Age and Gender Identification in Social Media

James Marquardt[1], Golnoosh Farnadi[2,3], Gayathri Vasudevan[1], Marie-Francine Moens[3], Sergio Davalos[1], Ankur Teredesai[1], Martine De Cock[1,2]

[1]Center for Data Science, University of Washington Tacoma, WA, USA
[2]Dept. of Appl. Math., Comp. Science and Statistics, Ghent University, Belgium
[3]Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
{jamarq, gvasu, sergiod, ankurt, mdecock}@uw.edu
golnoosh.farnadi@ugent.be
sien.moens@cs.kuleuven.be

Abstract. This paper describes the submission of the University of Washington's Center for Data Science to the PAN 2014 author profiling task. We examine the predictive quality in terms of age and gender of several sets of features extracted from various genres of online social media. Through comparison, we establish a feature set which maximizes accuracy of gender and age prediction across all genres examined. We report accuracies obtained by two approaches to the multi-label classification problem of predicting both age and gender; a model wherein the multi-label problem is reduced to a single-label problem using powerset transformation, and a chained classifier approach wherein the output of a dedicated classifier for gender is used as input for a classifier for age.

Keywords: Gender identification · Age prediction · Multi-label classification · Text mining

1 Introduction

There exist many applications which benefit from reliable approaches for inferring age and gender of users in social media. Such applications exist across a wide array of fields, from personalized advertising to law enforcement to reputation management. Text posts represent a large portion of user generated content, and contain information which can be relevant in discovering undisclosed user attributes, or investigating the truthfulness of self-reported age and gender.

A common approach of uncovering hidden user attributes in social media is to model the writing habits of users by extracting various features from texts they have posted. This approach, however, suffers from the inability of models generated from one genre of social media to be successfully applied to other genres in some cases.

In this work we address the issue of profiling authors of online textual media by selecting a feature set shown to have relatively high predictive power in terms of age and gender accuracy across multiple media genres. In the design of our system, we have taken into account the possibility of text to be classified as either a blog post, tweet, hotel review, or collection of social media posts. Our

system makes an informed guess about the gender of the author (male or female) as well as his or her membership in significant age brackets (18-24, 25-34, 35-49, 50-64, 65+) as determined by the organizers of the PAN 2014 author profiling task.

Additionally, we observe the accuracies gained by two approaches to the multi-label classification problem of simultaneously identifying both age and gender. We apply our feature set to one model wherein the multi-label problem is reduced to a single-label problem using powerset transformation, and one wherein the output of a single classifier is used as the input for a second classifier.

This paper is structured as follows. After reviewing related work in Section 2, we describe the data set and several preprocessing steps in Section 3. In Sections 4 and 5 we give a description of the features that we extract and our two approaches to multi-label classification, the results of which are discussed in Section 6.

2 Related Work

There exists a large amount of work in the area of age and gender prediction using textual data from social media, a generous portion of it having been completed in response to the PAN 2013 author profiling task [11]. Other recent work into author profiling has demonstrated the ability to infer the hidden attributes of authors of social media with accuracies in excess of 91% for attributes such as gender [15]. Works such as these, however, tend to focus on collections of lengthy text posts. Similar work has been done on inferring latent user attributes such as gender, age, regional origin, and political orientation from much shorter social media posts, such as Netlog chat messages [9] and Twitter microblogs [12]. It is interesting to note that in these works age identification is often treated as a binary classification problem (e.g. distinguishing between users who are below 30 and users who are above 30), while in the PAN 2014 task age prediction is defined as a more challenging multi-class prediction problem with five classes (18-24, 25-34, 35-49, 50-64, 65+).

For the aforementioned PAN 2013 task, investigation into inferring gender and membership in one of three age groups (13-17, 23-27, and 33-47) was conducted by 21 teams. The approaches taken by the different groups varied widely in terms of both feature sets and classification approaches [11]. Notable observations of these works include the relative lack of predictive utility of n-gram based models, as well as the high level of accuracy achieved by a group using class similarity based features [5]. Important differences between the PAN 2013 and PAN 2014 author profiling tasks are that the number of age groups has been increased from three to five, and that the different genres of social media text in the challenge has increased from a single genre in 2013 to four genres in 2014.

