Learning Multimodal Gender Profile using Neural Networks

Carlos Pérez Estruch **Roberto Paredes** **Paolo Rosso**
Pattern Recognition and Human Language Technology (PRHLT) Research Center
Universitat Politècnica de València, València, Spain
{carprees, rparedes, prosso}@prhlt.upv.es

Abstract

Gender identification in social networks is one of the most popular aspects of user profile learning. Traditionally it has been linked to author profiling, a difficult problem to solve because of the little difference in the use of language between genders. This situation has led to the need of taking into account other information apart from textual data, favoring the emergence of multimodal data. The aim of this paper is to apply neural networks to perform data fusion, using an existing multimodal corpus, the NUS-MSS data set, that (not only) contains text data, but also image and location information. We improved previous results in terms of macro accuracy (87.8%) obtaining the state-of-the-art performance of 91.3%.

1 Introduction

Nowadays we are experiencing a, more than remarkable, change and growth of technology. Emergence of online social networks has reinvented the form of communication and has taken it to another level. However, it has not been beneficial only to users, since companies and researchers have obtained access to tons of data (aka big data) generated in social media.

This information could be used in multiple ways. There is a growing interest in the search of "who" created the contents, but not exactly as a complete individual identity, rather as a collection of general user profiles traits (gender, age, personality...) in order to construct user groups.

Being able to infer the profile of a user can be beneficial for tasks such as security and marketing. In marketing, we may take advantage of this information, for example, for personalized products recommendations making the selection easier. With respect to security, it is important to have an idea about who could have written a potential threat.

In this paper we aim to approach how to identify the gender of the users taking into account multimodal data. Therefore, the proposed neural networks use different fusion strategies in order to combine the different modalities.

The paper is organized as follows. Section 2 describes related works on gender identification. In Section 3 we present our neural network architectures. In Section 4 we describe the data set, the used techniques and discuss the results. Finally, in Section 5 we draw some conclusions and discuss future work.

2 Related work

The gender identification task has traditionally been related to author profiling. Pennebaker et al. (2003) made some initial approaches to asses how the variation of linguistic characteristics in a text can help in the finding of gender and age of an author with respect to others. Argamon et al. (2003) explored the difference in writing style between male and female, in a large subset of the British National Corpus. Koppel et al. (2002) inferred in the search of gender in a less formal corpus, using a combination of lexical and syntactic features with 80% of accuracy. More related to our study, Schler et al. (2006) analyzed many blogs also looking for differences in writing style between gender and age groups with results closes to 80% of accuracy. Also, Argamon et al. (2009) worked with anonymous texts from blogs using two types of basic features: content-based and style-based features. From a gender perspective, it is interesting to see that with the content-based features they obtained better results (75.1%), 4%

Proceedings of Recent Advances in Natural Language Processing, pages 577–582,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_075

more than with the style features. With the combination of style and content features they gained one point of accuracy.

Burger et al. (2011) is one of the first works about gender identification in Twitter. They used a big data set of 184,000 users of which approximately 18,000 are for test set. They obtained very good results: 91.8% of accuracy. Notice that they used names of people which is a very discriminating information about gender. Using only textual information of tweets they reached 74.5% of accuracy.

Since 2013 tasks on author profiling have been organized at the PAN lab of CLEF (Rangel et al., 2013). The focus is on addressing the problem from a perspective exclusively related to the processing of text. The work of Farseev et al. (2015) is possibly the first preliminary study in author profiling from the multimodal perspective with very interesting results: 87.8% in terms of macro accuracy.

3 Model Description

In this work we developed different neural network topologies that operate under a multimodal approach. One of the first questions that can arise when starting to work with multimodal data is how to fuse the data. We can distinguish between two basic types of fusion strategies: early fusion and late fusion.

The idea of the early fusion approach (Figure 1a) is to concatenate all data sources into a new larger vector and feed it directly to the network without doing any previous single source classification (one learning phase). From the perspective of a basic neural network could be more difficult to make a good representation of the data if the different sources have different scales. This can result in a normalization-dependant model, precisely for reducing the variance of the data, and therefore for processing them in a common representation space.

