Age, Gender and Personality Recognition using Tweets in a Multilingual Setting

Notebook for PAN at CLEF 2015

Mounica Arroju[1], Aftab Hassan[1], Golnoosh Farnadi[2,3]

[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
{mounica,aftabh}@uw.edu
golnoosh.farnadi@ugent.be

Abstract User generated text on social media sites is a rich source of information that can be used to identify different aspects of their authors. Proper mining of this content provides an automatic way of identifying users which is very valuable for applications that rely on personalisation. In this work, we describe the properties of our multilingual software submitted for PAN2015 which recognizes the age, gender and personality traits of Twitter users in four languages, namely, English, Spanish, Dutch and Italian.

Keywords: Gender identification . Age prediction . Big 5 personality prediction. Multilingual prediction

1 Introduction

Author profiling deals with deciphering information about the author from the text that he/she has written. This helps in identifying aspects about the person such as his age, gender and personality traits. The field of author profiling is a problem of growing importance and has several applications in today's world, such as personalized advertising, law enforcement and reputation management among many others.

User generated contents in social media such as Twitter provide a valuable source of information for the task of author profiling. 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 [16], however most of the author profiling models in inferring age and gender worked with English speaking users or by leveraging lengthy texts (i.e., at least 1000 words). Similar works to ours which use shorter texts for age and gender prediction based on tweets are [11] and [15].

Similarly, studies have been published on automatically recognizing the personality traits of users of social media platforms based on their user generated content. Twitter has been used for the task of automatically predicting the personalities of the users, as well as for user behavior analyses [13,6]. For instance, Quercia et al. [13] found that extroverts and emotionally stable people are popular as well as influential users on Twitter. Golbeck et al. [6] used profile information from Tweeter as features when

training the machine learning algorithms to predict scores on each of the five personality traits that were predicted within 11% - 18% of their actual value. To the best of our knowledge, similar to the task of age and gender prediction, most of these studies used English user generated content to extract features from the textual content for inferring users' personality traits, however most of the social media site are multilingual and there is a need to study multilingual models.

In this work, we propose predictive models to identify age, gender and personality traits of the Twitter users using their tweets in a multilingual setting where tweets are in English, Spanish, Italian and Dutch. In the subsequent sections, we explain our approach in more details. Section 2 describes the data set used. In Section 3, we describe the pre-processsing strategies that were performed on the data and then explain the two models created for age and gender identification and for personality prediction. Finally, in Section 4, we present results, summarize them and make conclusions.

2 Problem formulation

This year's PAN author profiling task is based on data taken from Twitter [14]. Table 1 provides details about this dataset's characteristics. The task is to predict a user's age, gender and the big 5 personality traits (openness, conscientiousness, stability, aggression and extroversion), given their tweets. The dataset contains tweets in four languages, namely, English, Spanish, Dutch and Italian. For the tweets in English and Spanish, the task is to predict age, gender and personality scores while for the tweets in Dutch and Italian, since age is not provided, the task is to predict gender and personality scores.

Table 1. Characteristics of the training dataset

Statistic	English	Spanish	Dutch	Italian
# Female users	41	25	2	12
# Male users	38	36	5	12
Avg. tweets per male user	189	201	195	190
Avg. tweets per female user	194	195	202	202
Avg. length of tweets per user	72	73	74	72
Majority age group	18-24	25-34	Unknown	Unknown
Avg. Openness score	0.24	0.18	0.3	0.22
Avg. Conscientiousness score	0.17	0.26	0.12	0.17
Avg. Agreeableness score	0.12	0.11	0.15	0.23
Avg. Emotional Stability score	0.16	0.18	0.27	0.16
Avg. Extroversion score	0.12	0.09	0.32	0.20

Personality is a fundamental differentiating factor of human behavior. Research in the psychology literature has led to a well established model for personality recognition and description, called the *Big Five Personality Model*. In this study we use the big five model to recognize users' personality traits from their tweets. According to the big five model, five traits can be summarized in the following way [3]: *Openness to experience* (Openness) is related to imagination, creativity, curiosity, tolerance, political liberalism,

and appreciation for culture. *Conscientiousness* measures preference for an organized approach to life in contrast to a spontaneous one. *Extroversion* measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions. *Agreeableness* relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. *Emotional Stability* (reversely referred to as Neuroticism) measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression.

As shown in Table 1, the data is split fairly evenly between male and female users. Moreover, except for Spanish tweets, females on average have more tweets compared to males. Among English users, the majority age group of users is 18-24 whereas for Spanish users, it is 25-34. The age group information is not provided for Dutch and Italian tweets. And, across all languages, the average length of a tweet is between 72-74 characters. Interestingly, on average Dutch users have higher scores for *Extroversion*, Spanish users have higher scores of *Conscientiousness*, English speaking users have higher scores of *Openness* and Italian users have higher *Agreeableness* scores.

