Gender Identification in Twitter using N-grams and LSA

Notebook for PAN at CLEF 2018

Saman Daneshvar$^{[0000-0001-8780-7955]}$ and Diana Inkpen$^{[0000-0002-0202-2444]}$

School of Electrical Engineering and Computer Science, University of Ottawa, Canada
{saman.daneshvar,diana.inkpen}@uottawa.ca

Abstract In this paper, we describe the participation of the Natural Language Processing Lab of the University of Ottawa in the author profiling shared task at PAN 2018. We present our approach to gender identification in Twitter performed on the tweet corpus provided by CLEF for the task. Our approach takes advantage of textual information solely, and consists of tweet preprocessing, feature construction, dimensionality reduction using Latent Semantic Analysis (LSA), and classification model construction. We propose a linear Support Vector Machine (SVM) classifier, with different types of word and character n-grams as features. Our model was the best-performing model in textual classification, with the accuracy of 0.8221, 0.82, and 0.809 on the English, Spanish, and Arabic datasets respectively. Considering the combination of textual and image classification, and all three datasets, our model ranked second in the task.

Keywords: author profiling · user modeling · gender detection · natural language processing · Twitter · social media

1 Introduction

The rise of social media in the past decade, has introduced new forms of social interaction. With more than a billion daily active users[1] and millions of posts posted every hour[2] on social networking services, namely Facebook and Twitter, social media has become an invaluable source of data for researchers. The increasing amount of unstructured textual data generated on the social media has thrived the need for Natural Language Processing to extract useful information with various applications ranging from security and defense (forensics), marketing, and personalization to research in psychology and sociology [8].

It is possible to learn some traits of users, such as gender, age, native language, and personality, based on what they share on the social media. In the author profiling shared task at PAN 2018 [15], the focus is on multimodal gender identification in Twitter, by leveraging the textual and image tweets that a user has posted. This year's task includes three languages: English, Spanish, and Arabic. Gender is framed as a binary classification problem.

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

[2] https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute

The rest of this paper is structured as follows: Section 2 reviews some of the related work in author profiling, and their approaches towards this task. In section 3, we describe the Twitter corpus of the author profiling task at PAN 2018. In section 4, we discuss our proposed system, including our preprocessing steps, the features that we extracted from the corpus, and the classifier trained on those features. We have included all details necessary to reproduce our results. In section 5, we investigate the use of the frequency of offensive expressions as features, and also demonstrate the effect of dimensionality reduction on the accuracy of our model for the English dataset. Section 6 reports the accuracy scores of our final model in cross-validation experiments and on the official test set of the task, and highlights the importance of paying attention to confidence intervals. Section 7 contains a brief discussion and conclusion.

2 Related Work

Pennebaker *et al.* [12] explored the link between language and demographic, psychological, and social traits of a person, including gender. Along with his colleagues, he developed the Linguistic Inquiry and Word Count (LIWC) tool, which exploits word counts and word categories to identify traits of authors.

Koppel *et al.* [10] classified gender on a genre-controlled corpus of formal written documents (e.g., news papers and books) with an accuracy of about 79 percent. They started with a total of 1,081 features to represent the corpus, including part-of-speech (POS) n-grams and a list of 405 function words. They then performed feature selection, and trained their classifier on the most important features.

Argamon *et al.* [2] probed the variation of writing styles between male and female on a corpus similar to above-mentioned. They reported pronouns (I, you, she, her, their, myself, yourself, herself) as strong female indicators, and determiners (a, the, that, these) and quantifiers (one, two, more, some) as male indicators.

Schler *et al.* [18] obtained an accuracy of 80 percent on gender classification on a corpus of 71,000 blogs from *blogger.com*. They used some stylistic features, including POS and hyperlinks, along with the 1000 unigrams with the highest information gain. They adopted the Multi-Class Real Winnow (MCRW) learning algorithm, which they claimed to be more efficient than SVM with comparable results. They reported that male bloggers write more about politics, technology, and money, while female bloggers share more about their personal lives.

Burger *et al.* [5] undertook gender classification of Twitter users and achieved 76% accuracy, when trained their model only on tweets, using word unigrams and bigrams and character 1- to 5-grams as features. They adopted the Balanced Winnow2 learning algorithm, which they claimed to have a better combination of accuracy, speed, and robustness than Naive Bayes and linear SVM. Furthermore, they assessed the performance of 130 human annotators on Amazon Mechanical Turk, annotating the gender of Twitter authors solely based on their textual tweets, and reported that only 5% of the human annotators were able to out-perform the classification model.

