Author Profiling of Twitter Users
Notebook for PAN at CLEF 2015

Roy Bayot, Teresa Gonçalves, and Paolo Quaresma

Universidade de Évora
roybayot@gmail.com, tcg@uvera.pt, pq@di.uevora.pt

Abstract In this paper, we focused on profiling authors on age, gender, and five personality traits. The corpus consists of anonymized twitter posts categorized into 4 different languages. Our proposed approach was to use a combination of *tfidf*, function words, stylistic features, and text bigrams, and used an SVM for each task.

1 Introduction

Author profiling from text has been an interesting topic recently because of the increase in the availability of texts. This is mostly because of the internet where text is one of the forms of communication. This could be present in blogs, websites, customer reviews, and even twitter posts.

While author anonymity has been present mostly in the web, using profiling can be useful, especially in aspects such as marketing, advertising, as well as security. Profiling mainly uses such text to determine certain aspects of the author such as age, gender, and certain personality traits. The idea is that certain topics or word usage comes are affected by such aspects. For instance, talking about bands or any trending music at the time would be a topic for teenagers. This is not always easy since some people can always think not on their age, and that would affect the writing. Some people can write fiction and it can be that the text was written from the perspective of someone with a different personality type.

However PAN is making an effort in this aspect. In this year's edition of PAN for author profiling, the task is specific to author profiling of twitter users in 4 languages - english, dutch, italian, and spanish. The tasks include profiling for age, gender, and the big five personality traits - agreeability, conscientiousness, extrovertedness, openness, and stability [6].

Some approaches have been used previously that are similar. For instance, in [2], they used 405 function words, a list of ngrams part of speech tag where they used 500 most common ordered triples, 100 common ordered pairs, and all single tags, to categorize text by gender. In [7], both style-based features (POS tags, function words, blog words, and hyperlinks) and content-based features (content words and hand-crafted LIWC) were used to classify by age and gender. In the previous year, PAN also had ran author profiling but on different sources, not just tweets. In [3], the method used to represent terms in a space of profiles and then represent the documents in the space of profiles and subprofiles were built using expectation maximization clustering. In [4],

ngrams were used with stopwords, punctuations, and emoticons retained, and then idf count was also used before being processed with 5 different classifiers. Liblinear logistic regression returned with the best result. In [9], different features were used that were related to length (number of characters, words, sentences), information retrieval (cosine similarity, okapi BM25), and readability (Flesch-Kincaid readability, correctness, style). This was used on 7 different classifiers. Another approach is to use term vector model representation as in [8]. For the work of Marquardt et. al in [5], they used a combination of content-based features (MRC, LIWC, sentiments) and stylistic features (readability, html tags, spelling and grammatical error, emoticons, total number of posts, number of capitalized letters number of capitalized words).

Since this is the first attempt at a submission to PAN, we opted to take a simpler approach of using *tfidf*, function words, some stylistic features, and text bigrams.

2 Methodology

For a first submission to this task, we decided to use the same approach for all the tasks. The method we used is more or less straightforward - basic feature extraction, concatenating the different features, then use the combined features for classification or regression, and use 10 fold cross validation.

2.1 Features Vector Creation

There are four main feature types used in this submission and each processed separately. The first would be the *tfidf* features. Term frequency-inverse document frequency or *tfidf* is one of the most common features obtained.

Before running the feature extraction for *tfidf*, preprocessing was done to the tweets obtained. For this task, all tweets from a single person were concatenated. Numbers were removed, and turned into lower case equivalents. Then stopwords from the NLTK toolkit [1] were removed from the set of words. Finally, the resulting words were used to find at *tfidf* vector representation through the scikits-learn python library. The vector was set to 10000 and discard the excess based on the document frequency. The defaults were chosen for the vectorizer. It should also be noted some of the tfidf representations did not maximum of 10000 in terms of dimensions.

