Language Resources and Evaluation
https://doi.org/10.1007/s10579-022-09624-1

ORIGINAL PAPER

Register identification from the unrestricted open Web using the Corpus of Online Registers of English

Veronika Laippala[1] · Samuel Rönnqvist[1] · Miika Oinonen[1] · Aki-Juhani Kyröläinen[1] · Anna Salmela[1] · Douglas Biber[2] · Jesse Egbert[2] · Sampo Pyysalo[1]

Accepted: 4 October 2022

Abstract

This article examines the automatic identification of Web registers, that is, text varieties such as news articles and reviews. Most studies have focused on corpora restricted to include only preselected classes with well-defined characteristics. These corpora feature only a subset of documents found on the unrestricted open Web, for which register identification has been particularly difficult because the range of linguistic variation on the Web is known to be substantial. As part of this study, we present the first open release of the Corpus of Online Registers of English (CORE), which is drawn from the unrestricted open Web and, currently, is the largest collection of manually annotated Web registers. Furthermore, we demonstrate that the CORE registers can be automatically identified with competitive results, with the best performance being an F1-score of 68% with the deep learning model BERT. The best performance was achieved using two modeling strategies. The first one involved modeling the registers using propagated register labels, that is, repeating the main register label along with its corresponding subregister label in a multilabel model. In the second one, we explored how the length of the document affects model performance, discovering that the beginning provided superior classification accuracy. Overall, the current study presents a systematic approach for the automatic identification of a large number of Web registers from the unrestricted Web, hence providing new pathways for future studies.

Keywords Web register identification · Web-as-corpus · Deep learning · Document classification

✉ Veronika Laippala
veronika.laippala@utu.fi

[1] University of Turku, Turku, Finland

[2] Northern Arizona University, Flagstaff, AZ, USA

Published online: 26 October 2022

1 Introduction

The automatic identification of Web genres or registers, such as news articles, blogs and advertisements (Biber, 1988), has been a challenge for natural language processing (NLP) methods since the introduction of early Web corpora (Baroni et al., 2009). Web corpora are commonly composed of billions of words that are automatically crawled from the Web; these corpora have been used for significant advances in a large number of NLP tasks, ranging from word-level analyses (Mikolov et al., 2013a; Pennington et al., 2014) to analyzing meaning in context (Peters et al., 2018). However, register information is not typically available for these resources, which can affect their automatic processing (Maharjan et al., 2018; van der Wees et al., 2015; Webber, 2009). The inclusion of information pertaining to register and register variation is important for this type of documents because a register is one of the most important predictors associated with linguistic variation (Biber, 2012). Thus, the inclusion of register information can benefit not only tasks in NLP, but also increase the utility of Web corpora in general.

Automatic Web register identification is exceptionally challenging because of two issues. First, the Web displays a wide range of linguistic variation, and there are no annotated language resources that represent this variation in its totality. Indeed, the Web register corpora used in previous studies have usually consisted of predetermined, restricted sets of registers and documents that have been selected to represent these categories. Thus, register identification systems have targeted only these artificially restricted subsets. However, the unrestricted open Web has a much wider range of linguistic variation than these restricted datasets that are based on predetermined categories. On the unrestricted Web, there are no gatekeepers to ensure that documents follow the norms or traditions concerning the guiding principles of specific registers; thus, not all documents have the distinct characteristics of a single register or any register at all (Egbert et al., 2015; Santini, 2007). Second, linguistic analyses have shown that the number of registers on the Web is very high and that the registers can also display a wide range of inner variation (Biber & Egbert, 2018; Biber et al., 2020). All these properties make automatic register identification more difficult for the unrestricted open Web than for manually compiled, restricted corpora composed of preselected and well-defined registers.

Existing Web register identification systems have tackled the range of linguistic variation by modeling registers based on proportions of dimensions (such as *argumentative*) that reflect the characteristics of individual registers (Sharoff, 2018); by approaching the task as an open-class problem, in which not all documents need to fit predetermined categories (Pritsos & Stamatatos, 2018); and by suggesting to analyze registers in a continuous space, where *hybrid* documents combine the characteristics of several registers (Biber & Egbert, 2018; Biber et al., 2020). Furthermore, Egbert et al. (2015) created CORE, the first corpus with manual register annotations based on an unrestricted sample of the freely available English-speaking Web. CORE covers the unrestricted range of registers

and linguistic variation found on the open, searchable Web. Thus, it allows for the development a robust register identification system that is not artificially restricted to a subset of documents and registers found on the Web. Because of the difficulty of the task, however, the accuracy of register identification using CORE has remained low (Biber & Egbert, 2016b).

In the current article, we study the automatic identification of Web registers in CORE using state-of-the-art deep learning methods. As the first contribution, we also release the full CORE corpus under an open license. CORE consists of nearly 50,000 documents. This number of documents provides an excellent basis for modeling registers on the unrestricted open Web. The register taxonomy covers eight main registers and tens of subregister categories and was developed during the annotation process so that it covers the full extent of registers and subregisters in the data (Egbert et al., 2015). The annotation was performed so that each document was annotated by four individual coders, who independently assigned each document a main register label and, if possible, a subregister label that described the document in more detail. For some documents and register combinations, the coders' votes were systematically split between several categories; for instance, splits between the categories *Narrative* and *Opinion* were frequent. Instead of mere disagreement, the most frequent splits reflected the fact that the documents displayed the characteristics of several registers (Biber & Egbert, 2018). This could be seen also in the linguistic analysis of these hybrid documents (Biber & Egbert, 2016a). The CORE dataset is available for download at https://github.com/TurkuNLP/CORE-corpus.

The second contribution of the present study is related to the modeling of Web registers themselves. The classification of Web registers is known to be a difficult task when the data are not based on a restricted set of register categories. We explore various state-of-the-art text classification methods, including the deep transfer learning model BERT, which has provided improvements to a large number of NLP tasks (Devlin et al., 2018); the library fastText (Joulin et al., 2017); and a convolutional neural network (CNN) classifier (Kim, 2014). In the present study, we demonstrate that the automatic identification of registers in CORE is not only feasible, but it also provides new possibilities for gaining a better understanding of the variation associated with the data.

As the third contribution, we further investigate how to best model the range of variation attested by the registers in CORE. Web documents do not always exhibit properties associated with a single Web register, resulting in hybridism that features the characteristics of several registers simultaneously. Another extreme end of this variation is displayed by documents that do not belong to any register. For these purposes, we experiment with different classification settings: multiclass, where a document can only belong to a single class, and multilabel, where a document can belong to several classes. Furthermore, we test whether it is useful to *propagate*, that is, repeat, a document's main register label (such as *Narrative*), when the document already has a subregister label of the same register category (such as *Sports report*).