3 Methodology

In this section, we describe two multilingual predictive models that we use in our submission. We build a multilingual model for identifying age and gender of users and a multilingual model for predicting their personality traits. We report the results from the train set and then the results obtain from the test runs of the task on TIRA [7]. Identifying gender is a binary classification task with Male, Female as class labels. Prediction of age is a multi-class classification with class labels of [18-24], [25-34], [35-49] and [50-xx]. In the case of personality traits, it is a regression task where we have to provide a personality score between -0.5 to 0.5 for each of the five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism).

We compare the performance of our model with a majority baseline in the case of age and gender classification, and a mean value baseline in the case of personality prediction using the train data set. In the majority baseline, the majority class occurring in the data set is assumed to the predicted class for all instances in the test data set. In the mean value baseline, the mean of the personality scores is assumed to be the number conveying the personality trait of for all instances in the test data set.

3.1 Feature Extraction

Author profiling in social media differs from other platforms as the language that has been used by social media users is informal, unstructured and noisy. A common approach of inferring users' attributes in social media is to model the writing habits of users by extracting various features from texts they have posted. However, the primary challenge associated with dealing with textual data and tweets in particular is that people don't always spell words correctly, and sometimes even intentionally spell words incorrectly owing to the informal nature of Twitter. For example, in some tweets, people spell the word 'Friday' as 'Fridayyyy!!' to convey excitement. In such situations, tokenizing and identifying words becomes challenging. Furthermore, social media users use their own vocabulary to express their thoughts or feeling, thus extracting

vocabulary-based or grammar-based features may not work efficiently for these platforms. Furthermore, social media users use multiple languages to express their opinion. This makes it impossible to apply knowledge derived from one language by extracting language dependent features, onto another language. Many dictionary-based features such as emotion categories such as NRC [12] or psycholinguictic database such as MRC [18] are only available in English.

A common feature set that has been shown to be effective in predicting age and gender of authors according to many recent studies on Twitter is n-gram features such as [15,2]. Since in this study, we deal with a multilingual data from Twitter, we use n-gram features for modeling age and gender of users. Similarly, one of the most effective features for the task of personality prediction that has been used in many recent studies in social media such as [5,10] is based on the Linguistic Inquiry and Word Count software (LIWC) [8]. Since LIWC dictionaries are available in multiple languages, in this study for the task of personality prediction we leverage LIWC features.

3.2 Data Pre-processing

The data was given in the form of xml files containing tweets for several users. Prior to any modelling, we apply the following set of preprocessing steps to all documents. We extract all tweets belonging to one user and tokenize the words from the tweets. We count the number of user mentions, links and the hashtags to use them as features for the regression tasks. We apply the tweet tokenizer [9] to tokenize the words and removed the html tags, punctuation, username mentions, hashtags and emoticons. In our models, we represent each user by combination of all tweets belongs to the user and the extracted features from the combined tweets.

3.3 Multilingual Age and gender identification model

After the pre-processsing steps, we extract the word n-gram (uni, bi and trigram) features. For the classification tasks, we use only one set of features since according to the recent literature such as [17] simple word n-grams give better results for various languages. We apply a linear model with stochastic gradient descent (SGD) learning [1], which showed the best results in our initial analysis using the training data. SGD is a machine learning approach that iteratively optimizes the gradient descent and updates the model with each training example. The parameters used for this model are listed in Table 2.

Table 2. Parameters for Age and Gender Multilingual Model

parameter	loss	penalty	alpha	no. of iterations	random state
value	hinge	l2	1e-3	5	42

The results in Table 3 and Table 4 indicate that our simple model using only-ngrams, outperforms the majority baseline in inferring gender and age of twitter users for all four languages.

Table 3. Accuracy of gender classification using 10-fold cross validation for all four languages. Baseline is majority baseline and values outperforming the baseline are shown in bold.

Majority Baseline	Multilingual SGD model			
	English	Spanish	Dutch	Italian
0.50	**0.69**	**0.73**	**0.68**	**0.62**

For the task of gender prediction, we obtain better results for the Spanish users following by the English, Dutch and Italian users (shown in Table 3).

Table 4. Accuracy of age classification using 10-fold cross validation for English and Spanish. Baseline is majority baseline and values outperforming the baseline are shown in bold.

English		Spanish	
Majority Baseline	SGD	Majority Baseline	SGD
0.44	**0.69**	0.46	**0.48**

In inferring users' age, our model performs better for the English tweets compared to the Spanish tweets (shown in Table 4).

