Cross-lingual gender prediction with multi-lingual embeddings and linguistic features

Wolfgang Tessadri, B.A.
IMS
Universität Stuttgart
st163552@stud.uni-stuttgart.de

Faizan E Mustafa, B.Sc.
IMS
Universität Stuttgart
st170388@stud.uni-stuttgart.de

Abstract

Most systems for gender profiling have focused on mono-lingual use-cases. In this paper we implement two systems for cross-lingual gender prediction and compare them: One system is based on linguistic features while the other leverages a multi-lingual embedding approach (XLM-RoBERTa). Moreover, we analyse which linguistic properties are most predictive for gender profiling across languages. We find that XLM-RoBERTa performs best with accuracy scores of up to 0.87. Classification on top of linguistic features did not consistently generalize cross-lingually. However, linguistic feature analysis supported previously observed divergences of male and female language use across languages.

The resources described in this paper can be found in this GitHub repository.

1 Introduction

With constantly growing amounts of publicly available textual information in recent years, also the task of author profiling has increasingly gathered attention: As meta-data about the author of digital documents most often is not or only sparsely available, author profiling deals with "deciphering information about the author from the text that he/she has written" (Arroju et al., 2015, p. 22). Among others, systems for inferring the authors' gender (Verhoeven et al., 2016; Cheng et al., 2011), age (Johannsen et al., 2015; Peersman et al., 2011), and personality traits (Arroju et al., 2015; Verhoeven et al., 2016) have been developed. Most often, however, systems work on mono-lingual data. An exception to this are Van der Goot et al. (2018), who choose a cross-lingual perspective on gender classification: Using language-independent abstract features like binned word frequency, word length and consonant-vowel sequences they train a SVC model classifying the gender of Twitter users across five languages. Even though the authors find this technique of bleaching features to be "surprisingly effective" (Van der Goot et al., 2018, p.387) it has a major drawback: Due to the abstract nature of features gender divergences are not interpretable from a more linguistically motivated point of view. In the light of the above, in this paper we pursue three main objectives: (i) We develop linguistically motivated features and evaluate a linguistic feature classifier (LFC) for cross-lingual gender prediction of Twitter users. (ii) We evaluate which features are most relevant for identifying gender across the languages investigated. (iii) We implement a state-of-the-art (SOTA) system for gender profiling, which uses the most recent version of the cross-lingual transformer-based XLM-RoBERTa model (Conneau et al., 2020), and compare it to the linguistic feature classifier.

2 Methods

A crucial difference of the classifiers compared in this paper is their fundamentally different feature input:

1. Linguistic features (LFC): Koppel et al. (2002) as well as Argamon et al. (2003) were among the first to show that the use of POS-tags in English differs between male and female writers and provides valuable hints for gender classification. In the past the application of POS features and other linguistic features to a cross-lingual gender profiling scenario was hindered by different, language-specific tagsets, which required transformation processes for each language pair.

Cross-lingual feature extraction, however, has been greatly eased by the Stanza initiative and pertaining Python package (Qi et al.). In the spirit of the Universal Dependency framework (UD; see e.g. McDonald et al. 2013) Stanza allows for parallel cross-lingual analysis of syntactic, morphological and morpho-syntactic information. Based on this functionality we extract and evaluate four types of linguistic features: (i) (Universal) POS-tags, (ii) morphological properties, (iii) (Universal) Dependency relations, (iv) syntactic complexity[1].

2. Embeddings (XLM-RoBERTa): The other approach we test, applies a transfer learning idea. Howard & Ruder (2018) first showed that effectively fine-tuning language models can achieve SOTA performance for six text classification tasks. Transformer-based language models also showed promising results for tasks requiring cross-lingual transfer. Recently, XLM-RoBERTa was introduced to the field by Conneau et al. (2020), which learns unsupervised cross-lingual representations for one hundred languages and achieves SOTA results for various NLP classification tasks without hurting per-language performance. Thus, XLM-RoBERTa also promised good generalizability for cross-lingual gender profiling.

3 Experiments

3.1 Experimental setting

To perform our experiments we used the Twisty corpus provided by Verhoeven et al. (2016). The corpus contains gender and personality trait annotations for over 18.000 Twitter users writing in six different languages. Out of these six languages we selected four, pairwise closely related languages, i.e. German-Dutch as well as French-Italian, to test for effects of language affinity. Table 1 gives a short overview over the specifications of the data used in this paper.