3 Data Set and Preprocessing

The data used for the training of our system consists of different data sets that cover four online media genres: blogs, Twitter feeds, hotel reviews, and unspecified social media posts[1]. A corpus of each genre is present in both English and Spanish, with the exception of hotel reviews which is only present in English. Documents in the corpus consist of a collection of posts made by a single user.

All corpora used for the training of our final models are balanced in terms of authorship by each gender, and additionally by gender within each age group. However, each corpus displays imbalance in terms of age representation. Additionally, the level of representation of age groups differs between corpora. The proportion of each age group in all corpora within the training set is presented in Table 1.

Table 1: Proportion of each age group for all corpora

	English						Spanish					
	Age Group					Doc	Age Group					Doc
Genre	18-24	25-34	35-49	50-64	65-xx	#	18-24	25-34	35-49	50-64	65-xx	#
Blogs	4.1	40.8	36.7	15.7	2.7	147	4.6	29.5	47.7	13.6	4.6	88
Twitter	6.5	28.8	42.5	19.6	2.6	306	6.7	23.6	48.3	18.0	3.4	178
Social Media	20.0	27.1	29.0	23.7	0.2	7746	25.9	33.5	25.5	12.6	2.5	1272
Reviews	8.7	24.0	24.0	24.0	19.3	4160	-	-	-	-	-	-

Prior to any model training or testing, we apply the following set of preprocessing steps to all documents.

Firstly, we eliminate on average approximately 71% of each document file prior to extracting features. First, we disregard all file contents not determined to be text from a user post. This enables us to ignore characters belonging to XML tags, as our primary source of features is the text written by an author. To do this, we note that all user posts lie within the unparsed data tags of the source .xml file. We disregard any text not within these tags. From this text, we then discard any HTML, making note of occurrences of specific tags for our feature vectors (see Section 4).

Once the above step is completed, we eliminate portions of posts determined to be generated by spambots. Spam posts are determined to be those which contain a large amount of the % character; most likely due to an attempt to obfuscate spam lexicon words. This step removed a combined 0.7% of the text across all corpora.

As a preprocessing step specific to the Twitter corpora, we eliminate all posts determined to be retweets, as text in retweet posts is not the product of the poster, hence not a reliable source to determine his age and gender. It

[1] http://pan.webis.de/

was determined that 1.8% of the English tweets were retweets, and 2.0% of the Spanish tweets were retweets.

4 Feature Extraction

To create our feature space for age and gender inference we extract several different categories of features, drawing inspiration from related work, such as LIWC features [14, 15], sentiment features [6], and emoticons [12]. Features were selected by evaluating the contribution to accuracy for each feature. All features, with the exception of psycholinguistic features, are extracted from both English and Spanish texts.

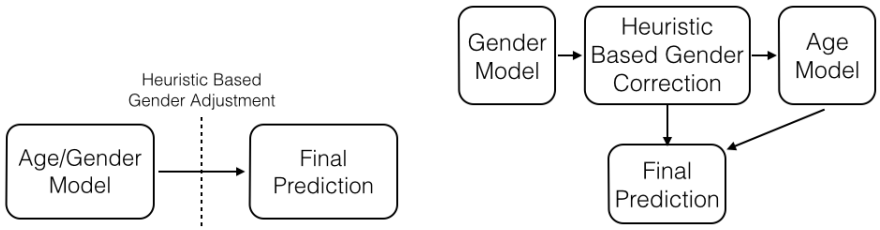
1. *Content-based features*
   - **MRC Features** We extract 14 features from the MRC psycholinguistic database [3] for the English data sets. These features capture information about frequency of words that connote psycholinguistic concepts such as familiarity, concreteness, and imagery.
   - **LIWC Words** We utilize the Linguistic Inquiry and Word Count dictionary (LIWC) [10] in order to extract 68 of our features. By using the LIWC dictionary, we developed our own software to determine the frequency of words that can be categorized as motion, anger, or religion based, along with a score for other categories. We observed these features to be particularly useful for classifying age and gender in our hotel reviews corpus.
   - **Sentiment** Using the SentiStrength tool [17], we extract features concerning the number of sentences expressing either positive, negative, or neutral sentiments. The tool calculates a sentiment value for each word, with positive values corresponding to positive sentiment, negative for negative sentiment, and zero for neutral sentiment. For each sentence in the document, we take the sum of the sentiment values for each word to find the sentence's overall sentiment; positive for positive, negative for negative, and zero for neutral. We then sum the number of sentences in the document belonging to each sentiment category to extract three separate features.
2. *Stylistic features*
   - **Readability** Six features are extracted that act as a measurement of readability for each document. We extract the average number of words per sentence, the number of sentences, and the number of characters. Additionally, we calculate the Automated Readability Index (ARI) [16], the Coleman-Liau Index (CLI) [2], and the Rix Readability Index (RIX) [1] of each document.
   - **HTML Tags** For every document, we determine the number of uses of various HTML tags to extract five features. In particular, we look for incidents of links, images, bold, italics, and lists.

