Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification

Taja Kuzman
taja.kuzman@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Nikola Ljubešić
nikola.ljubesic@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

This study analyses the impact of several types of linguistic features on the task of automatic web genre identification applied to Slovene data. To this end, text classification experiments with the fastText models were performed on 6 feature sets: original lexical representation, preprocessed text, lemmas, part-of-speech tags, morphosyntactic descriptors, and syntactic dependencies, produced with the CLASSLA pipeline for language processing. Contrary to previous work, our results reveal that the grammatical feature set can be more beneficial than lexical representations for this task, as syntactic dependencies were found to be the most informative for genre identification. Furthermore, it is shown that this approach can provide insight into variation between genres.

KEYWORDS

language processing, linguistic features, automatic genre identification, web genres, Slovene

1 INTRODUCTION

Automatic genre identification (AGI) is a text classification task where the focus is on genres as text categories that are defined based on the conventional function and/or the form of the texts. In text classification tasks, texts are generally given to the machine learning models in form of words or characters that are then further transformed into numeric vectors by using bag-of-words representations, or word embeddings created by training deep neural networks on the surface text. However, recent development of tools for linguistic processing for numerous languages, including Slovene, allows transformation of the original running text into various other sets of features to which further transformation into numeric representations can be applied. By learning on these linguistic sets, we get insight into the importance of features that cannot be analysed separately when given the running text, i.e., word meaning, function of a word, and its relation to other words.

When previous work compared importance of various textual feature sets on the performance of the models in automatic genre identification, lexical features, i.e., word or character n-grams, mainly provided the best results ([6], [7]). However, it was noted that by learning on lexical features, the models could learn to classify texts based on the topic instead of genre characteristics, and would not be able to generalize beyond the dataset.

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia

As learning on lexical features can introduce bias towards topic, Laippala et al. (2021) recently experimented with combining lexical with grammatical features, which are represented as part-of-speech tags, conveying information on the word type (e.g., noun, verb). This showed to yield better results than using solely lexical features, and provided more stable models, i.e., models that are able to generalize beyond the training data. Furthermore, their analysis revealed that the importance of feature sets varies between genre categories, and that while some are most efficiently identified when learning on lexical features, others benefit more from grammatical representations.

However, these experiments were in past mostly performed on English datasets. This article is the first to analyse the impact of various feature sets on automatic genre identification applied to Slovene data. This research was made possible by the recent development of the first Slovene dataset, manually annotated with genre, as well as the creation of state-of-the-art language processing tools for Slovene. To compare textual representations, additional feature sets were created from a selection of texts annotated with genre, presented in Section 2, by using common preprocessing methods and language processing (see Section 3). Thus, in this paper, 6 textual representations are compared: 1) original, running text that we consider as our baseline, 2) preprocessed text, i.e. lowercase text without punctuation, digits and stopwords, 3) lemmas, i.e. base dictionary forms of words, 4) part-of-speech (PoS) tags, i.e. main syntactic word types (e.g., noun, verb), 5) morphosyntactic descriptors (MSD), i.e. extended PoS tags which include information on morphosyntactic features (e.g., number, case), 6) syntactic dependencies, i.e. types of dependency relations between words (e.g. subject, object). The feature sets are compared based on their impact on the performance of the fastText models on the automatic text classification task. The results of the experiments, presented in Section 4, give insights into the role of linguistic feature sets on this task and the differences in performance between genre categories.

2 DATASET

For performing experiments in automatic genre identification, the Slovene Web genre identification corpus GINCO 1.0 [2] was used. The dataset consists of the "suitable" subset, annotated with genre, and the "not suitable" subset that comprises texts which can be deemed as noise in the web corpora, e.g., texts without full sentences, very short texts, machine translation etc. In this research, only the "suitable" subset, containing 1002 texts, was used.

The GINCO schema consists of 24 genre labels. However, previous experiments, performed with the fastText model on the entire dataset, showed that the model is not potent enough to differentiate between a large number of labels that are mostly represented by less than 100 texts, reaching micro and macro

Table 1: The original GINCO categories (left) included in the reduced set, and the reduced set of labels (right), used in the experiments, with the total number of texts (later divided between the train, dev and test split) in the parentheses.

