Segmenting Target Audiences: Automatic Author Profiling Using Tweets.

Notebook for PAN at CLEF 2015

Maite Giménez, Delia Irazú Hernández, and Ferran Pla

Univesitat Politècnica de València
{mgimenez, dhernandez1, fpla}@dsic.upv.es

Abstract This paper describes a methodology proposed for author profiling using natural language processing and machine learning techniques. We used lexical information in the learning process. For those languages without lexicons, we automatically translated them, in order to be able to use this information. Finally, we will discuss how we applied this methodology to the 3rd Author Profiling Task at PAN 2015 and we will present the results we obtained.

1 Introduction

The exponential growth of social networks has led to new challenges in the study of Natural Language Processing (NLP). In literature, we could find extensive work done in order to understand normative texts. Social profiling is a less explored topic, even though its study is relevant also to other sciences as: marketing, sociology, etc. [3,1,8]

This paper explores how to define user profiles using classic techniques of NLP. Corpora have been created compiling tweets in different languages. Twitter [15] is a microblogging service which, according to latest statistics, has 284 million active users, 77% outside the US that generate 500 million tweets a day in 35 different languages. That means 5.700 tweets per second and they had peaks of activity of 43.000 per second. This numbers justify the great interest in the automatic processing of this information.

1.1 Task Description

This task will address author profiling. Unlike user identification, author profiling does not try to identify author's identity. Author profiling tries to determine author's features as demographic features or personality traits.

In the literature this problem has been addressed with medium or long texts. In a speech it is more likely to find likely to find statistically significant features which identify the author. However, we worked with short text from *Twitter*. Statistical methods used require huge amount of data to properly train the models. Therefore, convergence is a problem in systems trained with short texts.

Author profiling competition was proposed by PAN 2015. A detailed explanation could be found in the overview paper of the task [11]. We have tackled this task using NLP techniques and machine learning (ML).

The remainder of this paper is organized as follows. Section 2 covers briefly the state of the art, section 3 describes the corpus, section 4 presents in detail the methodology we used and section 5 presents the experiments we have developed. Sections 6 and 7 discusse our results and the future work in order to improve them.

2 State of the Art

Author profiling task is a research area for disciplines such as: linguistics, psychology or marketing.
Task complexity made it unfeasible. However, since 2000 technology begun to be mature enough to tackle this task. Early works [4,14] only studied gender and age. Lately, new psychological features had been tackle [2]. Pennebaker *et al* work [10] linked the language with author's psychological features .

Since 2013 author profiling contest is held by the PAN. Participants of previous editions [13,12] used stylistic features, like: term frequency, POS, stop words, and content features, such as: n-grams, sets of words, lists of entities. They used those features to train systems based on support vector machines (SVM), decision trees, Naïve Bayes, etc. If we analyze the accuracy obtained in previous years we will notice how relevant is the nature of texts of the corpus. They achieved around 40 % accuracy predicting gender and age using data from Twitter, however accuracy falls to 25% using hotel reviews.

In this edition of the PAN, task has been extended. Participants should identify age, gender and personality traits as we described in section 1.1.

3 Corpora Description

We start our task studying the corpora. This will allow us to select the best methodology for the task.
Multilingual corpora were provided by task organizers. Corpora contain 14166 tweets from 152 English authors, 9879 tweets from 100 Spanish authors, 3687 tweets from 38 Italian authors and 3350 tweets from 34 Dutch authors.

Tweets were balanced by gender and unbalanced by age. There were much more tweets from users whose age range between 25-34. Nevertheless, according to Twitter's statistics, it is a safety guess to assume that age distribution is representative of the reality.

Then, we studied the vocabulary of each language. We removed punctuation signs and stop words to perform this study. We tokenized words in order to obtain the vocabulary. Consistently, most frequent words were words used in Twitter such as: RT, HTTP,

username, via and abbreviations. We followed our work, studying vocabulary distribution between age and gender for every language. Table 1 shows the most frequent words set for gender and age both for English and Spanish languages.