As is evident, Dutch and French comprise more than twice as many users and tweets compared to German and Italian. Generally, male and female class distributions are relatively balanced, except for Italian, where female users clearly predominate. This had to be considered for analysis. The number of average tweets per user is comparable across languages. However, there were significant intra-language differences: The number of tweets per user ranged from 1 to over 3200 tweets. Thus, in the computation of most feature dimensions listed in Appendix A we decided to use relative frequencies instead of absolute counts.

	Nr. users	Nr. tweets	Avg. Nr. tweets/user	Dist. M/F
DE	351	798303	2274	48/52
FR	1135	2204100	1942	44/56
IT	410	778011	1898	35/65
NL	899	1847154	2055	49/51

Table 1: Specifications of Twisty Data used in this paper.

For our experiments we split the users in each language into a train and test set at the ratio of 4:1. Additionally, in case of the embedding systems 10% percent of the train set were used for validation[2]. Based on this data we train two types of classifiers:

For the LFC we use the sklearn maximum entropy classifier with default settings[3].

While in case of the LFC we used all tweets of a user for feature extraction, in case of the embedding system averaging all word embeddings for all words in tweets of a user to generate a user representation is likely to hurt the classification system's performance. Thus, we decided to filter out the most informative tweets. To do so, we first extracted the 4000 words most associated with gender classes using χ^2 feature selection. Then we sorted tweets based on the frequency of these words in a tweet and selected the 30 most informative tweets, which were then used to generate a user representation with XLM-RoBERTa. To create this model we adapted code[4] using the Hugging face transformer library (Thomas Wolf et al., 2019). In order to adapt the language model to our classification task, we placed a classification layer on top and fine-tuned the model using our corpus data.

3.2 Results

In this section we compare performance results for the two described classifiers. What has to be borne in mind for result interpretation is that, especially for Italian, class distribution is rather imbalanced. For this reason, in contrast to most papers in the

[1] Appendix A shows how we instantiated each of these feature types.

[2] See Appendix B for Nr. train/test users per language.

[3] lbfgs solver, L2 regularization with reg. term = 1

[4] The code was taken from Abhishek Thakur on GitHub.

field, we report on micro as well as macro F1. We assume classification to be successful if the average of micro and macro F1 is higher than the same score with simple majority class assignment (see Appendix C). Table 2 shows prediction results for the LFC.

The table shows classification performance for all combinations of train and test languages. "Target Train" refers to the scenario where the classifier is trained on the target language and tested on all the other languages, while with "Target Test" the classifier is trained on all except the target language and tested on the target language. In case of the last cell "Target Test/Target Train" the classifier is trained and tested on all languages at once.

Regarding intra-language performance (italics) the classifier reached F1 micro/macro scores of around 0.72, which is in line with performance scores previously reported in the field. Scores drop significantly for cross-linguistic scenarios: The classifier surpasses the majority threshold only in 11 out of 20 possible cross-linguistic scenarios (boldfaced). Interestingly, this is always the case in combinations of related languages, with the combination French/Italian performing best.

Table 3 shows the results achieved with XLM-RoBERTa. It is evident that this model clearly prevailed over the LFC surpassing the majority threshold in all possible scenarios, which stands for a consistent cross-lingual generalization. Moreover, and contrary to the LFC, micro and macro F1 in case of XLM-RoBERTa most often are close, which speaks for a more balanced performance over classes. The only cross-lingual scenario where the LFC comes close to XLM-RoBERTa performance is if trained on Italian and tested on the remaining languages.

4 Discussion

The results presented in the previous section clearly showed the prevalence of the embedding system over the LFC. XLM-RoBERTa reached cross-lingual performance levels, which the LFC did not even reach intra-lingually. This is especially prominent when testing on German, where XLM-RoBERTa consistently reached F-scores over 0.8. However, even though performance was not completely convincing in case of the LFC, the classification on top of linguistic features led to an interesting result not observed with the embedding model: A tendency of similar gender marking for related languages was found, even if this relationship was much more pronounced for Italian-French as compared to German-Dutch.

A question which has not yet been addressed is which linguistic features were most predictive across languages. To evaluate this we took a closer look at the maximum entropy model trained on all languages and analysed which features were assigned the highest positive (male was encoded as category 1) and negative (female was encoded as 0) weights. Table 4 shows which fifteen features were found most predictive for gender profiling across languages.

The first thing to be noticed is the prevalence of POS features: 20 out of the 30 most predictive features belong to the POS feature set. By contrast only 7 features refer to dependency and 3 to morphological properties. Given the fact that the dependency feature set contains over 800 possible features while the morphological feature set consists of 11 possible different features, this indicates that POS tags are most distinctive in gender profiling followed by morphological properties.

