SU@PAN'2015: Experiments in Author Profiling

Yasen Kiprov[1], Momchil Hardalov[2], Preslav Nakov[3], and Ivan Koychev[4]

[1] Sofia University "St. Kliment Ohridski", Bulgaria
yasen.kiprov@gmail.com
[2] Sofia University "St. Kliment Ohridski", Bulgaria
momchil.hardalov@gmail.com
[3] Qatar Computing Research Institute, HBKU
pnakov@qf.org.qa
[4] Sofia University "St. Kliment Ohridski", Bulgaria
koychev@fmi.uni-sofia.bg

Abstract. We describe the submission of the Sofia University team for the Author Profiling Task, part of the PAN 2015 Challenge. Given a set of writing samples by the same person, the task asks to predict some demographical information such as age and gender, as well as the personality type of that person. We experimented with SVM classifiers using variety of features extracted from publicly available resources, achieving the second-best score for Spanish out of 21 submissions, and the sixth-best for English out of 22 submissions.

Keywords: author profiling, text mining, machine learning, PAN 2015.

1 Introduction

Social media applications such as Facebook and Twitter have hundreds of millions of users who share information at an unprecedented scale. Naturally, this has attracted business and research interest from various fields including marketing, political science, and social studies, among others. Initially, the primary research and practical interest was in the opinion and the sentiment users express, e.g., towards products such as *iPhone6*, policies such as *ObamaCare*, and events such as *Pope's visit to Palestine*.

Soon, companies realized that knowing the overall public opinion towards their products was not enough; for marketing purposes, it was also important to be able to break this opinion by demographic factors such as gender and age. Similarly, public policy and security experts wanted to further know the native language and some personality traits of the users expressing particular opinions. The personality traits were also important for human resources experts when considering to hire somebody in a team: they wanted to make sure the person's personality would fit well in the target team.

While some demographic information was sometimes directly extractable from the public user profiles, there was no guarantee it was correct and up-to-date, which motivated research in trying to predict it automatically from the text of the messages posted by a given user.

Early work on detecting sentiment focused on newswire text [3, 13, 21, 30]. As subsequently research turned towards social media, it became clear that this presented a number of new challenges. Misspellings, poor grammatical structure, emoticons, acronyms, and slang were common in these new media, and were explored by a number of researchers [4, 5, 11, 14, 15, 19, 20]. Later, specialized shared tasks emerged, e.g., at SemEval, the International Workshop on Semantic Evaluation, [18, 27, 28], which ran in 2013–2015 and compared systems by participating teams against each other in a controlled environment using the same training and testing datasets.

A similar research trend followed with respect to author profiling. While there were several publications that were trying to predict some demographical information such as gender, age, and native language [2, 23], as well as the personality type, and to perform author profiling in general [1], the real push forward was enabled by specialized competitions such as the PAN shared tasks on author profiling [22, 26], which ran in 2013–2015.

Below we discuss the participation of the Sofia University team, registered as *kiprov15*, in the 2015 edition of the task, which ran as part of the PAN 2015 [25],[5] the 13th evaluation lab on uncovering plagiarism, authorship, and social software misuse. The task focused on predicting an author's demographics (age and gender) and the big five personality traits [9] (agreeable, conscientious, extroverted, open, stable) from the text of a set of tweets by the same target author.

The task was offered in English, Spanish, Dutch and Italian, but we participated with a system for the first two languages only. We experimented with SVM classifiers using variety of features extracted from publicly available resources, achieving the second-best score for Spanish out of 21 submissions, and the sixth-best for English out of 22 submissions.

We built our system from scratch, as a M.Sc. class project; our research, the lessons we learned and some observations about the topic and potentially relevant references, datasets, and resources are presented in the next sections. We should note that we have saved a lot of time by reusing the GATE infrastructure and by performing feature extraction on top of it. We focused most of our efforts on feature engineering: we implemented some previously-proposed features, and we further analyzed the training data in an attempt to design some new ones.