GINCO	Reduced Set
News/Reporting Opinionated News	News (198)
Information/Explanation Research Article	Information/Explanation (127)
Opinion/Argumentation Review	Opinion/Argumentation (124)
Promotion Promotion of a Product Promotion of Services Invitation	Promotion (191)
Forum	Forum (48)

F1 scores of 0.352 and 0.217 respectively (see [3]). Therefore, to be able to infer any meaningful conclusions, this article focuses only on the most frequent genre labels, created by merging some labels. Instances of less frequent labels that could not be merged, namely *Instruction*, *Legal/Regulation*, *Recipe*, *Announcement*, *Correspondence*, *Call*, *Interview*, *Prose*, *Lyrical*, *Drama/Script*, *FAQ*, and the labels *Other* and *List of Summaries/Excerpts*, which can be considered as noise, were not used. To focus only on the instances that are representative of their genre labels, texts that were manually annotated as hard to identify (parameter `hard`) were not used in the experiments. Furthermore, paragraphs that were deemed to be noise in the text, e.g., cookie consent text, and were marked by the annotators with the `keep` parameter set to `False`, were left out of the final texts.

Thus, the final set of labels, used in the experiments, shown in Table 1, consists of 5 genre categories, *Information/Explanation*, *News*, *Opinion/Argumentation*, *Promotion* and *Forum*. As shown in the Table, the dataset is imbalanced, with *News* and *Promotion* being the most frequent classes, consisting of almost 200 instances, while *Forum* is the least represented class, consisting of about 50 texts. The subset, consisting of 688 texts in total, followed the original stratified split of 60:20:20, encoded in the GINCO 1.0 dataset, and the models were trained on the training set, tested on the test set, while the dev split was used for evaluating the hyperparameter optimisation.

3 FEATURE ENGINEERING

Feature engineering is a process of identifying features that are most useful for a specific task with the goal of improving performance of a machine learning model. In text classification experiments, basic preprocessing methods are often used to reduce the number of unique lexical features (words or characters) without losing much information which could provide better results. To test whether preprocessing the text improves the results for this task, the first additional feature set was created by preprocessing the running text as extracted from the GINCO dataset. Preprocessing consisted of the following steps: converting text to lowercase, and removing digits, punctuation and function words known as stopwords, e.g., conjunctions, prepositions etc.

In addition to this, various linguistic representations were created by applying linguistic processing to the texts, and replacing words with corresponding lemmas or grammatical tags. The language processing was performed with the CLASSLA pipeline [5]. The following text representations were produced: lexical feature set, consisting of lemmas, and three grammatical feature sets: part-of-speech (PoS) tags, morphosyntactic descriptors (MSD), and syntactic dependencies. The realisation of the created feature sets is illustrated on an example sentence in Table 2.

4 MACHINE LEARNING EXPERIMENTS

4.1 Experimental Setup

The experiments were performed with the linear fastText [1] model which enables text classification and word embeddings generation. The model is a shallow neural network with one hidden layer where the word embeddings are created and averaged into a text representation which is fed into a linear classifier. The model takes as an input a text file where each line contains a separate text instance, consisting of a label and the corresponding document. Thus, for each feature set, appropriate train, test and dev files were created, and the model was trained on each representation separately [1]. To observe the dispersion of results, five runs of training were performed for each feature set. To measure the model's performance on the instance and the label level, the micro and macro F1 scores were used as evaluation metrics.

The hyperparameter search was performed by training the model on the training split of the baseline text and evaluating it on the dev split. The automatic hyperparameter optimisation provided by the fastText model did not yield satisfying results, as three runs of automatic hyperparameter optimisation produced very different results in terms of proposed optimal hyperparameter values and yielded micro F1 0.479 ± 0.02 and macro F1 0.382 ± 0.06. Therefore, we continued searching for optimal hyperparameters by manually changing one hyperparameter at a time

[1]The code for data preparation and machine learning experiments is published here: https://github.com/TajaKuzman/Text-Representations-in-FastText.

Table 2: An example of the feature sets used in the experiments.

Feature Set	Example
Baseline - Running Text	V Laškem se bo v nedeljo, 21.4.2013 odvijal prvi dobrodelni tek Veselih nogic.
Preprocessed Baseline	laškem nedeljo odvijal dobrodelni tek veselih nogic
Lemmas	v Laško se biti v nedelja , 21.4.2013 odvijati prvi dobrodelen tek vesel nogica .
PoS	ADP PROPN PRON AUX ADP NOUN PUNCT NUM VERB ADJ ADJ NOUN ADJ NOUN PUNCT
MSD	Sl Npnsl Px——y Va-f3s-n Sa Ncfsa Z Mdc Vmpp-sm Mlomsn Agpmsny Ncmsn Agpfpg Ncfpg Z
Dependencies	case nmod expl aux case obl punct nummod root amod amod nsubj amod nmod punct

and conducting classification experiments. The optimum number of epochs revealed to be 350, the learning rate was set to 0.7, and the number of words in n-grams to 1. For the other hyperparameters, the default values were used. Manual hyperparameter search revealed to be considerably more effective than automatic optimisation, as it yielded the average micro and macro F1 scores of 0.625 ± 0.004 and 0.618 ± 0.003 respectively, which is in average 0.15 points better micro F1 and 0.24 points better macro F1 compared to the results of automatic optimisation.