The second would be the stylistic features. We only detected for the presence of absence of certain characters or combination of characters. This includes the following characters and combinations - "#", "@username", "http://", ":)", ";)", "o_O", "!","!!", "!!!", ":(". This is by no means exhaustive and was just an initial set. The octothorpe was to indicate if there was hashtag. The "@username" was used in case the user tags other twitter users. Normally, this will be of a twitter handle but since it was anonymized, we used this instead. The set ":)", ";)", "o_O", and ":(" just check of some sort of emotion. And finally, the exclamation points could indicate possible surprise intensity of a statement, which usually happens in the internet.

The third would detecting for function words. Function words are informative words that could be used to discriminate between classes. These were obtained by using all instances in the training data and was used to create a decision tree. And the most

informative features were obtained with entropy as the criteria. The succeeding tables at show the words/characters that obtained as function words.

	age
dutch	"zit","heel","best","geeft","idee","nooit","weer","binnen","goed","avond", "bijwerken","dag","laatste","man","voelt","hart","toekomst","boeit","dh", "feestje","ging","meisje","morgen","muzikanten","onderweg", "onderzoeksjournalistiek","onzin","proficiat","ten","verdient","verzuurde","werkt"
english	"co","wanna","us","haha","username","fitbit","et","bowl","academia","bitch", "happened","even","year","reach","free","times","speech","top","add","social", "think","nothing","financial","pop","inspiring","lil","complicated","aa"
italian	"domani","fa","poi","pezzo","immagini","quel","ultimo","binari","bravo", "foto","is","sentito","stato","pi","seguire","borgo","elected","federico", "riusciamo","super","tassoni","agendadigitale","casalinga","cc","de","dio", "eccomi","esempio","novit","oscena","pard","piazza","preso","pu","rispetto","yg"
spanish	"http","ma","dijo","momento","cil","as","buenos","mala","bieber", "falta","buscan","facebook","info","todas","favor","cula","nom","ofpbmahc"

Table 1. Function words for age task.

	gender
dutch	"username","goed","bent","saai"
english	"close","love","mention","co", "wife","lanka","believe","video", "cute","phone","le","day","urban", "round","thank","bird","wouldn","aa"
italian	"co","campagna","ottimo", "conoscessi","voci"
spanish	"vida","alguien","corrupci", "ciudades","si","temprano", "puro","meta","foto","dio"

Table 2. Function words for gender task.

For the personality tasks, the decision tree was made in such a way that the output was framed as a classification problem. Instead of having continuous numbers from -0.5 to 0.5, we used discrete numbers from -0.5 to 0.5 with an interval of 0.1. The words for personality tasks were shown in the tables 3-7.

Finally, we also add text bigrams. This was to possibly capture some structure in the input texts.

	extroverted
dutch	"dingen","blijft","bijna","mr","zeker","vallen","doet","xkwktrd","zoek"
english	"co","username","million","liked","facebook","last","better","de","music", "around","let","book","happy","friends","used","inside","really","di","work", "google","opinion","phd","racist","things","forget","via","need","nice","http", "application","slides","sign","sun","sell","years","latest","starbucks","jullie", "interesante","minute","screen","model","shirt","ziglar"
italian	"design","hotel","ore","dopo","oppure","ariosto","scaccia","son","date"
spanish	"xico","alguien","escribir","tambi","nueva","pe","gusto","http","comen", "mujeres","fico","toda","quiero","sue","aunque","ahora","chistes","mano", "ser","luz","verdad","dar","hoy","cticas","che","suicidio","portugal", "recuerdo","responsabilidad"

Table 3. Function words for extoverted task.

	stable
dutch	"username","snel","misschien","ergens","blijft","namelijk", "jaar","vrijdag","terwijl","hashtag","interviewee"
english	"like","re","god","computer","cause","android","follow","waiting", "well","school","ever","rock","part","photo","want","years","mind", "need","bring","original","says","back","colleagues","last","finally", "bu","according","experience","work","real","sour","sometimes", "many","savigny","play","st","silly","similar","birthday","dz", "holds","today","gerrard","middle","song","ve"
italian	"co","design","sostenibile","andare","me","esempio","at","buone", "semplicissima","incapace","tv"
spanish	"amigos","is","quiero","ja","despertar","noches","buenos","ah","mayor", "quieres","bado","iphone","est","culo","sesi","cient","pel","you","sab","internet", "torno","tardando","podemos","tampoco", "nnjutigybf","corriendo","va", "acompa","hacer","papaya","vas","bonitas"

