Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of Registers

Liina Repo[*†] **Valtteri Skantsi**[*°†] **Samuel Rönnqvist**[*] **Saara Hellström**[*]
Miika Oinonen[*] **Anna Salmela**[*] **Douglas Biber**[‡] **Jesse Egbert**[‡]
Sampo Pyysalo[*] **Veronika Laippala**[*]

[*]University of Turku [°]University of Oulu [‡]Northern Arizona University
[*]{tlkrep,valtteri.skantsi,saanro,sherik,mhtoin,annsaln,sampo.pyysalo,mavela}@utu.fi
[°]{valtteri.skantsi}@oulu.fi [‡]{douglas.biber,jesse.egbert}@nau.edu

Abstract

We explore cross-lingual transfer of register classification for web documents. Registers, that is, text varieties such as blogs or news are one of the primary predictors of linguistic variation and thus affect the automatic processing of language. We introduce two new register-annotated corpora, FreCORE and SweCORE, for French and Swedish. We demonstrate that deep pre-trained language models perform strongly in these languages and outperform previous state-of-the-art in English and Finnish. Specifically, we show 1) that zero-shot cross-lingual transfer from the large English CORE corpus can match or surpass previously published monolingual models, and 2) that lightweight monolingual classification requiring very little training data can reach or surpass our zero-shot performance. We further analyse classification results finding that certain registers continue to pose challenges in particular for cross-lingual transfer.

1 Introduction

Text genre or *register* (Biber, 1988), such as discussion forum, news article or poem, is one of the most important predictors of linguistic variation (Biber, 2012). Thus, register affects crucially also the automatic processing of language (Mahajan et al., 2015; Webber, 2009; Van der Wees et al., 2018). Yet, despite its importance, register information is not available in web-crawled datasets that are widely used e.g. for pre-training language models in modern NLP. This is a challenge, as better structured language resources would also enable more detailed understanding and more sophisticated use of this data.

While web register identification would allow better realization of the potential offered by web-crawled datasets, most previous web register identification studies have been limited by skewed datasets, low performance, and near-exclusive focus on English. For example, Asheghi et al. (2014) and Pritsos and Stamatatos (2018) reported comparatively strong results, but their evaluations were based on datasets representing only a subset of the registers found online. With the CORE corpus, Egbert et al. (2015) were the first to present a dataset featuring the full extent of registers found on the open, searchable English web. While Biber and Egbert (2016b) demonstrated the possibility of automatic register classification using Stepwise Discriminant Analysis, improvements in modeling and more efficient methods remained necessary in order to reach practical levels of performance.

A challenge in modeling web registers is that web documents drawn from the unrestricted web do not always fit discrete classes but could rather be described in a continuous space (Biber and Egbert, 2018; Sharoff, 2018). Not all documents have clear characteristics of one single register, or even any register at all. This has shown also in relatively low inter-annotator agreement for web register annotation (Crowston et al., 2010).

Very recently, however, the advances brought to NLP by neural networks have shown that registers can be identified also in a corpus featuring the full range of online language variation (Laippala et al., 2020a). Laippala et al. (2019) extended the possibilities of web register identification beyond English by presenting an online register corpus on Finnish (FinCORE) and demonstrating that web registers can be modeled also in a cross-lingual setting.

In this paper, we substantially extend on this early work on cross-lingual web register identification through the following contributions: 1) we

[†] The marked authors contributed equally to this paper.

Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, pages 183–191
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

General register category	English	Finnish	French	Swedish
NA Narrative	36.46 %	34.95 %	22.33 %	28.32 %
IN Informational description	19.24 %	17.03 %	20.74 %	27.68 %
OP Opinion	16.23 %	15.23 %	6.33 %	6.60 %
ID Interactive discussion	6.77 %	6.29 %	8.03 %	3.57 %
HI How-to/Instructions	3.08 %	6.47 %	3.08 %	2.80 %
IP Informational persuasion	2.75 %	20.04 %	24.15 %	16.82 %
LY Lyrical	1.32 %	0.00 %	0.33 %	0.14 %
SP Spoken	1.21 %	0.00 %	0.83 %	0.14 %
Empty	1.20 %	0.00 %	0.00 %	0.00 %
Hybrids	11.74 %	0.00 %	14.19 %	13.93 %
Total	**48452**	**2226**	**1818**	**2182**

