SiKDD 2022 conference, Information Society 2022
October 10, 2022, Ljubljana

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

Taja Kuzman and Nikola Ljubešić

Jožef Stefan Institute, Jožef Stefan International Postgraduate School

macocu

Co-financed by the Connecting Europe Facility of the European Union

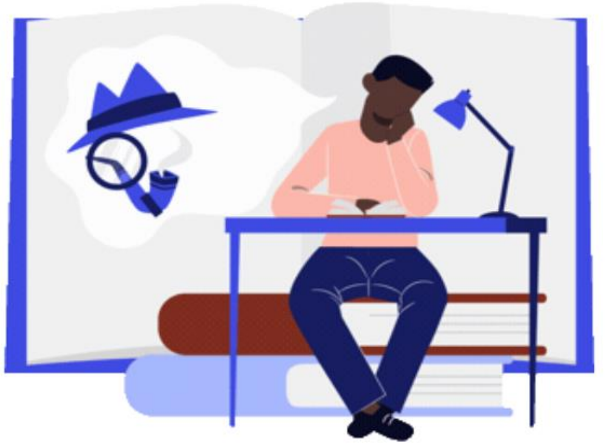
MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA SLOVENIJA LJUBLJANA 2004

Introduction

- Automatic genre identification = text classification task, focused on genres

- Genres = text categories based on author's purpose, function and form of the document
    - news article, recipe, legal texts, etc.

Goal



Various information is hidden in words in the running text: meaning, word function, relation with other words ...

Which of these signals are the most informative for identifying genres?

Method

- Text classification experiments
  - ML model: linear fastText classifier

  - Dataset: [Slovene Web genre identification corpus GINCO 1.0]
    - 5 largest genre classes used in the experiments
    - 688 texts, 60:20:20 training-dev-test stratified split
    - 6 representations, separate training and testing on each representation

Genre classes

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)

Feature sets

- 6 training and test datasets – 6 feature sets:
  1. **baseline text**: original running text as extracted from the dataset,
  2. **preprocessed text**: lowercase text without punctuation, digits and stopwords,
  3. **lemmas**: base dictionary forms of words,
  4. **part-of-speech (PoS) tags**: main word types (noun, verb)
  5. **morphosyntactic descriptors (MSD)**: extended PoS tags with information on number, case, person ...
  6. **syntactic dependencies**: types of dependency relations between words (subject, object)

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

Input to the classifier

upos-fasttext.train

```
__label__Information/Explanation NOUN NOUN ADJ NOUN
NOUN VERB DET NOUN AUX AUX ADJ NOUN NOUN ADP NOUN
PUNCT NOUN AUX PRON VERB ADP NOUN ADJ NOUN ADJ NOUN
ADP NOUN PUNCT CCONJ ADP NOUN CCONJ NOUN ADJ NOUN
CCONJ CCONJ ADP NOUN ADJ ADP ADJ NOUN ADP NOUN NOUN
PUNCT NOUN VERB DET NOUN AUX ADP NOUN NOUN ADJ NOUN
ADP ADJ NOUN NUM PUNCT NUM VERB NOUN ADP NOUN CCONJ
NOUN NOUN PROPN PUNCT NOUN AUX VERB NUM ADJ NOUN
PUNCT
__label__Promotion NOUN PROPN CCONJ NOUN PROPN NOUN
PROPN PRON AUX VERB NOUN ADJ NOUN NOUN NOUN PROPN
PUNCT SCONJ AUX ADP NOUN NOUN PROPN VERB NOUN PUNCT
SCONJ ADP ADJ NOUN ADP PROPN ADP PROPN PUNCT VERB
ADJ NOUN ADP NOUN NOUN NOUN PUNCT NOUN PUNCT NOUN
PUNCT NOUN PUNCT NOUN PUNCT NOUN ADP NOUN CCONJ
NOUN PUNCT ADJ NOUN PUNCT ADP DET NOUN VERB NUM ADJ
NOUN PUNCT ADP PRON ADJ NOUN NOUN PROPN SCONJ ADJ
ADJ NOUN PUNCT ADJ NOUN NOUN PROPN AUX NOUN ADJ
NOUN ADP PROPN ADP NOUN ADJ NOUN CCONJ ADJ NOUN ADJ
ADP DET NOUN ADP NOUN ADJ NOUN PUNCT ADJ PUNCT ADJ
NOUN PUNCT PUNCT ADV AUX PRON ADP NUM ADP NUM NOUN
NUM VERB NOUN CCONJ ADJ NOUN ADP PROPN PUNCT VERB
AUX NOUN DET ADJ NOUN NOUN ADJ ADP DET NOUN PUNCT
```