To analyse whether our choice of technology is the most appropriate one, we compared the performance of the fastText model, which uses the hyperparameters mentioned above, with the performance of various non-neural classifiers, commonly used in text classification tasks: dummy majority classifier which predicts the most frequent class to every instance, support vector machine (SVM), decision tree classifier, logistic regression classifier, random forest classifier, and Naive Bayes classifier. We used the default parameters for the classifiers. The models are compared based on their performance on the baseline text which was transformed into the TF-IDF representation where necessary. As shown in Table 3, fastText outperforms all other classifiers with a noticeable difference especially in the macro F1 scores, reaching 17 points higher scores than the next best classifier, the Naive Bayes classifier.

Table 3: Micro and macro F1 scores obtained by various classifiers, trained and tested on the baseline text.

Classifier	Micro F1	Macro F1
Dummy Classifier	0.24	0.08
Support Vector Machine	0.49	0.33
Decision Tree	0.34	0.35
Logistic Regression	0.52	0.38
Random Forest classifier	0.51	0.41
Naive Bayes classifier	0.54	0.42
FastText	**0.56**	**0.59**

4.2 Results of Learning on Various Linguistic Features

To explore the role of various textual representations on the automatic genre identification of Slovene web texts, we conducted text classification experiments with the fastText models on 6 feature sets:

- three lexical sets: a) baseline text, i.e., the original running text, b) preprocessed baseline text, i.e., baseline text converted to lowercase and without punctuation, digits and function words, c) lemmas, i.e., words reduced to their base dictionary forms;
- three grammatical sets: a) part-of-speech (PoS), i.e., main word types, b) morphosyntactic descriptors (MSD), i.e., extended PoS tags, c) syntactic dependencies, i.e., types of words defined by their relation to other words.

First, by comparing the baseline representation and the preprocessed representation, we aimed to determine whether common preprocessing methods can improve the results in the AGI task. As shown in Table 4, the results reveal that applying preprocessing methods improves the performance, especially on the micro F1 level. Analysis of the F1 scores obtained for each label in Figure 1 reveals that preprocessing especially improves the identification of *Promotion* and *News*. The two labels are the most frequent genre classes in the dataset which explains larger improvement of the micro F1 scores. If we compare the baseline text and the preprocessed text to the third lexical set, i.e., lemmas, the results show that by using lowercase words, reduced to their dictionary base form, the performance is further improved, although only slightly, as can be seen in Table 4.

Table 4: Average micro and macro F1 scores obtained from five runs of training and testing on each representation separately.

Representation	Micro F1	Macro F1
Baseline Text	0.560 ± 0.00	0.589 ± 0.00
Preprocessed Baseline	0.596 ± 0.00	0.597 ± 0.00
Lemmas	0.597 ± 0.01	0.601 ± 0.00
PoS	0.540 ± 0.01	0.547 ± 0.01
MSD	0.563 ± 0.01	0.536 ± 0.02
Dependencies	**0.610** ± 0.00	**0.639** ± 0.00

Secondly, we compared various lexical and grammatical feature sets, obtained with language processing tools. In previous work, which analysed English genre datasets, lexical features yielded better results than grammatical feature sets ([4], [6], [7]). Our results revealed that this conclusion holds also for Slovene when training on part-of-speech tags. Similar conclusion can be made for the extended part-of-speech tags (MSD) which only slightly improve the micro F1 scores compared to the baseline while there is a decrease in the macro F1 scores (see Table 4). However, the third grammatical feature set, consisting of tags for syntactic dependencies, which was not used in previous work, significantly outperformed the baseline text and all other feature sets. As shown in Figure 1, the improvement is especially noticeable for the categories *Forum*, *Opinion/Argumentation* and *News*. By learning on the dependencies instead on lexical features, the model learns from the structure of the sentences in the text, i.e., the syntax, instead of word meanings that can be more related to topic than genre, which could be the reason why this representation was revealed to be the most beneficial for the task.

As in previous work (see [4]), the experiments have revealed a dependence between the text representation and performance on specific genre labels, which is illustrated in Figure 1. The results show that *Promotion* and *Information/Explanation* can be most successfully identified when learning purely on the meaning of the words, i.e., on lemmas. In contrast to that, for identifying *News*, grammatical representations are more useful than lexical ones. Similarly, *Opinion/Argumentation* benefits more from grammatical feature sets than lexical representations, except in case of the MSD tags which significantly decreased the results for this class, yielding F1 scores below 0.3. Interestingly, although *Forum* is the least frequent label, its features seem to be the easiest to identify in the majority of representations. This genre benefits the most from learning on syntactic dependencies tags, which yielded F1 scores of almost 0.9.