Table 1: Proportional register distribution and total number of documents in CORE, FinCORE, FreCORE and SweCORE. Hybrids include all documents annotated with several register labels, and Empty refers to documents not assigned any label.

introduce manually annotated web register datasets for two new languages, French and Swedish, 2) we demonstrate competitive performance for cross-lingual transfer of a register classification model from English to other languages in a zero-shot setting, and 3) we analyze zero-shot vs. monolingual training for register classification and remaining challenges in both. In particular, using Transformer-based pre-trained language models, we show that a zero-shot cross-lingual approach outperforms monolingual results achieved by a previously proposed state-of-the-art method for all the three language pairs (En-Fr, En-Sv, and En-Fi), and that strong monolingual performance can be achieved with limited training data.

2 Data

We use four register-annotated corpora representing the unrestricted open web: the English CORE and Finnish FinCORE, which have been introduced in previous work (Egbert et al., 2015; Laippala et al., 2019), and two new corpora, FreCORE for French and SweCORE for Swedish. These novel datasets are released under open licences together with this paper.[1] With these new resources, the possibilities for web register identification expand substantially.

FreCORE and SweCORE are random samples of the 2017 CoNLL datasets (Ginter et al., 2017) originally drawn from Common Crawl. Both datasets were deduplicated using Onion (Pomikálek, 2011) with 0.7 threshold and n-gram length of 5. All material not belonging to the body of text, such as boilerplate, was removed. Titles, however, were preserved. The cleaning and pre-processing steps follow the procedure suggested in Laippala et al. (2020b). The register annotation of the datasets was conducted individually by two trained annotators with a linguistics background. Uncertain cases were discussed and resolved together with an annotation supervisor. The inter-annotator agreement, counted prior to the discussions, was 78% F1-score for FreCORE and 84% for SweCORE. This can be considered as a lower bound.

All datasets are similarly annotated across languages, and they all apply the same hierarchical register class taxonomy originally introduced for CORE. It includes eight main registers (e.g., Narrative) and approximately 30 sub-registers (e.g., News report within Narrative). The main and sub-register categories are illustrated in the appendix. When a document shares characteristics of several registers, it can be assigned several labels both at the main and sub-register level. These documents are called *hybrids*. As our focus in this paper is on general register categories, we initially pre-process all four corpora to remove the more specific sub-register labels.

The general register categories and their distributions as well as the average document length and standard deviation for all classes are presented in Table 1 and Table 2, respectively. The register class Empty consists of texts whose register the annotators could not agree on. Due to the very small number of each type of hybrid label combination in the data, in Tables 1 and 2, the class Hybrids includes all documents that have more than one label. Table 1 reveals that the register distributions in the four languages are broadly similar, featuring Narrative, Informational description, and

[1] Available at https://github.com/TurkuNLP/Multilingual-register-corpora

Register	English		Finnish		French		Swedish	
	mean	std.	mean	std.	mean	std.	mean	std.
NA	1081	2490	649	2170	623	2284	602	2461
IP	1066	3370	301	391	325	493	426	2225
IN	1353	3373	989	4755	1446	9688	323	626
OP	1595	4021	739	1188	857	1835	1055	1825
HI	1007	1402	277	285	623	1130	437	508
ID	1079	4042	2017	8907	970	1579	577	885
LY	468	1114	-	-	387	314	263	225
SP	2047	3335	-	-	999	939	525	178
Empty	13345	3215	-	-	-	-	-	-
Hybrids	1290	3141	-	-	1170	3296	859	1207
All	1083	2747	713	3295	703	3900	482	1446

Table 2: Average length (number of words) and standard deviation of Finnish, French, Swedish and English documents.

hybrids among the four most frequent categories. The top four also include Informational persuasion in FinCORE, FreCORE, and SweCORE, while in CORE this label is relatively infrequent. Additionally, Opinion is notably more frequent in CORE and FinCORE than in FreCORE and SweCORE. These differences may reflect differences in data compilation. Table 2 shows that, on average, English documents are longer than documents in other languages, whereas Swedish documents tend to be shortest. Overall the number of words in a document in most of the classes show large variation, with the longest documents containing tens of thousands of words.