Author profiling has been undertaken as a shared task at PAN annually since 2013. The best performing team at PAN 2017 used word unigrams and bigrams and character 3- to 5-grams as features [3], however some teams experimented with word and character n-grams with values of n up to 3 and 7 respectively [16]. Term frequency–inverse document frequency (tf-idf) was a common feature transformation technique. POS tagging, lemmatization and stemming were also investigated with mixed results.[3] The best performing team at PAN 2015 [14] adopted Latent Semantic Analysis (LSA), and reported a 4% increase in accuracy in their English dataset, compared to bag-of-words (BOW).[1]

3 Dataset Description

The training dataset of the author profiling task at PAN 2018 consists of tweets and images of three groups of authors (i.e., Twitter users):

- English: 3,000 authors
- Spanish: 3,000 authors
- Arabic: 1,500 authors

The dataset is balanced and labeled with gender, in every group. Gender annotation has been done based on the name of the users, their profile photos, description, etc.[16,15]

For each author (Twitter user), a total of 100 tweets and 10 images were provided. Authors were coded with an alpha-numeric author-ID.

According to PAN, this year's dataset was a subset of the PAN 2017 dataset for the author profiling task, and was collected in 2017. For each author, the last 100 tweets had been retrieved from the user's timeline. As a result, the time frame of the tweets might vary from days to months, depending on how frequently a user tweets. According to [16] none of the tweets are retweets.

As for the images, according to PAN, they retrieved all the images posted by each user, taking into account at least the last 1,000 tweets of the user. They then randomly selected 10 of the images. As a result, the time frame of the images does not necessarily match that of the textual tweets.

For training our classification model, we only took advantage of the tweets, and not the images.

4 Feature Construction and Classification Model

4.1 Preprocessing

For preprocessing we used the *TweetTokenizer* module from the Natural Language Toolkit (NLTK) library [4] along with some additional procedures using regex. We performed the following preprocessing steps:

1. Replaced the linefeed characters with <*LineFeed*>.
2. Concatenated all 100 tweets of each author into one string, with an <*EndOfTweet*> tag added to the end of each tweet.
3. Lowercased the characters

4. Trimmed the repeated characters: Replaced repeated character sequences of length 3 or greater with sequences of length 3
5. Replaced URLs with <*URLURL*>
6. Replaced @*username* mentions (i.e., Twitter handles) with <*UsernameMention*>
7. Removed punctuations: Although we did not remove the punctuations in our preprocessing function, scikit-learn[11]'s *TfidfVectorizer* function completely ignores punctuation.[3]
8. Stop words were detected by document frequency and removed. Any n-gram that occurred in all documents was considered a stop word and was ignored.

4.2 Features

To find the best values for parameters, we performed several hyper-parameter tuning experiments—some manually and some using the scikit-learn's grid search function. As the three languages possess separate datasets and need separate classification models, we treated them as three independent tasks, and tuned their hyper-parameters separately. We ended up with two different sets of values: one for the English dataset, and another for both the Spanish and Arabic datasets.

In short, the feature set we used for the English dataset had two differences:

1. It benefited from word trigrams in addition to other n-grams
2. Dimensionality reduction using LSA was performed on its n-grams

We used the following n-grams features in our final model:

- Word unigrams, bigrams and trigrams for the English dataset
- Word unigrams and bigrams for the Spanish and Arabic datasets
- Character 3- to 5-grams for all three datasets

Both word- and character-level n-grams for all three languages used the following parameters (for the feature sets):

- Term frequency–inverse document frequency (tf-idf) weighting
- Sub-linear term frequency scaling, which uses $1 + log(TF)$ instead of TF.
- Minimum document frequency = 2: Terms with a document frequency strictly lower than 2 would be ignored.
- Maximum document frequency = 1.0 (100%): Terms that occur in all documents would be ignored.

In addition to this, we performed topic modeling using latent semantic analysis (LSA) only on the English dataset. Linear dimensionality reduction was done by means of truncated singular-value decomposition (SVD), using the *TruncatedSVD* function from the scikit-learn library [11].

LSA was first introduced in 1988 [6] as a technique for improving information retrieval (IR) by reducing the dimensionality of the IR problem. LSA is an unsupervised

[3] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

learning technique. It starts with the tf-idf matrix, performs singular-value decomposition (SVD) on it, and represents the documents and terms in a reduced-rank space. This reduced space is referred to as the "semantic" space, as it captures the relationships among words and documents.[7]

In the final model, we set the number of dimensions to 300, which seems to capture the most useful information of the text [20].

Additional features, which did not make their way into the final model, are discussed in section 5.

4.3 Learning algorithm

We examined different classifiers, namely Naive Bayes, logistic regression, and support Vector Machine (SVM) with a linear kernel. We obtained the best results from linear SVM, with the default parameters of scikit-learn's implementation.

5 Additional Features

In this section, we explain our experiments with using the frequency of offensive expressions as features, on the English dataset. Since these features had a negative impact on the accuracy of the model, we did not use them in our final model.