Finally, as the fourth contribution, we investigate those challenges related to the modeling of longer Web documents. Issues pertaining to the length of the document have been demonstrated to impose challenges for text classification systems (Worsham & Kalita, 2018; Xiao & Cho, 2016; Zhang et al., 2015). Furthermore, this also

applies to deep language models because they are often trained on parts of texts to avoid a significant increase in the computational cost. In the current study, we operate with BERT models limited to simultaneously processing parts of text containing 512 tokens. This may affect the discriminative performance of the models. Thus, we examine how the performance of BERT varies when trained and tested on different parts of the document. Additionally, we examine the ways to combine the predictions of BERT at the document level to improve the performance. Although there may be several different motivational pathways that give rise to the importance of a particular part of text impacting model performance, we specifically focus on those language-related issues that can be modeled with BERT. We will provide a more in-depth discussion of this matter in the general discussion and conclusion.

We present previous studies on automatic identification of Web registers in Sect. 2 and previous work in deep learning methods for text classification in Sect. 3. Then, Sect. 4 presents CORE and its register categories, while Sect. 5 discusses the specific methods used in this study. In Sect. 6, we evaluate the different deep learning classifiers and compare the results based on multiclass and multilabel settings and on the propagated and nonpropagated label versions. We found that BERT, a multilabel setting, and the propagated data version provided the best performance. Finally, in Sect. 7, we present the experiments on different parts of the document, showing that the beginnings of a document provide the highest predictive power.

2 Web registers and their identification

The automatic identification of a Web registers presents both theoretical and practical challenges. In this section, we discuss how these issues have been addressed in previous studies. We start by discussing the theoretical challenges related to Web registers and their identification, including the ways in which these text varieties have been defined in previous studies and how, for example, loosely defined situational characteristics and conventionalization can affect their distinctiveness. Second, we present practical solutions and how previous register identification studies have tackled these challenges.

2.1 Theoretical considerations

The difficulty of modeling language use on the Internet and mapping its linguistic variation to functionally motivated categories is already evident from the range of terminological choices proposed in previous research [see the discussion in Santini et al. (2011a)]. One of the fundamental questions pertains to the distinction between the terms *genre* and *register*. Although they are often used interchangeably, the terms have theoretical differences [see (Lee, 2002; Sharoff, 2018) for reviews]. The term *genre* is typically used in studies concerned with text classification (Petrenz & Webber, 2011; Pritsos & Stamatatos, 2018; Santini et al., 2011b; Sharoff, 2018), and it is preferred in discourse analysis (Halliday, 1985; Miller, 1984; Swales, 1990). In contrast, the term *register* is typically used in studies related to the fields of text and

corpus linguistics. Register can be defined as a text variety with specific situational characteristics, including communicative purposes (Biber et al., 1998; Biber & Conrad, 2009). The situational context forms the basis of different register categories, along with their pervasive linguistic characteristics that are functionally associated with the situation. We also adopt this perspective in the current study.

The challenge with modeling Web registers is that not all registers are equally well defined when it comes to both their situational and linguistic characteristics. For instance, encyclopedia articles-as exemplified by Wikipedia articles-tend to be written in a very specific situation with well-defined communicative objectives, whereas other registers, such as different blogs or persuasive texts, may display a wider range of communicative objectives and other situational characteristics. This variation may also affect the distinctiveness of the linguistic characteristics associated with the registers and, thus, the extent to which they can be automatically identified.

In addition to the situational characteristics, another aspect that can influence the distinctiveness of Web registers is the degree of their conventionalization and presence of conventionalized linguistic patterns. For instance, these could be elements such as the date and name of the recipient used in business letters or forum discussions. These elements can facilitate efficient and robust communication (Gibson et al., 2019; Jaeger & Tily, 2011) and further contribute to the distinctiveness of the register. In the register framework, these conventionalized elements are categorized as genre markers because they are not as pervasive as register markers (Biber & Conrad, 2009, pp. 69–71). However, we do not pursue this distinction because both kinds of markers can contribute to the distinctiveness and discrimination of documents in automatic register identification.

Much like the distinctiveness of situational and linguistic characteristics, the degree of conventionalization can vary across registers. It is upheld not only by culturally shared properties, but it is also affected and guided through institutional practice over time (Biber & Conrad, 2009; Görlach, 2004; Swales, 1990). For example, practices in educational systems provide a strong and beneficial effect on language comprehension (Kyröläinen & Kuperman, 2021) and the means for upholding conventionalized linguistic patterns, for example, in the structuring of research articles and technical reports. Additionally, editorial conventions can contribute to the degree of conventionalization associated with registers, such as magazine articles and news reports. On the other hand, other registers, such as personal blogs, can feature much more loosely defined conventions and different practices to establish them-this register is likely to be primarily guided through the linguistic knowledge of the author rather than strict guidelines or institutionalized practices (Degaetano-Ortlieb & Teich, 2022).

Loosely defined conventions and fluctuations in the communicative situation can give rise to hybridism, where a given document may simultaneously display the characteristics of several registers. This also appears in Web registers, as we discuss in the next section. Additionally, this can be reflected in the naming of a register (Görlach, 2002, 2004; Santini, 2007). When a specific register is strongly conventionalized and clearly recognizable, such as a poem or song lyric, it is also easy to name, whereas loose conventions and situational parameters can result in vague

names (Kwaśnik et al., 2006; Rosso, 2008; Santini, 2008). This issue with naming also concerns the registers included in CORE and may also affect their automatic identification (see Sects. 4.1 and 6).

2.2 Practical solutions

The complex relationship between Web documents and registers has consequences for their automatic identification. It also affects the language resources that can be used to provide an avenue for better understanding their structure. Furthermore, this relationship imposes challenges for the development of robust systems for register identification (Petrenz & Webber, 2011; Sharoff et al., 2010). Although most of the Web corpora that are typically utilized tend to be relatively small-both in size and in coverage of Web registers, only representing selected registers found on the Internet (Asheghi et al., 2016; Santini, 2011; Meyer zu Eissen & Stein, 2004; Vidulin et al., 2009), there are certain collections that provide a large inventory of categories, such as the KRYS 1 corpus consisting of 70 genres (Berninger et al., 2008). Given this, the performance of the systems used to automatically identify registers tend to convey that the Web registers are relatively well discriminated. For example, Asheghi et al. (2014) used the Leeds Web Genre Corpus, which consists of 15 categories; they achieved an accuracy of 78.88% for plain texts, while Pritsos and Stamatatos (2018) achieved an F1-score of 79% using two corpora with seven and eight categories, respectively. Furthermore, using structural Web page information, such as boilerplates, has been reported to improve the results in some cases (Asheghi et al., 2014; Madjarov et al., 2019).