3 Experimental setup

The architectures and models we are using are presented below.[2] We perform multi-label document classification, where each document can have zero, one, or several register labels. The experiments are divided into 1) a monolingual setup with training and evaluation on Finnish, French, Swedish, and English (as reference), and 2) a zero-shot cross-lingual setup with training on English and evaluation on the other languages.

BERT, Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) is a state-of-the-art deep bidirectional language model pretrained on large unlabelled corpora. BERT's architecture is a multi-layer Transformer encoder that is based on the original Transformer architecture introduced by Vaswani et al. (2017). We use cased BERT models (TensorFlow versions) through the Huggingface Transformers library (Wolf et al., 2020) with the following language-specific models: the original English BERT, Finnish FinBERT (Virtanen et al., 2019), French FlauBERT (Le et al., 2020) and Swedish KB-BERT (Malmsten et al., 2020). Additionally, we use Multilingual BERT (mBERT) (Devlin et al., 2019), which was pretrained on monolingual Wikipedia corpora from 104 languages with a shared multilingual vocabulary.

XLM-RoBERTa (XLM-R, Conneau et al. (2020)) is a multilingual language model that follows the Cross-lingual Language Modeling (XLM) approach (Conneau and Lample, 2019) and is based on the RoBERTa model (Liu et al., 2019), which shares the architecture of BERT. The authors argue that XLM and mBERT are undertuned and that the improved and prolonged training procedure of RoBERTa in combination with more data – on average two orders of magnitude more for low-resource languages – is key to improving cross-lingual performance. XLM-R is trained on 2.5TB of filtered Common Crawl (Wenzek et al., 2020) data comprising of monolingual texts in 100 languages. It is claimed to be the first multilingual model to outperform monolingual models, as well as Multilingual BERT in a number of experiments (Conneau et al., 2020; Libovický et al., 2020; Tanase et al., 2020).

We also apply a **CNN** (Convolutional Neural Network) based architecture following Kim (2014), as our baseline model. We modify the cross-lingual CNN used by Laippala et al. (2019) to a multi-label setting. We use the multilingual word vectors introduced by Conneau et al. (2018). The CNN employs a convolution layer with ReLU activation,

[2] The code is available at: `https://github.com/TurkuNLP/Multilingual-register-corpora`

	Monolingual					Cross-lingual				
Model	Train-	Dev		Test		Train-	Dev		Test	
	Test	F1 (%)	Std.	F1 (%)	Std.	Test	F1 (%)	Std.	F1 (%)	Std.
CNN	Fi	59.04	(0.67)	58.04	(1.02)	En-Fi	40.53	(1.11)	41.56	(0.20)
mBERT	Fi	65.91	(0.85)	64.83	(1.16)	En-Fi	51.02	(2.92)	50.21	(0.74)
XLM-R large	Fi	76.25	(0.45)	**73.18**	(1.35)	En-Fi	**61.60**	(2.01)	**61.35**	(1.26)
FinBERT	Fi	**76.28**	(1.23)	72.98	(0.74)					
CNN	Fr	59.78	(1.10)	58.14	(1.10)	En-Fr	46.44	(0.51)	46.78	(1.80)
mBERT	Fr	70.74	(1.67)	68.66	(0.63)	En-Fr	56.73	(1.54)	55.04	(0.66)
XLM-R large	Fr	**77.38**	(0.51)	**76.92**	(0.24)	En-Fr	**65.66**	(0.52)	**64.27**	(1.58)
FlauBERT large	Fr	73.93	(0.93)	72.56	(1.40)					
CNN	Sv	69.43	(0.56)	67.89	(1.01)	En-Sv	43.74	(0.82)	43.78	(1.00)
mBERT	Sv	76.91	(0.45)	76.43	(0.46)	En-Sv	62.37	(0.82)	62.53	(0.78)
XLM-R large	Sv	**82.61**	(0.37)	**83.04**	(0.62)	En-Sv	**70.49**	(0.58)	**69.22**	(1.66)
KB-BERT	Sv	80.15	(0.50)	80.75	(0.09)					
CNN	En	64.56	(0.78)	64.03	(0.30)					
mBERT	En	72.80	(0.21)	73.06	(0.09)					
XLM-R large	En	**75.80**	(0.12)	**75.68**	(0.05)					
BERT large	En	74.01	(0.42)	74.07	(0.28)					