In contrast with the previous one, the late fusion (Figure 1b) is based on performing the fusion after the single source classification is done, at the decision level. We wanted to avoid the use of complex (and usually overmuch heuristic) fusion operations, and in order to address this problem without this disadvantage we developed from scratch a fusion neural network. This modification of the basic multilayer perceptron consists of as many softmax layers as needed, one for each data source (single source learning) and a last softmax layer that modifies the weights of all the network. As we can see, we concatenated each linear output of each single source sofmax layer. Finally, we added one last hidden layer after the concatenation of the outputs, and before the last output in order to discover the last discriminating information of the concatenation vector.

In this case, we consider that one iteration of the network finishes when the 6 learning phases (forward + backward) have been completed. The order in which the single source learning phases are applied should not influence the final result. In our case the last learning phase was the one that modifies the weights of all the network (in Figure 1b, `N: Out[2]`).

4 Evaluation

In this section we describe the data set, the data preprocesing and all the techniques used to develop the models. Finally, we present and discuss the results.

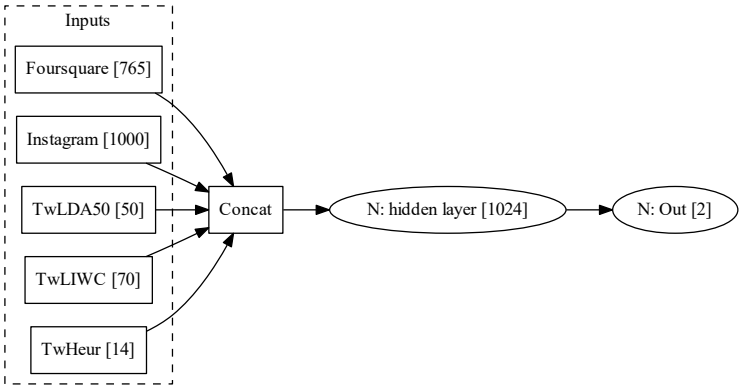
4.1 Data Set

The data set used in this work was obtained from Farseev et al. (2015). The NUS-MSS data set contains processed social media data from three different cities (Singapore, London and New York) and from three different social networks (Foursquare, Instagram and Twitter) for users of each city. We have focused our research on Singapore users. Following, we analyze each source separately:

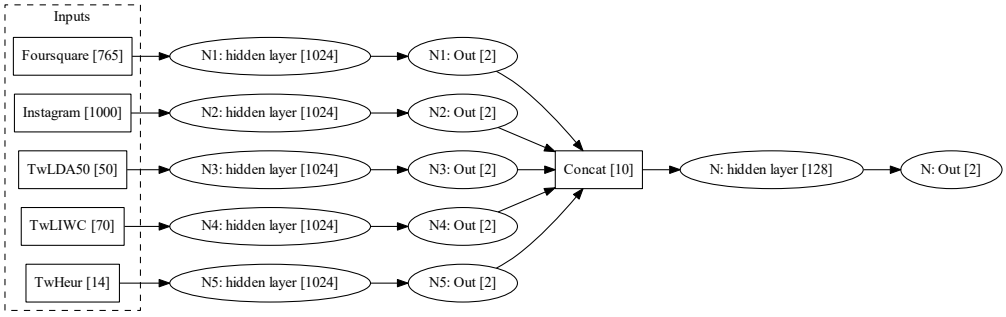
- *Foursquare*[1]: We have an user mobility profile as a count (checkins) of visited places. It follows the Foursquare category hierarchy[2] of Singapore, plus more than 150 extra categories. In total we have vectors with 764 components of mobility for each user.

- *Instagram*[3]: This is a preprocessed corpus constructed from users original uploaded images, mapped to a 1000-dimensional (ImageNet labels (Deng et al., 2009)) vector forming an image concept dictionary per user.

[1] https://es.foursquare.com/
[2] https://developer.foursquare.com/categorytree
[3] https://www.instagram.com/



(a) Early fusion architecture



(b) Late fusion architecture

Figure 1: Artificial neural network data fusion strategies. In (a) we can see the basic early fusion topology. In (b) we have the late fusion neural network. Note that the number of hidden layers depends exclusively on the problem to solve.

- *Twitter*[4]: Textual information had been processed in 3 ways: Latent Dirichlet allocation (LDA) 50-dimensional vectors with latent topic space information (Blei et al., 2003); 70-LIWC features vectors (Pennebaker et al., 2001); vectors with 14 manually defined characteristics, "heuristically-inferred features" (Farseev et al., 2015), with for example the number of hashtags, number of emoticons, number of tweets, among others, for each user.