Most interestingly, two associated morphological features, i.e. feminine and masculine gender ratio are very highly ranked with feminine gender ratio being the best female predictor overall. These features quantify to which degree the words used by a user belong to one or the other gender. The results now indicate that biological gender is correlated with an increased usage of the corresponding grammatical gender with women using feminine gender and, consequently, men using masculine gender more often[5]. The third morphological feature, i.e. first person ratio is correlated with female language use and indicates that female Twitter users use first-person pronouns more frequently. This is in line with the findings of Argamon et al. (2003), who observe a significantly increased use of first-person pronouns for female writers, which is also supported by the fact that 3 out of 9 female POS features contain the pronoun tag PRON. While Argamon et al. (2003) perform a mere monolingual analysis using the British National Corpus, Johannsen et al. (2015) investigate gender-related syntactic variation across eleven languages. Their results suggest that "the use of numerals and nouns is significantly correlated with men, while pronouns and

[5]These observations have to be taken with a grain of salt as Dutch does not distinguish between male, female and neuter gender but only between common and neuter.

		TEST				
		DE	FR	IT	NL	Target Train
T	DE	*0.72/0.72*	0.43/0.34	0.35/0.26	**0.47/0.43**	0.43/0.42
R	FR	**0.51/0.36**	*0.75/0.74*	**0.61/0.58**	0.49/0.37	**0.53/0.49**
A	IT	**0.59/0.54**	**0.56/0.54**	*0.70/0.66*	**0.61/0.56**	**0.58/0.56**
I	NL	**0.52/0.48**	0.47/0.42	0.35/0.29	*0.76/0.76*	0.45/0.40
N	Target Test	**0.58/0.54**	0.49/0.43	**0.71/0.63**	0.48/0.32	*0.73/0.73*

Table 2: Performance of the LFC (micro/macro F1).

		TEST				
		DE	FR	IT	NL	Target Train
T	DE	*0.90/0.90*	**0.64/0.68**	**0.60/0.65**	**0.75/0.76**	**0.68/0.71**
R	FR	**0.87/0.87**	*0.84/0.85*	**0.74/0.77**	**0.70/0.71**	**0.76/0.76**
A	IT	**0.80/0.80**	**0.67/0.69**	*0.61/0.66*	**0.67/0.68**	**0.70/0.70**
I	NL	**0.82/0.82**	**0.70/0.72**	**0.58/0.59**	*0.74/0.74*	**0.70/0.71**
N	Target Test	**0.83/0.83**	**0.66/0.70**	**0.73/0.76**	**0.74/0.74**	*0.81/0.82*

Table 3: Performance of the XLM-RoBERTa classifier (micro/macro F1).

Male	**Female**
PROPN, NUM, PUNCT	feminine gender ratio
AUX, ADV, DET	PRON, ADJ, PUNCT
NOUN, PUNCT, ADV	PRON, NOUN, AUX
NOUN, PRON, PUNCT	VERB, SYM
conj, flat:foreign, flat:foreign	ADP, VERB, ADP
DET, ADV	punct, punct, flat:name
masculine gender ratio	NOUN, PUNCT, ADJ
SCONJ, DET	advmod, conj, root
punct, flat, nmod	NOUN, ADP, AUX
mark, advcl, conj	nmod, appos, appos
ADP, ADV, ADP	PUNCT, ADV, SCONJ
X, X, NUM	VERB, SCONJ, PRON
DET, NOUN, DET	punct, aux:tense, root
NUM, PUNCT	first person ratio

Table 4: The 15 most predictive linguistic features for cross-lingual gender profiling.

verbs are more indicative of women" (Johannsen et al., 2015, p.107). While in our work we could not find an increased use of nouns for men, also for our data the VERB tag is more frequent in female features while the tags DET and NUM appear exclusively in male features.

5 Summary

The present paper compared two different systems for cross-lingual gender profiling. It was shown that XLM-RoBERTa generalizes best across languages and significantly outperforms the LFC. Nevertheless, the evaluation of linguistic features most relevant for gender profiling led to interesting results: We found that women use first-person pronouns more frequently and show an increased use of feminine gender marked words. Men cross-lingually use determiners and numerals more frequently, while the same is true for verbs and pronouns for women.

6 Future work

While the present work showed the feasibility of cross-lingual gender profiling via embedding based systems, we assume that the linguistic feature classifier has still not reached its full potential. A possible improvement regards the design of the dependency features. While in this paper we simply took sequences of consecutive relations into account, Johannsen et al. (2015) choose a different strategy: The authors took into account subtrees of up to three POS tokens with the relations linking them. These "treelets" might capture local syntactic variation better and, thus, lead to a higher predictive power.