– **Spelling and Grammatical Errors** Using the jLanguageTool [7], we extract the count of spelling and grammatical errors in each document. We then normalize these two features by dividing by the number of words in the document.
– **Emoticons** Through use of a fairly simple regular expression (i.e., $(? :: |; | =)(? : -)?(? :—|D|P))$, we extract a single feature to denote the frequency of emoticons used in each document e.g., :), ;).
– **Other Features** In our model, we consider a document to be the collection of all posts from a single user. As such, one feature we extract is the total number of posts by a user. Other features that are extracted are the number of capitalized letters and the number of capitalized words.

In addition to the features mentioned above, we employ a system of heuristics based adjustment for gender prediction using a customized lexicon of phrases. Inspired by the work in [12], we created a collection of n-grams which signify distinguishing properties of a specific gender. For example, the phrases 'my wife' and 'my girlfriend' are more likely to be used by men while it is highly probable that the expressions 'my husband' and 'my handbag' were written by a woman. The lexicon contains 20 English phrases (13 female; 7 male) and 16 Spanish phrases (9 female; 7 male). During classification, we label the gender of a document that contains a phrase in this collection to be the gender associated with said phrase.

5 Models and Evaluation

For each genre and language combination, we train two different models with the features from Section 4: one model based on label powerset transformation (LP), and one model based on the idea of classifier chains (CC). In both cases we use SVM as the underlying learning algorithm, and we evaluate the accuracy of the models using the scikit-learn [8] implementation of Liblinear SVM [4]. For comparison purposes, we use a simple majority class baseline model (see Table 2).



(a) Label powerset approach (LP) (b) Chained classifiers approach (CC)

Fig. 1: Two approaches to the multi-label classification problem of predicting age and gender

Label powerset transformation (LP) [18]: LP turns a multi-label classification problem into a single label one by unifying labels. We combined the separate gender and age labels into 10 gender-age labels (e.g., female-18-24, male-18-24, female-25-34 etc.), and trained an SVM classifier to distinguish between these 10 classes. Heuristic based adjustment of gender is performed after the initial prediction is made, leading to the final prediction, as shown in Figure 1.a. The results of this approach are shown in Table 2.

Classifier chains (CC) [13]: To consider the dependency between labels, CC approaches utilize two single label classifiers in which the prediction made by the first is used as a feature in the second. We created a classifier for gender and a classifier the five age classes that utilizes the inferred gender as a feature in the model. We choose this ordering of classifiers due to experimental observations indicating gender is a more useful feature in inferring age than age is for gender. Again, heuristic based adjustment of genders is performed, although in this case the adjustment occurs such that the age classifier will receive a "corrected" gender as input as shown in Figure 1.b. Table 2 presents the results of applying the CC approach.

