Language-independent Gender Prediction on Twitter

Nikola Ljubešić
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Darja Fišer
Faculty of Arts
University of Ljubljana
Aškerčeva cesta 2
1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

In this paper we present a set of experiments and analyses on predicting the gender of Twitter users based on language-independent features extracted either from the text or the metadata of users' tweets. We perform our experiments on the TwiSty dataset containing manual gender annotations for users speaking six different languages. Our classification results show that, while the prediction model based on language-independent features performs worse than the bag-of-words model when training and testing on the same language, it regularly outperforms the bag-of-words model when applied to different languages, showing very stable results across various languages. Finally we perform a comparative analysis of feature effect sizes across the six languages and show that differences in our features correspond to cultural distances.

1 Introduction

Gender prediction is a well-established task in author profiling, useful for a series of downstream analyses (Schler et al., 2006; Schwartz et al., 2013; Bamman et al., 2014) as well as predictive model improvements (Hovy, 2015). Most existing work on predicting gender focuses on exploiting the linguistic production of the users (Koppel et al., 2003; Schler et al., 2006; Kucukyilmaz et al., 2006; Burger et al., 2011; Miller et al., 2012; Rangel et al., 2016), just rarely using non-linguistic information such as metadata (Plank and Hovy, 2015) or visual information (Alowibdi et al., 2013).

In this paper we investigate the possibility of predicting gender of a Twitter user regardless of the language used in his or her tweets. We perform our experiments on an existing dataset of Twitter users speaking six different languages that were manually annotated for their gender. Our language-independent gender predictor relies on general linguistic features, such as the usage of punctuation, and non-linguistic features calculated from Twitter metadata, such as the user interaction in the form of replying, retweeting and favoriting, time of posting, color choices, client usage etc.

The potential of a language-independent procedure for gender prediction is substantial both for the field of natural language processing where using extra-linguistic variables is currently gaining momentum, as well as disciplines from social sciences and the humanities working with user-generated content, where such factors have a long tradition. We believe that building such language-independent procedures is the only tractable way of moving forward given the number of different languages used in social media and the existence of training data only for a few high-density languages.

In the next section we briefly describe the dataset we performed our experiments on, in Section 3 we describe our language-independent features, in Section 4 we give the experimental setup of our gender prediction experiments, while in Section 5 we present the gender prediction results, as well as a series of analyses of the feature spaces across languages. In Section 6 we give some conclusions and directions for further research.

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 1–6,
Vancouver, Canada, August 3, 2017.

2 The Dataset

In our experiments we fully rely on the TwiSty corpus (Verhoeven et al., 2016) which was developed for research in author profiling. It contains personality (MBTI) and gender annotations for a total of 18,168 authors posting in German, Italian, Dutch, French, Portuguese or Spanish. The manual gender annotations in the TwiSty corpus are based on the user's name, handle, description and profile picture and follow the performative view of gender, i.e., that gender is discriminated by performances that respond to societal norms or conventions (Larson, 2017). The corpus is distributed in the form of Twitter user IDs and specific tweet IDs of that user.

In this work we use only the user IDs and their gender and language annotations to collect timelines of users through the Twitter API. For each user we collect up to 3,200 tweets (API restriction) and discard users with less than 100 tweets. By doing so we collected 45 million tweets for 16,156 users across the six languages.

3 The Features

In this section we present the 51 user-level features which we consider to be good feature candidates for language-independent gender prediction. These features follow one of the four following feature types:

- `perc` - percentage of user tweets satisfying a condition (like the percentage of tweets containing emojis)
- `mean` - mean of a continuous tweet-level variable (like the mean of the posting hour)
- `med` - median of a continuous tweet-level variable
- `var` - variance of a continuous tweet-level variable
- `user` - variables derived from user-level metadata (such as the average number of tweets published daily)

Following the `perc` type, we define the following features: usage of various clients for posting the tweets (Android, iOS, web), presence of specific textual elements (emojis, emoticons, URLs, hashtags, mentions, commas, ellipses, question-marks, exclamation marks) and criteria depending on tweets' metadata (replies, posting during working hours, posting during weekends, truncated tweets, favorited tweets, quotes, retweeted tweets).

By following the three types, `mean`, `med` and `var`, we encode the following distributions in our feature space: retweet count, favorite count, posting hour, day of week the tweet was posted and tweet length.