In contrast to these restricted language resources, the guiding principle of CORE (Egbert et al., 2015) is to capture a large, unrestricted sample of English Web documents and their registers. A detailed description of the registers in CORE is provided in Sect. 4.1. Given the nature of the data contained in CORE, it is reasonable to state that the automatic identification of the Web registers can be a challenging endeavor. To this effect, Biber and Egbert (2016b) reported an average precision of 34% and recall of 39.6% on a subset of CORE documents consisting of 20 subregister categories. When the method was applied to the full corpus, the precision and recall were 26.9% and 28.6%, respectively. More recently, Laippala et al. (2021) used linear support vector machines to model 26 subregisters in CORE. The model was trained on lexico-grammatical features derived directly from the parsed documents, achieving an F1-score of 74.5%.

The commonality of these studies utilizing CORE is that the registers were taken as discrete categories, in which each document belonged to only one register (Biber & Egbert, 2016b; Laippala et al., 2021). However, Biber et al. (2020) suggested that registers could be analyzed in a continuous space instead of discrete categories to better capture hybridism. This idea has been supported by many Web register studies (Görlach, 2004; Santini, 2007). Some attempts have also been made to compare these two approaches. For instance, (Vidulin et al., 2009) and (Madjarov et al., 2019) compared multiclass and multilabel classification settings in Web register identification. Regarding the former, there are multiple register categories, but

a given document is associated only with one register category, whereas in the latter case, a given document may be associated with more than one register category. Madjarov et al. (2019) concluded that a multilabel setting was preferable because it achieved better performance in discriminating between the registers. A similar strategy was also suggested by the functional text dimensions approach presented by Sharoff (2018). Instead of discrete register categories, this approach analyzes registers in terms of their similarity to prototypical categories. These are represented by dimensions that reflect general functional categories, such as *informative reporting* and *argumentation*.

3 Deep learning methods for text classification

Previous studies on the automatic identification of registers have mostly applied traditional supervised machine learning methods, such as linear support vector machines (Asheghi et al., 2014; Pritsos & Stamatatos, 2018). In these studies, modeling has been based on presenting the registers with a set of features considered to reflect the categories. For instance, the bag-of-words model is based on a simple (normalized) frequency of each dataset word in each dataset document. The present study, however, focuses on deep learning methods and semisupervised transfer learning. In our approach, we utilize neural network architectures based on generalized language models that have been pretrained on large amounts of unannotated text. These generalized models can be fine-tuned for different downstream tasks, such as text classification and register identification, and are presented in the following section; following this, the challenges that they present when modeling long documents are discussed.

3.1 Pretrained language models and transfer learning

The use of pretrained models has allowed for substantial advancements in a broad range of NLP tasks involving text classification (Devlin et al., 2018), such as sentiment analysis (Hoang et al., 2019) and the recognition of hate speech (Mishra, 2019). The use of unannotated text in developing pretrained language models builds on one of the most fundamental ideas of NLP: the distributional hypothesis, according to which words used in similar contexts have a similar meaning (Firth, 1957). For decades, this hypothesis has been a major area of research in NLP (e.g. (Brown et al., 1992; Landauer & Dumais, 1997; Martin et al., 1998)).

Over the last decade, a substantial body of work has focused on the use of neural network models to create dense word representations that are predicted based on their contextual features. Following the distributional hypothesis, these representations, that is, vectors, capture aspects of meaning-semantically similar words are used in similar contexts, so they get nearby vectors (Turian et al., 2010). The skip-gram and continuous bag-of-words models proposed by Mikolov et al. (2013a) and implemented in the popular word2vec package have been particularly influential, demonstrating that prediction-based neural methods can create high-quality representations (Baroni et al.,

2014; Mikolov et al., 2013b) and serve as a basis for many subsequent studies in various downstream tasks in NLP [e.g. (Joulin et al., 2017; Levy & Goldberg, 2014; Pennington et al., 2014)].

Although methods such as word2vec can capture a wide range of usage patterns associated with word meaning, they are fundamentally limited in that they represent each word with a single vector, disregarding the differences in context, for example, polysemy and homonymy. A recent line of studies has demonstrated the applicability of deep learning models to create *contextualized* word representations that can accurately model the meaning of a word in context (Akbik et al., 2018; Devlin et al., 2018; Howard & Ruder, 2018; Peters et al., 2018). The BERT model by Devlin et al. employs the efficient transformer architecture (Vaswani et al., 2017), which consists of deeply stacked neural network layers where the representation of a word on a layer is built using a neural attention mechanism (Bahdanau et al., 2015).

Training a BERT model is partly analogous to that of earlier neural language models in that it is based on word prediction. However, the deeply bidirectional architecture allows for the modeling of words in contexts, which then enables the model to take, for example, polysemy into account. In addition to word prediction, which is typically applied to produce word vectors, BERT includes a next-sentence prediction objective. This further encourages the model to capture information about word relationships both within and across sentences. These properties can increase the performance of BERT-based systems in downstream tasks.

A challenge related to all BERT-based models is the complexity of the model and fact that they can usually operate with a maximum window size of only 512 tokens. Specifically, the transformer passes signals from each token position on the input side to all positions on the output side, which involves weighted combinations between position pairs using the attention mechanism and nonlinear transformations. Given that the cost of the attention mechanism is quadratic in the input sequence length, the limit of 512 tokens eases computational demand. The original BERT model is available in two sizes: BERT Base consists of 12 layers, and BERT Large consists of 24 layers.

The success of the original BERT model has resulted in a number of other transformer-based models, see for example https://huggingface.co/models. For instance, Multilingual BERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2020) provide cross-lingual language models that can be fine-tuned to model data in multilingual settings. These have also been applied successfully to register identification (Repo et al., 2021; Rönnqvist et al., 2021). Another line of modeling endeavor has focused on smaller and faster models, such as DistilBERT, which have been created to overcome the computational challenges related to the original version (Sanh et al., 2019). Finally, attempts have been made to process longer pieces of text than the standard 512 tokens. For instance, Longformer-the Long Document Transformer-can process sequences up to 4,096 tokens long (Beltagy et al., 2020).