Table 4. Function words for stable task.

	agreeable
dutch	"rt","terug","snel","bedankt","smh","terwijl","the","heerlijk","hallo"
english	"https","birthday","made","google","important","need","church","oh", "haha","early","hearts","personal","one","eat","girl","go","mo","ly", "facebook","amazing","keeping","speak","iv","secret","room","fate", "sit","married","background","sharedleadership","ward","anyone", "dream","succes","needs","views","annoyed","habit","walk"
italian	"bologna","via","twitter","style","co","sento","monti","disegni"
spanish	"sabes","cc","dif","quedan","username","despedida","estudiar", "vez","pesar","vamos","esperar","tambi","solo","sociales", "hacen","luego","ngelamaria","fin","acordaba","terror","ja", "bellas","firmad","fr"

Table 5. Function words for agreeable task.

	open
dutch	"hahaha","week","tijd","username","we","kaviaarbehandeling","jeeeej","can"
english	"love","time","years","http","goes","dreams","birthday","high","win","world", "wanna","digital","replies","would","women","ready","get","wall","point", "lot","project","mean","meet","right","people","page","season","bit","fall", "qenbj","er","looks","year","go",want","midnight","username","attention", "cold","like","little","psd"
italian	"qualcosa","anni","bel","ricerca","sangue","zagaria","sento","striati"
spanish	"puta","jajaja","interesante","luego","espa","esperar","dia","acuerdo", "grande","ma","amigo","siempre","sonrisa","haber","pista","buenos", "penlties","aburrida","burra","venes","pelotita","crisis","youtube", "social","hombres","plana","serie"

Table 6. Function words for open task.

	conscientious
dutch	"mag","fietsen","mn","dacht","zet","moddermanstraat"
english	"awesome","party","maybe","crazy","ff","using","thanks", "little","new","could","tears","long","thirty","saying","system", "find","wtf","one","someone","reason","john","lasting","re", "five","reat","http","via","thrones","words","furious","sjgy", "bout","thank","mini","qw","central","looks","playing"
italian	"design","ore","username","anni","sembra", "oppure","massimo","purtroppo","confermo"
spanish	"siempre","fer","cc","rtela","tico","corrupci","solo", "momento","mundo","mal","empleo","do","pone","va", "transici","veces","pa","escuchar","mayor","meses", "puede","ciento","andar","article","gt","moralmente", "preguntar","online"

Table 7. Function words for conscientious task.

2.2 Training and Testing

After features were extracted and concatenated, we used a linear SVM with a default relaxation parameter of 1. We used the scikits-learn library for this and used the SVM as an initial check for results.

3 Experiments and Results

3.1 Setups

Each of the different features were also individually used to classify or perform a regression. Some combinations of the features were also used. In tables 8-11, different tasks were done with *tfidf*, function words (FW), stylistic features(SF), and text bigrams(TB), as well as combinations of these.

	tfidf	FW	SF	TB	FW+TB	tfidf+FW+SF	tfidf+FW+SF+TB
gender	0.625	0.704	0.644	0.744	0.618	0.630	0.730
age	0.632	0.651	0.511	0.612	0.650	0.538	0.677
extroverted	-0.045	-0.026	-0.035	-0.045	-0.026	-0.058	-0.025
stable	-0.060	-0.048	-0.063	-0.055	-0.046	-0.067	-0.040
agreeable	-0.035	-0.030	-0.034	-0.042	-0.028	-0.047	-0.031
open	-0.033	-0.029	-0.027	-0.029	-0.023	-0.039	-0.024
conscientious	-0.027	-0.023	-0.023	-0.029	-0.020	-0.030	-0.020