3.4 Multilingual Personality prediction model

The words from the tokenized text which contribute to our prediction were matched against dictionaries of the Linguistic Inquiry and Word Count (LIWC) [8]. LIWC is a well-known text analysis technique which is widely used in psychology studies [8]. In this work instead of using LIWC software, we use the LIWC dictionaries which includes categories related to psychological processes (e.g., anger words such as *hate* and *annoyed*), relativity (e.g., verbs in the future tense), personal concerns categories (e.g., occupation such as *job* and *majors*), and linguistic dimensions (e.g., swear words).

LIWC as a text analysis software program extracts different counting features to measure the degree to which people use different categories of words across a wide array of texts, including emails, speeches, poems, or transcribed daily speech. We build a similar LIWC feature extractor using the LIWC dictionaries.

Each word in the LIWC dictionary may belong to more than one category, for example, word "baseball" belongs to categories 'Occup','School','Leisure','Sports' that we present it as a vector such as baseball: [Occup,Occup,Leisure,Sports].

We build LIWC feature vector as follows: we first compute the term frequency-inverse document frequency (TF/IDF) of all tokens in the dataset. Then, we search the corresponding LIWC dictionary based on the given language and make the LIWC vector for each token. Then, by adding the TF-IDF value of the token in each combined tweets text, we creat the corresponding LIWC feature vector to represent each user in our dataset.

According to [4], multivariate regression techniques such as Ensemble of Regressor Chains Corrected (ERCC) model performs well for the task of personality prediction.

Therefore, in our model once we make the feature vector, we use the ERCC Model for personality prediction. ERCC is a multivariate technique that let us leverage the prediction result for one personality trait to make a prediction for another.

The idea of ERCC is chaining single-target regression models. By choosing an order for the target variables (e.g., $O = (t_1 = Openness, t_2 = Conscientiousness, t_3 = Agreeableness, t_4 = EmotionalStability, t_5 = Extroversion)$), the learning model for each target variable t_j relies on the prediction results of all target variables t_i which appear before t_j in the list. For the first target variable, a single-target regression model predicts the value, then the input space for the next target variable is extended with the prediction results of the previous one and so on. Since in this model the order of the chosen chain affects the results, the average prediction result of r different chains (typically $r = 10$) for each target variable is used as the final prediction result. By using k-fold sampling, the prediction results of $\frac{k-1}{k}\%$ of the whole training set are used to expand the input space which increases the reliability of the predictions based on the training set. We choose $k = 10$ randomly selected chains and enumerate the results to report as a final predicted result. To get the results of the ERCC model, we used the implementation of this algorithms in Mulan [1]. The base learner of the algorithm in Mulan is the Weka decision tree algorithm. For further information we refer to [19].

Enhancements made to the LIWC Dictionary file for English To deal with the noisy and informal language of social media users, we manipulate the LIWC features to cover most of the written text from the Twitter users. Note that in our model, we apply this enhancement only for extracting English LIWC features, however it is possible to use the same technique for other languages.

Our dictionary file for English contains 2,320 entries. In order to enhance this word list, we implement a web crawler to find synonyms of each word in this word list and assign it to the same category as the original word. By doing this, we extend the LIWC English word list to 33,166 words and corresponding category information.

Table 5. MAE results of personality prediction using 10-fold cross validation for English Tweets. Baseline is mean baseline and best performing values are shown in bold.

Personality traits	Baseline	ERCC using enhanced LIWC	ERCC using LIWC
Openness	0.0848	**0.0811**	0.1022
Conscientiousness	0.0984	**0.1085**	0.1194
Agreeableness	0.1560	0.1337	0.1188
Emotional Stability	0.1848	**0.1824**	0.2621
Extroversion	0.2091	**0.2013**	0.3066

The web crawler works by looking up these words in the official thesaurus website [2]. It parses the web page and makes a list of all the synonyms of the word. For each of these synonyms, it assigns the same LIWC category as the base word. Upon enhancing

[1] http://mulan.sourceforge.net/

[2] http://www.thesaurus.com/

our word list, we get noticeable improvements in our personality predictions for the English language as shown in Table 5.

Table 6. MAE results of personality prediction using 10-fold cross validation for all four languages. Baseline is mean baseline and values outperforming the baseline are shown in bold.

Personality traits	English		Spanish		Dutch		Italian	
	ERCC	Baseline	ERCC	Baseline	ERCC	Baseline	ERCC	Baseline
Openness	**0.0811**	0.0848	**0.0811**	0.0951	0.0821	0.0676	**0.0821**	0.0978
Conscientiousness	0.1085	0.0984	0.1985	0.0995	**0.1095**	0.1157	**0.1585**	0.1757
Agreeableness	**0.1337**	0.1560	**0.1837**	0.1850	**0.1347**	0.1818	**0.1837**	0.1928
Emotional Stability	**0.1824**	0.1848	**0.1924**	0.1988	0.1894	0.1676	**0.1124**	0.1986
Extroversion	**0.2013**	0.2091	**0.2813**	0.4091	0.2013	0.2004	**0.2213**	0.2224

Table 6 presents the results of applying our multilingual personality prediction model on Twitter dataset. Our model outperforms the mean baseline for all five personality traits for Italian tweets, however for English and Spanish tweets our model outperforms the baseline for all traits except Conscientiousness. For the case of Dutch tweets, our model only outperforms in inferring Conscientiousness and Agreeableness.