English	18-24	username, HTTP, m, like, know, love, want, get, RT, 3, one, people, time.
	25-34	HTTP, username, via, m, w, NowPlaying, like, others, 2, Photo, new, pic.
	35-49	HTTP, username, via, new, Data, RT, New, Big, Life, m, data, Facebook.
	50-XX	username, HTTP, RT, via, know, 2, like, m, good, day, love, 3, time, new.
Spanish	18-24	username, HTTP, si, día, quiero, ser, 3, mejor, bien, vida, hoy, voy, ver.
	25-34	username, HTTP, q, si, vía, RT, d, Gracias, ser, ver, bien, día, va, hacer.
	35-49	username, HTTP, si, q, ví, RT, México, ser, hoy, Si, d, jajaja, Gracias, 1.
	50-XX	username, HTTP, q, RT, si, i, els, l, 2, 0, 1, Mas, d, amb, és, tasa, per. 2

English	Female	username, HTTP, via, m, like, love, know, RT, 3, get, want, one.
	Male	username, HTTP, m, via, like, RT, 2, new, w, NowPlaying, know.
Spanish	Female	username, HTTP, q, si, vía, ser, d, RT, vida, Gracias, ver, mejor, día.
	Male	HTTP, si, RT, ser, ver, q, d, hoy, d??a, xD, 1 va, bien.

Table 1. Most frecuent words set in corpora.

Finally, we studied hashtags. Hashtags are relevant in Tweeter, becasuse it is how users self annotate their tweets. We found out that 37.9 % of English tweets, 26.7 % of Spanish tweets, 59.9 % of Italian tweets and, 27.3 % of Dutch tweets have hashtags. It is interesting to highlight that English words are present in others corpora, due to the massive use of English in social media.

4 Methodology Description

Based on the briefly analysis presented in Section 3, we decided to apply machine learning algorithms in order to identify personality traits. We employed the Scikit-learn toolkit [9] in our analysis and experimental settings. In order to perform training process in our approach, we developed a novel function in the aforementioned toolkit (we consider this as one of our main contributions). This new function allows training a machine learning algorithm using both word lexicons and stylistic features. Furthermore, we automatically translated some lexicons originally developed for English to Spanish, Italian and Dutch. In our model we considered three subsets of features:

- Textual features. This set relies only on textual content (a lower casting process had been carried out). We took into account four configurations using different n-grams sizes: 1-3, 1-4, 1-6, 3-6 and 3-9
  - TF-IDF coefficients
  - Inter-word chars with TF-IDF coefficients

    - Intra-word chars with TF-IDF coefficients
    - Bag of words
- Stylistic features.
    - Frequency of words with repeated chars.
    - Frequency of uppercase words
    - Frequency of hashtags, mentions, URL and RT.
- Lexicon-based features. Using four different lexicons, we calculated a score for each one, by using the formula $\frac{1}{|W|}\sum_{w \in W} lexicon(w)$. In order to extract this information we removed the stop words.
    - Afinn [5]. This resource consists of a list of words with polarity values between the range -5 and +5.
    - *NRC* [7] . It is a polarity dictionary that gives us a real value that represents the polarity value for a word.
    - *NRC* hashtags. It consists of a list of positive and negative hashtags. We normalized the polarity values in this dictionary considering as a positive value +5 and as negative value -5.
    - *Jeffrey* [6]. This resource contains two different lists of words: positives and negatives. We computed two scores from this resource (positive and negative).

As we mentioned above, we decided to consider a machine learning experimental setting. We carried out different classification tasks, one for determining the gender of the author, a second for age's identification and for each one of the personality traits we applied a binary classification. At the end, our experiments consider seven different classifications tasks. We tested the following classification algorithms:

- Linear Support Vector Machine (all implementations in the toolkit were applied)
- Polynomial Kernel Support Vector Machine
- Naïve Bayes
- Descendent gradient
- Logistic Regression
- Random Forest

5 Experimental Work

We considered two appoaches to train our system. The first one joins all tweets for each user, therefore we will have a sample for each user. The second one uses each tweet as training sample. This last approach will reduce spatial sparsity.

As first step, we performed a preliminary experimental setting that considers the whole set of features and all the classifiers mentioned above. The well-known 10-fold cross validation was applied over the corpus. As evaluation measure the precision was chosen. These experiments allow us to compare the performance of our model using different configurations. For gender and age identification SVM was chosen, while linear regression was selected for dealing with personality traits. As a second experimental setting, the best ranked models were grouped in order to carry out a parameter adjustment. The features considered are: textual, stylistic and lexical based features.