A rather intuitive hypothesis is that males use profanities or offensive language more often than females. We investigated this hypothesis using the Flame dictionary of offensive language expressions [17]. The dictionary[4] consists of 2,648 expressions, each marked with a flame level from 1 (lowest) to 5 (highest).

To compare the information gain of the expressions in each flame level, we first trained our SVM classifier on the expressions of each flame level separately. The features were extracted with the following steps:

1. Count the number of occurrences of each expression in each document (all tweets of the user).
2. Transform the count matrix into a normalized tf or tf-idf representation, using the scikit-learn's *TfidfTransformer* function.

In the first set of experiments, we used the unnormalized count matrix (tf). In the second set of experiments, we performed l^2-normalization on the count matrix. With l^2-normalization, each row of the matrix (representing each document, i.e., the tweets of each author) is normalized to have a sum of squares equal to 1. The accuracy of the models was measured using 10-fold cross-validation.

Table 1 shows the results of these experiments. CI refers to the 95% confidence interval for the accuracy values of the 10-fold cross-validation, as discussed in section 6.

With the majority classifier baseline being 50%, we can see that the flame level 3–5 expressions contain nearly no discriminating information. We then combined the expressions of flame level 1 and 2, and finally, combined the expressions of all flame

[4] http://www.site.uottawa.ca/~diana/resources/

Table 1. Cross-validation scores for offensive expressions feature set, with and without normalization

	No norm.		l^2 norm.	
Flame level	Mean	CI	Mean	CI
1	57.67	5.32	59.67	5.99
2	61.72	6.39	63.00	5.88
3	52.83	5.17	53.89	5.79
4	54.28	8.56	54.56	9.79
5	52.56	4.51	51.44	4.90
1 & 2	61.17	7.36	65.33	5.43
All (1–5)	61.28	6.90	65.94	7.18

levels. The highest accuracy ($65.94\% \pm 7.18\%$) belongs to term frequency of all flame levels, with l^2 normalization.

We also performed the same experiments on the tf-idf matrix, with and without normalization, but the accuracy scores were hurt rather than improved.

Next, we combined the offensive expressions feature set with the word- and character-level n-grams, after performing LSA on them. We also performed LSA on the offensive expressions features, with different number of dimensions. The accuracy of the model was not improved in any of the experiments. The results are shown in table 2.

Table 2. Cross-validation scores for the combination of offensive expressions and word and character n-grams

Features	Mean	CI
N-grams	81.89	5.40
N-grams, LSA 300	82.83	5.50
N-grams, LSA 300 + Offensive expressions	82.72	6.80
N-grams, LSA 300 + Offensive expressions, LSA 300	82.61	6.87
N-grams, LSA 300 + Offensive expressions, LSA 10	82.67	4.79
N-grams, LSA 300 + Offensive expressions, LSA 5	82.78	5.04

6 Results

"You are only as good as your cross-validation results."

To prevent over-fitting on the training set, and to avoid setting aside a significant portion of the training set for validation, we evaluated the accuracy of our model using stratified 10-fold cross-validation, during all experiments. There is a bias-variance trade off in choosing the value of k in k-fold cross validation. $k = 5$ and $k = 10$ have been shown empirically to yield acceptable results.[9]

We set aside 40% of the task's training set for test. During all experiments, our model never saw this portion of the dataset and was trained only on the remaining 60% (i.e., 1,800 authors for each of the English and Spanish datasets and 900 authors for the Arabic dataset). With the exception of table 3, all the reported cross-validation scores have been performed on this 60% subset of the task's training set.

It goes without saying that the submitted model was trained on the whole training set and tested on the official PAN 2018 [19] test set for the author profiling task, on the TIRA platform [13].

Table 3 shows the 10-fold cross-validation scores for our final system performed on the whole training set (3,000 authors for the English dataset, and so on), as well as the accuracy score on the official test set. All numbers are expressed as percentages.

Table 3. Accuracy scores of 10-fold cross-validation and the official test set

	Cross-validation		Test set
Dataset	Mean accuracy	CI	Accuracy
English	82.73	5.22	82.21
Spanish	79.87	4.17	82.00
Arabic	81.53	6.40	80.90

CI refers to the confidence interval for the accuracy values of the 10-fold cross-validation. We calculated the 95% confidence interval as 2σ, where σ is the population standard deviation:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $n = 10$, x_i is the accuracy score of the i-th cross-validation run, and $\bar{x}$ is the mean of the 10 accuracy scores.

Loosely speaking, this means that 95% of the cross-validation scores lie within the range of $\bar{x} \pm CI$. As can be seen in Table 3, this range is quite considerable. This indicates the variance in the dataset, and suggests that testing the model on a single test set cannot be considered the ultimate judgment on its accuracy.