Gender labels were extracted from the Facebook[5] ground truth file provided in the data set.

For the total amount of collected users, we trained our model with the 3172 of which we had data in all three sources. The test set is composed of 222 *other* users.

4.2 Data Preprocessing

There are certain classifiers, such as random forest, that work well regardless of the initial data representation space, but this is not the case of neural networks. Neural network models can learn from non-standardized data, but they will not do it in an optimal way. In this case:

1. Instagram data was modified because the original uploaded data was scaled depending on the number of images for each user. We divided every user's 1000 image concept vector by their number of images to scale them to a more practical range of values.

2. We applied z-score normalization (Eq. 1) for every source (except for Foursquare data)

[4]https://twitter.com/
[5]https://www.facebook.com/

in order to accelerate the learning, making stochastic gradient descent convergence faster.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Foursquare vectors were difficult to train in our models because of their sparsity. Some categories, or "visited" places, are 0 for all users. It was not possible to apply z-score normalization to these categories (0 division in the scale operation). Finally, we decided to use this source without preprocessing after testing other types of normalization as for example min-max.

4.3 Training

4.3.1 Activation Function

We used ReLU (Eq. 2 and 3) as activation function for the hidden layers (Nair and Hinton, 2010). At present ReLU is performing better than other activation functions, such as sigmoid, in many scenarios.

$$ReLU(x) = f(x) = max(0, x) \quad (2)$$

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \quad (3)$$

He et al. (2015) introduced a modification of ReLU called PReLU (Eq. 4). They proposed to use a small learnable parameter to control the slope of the negative part, in order to avoid zero gradients:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ a_i x, & \text{if } x \leq 0 \end{cases} \quad (4)$$

We tested it, but in comparison with the basic ReLU function we did not see any significant difference in results, although the model converged faster. This can be explained since the trainable parameter acquires a higher value in the first epochs of the training, and therefore the model learns faster because of the highest weights modification.

Finally, we discarded this option because not a big improvement was observed and, on the contrary, it added more training parameters.

4.3.2 Weight Initialization

Weight initialization is linked to the activation functions used in the hidden layers. We used He et al. (2015) initialization for ReLU in all cases. We did not use any pretrained initialization.

4.3.3 Regularization Techniques

We detected overfitting in training, possibly due to the relatively few number of users in the training set. Overfitting is one of the most common problems in neural networks but there are some ways to prevent it.

Dropout (Srivastava et al., 2014) is an extended technique used in these cases. Dropout is based on the idea of leaving disabled a certain number of neurons in each iteration to simulate the train of several networks. The basis is that it is very difficult for the network to see the same data distribution twice and, therefore, the representation should be more distributed.

As we said, data sources were from different social networks, and from different modalities, that means that the vectors were in different representation spaces. Moreover, each source itself has data with a lot of variability. For this reason we decided to apply one interneuron normalization technique known as batch normalization (Ioffe and Szegedy, 2015).

Batch normalization reduces internal covariate shift of data and makes training faster. It can be used jointly with dropout, although there is a better alternative that does the same task as dropout, but with the advantage of data augmentation. We added Gaussian noise with a normal distribution before the scale and shift step (just after the normalization):

$$y_i \leftarrow \gamma(\widehat{x_i} + \widehat{x_i}\mathcal{N}(\mu, \sigma)) + \beta \equiv \text{BN}_{\gamma,\beta}(x_i) \quad (5)$$

where γ and β are the scale and shift learnable factors respectively, $\mathcal{N}(\mu, \sigma)$ is the random values generator which follows a normal distribution, and $\widehat{x_i}$ is the normalized activation over a mini-batch.

The two learnable parameters (γ and β, eq. 5) learn from different data in each epoch (very low probability that each value appears twice). It is important to know that if we use a too large deviation value, the model is going to learn from random data, and if we use a small parameter the modified values of the data will be too close to their representation without noise. For this reason, the best practice is to generate noise with 0 mean and standard deviation between 0.2 and 0.4. All results presented in this paper were obtained using Batch Normalization with Gaussian noise ($\mu = 0, \sigma = 0.3$).