6 Results

Table 2 shows the results in terms of accuracy obtained. First column shows the accuracy after tunning our system in development doing a ten fold evaluation, meanwhile second column shows the results we got testing our system against PAN test set.

		Accuracy	
		Dev.	Test
English	Gender	53.49 %	63.38%
	Age	55.29 %	59.86%
	Agreeable	23.7 %	17.54%
	Conscientious	20.8 %	18.19%
	Extroverted	20.85 %	17.70%
	Open	24.78 %	20.73%
	Stable	17.81 %	27.81%

		Accuracy	
		Dev.	Test
Spanish	Gender	56.9 %	62.5 %
	Age	46.58 %	56.82 %
	Agreeable	40.44 %	17.29%
	Conscientious	32.84 %	18.53 %
	Extroverted	36.98%	20.97%
	Open	39.55 %	16.17%
	Stable	29.05 %	24.40%

		Accuracy	
		Dev.	Test
Italian	Gender	61.63 %	69.44%
	Agreeable	43.28 %	16.24%
	Conscientious	52.67 %	12.47%
	Extroverted	45.65%	13.94%
	Open	42.20 %	20.21%
	Stable	46.15 %	25.33%

		Accuracy	
		Dev.	Test
Dutch	Gender	57.49%	71.88%
	Agreeable	42.33 %	17.05%
	Conscientious	49.82 %	13.92%
	Extroverted	46.37%	18.29%
	Open	43.69 %	1323%
	Stable	38%	17.85%

Table 2. Results obtained during development time and againts PAN's test.

Overall we obtained 0.6857 accuracy, achieving the 13th position over over 22 participants.

7 Conclusions and future work

In this paper we presented our partitipation in PAN author profiling competition. We used Natural Language Processing techniques to solve this task. We could find that accuracy obtained for personality traits is still low. User profiling is a hard task, especially when we are dealing with fine grained traits.

Our system performed acceptably for all languages and demographic traits studied. Poor gender identification has penalized our global results. Our results in development were over fitted when we adjust the parameters of our system. However, a strength of our system it is how it can be applied automatically adapted to new languages.

In the future, there are issues we should tackle such as how to deal with big data and real time. Twitter users generates huge amount of data and if we are able to process

it in real time our systems will improve its accuracy and it could have a huge impact in other areas as marketing. Moreover we plan to deal with slang which it is very present in social media and it has a deep impact in NLP tools as lexicons and part of speech taggers.

Finally, we will like to try new distributed representation of the data and new stylistic features. Distributed representation will reduce the spatial complexity which will reduce training time, and hopefully, it will improve the accuracy of our system.

Acknowledgments

This work has been partially funded by the projects, DIANA: DIscourse ANAlysis for knowledge understanding (MEC TIN2012-38603-C02-01) and ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (MEC TIN2014-54288-C4-3-R).

References

1. Alarcón-del Amo, M.d.C., Lorenzo-Romero, C., Gómez-Borja, M.Á.: Classifying and profiling social networking site users: A latent segmentation approach. Cyberpsychology, behavior, and social networking 14(9), 547–553 (2011)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
3. Boe, B.J., Hamrick, J.M., Aarant, M.L.: System and method for profiling customers for targeted marketing (2001), uS Patent 6,236,975
4. Corney, M., de Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: Computer Security Applications Conference, 2002. Proceedings. 18th Annual. pp. 282–289. IEEE (2002)
5. Hansen, L.K., Arvidsson, A., Nielsen, F.Å., Colleoni, E., Etter, M.: Good friends, bad news-affect and virality in twitter. In: Future information technology, pp. 34–43. Springer (2011)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM (2004)
7. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013). Atlanta, Georgia, USA (June 2013)
8. Orebaugh, A., Allnutt, J.: Classification of instant messaging communications for forensics analysis. Social Networks pp. 22–28 (2009)
9. 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)
10. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology 54(1), 547–577 (2003)
11. Rangel, 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 (2015)

12. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes) (2014)
13. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
14. 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)
15. Twitter: About twitter,inc. https://about.twitter.com/company (2014), accessed: 30-12-2014