Since the dataset is balanced, the accuracy of the majority classifier baseline is 50%. Therefore we can say that there is a substantial improvement over this baseline.

Table 4 shows the cross-validation scores for the word and character n-grams feature sets and their combination. As stated in section 4.2, word unigrams and bigrams were used for the Spanish and Arabic datasets, and the English dataset employed word trigrams as well. All three datasets had character 3- to 5-grams in their feature set.

Table 4. Cross-validation scores for word and character n-grams, as stated in section 4.2

	English		Spanish		Arabic	
Features	Mean	CI	Mean	CI	Mean	CI
Word n-grams	80.67	4.15	77.06	7.01	78.22	7.52
Character n-grams	81.33	4.80	77.11	5.30	77.89	7.40
Word and character n-grams	81.89	5.40	78.17	6.83	80.33	6.82

7 Discussion and Conclusion

For the author profiling task at PAN 2018, a relatively simple system using word and character n-grams and an SVM classifier proved strong. This underlines the importance of careful hyper-parameter tuning for the feature sets. For the English dataset, performing dimensionality reduction using LSA on the tf-idf matrix improved the accuracy by about 1%.

As for preprocessing, it is worth mentioning that removing the repeated character sequences helped improve the accuracy and lowered the number of features. This is among the preprocessing procedures that is specifically beneficial to Twitter corpora. Due to the nature of the social media, users often use repeated sequences to express their feelings, among other reasons.

While the normalized term frequency representation of the offensive expressions yielded an accuracy well above the random baseline, combining it with the n-grams features did not have a considerable effect on the accuracy. This can be due to the fact that the n-grams already contain the information that the offensive expressions feature set embodies.

SVM has long proven a strong performance; however, we believe that logistic regression is worth considering in similar experiments. Furthermore, deep neural networks may achieve outstanding results on larger datasets.

Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Ottawa for funding this research. Moreover, we thank the organizing committee of PAN, especially Paolo Rosso, Francisco Rangel and Martin Potthast for their encouragement and kind support.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Escalante, H.J.: INAOE's participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org (9 2015) 2
2. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text **23**(3), 321–346 (2003). https://doi.org/10.1515/text.2003.014 2
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR-WS.org (9 2017), http://ceur-ws.org/Vol-1866/ 2
4. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. OâĂŹReilly Media Inc. (2009) 4.1
5. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145568 2

6. Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, R.: Using Latent Semantic Analysis to Improve Access to Textual Information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 281–285. CHI '88, ACM, New York, NY, USA (1988). https://doi.org/10.1145/57167.57214, http://doi.acm.org/10.1145/57167.57214 4.2
7. Dumais, S.T.: Latent semantic analysis. Annual Review of Information Science and Technology **38**(1), 188–230 (2005). https://doi.org/10.1002/aris.1440380105, http://doi.wiley.com/10.1002/aris.1440380105 4.2
8. Farzindar, A., Inkpen, D.: Natural language processing for social media. Synthesis Lectures on Human Language Technologies **i**(2), 1 PDF (xix, 146 pages) (8 2015). https://doi.org/10.2200/S00659ED1V01Y201508HLT030, https://doi.org/10.2200/S00659ED1V01Y201508HLT030 1
9. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated (2014) 6
10. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing **17**(4), 401–412 (11 2002), http://dx.doi.org/10.1093/llc/17.4.401 2
11. 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 {P}ython. Journal of Machine Learning Research **12**, 2825–2830 (2011) 7, 4.2
12. 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 (2 2003). https://doi.org/10.1146/annurev.psych.54.101601.145041, https://doi.org/10.1146/annurev.psych.54.101601.145041http://arxiv.org/abs/1611.08945 2
13. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PANâĂŹs Shared Tasks: BT - Information Access Evaluation. Multilinguality, Multimodality, and Interaction. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 8685 LNCS, pp. 268–299. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_22 6
14. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org (9 2015) 2
15. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (9 2018) 1, 3
16. Rangel Pardo, F.M., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, vol. 1866. CLEF and CEUR-WS.org (9 2017), http://ceur-ws.org/Vol-1866/ 2, 3
17. Razavi, A.H., Inkpen, D., Uritsky, S., Matwin, S.: Offensive Language Detection Using Multi-level Classification. In: Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence. pp. 16–27. AI'10, Springer-Verlag, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13059-5_5, http://dx.doi.org/10.1007/978-3-642-13059-5_5 5

18. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. AAAI spring symposium: Computational approaches to analyzing weblogs **6**, 199–205 (2006) 2
19. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (9 2018) 6
20. Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A.: Latent semantic analysis. In: Proceedings of the 16th international joint conference on Artificial intelligence. pp. 1–14. Citeseer (2004) 4.2