4 Discussion and Conclusion

Table 7 presents the results of applying our model on the unseen test dataset reported by the PAN organizers. As seen in Table 7, our model obtains the best results in inferring age, gender and personality traits for the Italian users with an accuracy of 71%. This is inline with our results presented in Table 6 that our model outperforms other languages in inferring personality traits using only the training examples of the Italian tweets. Moreover, the worst performance is reported for the Dutch users, as we expect this behaviour according to the performance of our models based on the training examples. However, it is interesting that our model gets better results in inferring age for the test

Table 7. Comparison results of predicting age, gender and personality traits (*Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Con)*, *Emotional Stability (Ems)*, and *Openness (Open)*.) of English, Dutch, Spanish and Italian Twitter users using our trained model on the unseen test dataset on TIRA.

Age	Gender	Con	Ext	Agr	Openness	Ems	Global	RMSE
English								
0.70	0.77	0.1481	0.1636	0.1513	0.1584	0.2349	0.70	0.1713
Spanish								
0.69	0.75	0.1785	0.1980	0.1727	0.1469	0.2125	0.65	0.1817
Dutch								
-	0.53	0.1553	0.1573	0.1672	0.1103	0.2235	0.68	0.1627
Italian								
-	0.58	0.1345	0.1480	0.1520	0.1620	0.1941	0.71	0.1581

dataset (i.e., 70% and 69% for English and Spanish, respectively) compared to the train dataset (i.e., 69% and 48% for English and Spanish, respectively). Moreover, for the case of predicting the gender of users, our model performs better for predicting the gender of the Italian users in the training examples with 62% accuracy compared to the ones in the test dataset with only 58% accuracy.

In this work we have presented two multilingual models for inferring age, gender and personality traits of Twitter users. We applied our model on four different languages, namely, English, Spanish, Dutch and Italian. We leveraged n-gram features for the case of age and gender identification using stochastic gradient descent learner and LIWC features for the task of personality prediction with a multivariate regression model of Ensemble of Regressor Chains Corrected (ERCC). We obtain an average 68.5% accuracy for identifying users' attributes in four different languages. Aside from the work we have presented in this paper, there is clear potential for more fine grained models for the task of author profiling to deal with multilingual, noisy, short and informal social media user generated text.

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. Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT*, pages 177–187. Springer, 2010.
2. John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.
3. Paul T Costa and Robert R McCrae. The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook Of Personality Theory And Assessment*, 2:179–198, 2008.
4. Golnoosh Farnadi, Shanu Sushmita, Geetha Sitaraman, Nhat Ton, Martine De Cock, and Sergio Davalos. A Multivariate Regression Approach to Personality Impression Recognition of Vloggers. In *Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition (WCPR)*, pages 1–6, 2014.
5. Golnoosh Farnadi, Susana Zoghbi, MarieFrancine Moens, and Martine De Cock. Recognising personality traits using Facebook status updates. In *Proceedings of the Workshop on Computational Personality Recognition (WCPR), AAAI Press*, pages 14–18, 2013.
6. Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *IEEE Third International Conference on Social Computing (SocialCom) on Privacy, Security, Risk and Trust (PASSAT)*, pages 149–156, 2011.
7. Tim Gollub, Benno Stein, and Steven Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM, 2012.
8. Roger J. Booth James W. Pennebaker and Martha E. Francis. *Linguistic Inquiry and Word Count*. Taylor & Francis, 1999.

9. Michel Krieger and David Ahn. Tweetmotif: exploratory search and topic summarization for Twitter. In *Proceedings of the AAAI Conference on Weblogs and Social Media*. Citeseer, 2010.
10. FranÃğois Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–501, 2007.
11. James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 2014.
12. Saif Mohammad, Xiaodan Zhu, and Joel Martin. Semantic role labeling of emotions in tweets. In *Proceedings of the WASSA*, pages 32–41, 2014.
13. Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our Twitter profiles, our selves: Predicting personality with Twitter. In *IEEE Third International Conference on Social Computing (SocialCom) on Privacy, Security, Risk and Trust (PASSAT)*, pages 180–185, 2011.
14. Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2015.
15. Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 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.
16. H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.
17. Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. Overview of the author identification task at pan 2014. *analysis*, 13:31, 2014.
18. Michael Wilson. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10, 1988.
19. Eleftherios Spyromitros Xioufis, William Groves, Grigorios Tsoumakas, and Ioannis P. Vlahavas. Multi-label classification methods for multi-target regression. *CoRR*, 2012.