The remainder of this paper is organized as follows: Section 2 gives an overview of our approach, including a description of the preprocessing, the features, and the learning algorithm we used. Section 3 presents our experimental setup and the official results our system achieved. Section 4 discusses our results and provides some deeper analysis. Finally, Section 5 concludes and points to possible directions for future work.

[5] http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html

2 Method

We use the GATE framework [10] with various plugins to extract features from the input documents. We then use these features in an SVM classifier, which is implemented as a GATE plugin and uses the LibSVM library [8]. The whole system is wrapped in a Java console application and can be executed from the command line.

2.1 Preprocessing

We integrated a pipeline of various resources for text analysis that are already available in GATE such as a Twitter-specific tokenizer [6], a regular expression-based sentence splitter, a language identifier inplemented as a GATE plugin based on TextCat [7], the OpenNLP[6] POS tagger, and a Twitter POS Tagger [12] for English. We further implemented custom text processing components in order to handle emotions, elongated words, and punctuation. Finally, we integrated various dictionaries and lexicons to detect sentiment polarity, emotions, profanity, and personality-specific phrases in the tweets. We processed the documents using a pipeline with the following components:

1. Twitter tokenizer
2. RegEx sentence splitter
3. Language identifier
4. Language-specific feature extractors (POS tags)
5. Gazetteer lookups
6. Rule-based feature extractors
7. Classifiers

2.2 Features

Lexicon Features. For English, we used several lexicons, both manually-crafted and automatically generated:

- **NRC Hashtag Emotion Lexicon** [16]: 16,862 terms with emotion;
- **Bad words lexicon**: a combination of Google's "what do you love" profanity dictionary[7] and a manually assembled dictionary, with 874 terms in total;
- **World Well-Being Project Personality Lexicon** [29]: top and bottom 100 words for each of the five traits with their score, a total of 1000 words.

For each lexicon, we found in the tweets the terms that were listed in it, and then we added as features the total terms count, normalized by the total number of tweets. We instantiated separate features for the different lexicons.

[6] https://opennlp.apache.org/
[7] http://www.wdyl.com/

For the Hashtag Emotion Lexicon, we also calculated the total score of all matching terms for each emotion type: anticipation, fear, anger, trust, surprise, sadness, joy, and disgust.

For the World Well-Being Project Personality Lexicon, we calculated a total of 30 features, i.e., the following six features for each of the big five types (extroverted, stable, agreeable, conscientious, and open):

- Total positive terms score and count;
- Total negative terms score and count;
- Total terms score and count.

Twitter-specific Features. We used the following Twitter-specific features:

- **Letter case:** the number of lower-case, all-caps, and mix-case words;
- **Hashtags:** the number of hashtags;
- **URLs:** the number of URLs posted;
- **Retweets:** the number of retweets;
- **User mentions:** the number of mentions of users using the pattern `@username`;
- **User mentions start:** the number of tweets starting with a user mention;
- **Picture share:** the number of shared pictures using the pattern `[pic]`. This feature was eventually dropped because there were too few users sharing this type of content, and thus using it did not yield improvements.

All of the above counts are normalized by the total number of available tweets for the target user; so they could be viewed as "average number per tweet".

Orthographic Features. We used the following orthographic features:

- **Elonged words:** the number of words with a sequence of more than two identical characters;
- **Average sentence length:** the average length of a sentence.

Term-level Features. We used the following term-level features:

- n**-grams:** presence and count of unigrams and bigrams. This feature helps to find similar users based on their vocabulary overlap. For tokenizing the text, we used the GATE Twitter-specific tokenizer, which is aware of URLs, emoticons, Twitter tags, etc.
- **Vocabulary size:** the number of different words used by a user in all his tweets.
- **POS tagging:** We used a specialized POS tagger for tweets in English, TwitIE, which is available in GATE [10]; it uses the Penn Treebank tagset, but is optimized for tweets. For Spanish, we used OpenNLP, with pre-trained models again in the Penn Treebank tagset, but the results were not so accurate because of the tweet specifics. Using these toolkits, we performed POS tagging for both languages, and we extracted all POS tag types used in the tweet together with their frequencies as features.