The last feature type, `user`, is used to encode the following information: average daily number of tweets, overall number of tweets, number of tweets the user has favorited, number of followers, number of friends, the ratio of follower to friend numbers, number of lists the user is on, whether the user has a background image defined, whether the user has the default profile image, whether the user has a profile description, whether the user has a location defined, and red, green and blue color component intensity (two-digit hexadecimal code from the RGB color definition) of the user's text and background color.

4 Experimental Setup

In this section we outline the setup of our gender classification experiments, whose results we report in Section 5.1.

We train models based on standardized (zero mean, unit variance) language-independent features described in the previous section with support vector machines (SVMs) using a radial basis function (RBF) kernel and optimizing the γ and C hyperparameters via 5-fold cross-validation.

To have a reasonable point of comparison for our language-independent models, we built bag-of-words (BoW) models on a concatenation of all tweets of a user by using lowercased character 5-grams as features and an SVM with a linear kernel.

We use character 5-grams as they have proven in our initial experiments to yield better results than words or character n-grams of different length. We use a linear kernel and not the RBF one in these experiments as the number of features is much higher than the number of instances. We do not perform any input processing except lowercasing as we expect useful signal for the task to be present in non-alphabetic characters, URLs, hashtags, mentions etc.

The number of features in our BoW models ranges from 6.2 million for German to 51.2 million for Spanish.

We discriminate between in-language and

Lang	Inst. #	MFC	ILBoW	CLBoW	DE	IT	NL	FR	PT	ES
DE	376	36.63	77.91	61.26	69.37	**63.30**	**67.26**	**68.35**	65.59	**69.92**
IT	429	50.96	62.46	58.66	**66.98**	**63.91**	**66.76**	**63.73**	**63.47**	**66.12**
NL	933	34.59	80.68	61.55	62.10	**61.15**	68.02	**57.87**	59.64	**64.68**
FR	1207	41.78	78.70	56.61	**69.70**	**65.12**	62.68	67.47	**65.60**	**66.35**
PT	3572	43.97	85.26	53.18	**61.94**	**57.31**	57.23	**62.65**	69.51	68.12
ES	9639	41.13	83.04	57.99	**62.89**	55.80	**64.85**	**66.82**	67.27	71.47

Table 1: Gender classification results on the six languages (rows), columns encoding the testing language (Lang), number of instances (Inst. #) and the weighted F1 results on most-frequent class baseline (MFC), in-language bag-of-words (ILBoW), average cross-language bag-of-words (CLBoW) and the six language-independent models. Bold results outperform the corresponding BoW baseline.

cross-language experiments. In all in-language experiments we perform 5-fold cross-validation, while in cross-language experiments we simply apply the model from the training language on the test language dataset.

We use weighted F1 as our evaluation metric and the most-frequent class baseline as our weak baseline.

5 Results

In the first part of this section we report on the gender classification results while in the second part we perform a series of feature analyses.

5.1 Gender Classification

We report results on gender classification in Table 1. Each of the rows represents the evaluation on a specific language encoded in the first column. The second column contains the number of instances, i.e., users available per language. The next column encodes the most-frequent class baseline (MFC) while the two columns that follow contain the bag-of-words results, either in the in-language setting (ILBoW) or the cross-language setting (CLBoW) for which, due to space constraints, we report only the average results over the five different languages.

In the remaining six columns we report the results obtained with models based on the language-independent features trained on specific language datasets. If the training language is the same as the testing language, we report the 5-fold cross-validation results. The results given in bold are of those systems that perform better than the BoW model with the same training and testing language.

The first observation we make is that all the models outperform the MFC baseline significantly. In-language BoW models perform, as expected, in all cases better than the average cross-language BoW model. They also perform better than most language-independent models, the Italian one being an exception. In cases where the training and testing language differ, in most cases the models based on language-independent features outperform the BoW models. We can observe a positive effect of the training data size on most of the BoW models since in the three languages with less training data (first three rows) CLBoW models outperform the language-independent ones only in three (20%) settings, while for the last three languages this is the case in five (33%) settings.

Finally, the language-independent models show much more consistent results than BoW models in the cross-lingual setting with an average per-language variance of the cross-lingual experiments of 0.001 for language-independent models and 0.01 for BoW models.

5.2 Feature Analysis

To obtain a better understanding of the informativeness of specific features for the task at hand, we performed a univariate analysis of each feature in each language. On a scaled (zero mean, unit variance) dataset of each language, we ranked the features by the p-value of the Mann Whitney U test.[1] In Table 2 we present features ranked by the average rank throughout our six languages. Due to space constraints we present only the 30 highest ranked features.

Each feature in each language is quantified by the effect size of the gender-conditioned distributions which we simply calculate as the difference

[1]The p-value quantifies the probability that we falsely reject the null hypothesis that the two gender-conditioned samples were selected from populations having the same distribution.