Table 3: Monolingual and zero-shot cross-lingual classification results (N=3). Best results for each experiment shown in bold.

a max-pooling layer and sigmoid activation.

The French and Swedish data were divided into training, development and test sets using stratified sampling with a 50/20/30 split. For BERT-based models we used large model size when available to maximize model performance. We used the maximum sequence length of 512 tokens (with truncation at the end) and batch size of 7, and performed a grid search on learning rate (8e^{-6}–6e^{-5}) and number of training epochs (3–7). For the CNN, we performed a grid search on the kernel size (1–2), learning rate (1e^{-4}–1e^{-2}), and prediction threshold (0.4, 0.5, 0.6).

4 Results

In Table 3, we present the primary results on English, Finnish, French and Swedish monolingual classification with the models described in Section 3, as well as cross-lingual results with English as the source language and Finnish, French and Swedish as target languages. We report the mean and standard deviation of F1 over three repetitions.

In monolingual settings, XLM-R large performs competitively compared to monolingual models and clearly outperforms both mBERT and the CNN baseline. The lead of XLM-R over monolingual models is substantial in all cases except for the FinBERT model, where the two perform within one standard deviation of each other. Our results support the claimed competitiveness of XLM-R large with monolingual models, mentioned in Section 3.

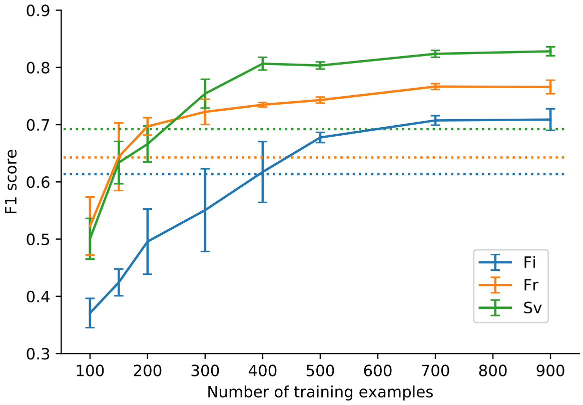

Figure 1: Monolingual performance when training with varying number of examples (solid lines) in relation to zero-shot cross-lingual performance when training on full English set (dotted lines). Error bars represent standard deviations (N=6).

English, Finnish and French BERT models achieve similar monolingual test results (73–74% F1-score), while the Swedish KB-BERT achieves the highest F1-score (81%). The Finnish classification task is seemingly easier due to smaller number of classes, nevertheless, other factors may cause the difficulty of the task to differ between languages. For instance, the measured human inter-annotator agreements at 78% (Fr) and 84% (Sv) F1-score (see Section 2) represent a theoretical upper bound for the classification task and reflect the tendency of