2.3 Classification/Regression

We used the above features and support vector machines (SVM) as implemented in LibSVM. We trained a separate classifier for each language and for each target category: one for age, one for gender. As the five personality traits (extroverted, stable, agreeable, conscientious, and open) were real values, we predicted them using support vector regression. For both classification and regression, we used the features as they were generated in the feature-extraction phase without any further scaling or normalization, as most of them were in reasonable ranges, e.g., because they were already normalized by the number of tweets.

Since the challenge contained documents from different languages and topics, we aimed to avoid as much as possible the use of language-specific markers. Note that most of our features are token-based.

For training, we used linear kernels, which are known to be sufficient for text classification: as we had a very large number of unigrams and bigrams, there was no need to use a kernel in order to make the two classes linearly separable.

3 Experiments and Evaluation

3.1 Experimental Setup

During development, we trained on training datasets provided by the organizers:

- pan15-author-profiling-training-dataset-english-2015-03-02
- pan15-author-profiling-training-dataset-spanish-2015-03-02

We then tested on the following datasets:

- pan15-author-profiling-test-dataset2-english-2015-04-23
- pan15-author-profiling-test-dataset2-spanish-2015-04-23

While developing the system, we randomly selected 15 percent of our training data to try out new features and to measure progress. For the official submission, we trained our model on all provided training data, and we tested on the test datasets.

3.2 Official Results

We were ranked 6th out of 22 submissions for English, and 2nd out of 21 submissions for Spanish. A summary of our official results is shown in Table 1, where we have included our GLOBAL and RMSE scores, as well as our ranking among the other participants in PAN-2015 Author profiling task. Table 2 shows our official scores for each of the individual categories: age, gender and each of the five profile traits.

Although we also experimented with Dutch and Italian, we did not have enough time to fine-tune our systems for them, and thus we eventually decided not to submit results for them.

Language	GLOBAL	RMSE	Ranking
English	0.7211	0.1493	6
Spanish	0.7889	0.1495	2

Table 1. Summary of our official results for English and Spanish tweets.

Language	Age	Agreeable	Both	Consc.	Extrov.	Gender	Open	Stable
English	0.7254	0.1411	0.5915	0.1318	0.1416	0.8451	0.1198	0.2123
Spanish	0.7841	0.1249	0.7273	0.1386	0.1625	0.9091	0.1334	0.1884

Table 2. In detail look at our official results for English and Spanish tweets.

4 Discussion

In our experiments above, we tried to solve the author profiling task as a standard n-gram based classification/regression problem, treating the target classes as nominal (age, gender) or as real-valued (the big five traits).

We started with the same types of features for both language, only adding dictionaries for English. Thus, our general method is in principle language-independent and applicable to any language for which data is available.

We further did a lot of extra feature engineering and selection for English, but we did not have time to do anything special in that respect for Spanish. Yet, we performed better for Spanish, which was a surprise for us.

Over the development of our final model, we tested various combinations of features, and we eventually excluded groups of features that did not actually improve the performance. One such feature group were n-grams with length more than 2; our observation is that 3-grams and 4-grams, although helpful for sentiment analysis [17] on corpora with comparable size, did not improve the performance for author profiling. It looks like the tweet discourse is less important than the actual vocabulary.

Most of the orthographic features and dictionaries did not improve significantly the big five score, although improving the age and gender accuracy. Removing all dictionaries drops the average age and gender F_1 score by 2.5 percent while not changing any of the big five RMSEs by more than 0.005.