Feature	Avg rank	DE	IT	NL	FR	PT	ES
perc_emoji	1.17	0.63	0.21	0.45	0.49	0.41	0.5
mean_retweet_count	11.5	0.09	0.03	0.09	0.38	0.27	0.22
red_back	12.0	0.24	0.09	0.13	0.23	0.38	0.42
perc_http	13.5	-0.21	-0.24	-0.25	-0.15	-0.27	-0.17
perc_ios	14.0	-0.23	-0.22	-0.09	-0.19	-0.09	-0.13
var_retweet_count	15.17	-0.1	0.05	0.1	0.11	0.03	0.04
perc_retweeted	15.33	-0.01	0.2	-0.2	0.2	0.26	0.17
perc_question	16.0	-0.35	-0.13	-0.1	-0.29	-0.14	-0.11
user_tweet_per_day	17.0	0.08	0.19	0.01	0.31	0.15	0.12
perc_emoticon	18.17	-0.23	-0.25	-0.17	-0.18	-0.24	-0.1
user_location	18.67	-0.17	-0.2	-0.21	-0.11	-0.17	-0.12
mean_hour	19.33	0.08	0.23	0.18	0.22	-0.1	-0.02
var_len_text	20.0	0.25	0.24	0.2	0.24	0.01	0.08
user_favour_count	20.33	0.06	0.09	0.02	0.1	0.02	0.06
user_tweet_count	20.33	0.03	0.2	-0.01	0.23	0.13	0.09
user_follow_friend_rat	21.5	-0.13	0.12	-0.05	-0.08	-0.04	-0.03
mean_favorite_count	21.5	0.16	0.09	0.02	-0.07	-0.02	-0.03
med_hour	22.0	0.13	0.23	0.17	0.2	-0.01	-0.07
green_back	22.17	0.2	0.04	-0.04	0.12	0.26	0.25
blue_back	22.33	0.27	0.06	-0.01	0.11	0.29	0.33
perc_is_quote	22.83	-0.04	0.17	-0.21	0.18	0.17	0.03
perc_favorited	23.33	0.31	0.16	-0.02	0.14	0.05	0.01
med_retweet_count	24.17	-0.08	-0.06	-0.09	0.16	0.06	0.04
var_favorite_count	24.17	-0.09	0.07	0.07	-0.12	-0.02	-0.03
var_hour	25.17	-0.11	-0.01	-0.1	-0.14	0.21	0.05
user_red_text	25.83	0.15	0.05	0.09	0.22	0.13	0.16
user_listed_count	28.33	-0.12	-0.09	-0.09	-0.17	-0.0	-0.07
perc_exclamation	28.83	0.26	0.09	0.49	-0.04	-0.04	0.14
var_day	29.17	0.09	0.1	-0.0	0.14	0.12	0.12
perc_hash	29.67	-0.11	-0.05	-0.03	-0.16	-0.09	-0.11

Table 2: Representation of 30 (out of 51) features with the highest average rank across languages. Each feature in each language is represented through the difference between feature means of the female and male subsets in a standardized dataset. Red encodes higher female mean, blue male.

in the mean of the female and the male subsample. A positive value therefore means that female users have a higher average value of that feature than male users, and vice versa. Let us repeat that these calculations were performed on scaled data, therefore these quantifications are comparable across variables. To simplify the reception of the data, we color the background of each cell either with red (female) or blue (male) with the color intensity corresponding to the effect size.

Such a feature representation enables a comparison of various features, as well as identical features across languages. Given the good results of the classification task presented in the previous subsection, we hypothesize that the effect sizes, and especially their signs, should correspond between languages.

This hypothesis is largely confirmed, especially on the highest ranked features. The three highest ranked features – percentage of emoji usage, mean retweet count and intensity of the red component in the background color – signal that the user is female across all the six languages. The two features that follow – percentage of tweets containing URLs and percentage of tweets sent from an iOS device – are indicative of the male gender, again, across all the languages. Among the top 20 features, 5.3 out of 6 features on average have an identical sign, while among the top 30 features this is the case for 5.1 features.

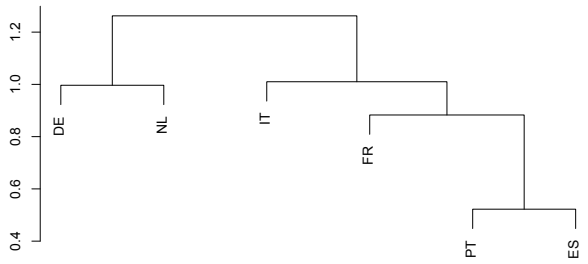

Figure 1: Dendrogram of the hierarchical language clustering. Each language is represented with feature effect sizes of all 51 language-independent features.