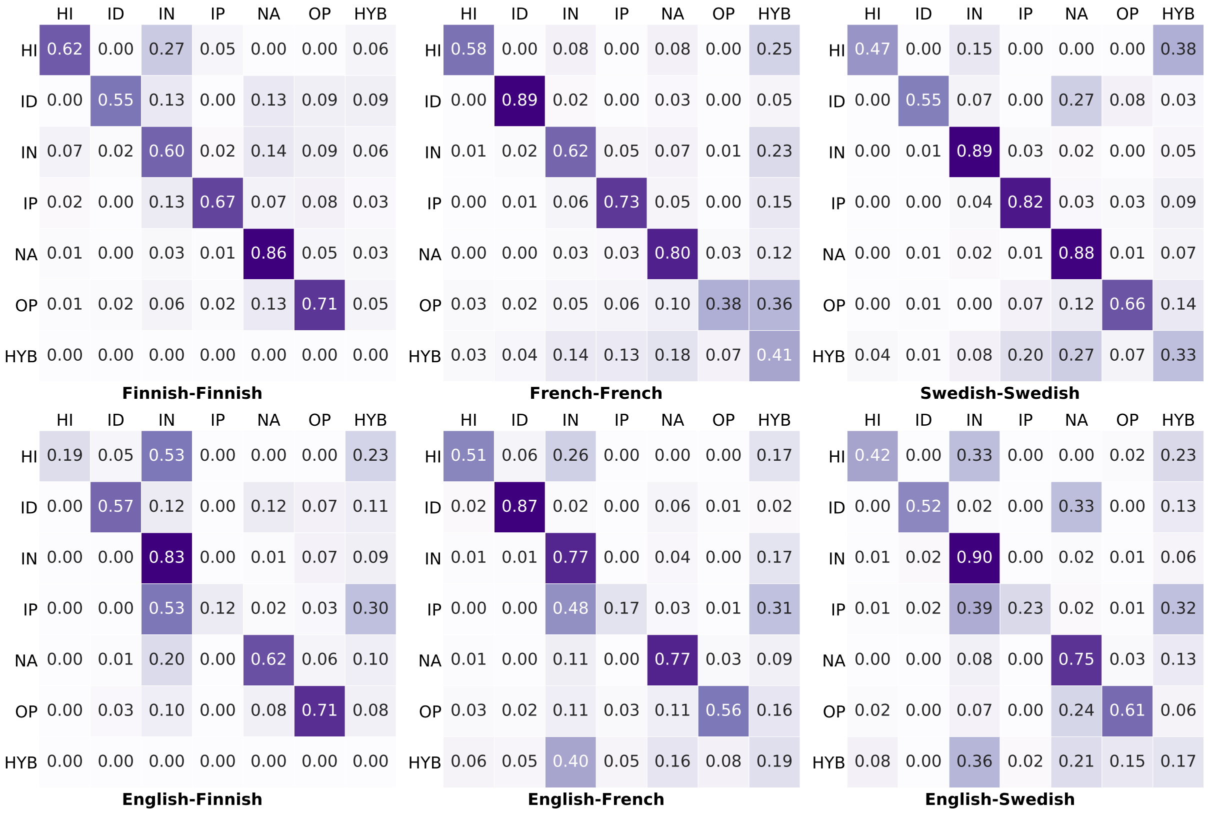

Figure 2: Confusion matrices presenting class label observations (rows) vs. class label predictions (columns) in monolingual (upper row) and cross-lingual (lower row) settings. The numbers and coloring represent the proportions of predictions per row. HYB is a combination of all hybrid cases with multiple labels.

Swedish being easier to classify; the level of agreement has not been reported for Finnish. Although not strictly comparable, our results clearly outperform the previous state-of-the-art results achieved with the CNN (Laippala et al., 2019) in terms of F1, which in turn outperforms Biber and Egbert (2016b), who used the same corpus but in multiclass setting.

Furthermore, Table 3 shows very strong zero-shot cross-lingual results with XLM-R large, with F1-scores in the 61–69% range. This represents a remarkably consistent relative decrease of 16.2–16.6% (11.8–13.8% absolute) from the monolingual scores of XLM-R. Its lead over mBERT increases from 6.6–8.4% absolute F1 to 7.8–11.4% in the cross-lingual settings, whereas its lead over the CNN goes from 15.1–18.8% to 17.5–25.4%. Most interestingly, the zero-shot XLM-R even beats the monolingually trained CNN baselines by a significant margin for Finnish and French, while its lead remains within a standard deviation for Swedish.

In Figure 1, we illustrate the effect of training monolingual XLM-R large models with varying train set sizes and compare the performance against the reported zero-shot performance. The optimal monolingual hyperparameter settings for each language are used, while training the model instances on 100–900 examples each. We see that zero-shot cross-lingual performance is surpassed already with about 150 training instances for French, 225 for Swedish and 400 for Finnish, while performance seems to converge around 500.