3.2 Challenges presented by document length

The capacity of BERT to work on a maximum window of 512 tokens at a time presents a challenge for register identification because documents tend to be longer in

length (see Sect. 4). Similarly, models based on simple frequency-based feature representations, such as the bag-of-words model, can have difficulties with longer documents because they are not always able to capture all the relevant information. Furthermore, longer documents can have characteristics that complicate their modeling. For instance, a document can include passages that may not display the distinctive characteristics of its register.

In a previous study, Worsham and Kalita (2018) compared a number of methods to tackle the challenges associated with the text length. Focusing on the classification of novels based on their topical area, such as sci-fi, adventures, love stories, detective and mystery stories, historical fiction, and Westerns, they examined how classification performance can be affected by different representations of the data, here ranging from entire novels modeled as bag-of-words to various partition methods, such as using the first or last 5,000 words of the document only. Their best classification performance-an F1-score of 81%-was achieved with XGBoost decision trees trained on bag-of-words features derived from the entire documents (Worsham & Kalita, 2018).

In the present study, we attempt to overcome the limitation of BERT when it comes to modeling documents with varying length by exploring similar strategies to those presented by Worsham and Kalita (2018). In particular, we investigate the effects of applying BERT to different parts of documents and how the predictions for these different document parts can be combined to extend the modeling to the document level. These are further discussed in Sect. 7.

4 Data

In this section, we present the CORE register categories and their distribution. Furthermore, we briefly discuss the annotation process and describe how the register labels were assigned based on the annotations.

4.1 Web registers in CORE

CORE is currently the largest collection of Web registers with manually annotated categories consisting of nearly 50,000 documents. A detailed description of CORE is presented in Biber and Egbert (2018) . The compilation, along with sampling of the data included in CORE, is described in detail in Egbert et al. (2015). In short, CORE was compiled by running Google searches of highly frequent English three-grams, then saving a fixed number of pages retrieved from each n-gram search, and randomly sampling from a larger set of retrieved documents. The original searches were made from November to December 2012. The goal of this compilation process was to ensure that the dataset covered an unrestricted sample of the searchable, open English Web. Importantly, CORE was not restricted to include only certain, preselected Web registers; it aimed to be a sample of documents that can be found when searching the Web.

The taxonomy representing the register categories in CORE was determined during the annotation process rather than before the corpus compilation. The idea behind this approach was to capture the unrestricted range of Web registers associated with the documents attested in CORE. A detailed linguistic description of each of the registers in CORE can be found in Biber and Egbert (2016a); Egbert et al. (2015) and Biber and Egbert (2018). Here, we provide a brief overview of the annotation process. The annotation was done via crowd-sourcing using Mechanical Turk. Each document was annotated by four coders based on a hierarchical decision tree starting with basic situational characteristics associated with the document, which first led to the main register and, then, if possible, to its subregister. The final register taxonomy was created during several rounds of iterative refining of the register taxonomy leading to the final taxonomy consisting of eight main registers and 47 subregisters, as shown in Table 1. Given this data-driven, iterative process, both broad and fine-grained registers emerged from the data that are, consequently, reflected in the register taxonomy. For example, Travel blog is a register category with a high degree of specificity that, nevertheless, was frequently attested in the data. On the other hand, Description of a thing is a broad category that covers multiple topics and communicative situations. It is worth pointing out that the data do not

Table 1 The register categories and their abbreviations

IN	INFORMATIONAL DESCRIPTION/EXPLANATION	NA	NARRATIVE	HI	HOW-TO/INSTRUCTIONAL
cm	Course materials	ha	Historical article	fh	FAQ about how-to
dp	Description of a person	ma	Magazine article	ht	How-to
dt	Description of a thing	ne	News report/blog	oh	Other
en	Encyclopedia article	on	Other narrative	re	Recipe
fi	FAQ about information	pb	Personal blog	ts	Technical support
ib	Information blog	sr	Sports report	LY	LYRICAL
lt	Legal terms and conditions	ss	Short story	ol	Other
oi	Other information	tb	Travel blog	po	Poem
ra	Research article	IP	INFORMATIONAL PERSUASION	pr	Prayer
tr	Technical report	ds	Description with intent to sell	sl	Song lyrics
OP	OPINION	ed	Editorial	SP	SPOKEN
ad	Advertisement	oe	Other	fs	Formal speech
av	Advice	pa	Persuasive article or essay	it	Interview
le	Letter to editor	ID	INTERACTIVE DISCUSSION	os	Other
ob	Opinion blog	df	Discussion forum	ta	Transcript of video/audio
oo	Other opinion	of	Other forum	tv	TV/movie script
rs	Religious blogs/sermons	qa	Question/answer forum		
rv	Reviews	rr	Reader/viewer responses		

The main registers are in capital letters

contain documents associated with social media. The register taxonomy used in this study has been described in detail in a number of publications, see (Biber & Egbert, 2016a, 2018; Egbert et al., 2015).

It is important to point out that some of the subregisters in the taxonomy were excluded from the study by Biber and Egbert (2016b) because of having a small number of documents or because of a lack of annotator agreement. However, these subregisters are included in the present study because we aimed to include the unrestricted range of registers attested to in CORE.

The annotation taxonomy in CORE also gives us the opportunity to include hybridism as part of the analysis-hybrid documents can be formed based on the frequent disagreements between coders. Linguistic studies have shown that these documents do not contain the distinct linguistic characteristics of a single register, but instead, they simultaneously have elements from several categories and display specific combinations of registers (Biber & Egbert, 2016a; Egbert et al., 2015). Therefore, it is highly unlikely that these disagreements would simply reflect the unreliability of the annotation process.

4.2 Assigning documents to register categories in a continuous space of register variation

Given the rich source of register information presented in CORE, it is possible to experiment with a number of different settings to derive the register labels from the annotations. This also relates to the question of how best to capture the linguistic variation associated with the Web registers and hybridism. While, from a theoretical perspective, it would be prudent to pursue a representation of the Web registers in a truly continuous space, there are, nonetheless, technical limitations that apply to contemporary classifiers. Therefore, we implemented the following settings in the present study: multiclass, multilabel and finally propagated versus non-propagated.