Table 8. English

	tfidf	FW	SF	TB	FW+TB	tfidf+FW+SF	tfidf+FW+SF+TB
gender	0.683	0.792	0.525	0.708	0.658	0.417	0.742
extroverted	-0.027	-0.016	-0.025	-0.015	-0.015	-0.022	-0.015
stable	-0.029	-0.034	-0.040	-0.022	-0.033	-0.046	-0.027
agreeable	-0.030	-0.037	-0.050	-0.020	-0.031	-0.039	-0.025
open	-0.023	-0.025	-0.029	-0.013	-0.015	-0.019	-0.009
conscientious	-0.014	-0.009	-0.016	-0.013	-0.012	-0.014	-0.009

Table 9. Dutch

3.2 Results from PAN

The results from PAN are summarized in the table below. The results were not as satisfactory as we had hoped.

	tfidf	FW	SF	TB	FW+TB	tfidf+FW+SF	tfidf+FW+SF+TB
gender	0.383	0.750	0.542	0.650	0.717	0.525	0.675
extroverted	-0.032	-0.028	-0.049	-0.029	-0.020	-0.031	-0.020
stable	-0.042	-0.041	-0.243	-0.049	-0.029	-0.045	-0.035
agreeable	-0.045	-0.043	-0.037	-0.045	-0.026	-0.019	-0.025
open	-0.011	-0.036	-0.061	-0.022	-0.015	-0.018	-0.016
conscientious	-0.024	-0.033	-0.056	-0.034	-0.023	-0.032	-0.021

Table 10. Italian

	tfidf	FW	SF	TB	FW+TB	tfidf+FW+SF	tfidf+FW+SF+TB
gender	0.690	0.560	0.570	0.640	0.560	0.620	0.650
age	0.520	0.580	0.390	0.540	0.610	0.470	0.630
extroverted	-0.039	-0.030	-0.044	-0.038	-0.026	-0.035	-0.027
stable	-0.065	-0.046	-0.060	-0.058	-0.040	-0.055	-0.038
agreeable	-0.025	-0.033	-0.049	-0.024	-0.027	-0.042	-0.027
open	-0.035	-0.036	-0.049	-0.038	-0.030	-0.034	-0.030
conscientious	-0.044	-0.034	-0.046	-0.038	-0.027	-0.031	-0.028

Table 11. Spanish

	global	rmse	age	agreeable	conscientious	extroverted	gender	open	stable	runtime
Dutch	0.6881	0.1863		0.1631	0.1978	0.1705	0.5625	0.1969	0.2031	00:00:37
English	0.5253	0.1958	0.5915	0.1634	0.1866	0.2137	0.5000	0.1844	0.2308	00:05:57
Italian	0.6644	0.1989		0.1820	0.2173	0.1928	0.5278	0.1676	0.2349	00:00:50
Spanish	0.5932	0.1773	0.5682	0.1593	0.1852	0.1853	0.6136	0.1540	0.2025	00:03:03

Table 12. Results from PAN

4 Conclusion and Recommendations

As a conclusion, much improvement still needs to be done for such tasks. For instance exploration of more features such as stylistic features. Other classifiers are also to be explored as well as parameter tuning. Possibly one mistake this year is to just get the combination that yields more better result over all than picking and choosing certain models to certain languages and tasks. It would have been better if the system was adapted to that.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
2. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
3. López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L.: Using intra-profile information for author profiling
4. Maharjan, S., Shrestha, P., Solorio, T.: A simple approach to author profiling in mapreduce
5. Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.F., Davalos, S., Teredesai, A., De Cock, M.: Age and gender identification in social media. Proceedings of CLEF 2014 Evaluation Labs (2014)
6. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: 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 (Sep 2015), http://www.clef-initiative.eu/publication/working-notes
7. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
8. Villena-Román, J., González-Cristóbal, J.C.: Daedalus at pan 2014: Guessing tweet author's gender and age
9. Weren, E.R., Moreira, V.P., de Oliveira, J.P.: Exploring information retrieval features for author profiling—notebook for pan at clef 2014. Cappellato et al.[6]