Previous studies have shown repeatedly that registers vary considerably in terms of how well they are linguistically defined and thus how well they can be automatically identified (Biber and Egbert, 2018, 2016a; Laippala et al., 2020a). For instance, while texts in the IN (Informational description) and NA (Narrative) classes, such as Encyclopedia articles and Sports reports, have very distinctive characteristics and can be identified with a very high reliability, others, such as Information blogs in the IN class or Advice in the OP (Opinion) class receive much lower scores.

Figure 2 presents confusion matrices on the predictions in monolingual and cross-lingual settings, using the best-performing model.[3] For the sake of simplicity, the multi-label predictions have been

[3] See appendix for class-specific F1 results.

collapsed into multi-class by including all hybrids under one label HYB in Figure 2. In the monolingual settings, we can see that particularly hybrids present a challenge. This is expected, as they feature characteristics of several registers. Additionally, while IP (Informational persuasion) and NA are predicted with high performance in all three languages, the other classes display more variation. For instance, ID (interactive discussion) reaches an F1-score of 90% (see appendix) in French monolingual setting, whereas in Swedish and Finnish it is frequently misclassified, most likely because of the small number of examples in the training data.

The hybrids are also frequently misclassified in cross-lingual settings. Interestingly, register classes also feature clear differences in the extent to which the cross-lingual transfer affects the identification performance. The register class IN tends to be predicted strongly in all zero-shot language pairs. This is probably due to the IN class including documents with strong cross-lingual signals. For instance, IN includes Encyclopedia articles (see appendix), such as Wikipedia texts, that tend to be very similar across languages.

While most of the non-hybrid classes experience a small drop in performance, the identification rate for IP and HI (How-to/Instructions) drops dramatically in cross-lingual settings in all language pairs. The decrease of IP can be linked to its smaller proportion in the English data (see Section 2), but the drops experienced by IP and HI can also reflect the variation displayed by registers across languages. Biber (2014) showed that registers, such as spoken texts, display functional similarities across languages, which obviously is needed for high-quality transfer in register identification. However, analyzing the English CORE registers, Laippala et al. (2020a) noted that some registers, such as many blogs, depend highly on lexical characteristics reflecting the discussion topics. These topics, however, may vary extensively between languages. This, again, may complicate the transfer learning for these classes.

5 Discussion and conclusions

Despite the many opportunities that reliable recognition of text register would introduce for the analysis and use of web documents and many efforts to address this task over the years, only limited progress has been made toward unrestricted web document register classification. Previous work has also focused almost exclusively on English.

In this study, we have introduced manual register annotation compatible with that of the large English CORE corpus for two languages previously lacking such a resource, namely French and Swedish. We also demonstrated that state-of-the-art multilingual neural language models can support zero-shot transfer of register annotations from English to a Germanic, Romance and Finnic language at levels of performance broadly comparable or better to previously published monolingual results on CORE.

Moreover, we demonstrated that small amounts of monolingual training data are needed to reach or surpass this level of performance, which attests that reliable register identification in a new language is readily attainable using current pre-trained language models. We further compared and analysed the results for monolingual and cross-lingual register classifiers, finding that certain registers as well as hybrid texts combining several register characteristics continue to pose challenges in particular for cross-lingual transfer. In future work, we will build on these results to extend multi- and cross-lingual modeling in order to create massive multilingual register-annotated web corpora.

Acknowledgments

We thank for the financial support of the Emil Aaltonen Foundation and Academy of Finland. We also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Rezapour Noushin Asheghi, Katja Markert, and Serge Sharoff. 2014. Semi-supervised graph-based genre classification for web pages. In *Proceedings of TextGraphs-9*, pages 39–47. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37.

Douglas Biber. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in contrast*, 14(1):7–34.

Douglas Biber and Jesse Egbert. 2016a. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.

Douglas Biber and Jesse Egbert. 2016b. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:3–36.

Douglas Biber and Jesse Egbert. 2018. *Register variation online*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *CoRR*, abs/1710.04087.