We based the assignment of the final register labels on the agreements between coders. Specifically, a two-way agreement between coders was required at a minimum to assign a given register label to a particular document. In other words, a document was given a register label whenever at least two coders agreed upon it, both at the main register and subregister levels. If this requirement was not met, a particular document was not associated with any register label. Hybrid documents are a logical consequence of this labeling strategy. Thus, unlike in Biber and Egbert (2016b), the formation of hybrid labels was not based on the frequency of the label combination. Instead, hybrids were formed whenever the coders' votes were split so that at least two coders agreed on a given register category. For example, if two coders agreed that a specific document denoted the Narrative register and the two other coders assigned the document to Opinion, the document was annotated as having both labels. This strategy allowed us to avoid imposing a frequency threshold. At the same time, despite being data driven, it may lead to the formation of hybrids that do not necessarily represent recurrent register combinations in the data.

The labels associated with the registers allowed us to form our settings for the register identification task. The first is the multiclass setting. In this setting, a given

document is always associated with one label, and the labels are mutually exclusive. In case of hybrids, each register combination, such as Narrative and Opinion, constituted a single label, that is, a register category of its own. This yielded a total of 460 labels, as shown in Table 4.

The second is the multilabel setting. In this setting, a given document could be associated with one or more labels, and the labels are independent and not mutually exclusive. For example, the label Narrative is "shared" with documents consisting of a single label, that is, Narrative, and hybrid, that is, Narrative and Opinion. This setting yielded a total of 56 labels, as shown in Table 4.

The final setting consists of the propagation of the register labels. The propagation of the register labels was achieved by assigning a main register label to a particular document when it was associated with a subregister label. For example, if a particular document was labeled as News article, the document was also labeled as Narrative because the label Narrative represents the main register category of News article. Thus, all the different subregisters belonging to the Narrative main category received the label Narrative as well. This propagation of the labels is shown in Table 2. Document 1 represents a document on which all the coders agreed; thus, the document was assigned to both the main and subregister labels. Documents 2, 3, and 4, on the other hand, represent those with disagreements. The final labels were assigned when a two-way agreement was found for both the main and subregister categories. The main register labels were propagated whenever a subregister label was assigned. Finally, document 5 represents a complete disagreement. Hence, a register label was not assigned to it.

We hypothesize that the propagation of the register labels can improve the classification performance because it can make explicit further structure in the data and it allows to increase the number of training examples for the main register categories. In some cases, however, subregisters belonging to different main registers can seem more similar than subregisters within a single main register. For instance, despite representing different main registers, a Personal blog in the Narrative main register and an Opinion blog in the Opinion main register can seem more similar than a Personal blog and a News article that are both in the Narrative category. To test this, we also experiment on a *nonpropagated* version of the dataset, where the main register labels were not assigned for documents with subregister labels. The assignment of nonpropagated register labels is shown in Table 3.

4.3 Document length and register distribution in CORE

The length of the documents may vary considerably in CORE, as visualized in Fig. 1. This is not insignificant in terms of register identification because length can be an important factor in the task (see Sect. 3). The most frequent document lengths are between 700 and 1000 words, and the shortest lengths are fewer than 100 words. Document lengths between 1000 and 10,000 words are still relatively common, and some individual texts exceed 10,000 words.

Also the registers are very unevenly distributed in CORE, reflecting its compilation strategy. The most frequent registers and their combinations for the propagated

Table 2 The assignment of the register label in the propagated setting

	Coder 1	Coder 2	Coder 3	Coder 4	Final labels
Document 1	NARRATIVE Sports report	NARRATIVE Sports report	NARRATIVE Sports report	NARRATIVE Sports report	NARRATIVE Sports report
Document 2	NARRATIVE News article	NARRATIVE News article	NARRATIVE Sports report	NARRATIVE Sports report	NARRATIVE News article, Sports report
Document 3	INFORMATIONAL DESCRIPTION Encyclopedia article	INFORMATIONAL DESCRIPTION Research article	NARRATIVE Description of a person	NARRATIVE Historical article	INFORMATIONAL DESCRIPTION, NARRATIVE
Document 4	INTERACTIVE DISCUSSION Discussion forum	INTERACTIVE DISCUSSION Discussion forum	HOW-TO INSTRUCTIONAL How-to pages	HOW TO INSTRUCTIONAL Technical support	INTERACTIVE DISCUSSION, HOW-TO INSTRUCTIONAL Discussion forum
Document 5	INTERACTIVE DISCUSSION FAQ about information	NARRATIVE News article	HOW-TO INSTRUCTIONAL FAQ about how-to	SPOKEN Interview	No labels

The main register labels are in capital letters

Table 3 The label assignment based on the coders' votes for the nonpropagated data version

	Coder 1	Coder 2	Coder 3	Coder 4	Final labels
Text 1	NARRATIVE	NARRATIVE	NARRATIVE	NARRATIVE	
	Sports report	Sports report	Sports report	Sports report	Sports report
Text 2	NARRATIVE	NARRATIVE	NARRATIVE	NARRATIVE	
	News article	News article	Sports report	Sports report	News article, Sports report
Text 3	INFORMATIONAL	INFORMATIONAL	NARRATIVE	NARRATIVE	INFORMATIONAL
	DESCRIPTION	DESCRIPTION	Description	Historical article	DESCRIPTION, NARRATIVE
	Encyclopedia article	Research article	of a person		
Text 4	INTERACTIVE	INTERACTIVE	HOW-TO	HOW TO	
	DISCUSSION	DISCUSSION	INSTRUCTIONAL	INSTRUCTIONAL	HOW-TO INSTRUCTIONAL
	Discussion forum	Discussion forum	How-to pages	Technical support	Discussion forum
Text 5	INTERACTIVE	NARRATIVE	HOW-TO	SPOKEN	no labels
	DISCUSSION	News article	INSTRUCTIONAL	Interview	
	FAQ about		FAQ about		
	information		how-to		