Table 2: Accuracy of age classification in all corpora for all models; highest results for language/genre is shown in bold

		English			Spanish		
Model	Genre	Total	Gender	Age	Total	Gender	Age
Baseline	Blogs	19.60	50.00	40.80	23.80	50.00	47.70
	Twitter	28.75	50.00	42.50	24.15	50.00	48.30
	Social Media	14.49	50.00	29.00	16.74	50.00	33.50
	Reviews	12.01	50.00	24.0	-	-	-
LP	Blogs	**23.12**	**68.71**	39.46	37.50	**80.68**	47.72
	Twitter	32.79	**71.15**	**46.89**	**33.71**	**74.72**	**48.31**
	Social Media	19.86	54.22	36.56	**26.10**	**64.62**	41.67
	Reviews	19.09	**65.46**	**29.83**	-	-	-
CC	Blogs	23.05	66.59	**42.86**	**38.71**	72.93	**47.73**
	Twitter	**33.44**	69.15	47.73	31.62	71.35	**48.31**
	Social Media	**20.16**	**57.39**	**36.78**	24.48	63.14	**41.75**
	Reviews	**19.25**	63.11	**29.83**	-	-	-

6 Discussion

As seen in Table 2, our models outperform the baseline for combined label prediction accuracy in for all corpora. It is notable that this is not necessarily the case for age prediction. Our powerset transformation models for English blogs, fail to beat the baseline for age prediction. In all cases where this occurs, how-

ever, the accuracy of gender prediction is high enough such that combined label accuracy still beats the baseline.

Comparison between accuracies obtained between the two models reveals little advantage to either in terms of combined label accuracy, with the difference in accuracies being as little as .07% for some models. However, it should be noted that the powerset transformation model outperforms those using chaining classifiers in terms of accuracy of gender prediction. In particular, accuracies for gender prediction using the Spanish blogs corpus were nearly 8% higher for the label powerset transformation model.

Although the combined label accuracies for all models outperform the baseline, the prediction accuracies across corpora vary wildly. For example, while accuracies achieved by the Spanish blogs model were 37.50%, the accuracies seen for the English reviews model were only 19.09% in the powerset transformation approach. This indicates that although the feature set used in our system will beat random labeling of gender and age, the degree to which it does so depends largely on the corpus being evaluated.

7 Conclusion

In this work we have presented a feature set with predictive power that can be extended across multiple genres of online textual media. We found that for our model to remain relatively stable across different genres, it requires multiple categories of features to be extracted. However, as seen in the English social media corpus, features that work well across many genres may not necessarily perform well on others. However, considering the relatively small feature vector size, the models' performance relative to the baseline helps establish its value. We also found the accuracy of predicting age gained by using a more complicated classification scheme such as chained classifiers to be negligible.

Acknowledgements

This work was funded in part by the SBO-program of the Flemish Agency for Innovation by Science and Technology (IWT-SBO-Nr. 110067).

References

1. Anderson, J. LIX and RIX: Variations on a little-known readability index. *Journal of Reading*, pages 490–496, 1983.
2. Coleman, M., Liau, T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.
3. Coltheart, M. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981.
4. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

5. López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Villatoro-Tello, E. INAOE's participation at PAN'13: Author profiling task. 2013.
6. Mukherjee, A., Liu, B. 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, 2010.
7. Naber, D. A rule-based style and grammar checker. Bielefeld University Bielefeld, Germany, 2003.
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
9. Peersman, C., Daelemans, W., Van Vaerenbergh, L. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.
10. Pennebaker, J. W., Francis, M. E., Booth, R. J. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.
11. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. Overview of the author profiling task at pan 2013. In *Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013*, pages 23–26, 2013.
12. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.
13. Read, J., Pfahringer, B., Holmes, G., Frank, E. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.
14. Schler, J., and Koppel, M., and Argamon, S., and Pennebaker, J. W. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.
15. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., Ungar, L. H. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 09 2013.
16. Senter, R. J., Smith, E. A. Automated readability index. Technical report, DTIC Document, 1967.
17. Thelwall, M., Buckley, K., Paltoglou, G. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.
18. Tsoumakas, G., Vlahavas, I. Random k-labelsets: An ensemble method for multi-label classification. In *Machine Learning: ECML 2007*, pages 406–417. Springer, 2007.