Kevin Crowston, Barbara Kwaśnik, and Joseph Rubleske. 2010. Problems in the use-centered development of a taxonomy of web genres. In *Genres on the Web*, pages 69–84. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66:1817–1831.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - Automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Veronika Laippala, Jesse Egbert, Douglas Biber, and Aki-Juhani Kyröläinen. 2020a. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language resources and evaluation*.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297. Linköping University Electronic Press.

Veronika Laippala, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi, and Sampo Pyysalo. 2020b. From web crawl to clean register-annotated corpora. In *Proceedings of the 12th Web as Corpus Workshop*, pages 14–22. European Language Resources Association.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490. European Language Resources Association.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. arXiv preprint arXiv:2004.05160.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Anuj Mahajan, Sharmistha Jat, and Shourya Roy. 2015. Feature selection for short text classification using wavelet packet transform. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 321–326. Association for Computational Linguistics.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden – making a Swedish BERT. arXiv preprint arXiv:2007.01658.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics.

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

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 1(13):65–95.

Mircea-Adrian Tanase, Dumitru-Clementin Cercel, and Costin-Gabriel Chiru. 2020. UPB at SemEval-2020 task 12: Multilingual offensive language detection on social media by fine-tuning a variety of BERT-based models. arXiv preprint arXiv:2010.13609.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. Curran Associates, Inc.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682. Association for Computational Linguistics.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2018. Evaluation of machine translation performance across multiple genres and languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4003–4012. European Language Resources Association.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

A Appendix

Tables 4 and 5 present the detailed results for the zero-shot cross-lingual and monolingual register classification experiments, respectively. Table 6 presents the register taxonomy with the main registers and their sub-registers.

	En-Fi		**En-Fr**		**En-Sv**	
	F1	**Std.**	**F1**	**Std.**	**F1**	**Std.**
HI	48.43 %	1.98 %	55.12 %	5.65 %	62.91 %	0.60 %
ID	69.79 %	6.06 %	87.48 %	2.05 %	52.05 %	3.07 %
IN	44.43 %	0.32 %	58.68 %	0.17 %	68.81 %	0.20 %
IP	52.79 %	5.72 %	53,57 %	2.53 %	51.45 %	1.88 %
LY	0.00 %	0.00 %	66.67 %	0.00 %	95.24 %	6.73 %
NA	77.85 %	0.80 %	75.18 %	0.32 %	78.36 %	0.70 %
OP	70.32 %	1.59 %	59.26 %	1.51 %	60.57 %	0.32 %
SP	0.00 %	0.00 %	79.08 %	7.53 %	0.00 %	0.00 %

Table 4: Class-wise F1-scores and standard deviations on cross-lingual experiments

	Fi-Fi		**Fr-Fr**		**Sv-Sv**	
	F1	**Std.**	**F1**	**Std.**	**F1**	**Std.**
HI	64.02 %	1.94 %	58.81 %	0.59 %	70.70 %	4.56 %
ID	66.18 %	3.54 %	90.37 %	1.58 %	60.48 %	3.21 %
IN	58.68 %	1.59 %	74.00 %	0.40 %	87.79 %	0.29 %
IP	75.74 %	2.34 %	80.02 %	1.04 %	81.75 %	1.10 %
LY	–	–	66.67 %	0.00 %	0.00 %	0.00 %
NA	82.38 %	0.98 %	77.02 %	1.16 %	86.66 %	0.69 %
OP	67.10 %	2.05 %	66.23 %	3.08 %	75.37 %	1.66 %
SP	–	–	65.28 %	1.96 %	0.00 %	0.00 %

Table 5: Class-wise F1-scores and standard deviations on monolingual experiments

Narrative
News report / news blog, sports report, personal blog, historical article, fiction, travel blog, community blog, online article
Informational description
Description of a thing, encyclopedia article, research article, description of a person, information blog, FAQ, course material, legal terms / condition, report, job description
Opinion
Review, opinion blog, religious blogs/sermon, advice
Interactive discussion
Discussion forum, question-answer forum
How-to/Instructions
How-to/instruction, recipe
Informational Persuasion
Description with intent to sell, news+opinion blog / editorial
Lyrical
Songs, poem
Spoken
Interview, formal speech, TV transcript

Table 6: All register classes. Main registers are shown in bold.