The main registers are in capital letters

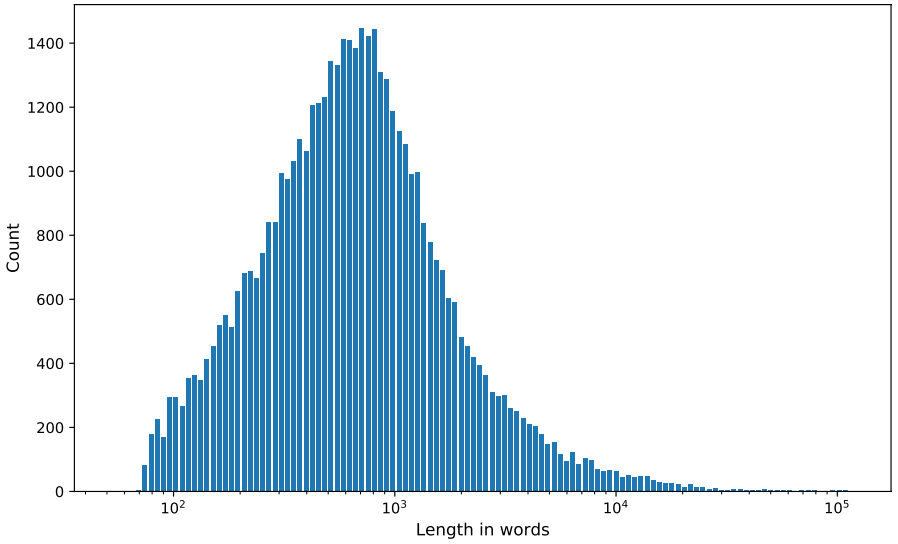

Fig. 1 Distribution of document lengths in CORE

data are described in Figs. 2 and 3. Figure 2 shows only the top-level labels, while Fig. 3 covers all the register labels and their combinations with more than 60 occurrences. The total number of documents in CORE is 48,452. Note that this differs slightly from the version applied in Biber and Egbert (2016b) because the preprocessing pipeline of the data was different.

The distribution of the main register labels is provided in Fig. 2. By far the most frequent label was Narrative (NA), followed by Informational description (IN) and Opinion (OP). The frequencies of Informational persuasion (IP) and How-to instructional (HI) were in the middle range, while Interactive discussion, Spoken, and Lyrical were among the least frequent labels. Additionally, the hybrid labels are presented in Fig. 2. The most frequently hybrids consisted of Narrative combined with Informational description (IN) and Opinion (OP). Additionally, the proportion of hybrid documents was large for Informational description (IN) and Opinion (OP). Approximately 20% of the documents formed hybrids with Narrative, and another 9% formed hybrids with the other one in the pair.

Figure 3 presents the distribution of the subregister labels. The most frequent subregister labels were as follows: News article, Sports report, Personal blog, Description of a thing, Information blog, Opinion blog, Review, Discussion forum, Description with intent to sell, and How-to. Furthermore, two kinds of hybrid combinations are shown in the figure: those consisting of labels under one main register and those combining several main register labels. Under one main category, the most frequent combinations were News report and Sports report, Personal blog and Travel blog, Discussion forum and Q-A forum and Description of a thing and Information blog. The situational characteristics of the registers in these combinations can be very similar, which can be used to motivate their hybridism. For example, it is likely that the difference between *News report* and

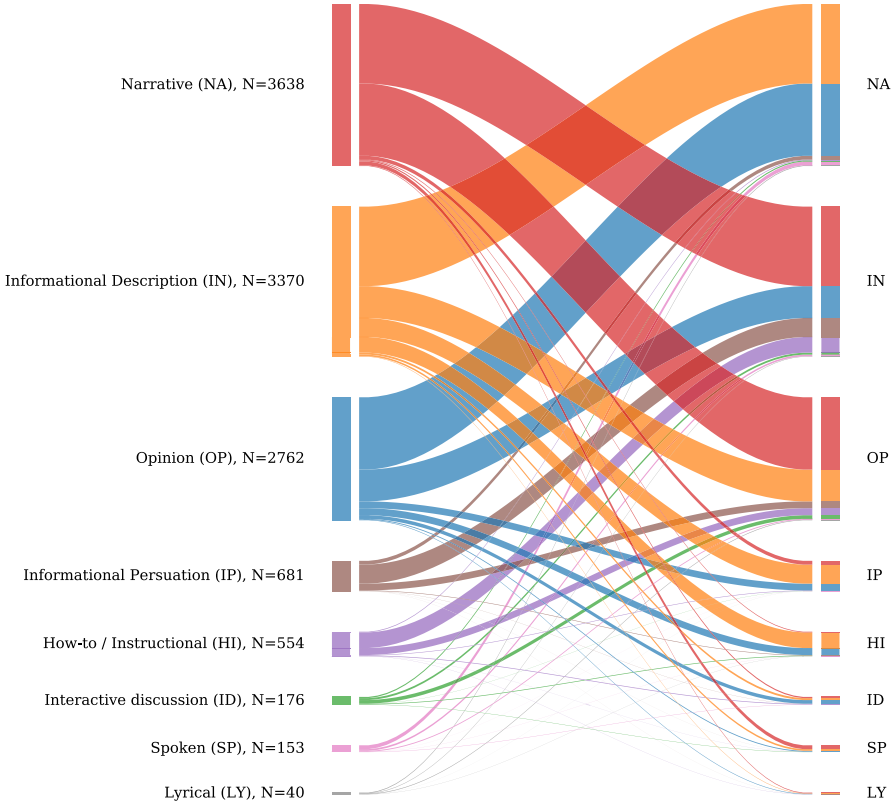

Fig. 2 The combinations of main-level registers in the propagated data. The thickness of a connection reflects the relative frequency with which the two labels co-occur

Sports report resides in topicality; often, the distinction can be fine-grained. On the other hand, hybridism is specifically reflected by the combinations of labels under different main registers. For instance, Opinion blog was frequently combined with News report or Personal blog, and News report was often combined with Description of a thing or Description of a person within the IN main category. Furthermore, Description of a thing (IN) co-occurred with Description with intent to sell (IP).

5 Methods

In this section, we describe the practical implementation of the experiments used in the current study. Additionally, the classifiers used to model register variation are discussed and presented. Specifically, the experiments were carried out using state-of-the-art classification methods.