baseline_text-fasttext.train

```
__label__Information/Explanation JEDILNIK Iskalnik
Poglavitni cilj projekta Najdi svojo službo je bil
aktivno sodelovanje šole z gospodarstvom, dijaki so
se seznanili s smernicami gospodarskega razvoja
kraške regije v prihodnosti, ter z razvojem in
potrebami obstoječih podjetij in zato k usmerjanju
mladih k ciljnemu izobraževanju za potrebe
delodajalcev. Projekt Najdi svojo službo je v
okviru razpisa Skriti zaklad v šolskem letu 2003/04
odobrilo Ministrstvo za šolstvo in šport Republike
Slovenije. Projekt je trajal dve šolski leti.
__label__Promotion Projekt INNOVAge in zavod Oreli
Zavod Oreli se je odzval povabilu Razvojnega c
entra Srca Slovenije, ki je v okviru projekta
INNOVAge imelo nalogo, da na študijski obisk v
Helsinke na Finskem, pripelje zunanje strokovnjake
na področju oskrbe starostnikov, teleoskrbe,
e-zdravja, eko-inovacij v stanovanjih in hišah,
prilagojenih starostnikom. V tem projektu sodeluje
13 evropskih partnerjev, med njimi Razvojni center
Srca Slovenije kot edini slovenski partner. Glavni
cilj projekta INNOVage je prenos dobrih praks v
Evropi na področju aktivnega staranja in
```

Experimental setup

- Prepared 6 training and test files (6 feature sets):
    - applied preprocessing methods → preprocessed text
    - applied linguistic processing with the CLASSLA pipeline → lemmas, POS, MSDs and syntactic dependencies
- Hyperparameter search for fastText (evaluation on dev split):
    - automatic hyperparameter optimization did not yield satisfactory results: very different hyperparameter values, 0.48 micro F1, 0.38 macro F1.
    - manual hyperparameter search: changing one hyperparameter at a time (epochs, learning rates, number of words in n-grams) → much better scores: 0.63 micro F1, 0.62 macro F1.
- Training and testing with fastText on each of the 6 feature sets.

Comparison with other models

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**

Results: Lexical representations

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

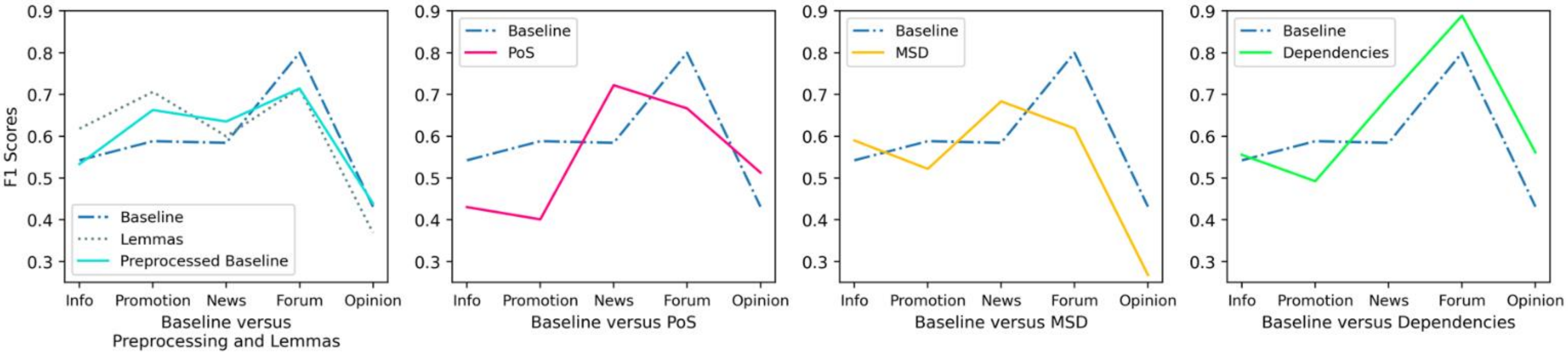
- Preprocessing improves results.
- Further improvements when using base dictionary forms of words (lemmas).

Results: Lexical versus Grammatical Features

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

- Syntactic dependencies provide the best results → model learns on the structure of the sentences in the text instead of word meanings (topic).

Results: Variation between genre classes



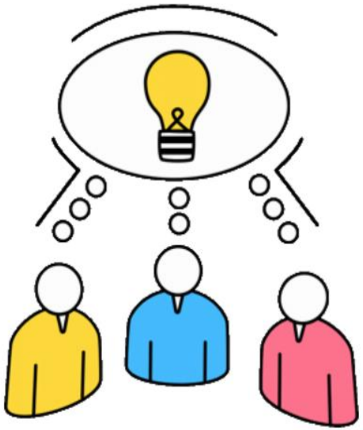
- Best feature set for:
 - Information/Explanation, Promotion: lemmas
 - News, Opinion/Argumentation, Forum: grammatical representations

- Forum – best scores, although it is the least frequent.

Conclusions

- The choice of textual representation does impact the results of automatic genre identification.
- POS tags result in worse performance than lexical features (as in previous work, performed on English).
- The most beneficial textual representation: syntactic dependencies (not studied in previous work).
- Variation between genres: some benefit more from lexical features, other from grammatical.

Further work



- Combining multiple features sets.
- Analysis of English and Croatian dataset: are characteristics of genres language-independent?
- Transformer-based models significantly outperform fastText (XLM-RoBERTa: 0.22-0.26 micro/macro F1 scores higher):

→ adapt classifier's heads so that syntactic information has larger impact on classification.

→ experimenting how outputs of different layers effect the classification results.

https://github.com/TajaKuzman/Text-Representations-in-FastText

Thank you!

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).

Co-financed by the Connecting Europe Facility of the European Union