We further tested the assumption that advertising and posting titles and news articles introduces noise regarding personality detection by removing all upper-case words from the n-grams. However, this worsened the RMSE by 0.02 on average.

We believe the LIWC [24] resources could improve the accuracy of our model. However, we decided to throw extra effort in expanding the training corpus with unsupervised data, to the point where LIWC features would be covered by our own, trained on enough data. It is easy to detect user personality indications on Twitter as many people tweet results from personality tests online; however, our investigation shows that most of them follow the Myers-Briggs indicator, which does not correlate well with the big five. Thus, we did not use any extra training data as we managed to find very few examples per language.

5 Conclusion and Future Work

We have described the system built by Sofia University's *kiprov15* team for the PAN-2015 Author Profiling task, which was ranked 2nd in Spanish and 6th in English, according to the official GLOBAL score.

We have made some interesting observations about the impact of the different features. Among the best feature groups were POS-tag counts and word unigrams and bigrams. These had the most sustainable performance over the provided test datasets.

Even though we managed to achieve good ranking for the two languages we made submissions for, we feel that there is a lot more to gain. For example, we would like to try using different word clusters, thesaurus, the Linguistic Inquiry and Word Count (LIWC) lexicons for different languages, named entity recognition and normalization, e.g., locations, dates, numbers, money, person names, etc.

In addition to adding extra features, we are interested in using the social media to generate more training examples. In particular, we would like to explore the way personality is expressed in Twitter and whether it is dependent upon language usage in general. For instance, would the model improve if we have different training sets for English-speaking users in USA vs. UK vs. Canada vs. Australia vs. India, etc.

6 Source Code

The project source code can be found on GitHub:
https://github.com/ykiprov/pan2015

References

1. Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, February 2009.
2. Shlomo Argamon and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17:401–412, 2003.
3. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC '10, pages 2200–2204, Valletta, Malta, 2010.
4. Luciano Barbosa and Junlan Feng. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Beijing, China, 2010.
5. Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, and Ricard Gavaldà. Detecting sentiment change in Twitter streaming data. *Journal of Machine Learning Research, Proceedings Track*, 17:5–11, 2011.

6. Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. TwitIE: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '13, pages 83–90, Hissar, Bulgaria, 2013. INCOMA Ltd.
7. William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, SDAIR '94, pages 161–175, Las Vegas, Nevada, 1994.
8. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
9. 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.
10. Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011.
11. Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Uppsala, Sweden, 2010.
12. Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '13, pages 198–206, Hissar, Bulgaria, 2013. Incoma Ltd.
13. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004.
14. Bernard Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, 2009.
15. Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM '11, pages 538–541, Barcelona, Catalonia, Spain, 2011.
16. Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255, Montréal, Canada, 2012.
17. Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 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, pages 321–327, Atlanta, Georgia, USA, 2013.
18. Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 312–320, Atlanta, Georgia, USA, 2013.
19. Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time se-

ries. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10, pages 122–129, Washington, DC, USA, 2010.
20. Alexander Pak and Patrick Paroubek. Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 436–439, Uppsala, Sweden, 2010.
21. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Philadephia, Pennsylvania, USA, 2002.
22. Francisco M. Rangel Pardo, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at PAN 2013. In *Working Notes for CLEF 2013 Conference*, Valencia, Spain, 2013.
23. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 37–44, New York, NY, USA, 2011.
24. James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.
25. F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. Overview of the 3rd authorprofiling task at PAN 2015. In *Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers*, Toulouse, France, 2015.
26. Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, Sheffield, UK, 2014.
27. Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 73–80, Dublin, Ireland, 2014.
28. Sara Rosenthal, Alan Ritter, Veselin Stoyanov, Svetlana Kiritchenko, Saif Mohammad, and Preslav Nakov. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 451–463, Denver, CO, USA, 2015.
29. Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin Seligman, and Lyle Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. In *PLOS ONE*, volume 8, page e73791. Public Library of Science, 09 2013.
30. Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.