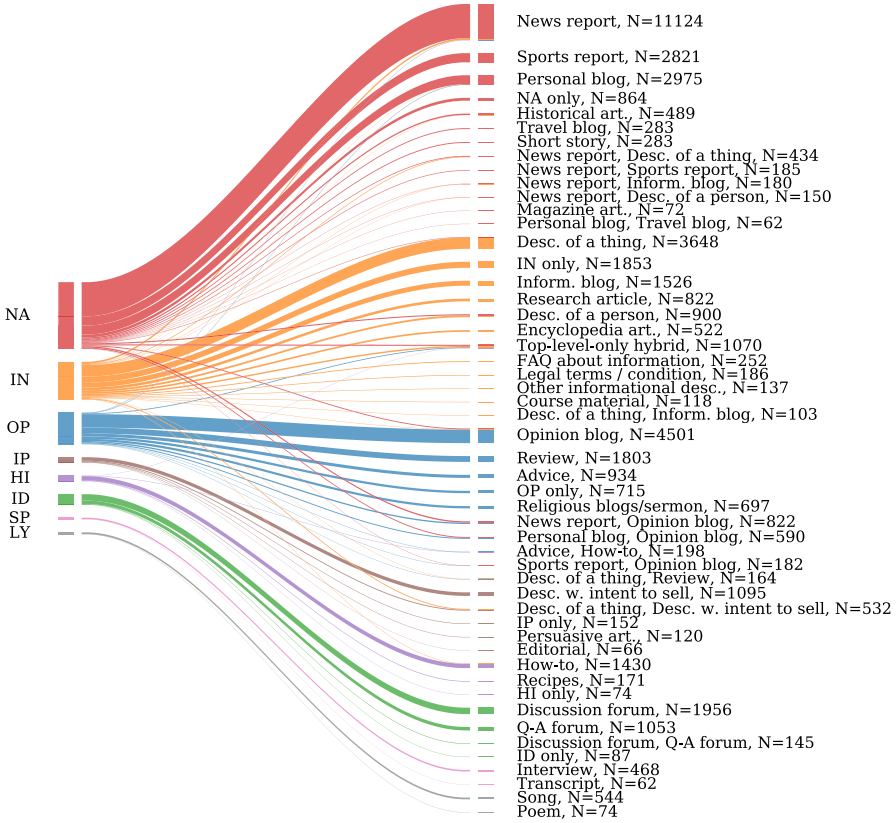

Fig. 3 The combinations of the main and subregisters in the propagated data. The thickness of the connection line reflects the relative frequency with which the main register label co-occurs with a particular set of subregister labels (excluding other main register labels) or with which it occurs individually (e.g., NA only)

5.1 Implementation and evaluation

We divided the CORE documents into training, development, and test sets, here with a 70%, 10%, and 20% split, respectively. The sampling was stratified based on the register labels. To evaluate the performance of the models, we report the precision, recall, and F1-score, with the F1-score being the balanced and harmonic mean of the first two measures. Additionally, we report the area under the precision-recall curve (PR-AUC). We report the results as micro-averages because the category distribution is very skewed. In the multilabel setting, the labels were evaluated separately, contrasting the multiclass setting in which each of the label combinations was evaluated as one unit. Also, the performance of the models was evaluated using the precision-recall curve (PR-curve) because these allowed us to examine the trade-off between precision and recall with different probability thresholds (Boyd et al., 2013). A grid search was used to find the optimal hyperparameters for

Table 4 The classification results for the different models and datasets (micro-average)

Model	Dataset	Labels	F1% (*SD*)	Precision% (*SD*)	Recall% (*SD*)	PR-AUC% (*SD*)
BERT	Propagated	56	68 (0.33)	71 (0.66)	65 (0.07)	75 (0.52)
Large	Nonpropagated	56	58 (0.21)	68 (1.23)	51 (0.37)	63 (0.96)
	Prop. multiclass	460	56 (0.12)	56 (0.12)	56 (0.12)	46 (0.06)
BERT	Propagated	56	67 (0.21)	69 (0.57)	66 (0.63)	75 (0.24)
Base	Nonpropagated.	56	56 (0.02)	63 (0.07)	51 (0.03)	59 (0.23)
	Prop. multiclass	460	55 (0.21)	55 (0.21)	55 (0.21)	46 (0.08)
fastText	Propagated	56	62 (0.01)	56 (0.02)	69 (0.02)	67 (0.02)
	Nonpropagated	56	52 (0.01)	53 (0.01)	52 (0.01)	53 (0.02)
	Prop. multiclass	460	53 (0.01)	48 (0.01)	58 (0.01)	55 (0.01)
CNN	Propagated	56	59 (0.26)	64 (0.77)	53 (0.31)	64 (0.32)
	Nonpropagated	56	45 (0.39)	59 (1.18)	36 (0.86)	48 (0.25)
	Prop. multiclass	460	41 (0.08)	64 (0.69)	30 (0.74)	42 (0.32)

The applied hyperparameters are denoted in Appendix 1

the models. Finally, the reported averaged performance metrics were based on the scores obtained from three different model trainings with the same hyperparameters to ensure the stability of the results.

5.2 Classifiers

In the experiments, we apply several neural network architectures. Additionally, we experiment with both multiclass and multilabel settings. The applied methods are described below.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a language model that builds on the Transformer neural network architecture (Vaswani et al., 2017). It was introduced by Devlin et al. (2018) , see Sect. 3.

We apply the BERT Base and BERT Large models in two settings: multiclass and multilabel. The multiclass setting follows the standard approach for text classification with BERT. The code is available at https://github.com/google-research/bert. The multilabel setting extends the original approach to allow for multiple labels to be predicted for each document. The code is available at https://github.com/haamis/neuro_classifier. Fine-tuning was done using backpropagation. Similar to the original setting (Devlin et al., 2018), we experiment with different numbers of iterations (epochs) over the training set, batch sizes, and learning rates.

Because of the restrictions imposed by the model architecture, our BERT experiments focus on the beginning of the documents. This also ensured that the model performance reported in this study is comparable across languages when modeling register categories based on the taxonomy presented in CORE. Currently, as part of an on-going project, corpora has been released for three languages that follow the register taxonomy presented in CORE (Repo et al., 2021; Rönnqvist et al., 2021). Additionally, we are in the process of releasing corpora for over 10 languages. For

this reason, we leave experiments with Longformer and similar model architectures for future studies. However, we specifically dedicate analysis for gaining a better understanding of the model performance of BERT in relation to the length of the document. For a detailed discussion and analyses, please refer to Sect. 7, where we explore a method for utilizing complete documents with BERT.

fastText is a simple and fast tool for text classification that builds on the word2vec models, adding subword information (Bojanowski et al., 2017) and support for word n-grams (Joulin et al., 2017) in an efficient feed-forward neural network model. It typically functions as a strong baseline (Joulin et al., 2017). Subword information is added by learning vectors for the character n-grams of words, in addition to a vector for a word token itself, and by summing all the vectors. This allows morphology to be handled more effectively. As a simple trick to efficiently incorporate some word order information, the tool uses bag-of-word n-gram features to capture very local word patterns. Text classification is performed using a linear classifier. In our experiments, we use pretrained English vectors of 300 dimensions. We optimize the learning rate, number of training epochs, and maximum length of word n-grams. The code is available at https://github.com/facebookresearch/fastText/.