4.3.4 Other Considerations

The learning rate was 0.001 in all cases. It may seem small but according to the data set size we preferred that weight updates were small to control better the training.

We used minibatch stochastic gradient descent (SGD) as optimizer. All parameters were selected applying 10 fold cross-validation, extracting each fold from the training set.

4.4 Results

In Table 1 we show our results together with the previously obtained ones (Farseev et al., 2015). We would like to highlight the improvements we obtained in terms of macro accuracy.

Gender (Macro acc.)	Baseline	Ours
Single source		
Location (Foursquare)	0.649	**0.714**•
LIWC text (Twitter)	0.716	0.737
Heuristic text (Twitter)	0.685	0.712
LDA 50 text (Twitter)	0.788	**0.806**
Img. concepts (Instagram)	0.784	**0.885**•
Multisource combinations (late fusion)		
LDA50 + LIWC	0.784	0.823
LDA50 + heuristic	0.815	0.805
Heuristic + LIWC	0.730	0.755
All text	0.815	0.822
Img. concepts + location	0.802	0.875•
All text + img. concepts	0.824	**0.900**•
All text + location	0.743	0.833•
All sources		
Complete early fusion	0.707	0.910•
Complete late fusion	0.878	**0.913**

Table 1: Macro accuracy results. In the first row we show Farseev et al. (2015) previous results obtained with random forest classifiers. The second row shows the results obtained with our neural network based approach. In bold the best performance on each social media source, the best multisource combination, and all sources. The values with (•) are significantly higher than the baseline (significance level 0.95).

As we can see, there is a high improvement of accuracy on the Instagram set as well as on the other single sources. This is, in part, because random forest classifiers do not have much capacity to discriminate between extensive data vectors that, in contrast, is one of the main advantages of the neural networks. The Instagram source is more related to the liking of a person, and is very helpful from a point of view of gender identification, because males and females (with exceptions) tend to upload very correlated and representative pictures to how they are.

In the case of fusing the LDA-50 characteristics with the heuristic data of the tweets, our result is worse than previous and similar to the LDA-50 single one. In general, the base operation of the late fusion network causes that, if there is contradictory information in the sources, the decision will be related to the most discriminating one.

	Male	Female
Male	97	32
Female	13	80

Table 2: Confusion matrix for single LDA50 model.

	Male	Female
Male	94	35
Female	11	82

Table 3: Confusion matrix for LDA50 + heuristic model.

In this case, the fusion model with the heuristic features contributed to classify better the females but it misclassified some of the males in contrast to single LDA (Tables 2 and 3). These problems do not affect in the same way other classifiers as random forest. A similar behaviour occurs when using only the location information with the image concepts.

Analyzing the results where we used all sources, our approaches worked very well for both fusion techniques. Also, we obtained a very good result employing early fusion, which means that using the right techniques, a neural network has the capacity to classify correctly gender regardless where the fusion takes place.

It is difficult to know which one of the two architectures is better because the macro accuracy difference is not statistically significant (t-test), and both models have a similar number of parameters ($\approx$ 2 million). With relation to its training speed and simplicity, it is clear that the early fusion one can be more efficient, although it worked well in part due to the normalization techniques.

When not using it, the training was more inconsistent than with the late fusion approach, and the accuracy result was worse.

5 Conclusions and Future Work

We developed and tested different neural networks models using multimodal data for a gender identification task with successful results: 91.3% of macro accuracy.

As future work, we will continue working in this direction, searching for the best way to combine and use data. The base fusion neural network models work well and it will be interesting to extend them to add, for example, convolutional neural networks for preprocessing of the raw images and texts, in order to solve these problems using deep learning approaches.

Acknowledgments

This work has been funded by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat R. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT & TALK - An Interdisciplinary Journal of Language, Discourse & Communication Studies* 23:321–346.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '11, pages 1301–1309.

Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting multiple sources for user profile learning: A big data study. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, New York, NY, USA, ICMR '15, pages 235–242.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, ICCV '15, pages 1026–1034.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*. JMLR, Lille, France, volume 37 of *ICML'15*, pages 448–456.

Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing 17(3), in press.* 17(4):401–412.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*. Omnipress, Haifa, Israel, ICML'10, pages 807–814.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *Working Notes of CLEF 2013 Conference and Labs of the Evaluation forum*. Valencia, Spain.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. pages 199–205.

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.