Regarding the use of emojis and emoticons, is it quite interesting that emojis are in all six languages preferred by the female gender while emoticons are preferred, again in all six languages, by the male gender. Male users tend to use more questionmarks, hashtags and share their location across all languages, while female users tend to produce more tweets per day, tweets of varying length, favorite more tweets and use more of the red color component in the tweet text, again, across all the languages.

Finally, given that there still is variation in our feature effect sizes across languages, we investigated whether this variation follows cultural differences between the speakers of the six languages. To investigate this matter we represented each of the six languages as a vector of the 51 effect sizes from Table 2 and performed agglomerative clustering of the six languages by using the Euclidean distance and the complete agglomeration method. The resulting dendrogram is presented in Figure 1.

The dendrogram shows that the difference between the features across languages corresponds to the linguistic as well as cultural distance of the cultures the languages are dominant in. We argue that the measured differences are mostly due to cultural differences as just the small number of punctuation-based variables, more precisely 4 out of 51, have any linguistic merit while the rest of the variables encodes other behavioral differences.

The two languages with the most similar feature effect sizes are Portuguese and Spanish, this cluster being expanded with French and then Italian. At a similar distance threshold point, German and Dutch are merged into one cluster.

Some of the variables that support such a clustering outcome are (1) the percentage of tweets that are retweeted which tends to be higher for male users in German and Dutch and for female users in the remaining languages, (2) the average posting hour that is higher for male Portuguese and Spanish users and female users in the remaining languages, (3) the average number of favorites per tweet which is higher for male users in French, Portuguese and Spanish and female users in the remaining languages (4) the percentage of tweets that are quotes which is higher among male users in German and Dutch and among female users in the remaining languages and (5) the variance of posting hour which is higher for female users in Portuguese and Spanish and for male users in the remaining languages.

6 Conclusion

In this paper we have presented a first run at the problem of language-independent gender identification among Twitter users. We have shown that with 51 language-independent features in the cross-lingual setting we regularly beat the bag-of-words baseline, and, furthermore, that the language-independent models have a ten times smaller F1 variance, which proves for our models to be more robust than the bag-of-words models, and therefore more reliably applicable to new languages.

We have analyzed the effect sizes of specific features among languages and have shown that our features regularly correspond across languages which also explains why the models work reliably across languages. By performing hierarchical clustering over languages represented through feature effect sizes we have shown that the difference in feature values across languages corresponds to the cultural distances of the speakers of those languages.

While the results presented in this paper are promising, there is a series of open questions that have to be explored. The most pressing one is the representativeness of users in the TwiSty corpus as they are Twitter users that have self-reported their personality test results. A way of measuring this representativeness is to apply these models to another gender prediction dataset. Further features should also be explored (network-based, image content etc.), as well as the potential of building additional language-independent author profiling models, such as age or educational level predictors.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, methods and tools for the understanding, identification and classification of various forms of socially unacceptable discourse in the information society" (J7-8280, 2017-2020).

References

Jalal S. Alowibdi, Ugo A. Buy, and Philip Yu. 2013. Language independent gender classification on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, New York, NY, USA, ASONAM '13, pages 739–743.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 1301–1309. http://www.aclweb.org/anthology/D11-1120.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 752–762. http://www.aclweb.org/anthology/P15-1073.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17:401–412.

Tayfun Kucukyilmaz, Berkant Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. 2006. Chat Mining for Gender Prediction. In *ADVIS*. Springer, volume 4243 of *Lecture Notes in Computer Science*, pages 274–283.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, pages 30–40. http://www.aclweb.org/anthology/W/W17/W17-1604.

Z. Miller, B. Dickinson, and W. Hu. 2012. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. *International Journal of Intelligence Science* 2(24).

Barbara Plank and Dirk Hovy. 2015. Personality Traits on Twitter or How to Get 1,500 Personality Tests in a Week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Lisboa, Portugal, pages 92–98. http://aclweb.org/anthology/W15-2913.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. AAAI, pages 199–205.

H A Schwartz, J C Eichstaedt, M L Kern, L Dziurzynski, S M Ramones, M Agrawal, A Shah, M Kosinski, D Stillwell, M E Seligman, and L H Ungar. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One* 8(9).

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, ELRA, Portorož, Slovenia.