Another worthwhile path to explore is to integrate a strategy for dealing with the fact that linguistic features were highly correlated. As Johannsen et al. (2015, p. 105) write, this can lead to a situation where "small and inessential variations in the dataset can determine which of the variables are selected to represent the group". The authors solve this problem by applying stability selection which trains a classifier multiple times with different sets of features and keeps only features which are consistently predictive over a certain number of models.

Lastly, this paper only compared the performance of linguistic features and embedding vectors independently. A combined system could have beneficial effects for the task, which is another promising direction to follow in the future.

7 Contributions

The present paper and the pertaining code were created in close collaboration of the authors. However, the authors in each phase had different focus points:

- Phase 1: In the first part of this seminar - the preparation phase - Faizan E Mustafa developed a Logistic Regression/Maximum Entropy Classifier and was responsible for implementing an evaluation procedure. Wolfgang Tessadri, in turn, created a pipeline to load and pre-/post-process the provided data (including a BOW feature extractor and a Sparse Vector class) and implemented a Perceptron classifier to be able to compare performances across classifiers.

- Phase 2: The second phase involved the realization of the cross-lingual systems described in this paper. In this phase, Faizan E Mustafa was responsible for putting into practice the XLM-RoBERTa system. Wolfgang Tessadri conceptualized potentially predictive linguistic features, implemented the linguistic feature extraction process as well as the classifier based on these features. In addition to the approaches described in the paper, we also implemented two additional models: a cross-lingual system using MUSE embeddings and a BOW model as a baseline which were not described here due to the page limit.

  Regarding the report and the code repository both authors participated evenly in their creation. However, Wolfgang Tessadri was responsible for conceptualizing and structuring the report as well as revising the final version, while Faizan E Mustafa took care of the final revision and structure of the GitHub repository.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346.

Mounica Arroju, Aftab Hassan, and Golnoosh Farnadi. 2015. Age, gender and personality recognition using tweets in a multilingual setting. In *6th Conference and Labs of the Evaluation Forum: Experimental IR meets multilinguality, multimodality, and interaction*, pages 22–31.

Na Cheng, Rajarathnam Chandramouli, and K. P. Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78–88.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, and Oscar Täckström. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Claudia Peersman, Walter Daelemans, and Leona van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Association for Computational Linguistics (ACL) System Demonstrations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Rob Van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 383–389.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637.

Appendices

A Linguistic features

Feature Type	Feature description	Example
i. POS tag		
unigrams	relative frequency of all POS-tags per user	NUM
bigrams	relative frequency of the 200 most frequent POS-tag bigram sequences per user	(VERB, NOUN)
trigrams	relative frequency of the 800 most frequent POS-tag trigram sequences per user	(DET, PROPN, VERB)
ii. Morphological properties		
word-lemma ratio	ratio of word tokens vs. word types per user	
indicative/subjunctive /imperative ratio	ratio of verbs in different grammatical moods per user	
first/second /third person ratio	ratio of personal pronouns in different grammatical persons per user	
feminine/masculine /neuter /common ratio	ratio of nouns/pronouns/adjectives with different grammatical gender per user	
iii. Dependency relations		
single path	relative frequency of alldependency relation labels per user	nsubj:pass
double path	relative frequency of the 200 most common two subsequent dependency relation labels per user	(nmod, obl)
triple path	relative frequency of the 800 most common three subsequent dependency relation labels per user	(det:poss, nsubj, root)
iv. Syntactic complexity		
average depth	average depth of all dependency trees of a user	
maximum depth	minimum depth of all dependency trees of a user	
minimum depth	maximum depth of all dependency trees of a user	
average length	average sentence length per user	
maximum length	maximum sentence length per user	
minimum length	minimum sentence length per user	

B Ratio of train and test users per language

	Nr. train users	Nr. test users
DE	280	71
FR	908	227
IT	328	82
NL	719	180

C Majority threshold

		TEST				
		DE	FR	IT	NL	Target Train
T	DE	*0.51/0.34*	0.57/0.36	0.65/0.39	0.52/0.34	0.56/0.36
R	FR	0.51/0.34	*0.57/0.36*	0.65/0.39	0.52/0.34	0.55/0.35
A	IT	0.51/0.34	0.57/0.36	*0.65/0.39*	0.52/0.34	0.54/0.35
I	NL	0.51/0.34	0.57/0.36	0.65/0.39	*0.52/0.34*	0.57/0.36
N	Target Test	0.51/0.34	0.57/0.36	0.65/0.39	0.52/0.34	*0.56/0.36*