML formatted conversation data. The XML data was converted to CSV, as it is easier to work with. Each XML file represented a collection of conversations involving a certain user.

Before saving a conversation to the CSV file, the text was cleaned by removing HTML tags, newlines, extra spaces, and non-word characters.

A conversation's class (truth label) was determined by the age group and gender to which it belonged to. As such, there were six classes ranging from "male-10's" (marked as 0) to "female-30s" (marked as 5). These class labels were generated and stored in the CSV file.

[2] https://www.cnn.com/2017/07/10/world/amanpour-j-k-rowling-interview/index.html

[3] https://www.latimes.com/local/california/la-me-profiler-wrongful-conviction-20170720-htmlstory.html

[4] https://github.com/saptarshi059/Advanced_NLP_Research_Project

5.2 Feed-Forward Network

A FFNN is perhaps the most rudimentary of all deep learning (DL) models. As such, we decided to use it as a starting point into DL methods on such a task. Our models were implemented using the popular DL framework *PyTorch* [5] for Python.

Both models ignored word order and treated conversation data as a bag-of-words (BOW). In one model, the words (unigrams) were represented using TFIDF features and in the other, by embeddings. For the latter, conversations could be viewed as bag-of-embeddings (BOE).

Apart from the BO(W|E) technique, each model used *Cross Entropy* as the loss function and *Adam* as the optimizer. Both networks were trained over 10 epochs with a 0.01 learning rate.

5.2.1 TFIDF Based Network

The first model was trained using TFIDF features extracted from the training data. Although simple, these features often provide interesting insights and serve as a nice starting point into text classification problems. We decided to keep the top 51 features as values higher/lower than this led to a decrease in classification accuracy. The features were obtained using *scikit-learn* (Pedregosa et al., 2011).

We created a simple three-layer FFNN consisting of two hidden layers and a single output layer. The input layer had as many nodes as the input feature vector's size (51) and the output layer had 6 nodes, each corresponding to a certain class a conversation could belong to. Each hidden layer had 100 dimensions and *ReLU* was used as their activation function.

5.2.2 GloVe Based Network

Embedding features tend to capture deeper semantic information about words and has become the go-to in terms of modern NLP. As a result, we decided to train a model using such features.

Instead of training our embeddings, we opted to use pre-trained GloVe embeddings (Pennington et al., 2014). From the assortment of pre-trained models available, the 42 billion tokens with 300 dimensions version was used. A smaller variant was initially chosen, but due to out-of-bounds vocabulary issues, a larger model was required.

Here, the idea was to tokenize a conversation into unigrams (BOW), replace each unigram with its corresponding embedding, and finally condense the BOE into a single representation that would then be passed on to the network.

The network consisted of a single linear layer and an *EmbeddingBag* layer, which condensed the BOE into a single vector by taking an element-wise average of the embeddings. Finally, the predictions were obtained by applying a simple linear transformation on the condensed input vector.

6 Experimental Results

As there was much more data for the EN version of the task, we decided to deploy our models on it only. We took such a decision since DL methods typically work well when they have a large dataset to train on.

The results from the majority classifier and the FFNN's on the final test data are shown in Table 1.

Classifier	Accuracy
Majority	0.29
FFNN-TFIDF	0.34
FFNN-GloVe	0.31

Table 1: Prediction Accuracy on Final Test Data.

From the results above, we see that our models offer a slight improvement over a simple majority classification method. However, even a 0.34 accuracy seems to be a positive result when compared with the top team in the competition [6] which achieved an accuracy of only 0.39. Such scores perhaps speak more of the data than the models as it might indicate how linguistically challenging the data is to process and thus why the models were performing so poorly.

Tables 2 and 3 describe the confusion matrices for the FFNN-TFIDF and FFNN-GloVe models respectively. The *true positives* for each class are highlighted. First of all, it should be mentioned that either model was not able to predict all of the classes. FFNN-TFIDF only predicted 2 of the 6 classes while FFNN-GloVe predicted 4. This behavior was expected since the training data provided was wildly imbalanced. The total number of authors in the 30's age group was 133,508 while the 20's and 10's group had 85,703 and 17,200. The test data was also imbalanced with 30's (14,408), 20's (9,174), and 10's (1,776). Thus, it was natural for our models to gravitate

[5] https://pytorch.org/

[6] https://pan.webis.de/clef13/pan13-web/author-profiling.html#results

towards the 30's group, as is evidenced by both matrices. Be that as it may, our GloVe model was still able to predict classes 1(male,20's) and 4 (female,20's), the next larger bracket. This just goes to show how effective embedding features are at capturing deeper contextual information. However, on account of predicting more classes, the model suffered from lower accuracy.

Finally, it was seen that for classes 2 (male,30's) and 5 (female,30's), the precision remained practically the same, 0.3 and 0.47 resp., for both models. This could perhaps indicate that both models were learning a similar representation for these classes.

Predicted / True	2	5	All
0	868	565	1433
1	7967	690	8657
2	11517	2689	14206
3	1085	377	1462
4	8089	1006	9095
5	8268	4806	13074
All	37794	10133	47927

Table 2: FFNN-TFIDF Confusion Matrix. [0:male,10's; 1:male,20's; 2:male,30's; 3:female,10's; 4:female,20's; 5:female,30's]

Predicted / True	1	2	4	5	All
0	2	1273	7	151	1433
1	45	8375	100	137	8657
2	24	13387	82	713	14206
3	7	1329	19	107	1462
4	38	8619	182	256	9095
5	13	11778	68	1215	13074
All	129	44761	458	2579	47927

Table 3: FFNN-GloVe Confusion Matrix.

7 Future Work

In this paper, we presented two models for the author profiling task. As evidenced by the results, there remains much more to be done. Viewing the accuracies in isolation (i.e. not in light of the competition), 0.34, and 0.31 is simply not enough to warrant commend. Thus, we are interested in investigating other feature classes and ways in which our proposed architectures could be improved. We also need to figure out a way to overcome the class imbalance hurdle since we would at least want our models to be predicting all of the available classes.

The models proposed in this work are quite simple and as such, we are curious to see how well more sophisticated architectures like RNN's or LSTM's perform. Also, with Google's BERT achieving state-of-the-art performance on a range of GLUE [7] and SuperGLUE [8] tasks, it would be another point of exploration.

References

Yuridiana Alemán, Nahun Loya, Darnes Vilariño, and David Pinto. 2013. Two Methodologies Applied to the Author Profiling Task. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*. CEUR-WS.org.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123.

Fermín L Cruz, R Rafa Haro, and F Javier Ortega. 2013. Italica at pan 2013: an ensemble learning approach to author profiling—notebook for pan at clef 2013. In *Forner et al.[8]*. Citeseer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013a. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013b. Pan13 author profiling.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics.

[7] https://gluebenchmark.com/

[8] https://super.gluebenchmark.com/