5 CONCLUSIONS

In this paper, we have investigated the dependence of automatic genre classification on the lexical and grammatical representation of text. Our experiments, performed on three lexical and three

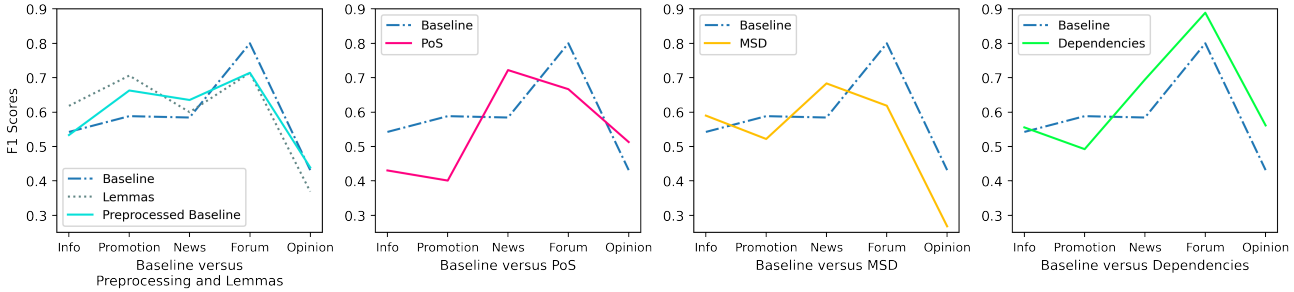

Figure 1: The impact of various linguistic features on the F1 scores of genre labels (*Information/Explanation, Promotion, News, Forum* and *Opinion/Argumentation*).

grammatical feature sets, revealed that the choice of textual representation impacts the results of automatic genre identification. Similarly to previous work, it was revealed that part-of-speech features give worse results than lexical features. However, a grammatical feature set, consisting of syntactic dependencies, that has not been studied in previous work, revealed to be the most beneficial for the automatic genre identification task. Furthermore, the experiments revealed variation between genres regarding the impact of feature sets on the F1 scores of each label. While some genres, such as *Promotion*, benefit more from learning on lexical features, others, such as *Opinion/Argumentation*, benefit more from grammatical representations.

However, it should be noted that this study has been limited to the 5 most frequent genre labels, as the previous experiments showed that the fastText model is not potent enough to identify other categories represented by a small number of instances ([3]). Thus, the results of these experiments give insight into which linguistic features are the most important for differentiating between the five most frequent genres, not for identifying the 24 original labels that encompass all the genre variation found on the web, and include noise. This is why we plan to continue genre annotation campaigns to enlarge the Slovene genre dataset, which would allow extending the analysis to all genre labels. In addition to this, as we are interested in cross-lingual genre identification, in the future, we plan to analyse the importance of linguistic feature sets on the Croatian and English genre datasets to analyse whether the characteristics of genre labels are language independent.

The fastText model was revealed to be useful for the analysis of the impact of linguistic features on the AGI task, however, previous work on automatic genre identification using the GINCO dataset revealed that if the aim of the research is to create the best-performing classifier and not to analyse the impact on the performance, the Transformer-based pre-trained language models are much more suitable for the task ([3]). This was also confirmed by our experiments on the running text, where the base-sized XLM-RoBERTa model reached micro and macro F1 scores 0.816 and 0.813, which is 22–26 points more than the fastText model. Based on the findings from this paper, one of the reasons why the Transformer models perform better could also be that the Transformer text representations incorporate information on syntax as well. In the future, we plan to investigate this further, adapting the classifier heads so that the syntactic information has a larger impact on the classification than the lexical parts of the representation.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

REFERENCES

[1] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[2] Taja Kuzman, Mojca Brglez, Peter Rupnik, and Nikola Ljubešić. 2021. Slovene web genre identification corpus GINCO 1.0. Slovenian language resource repository CLARIN.SI. (2021). http://hdl.handle.net/11356/1467.

[3] Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1584–1594. https://aclanthology.org/2022.lrec-1.170.

[4] Veronika Laippala, Jesse Egbert, Douglas Biber, and Aki-Juhani Kyröläinen. 2021. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language resources and evaluation*, 1–32.

[5] Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, (August 2019), 29–34. DOI: 10.18653/v1/W19-3704. https://www.aclweb.org/anthology/W19-3704.

[6] Dimitrios Pritsos and Efstathios Stamatatos. 2018. Open set evaluation of web genre identification. *Language Resources and Evaluation*, 52, 4, 949–968.

[7] Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: evaluating genre collections. In *LREC*. Citeseer.