CNN-convolutional neural networks (Kim, 2014)-are a relatively lightweight and popular architectures for building classifiers. In NLP, CNNs have been shown to be particularly effective at modeling local patterns because they learn to recognize short sequences as features very flexibly. These are then pooled together into useful representations of full text sequences by, for example, point-wise maximum functions.

We apply a CNN similar to that used by Laippala et al. (2019) for multilingual register classification to serve as a baseline for our experiments. The network is initialized with pretrained English word embeddings and employs a convolution layer with ReLU activation, a max-pooling layer, and a decision layer with softmax/sigmoid activation for the multiclass/multilabel settings. The word embeddings are frozen, their number is capped at 100,000, and we use 128 convolution filters. We optimize the window size and learning rate, available at https://github.com/mavela/Multiling-Multilabel-CNN for multi-label and https://github.com/TurkuNLP/multiling-cnn for the multi-class.

6 Evaluation

In this section, we evaluate the different classifiers for identifying the CORE registers. We include the propagated and nonpropagated data versions and experiment with both multiclass and multilabel settings (see Sect. 4.2 for a description).

This section starts by comparing the performance of the classifiers. Then, it evaluates the results in detail by focusing on register-specific differences. Linguistic studies have shown that registers vary considerably in terms of how well they are defined (Biber & Egbert, 2018; Biber et al., 2020; Laippala et al., 2021). This affects the extent to which they can be automatically identified; consequently, the register classes should be considered individually in the evaluation of the model.

6.1 Classifier performances

Table 4 presents the classification results across different classifiers and dataset versions. First, the highest F1-score of 68% and PR-AUC of 75% were achieved using the propagated data version and BERT Large. BERT Base achieved nearly similar results, whereas the performances of fastText and CNN were clearly lower. Although these scores are not directly comparable to the results presented in previous studies, they can be considered very competitive; for example, Biber and Egbert (2016b) reported a precision of 26.9% and recall of 28.6%. The data in their study was based on CORE and used a multiclass setting, even though the register labels and their assignment strategy were different. Another point of comparison between model performances can be made against Asheghi et al. (2014), who reported an accuracy of 78.88% using the Leeds Web Genre corpus with 15 balanced register categories from a restricted sample of the Web.

Here, we focus on discussing the best results achieved with BERT Large. As already demonstrated by Madjarov et al. (2019), the multilabel setting is better for learning the mapping between the linguistic characteristics and their associated register categories. This is evident from our experiments as well. The number of labels tended to increase significantly with this type of data when operating in a multiclass setting. This increase became drastic when operating with datasets such as CORE that feature an unrestricted sample of Web documents.

In the current study, a primary point of interest is the difference between propagated and nonpropagated versions of the data. The former provided substantially better discrimination power between the register labels. While both versions of the data contain the same number of register labels, they do deviate from each other in terms of the number of examples per label in training data—in the propagated version, the main register label was repeated with each subregister label, whereas in the nonpropagated version this as not the case. This seems to suggest that improvements could be gained in this task by increasing the size of the training data. In particular, the high precision and low recall of the nonpropagated data indicated that the coverage could be increased by propagating the main register labels. Furthermore, it is also possible that propagation brings forth further structuring present in the data. Although certain subregisters can be associated with a different main register label, the structure resulting from propagation may provide additional information that BERT can then utilize in learning the mapping between the linguistic characteristics and register label. Specifically, in the propagated version of the data, improvements in the classification performance were also supported by the learning curve of the classifier presented in Fig. 4 because the learning curve flattened first at 30% and then at 70% of the training data, showing that the size of the corpus is sufficient for this task.

To further analyze the classification performance of BERT Large, a precision-recall curve with AUC is presented in Fig. 5. The curve follows a typical pattern: first, the precision was very high for documents predicted with a high probability, but it started to gradually decrease as the probability dropped.

Finally, the task of automatically identifying the register associated with Web documents is not easy. Based on the results that we have presented in this section,

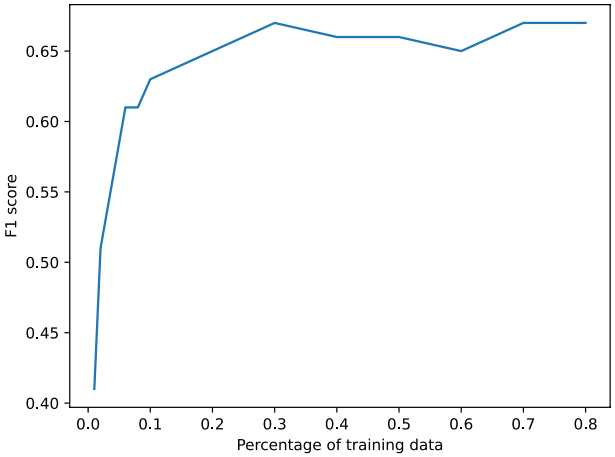

Fig. 4 The relationship between F1-score and the size of the training data in proportion

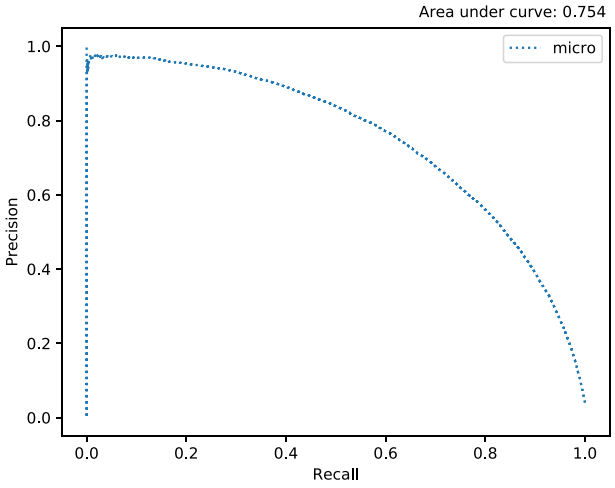

Fig. 5 The PR curve for the propagated multilabel data based on BERT Large

improvements in the classification of the Web registers would require methodological advances, such as more sophisticated machine learning techniques or features rather than an increased size of the training dataset. At the same time, with an F1-score of nearly 70%, our results have demonstrated that the Web registers can be identified even in an unrestricted corpus, at least up to a certain degree. These results pertain to identifying the registers globally. However, we know that Web registers and those specifically associated with Web documents can display significant differences in terms of their situational, linguistic, and conventional characteristics (see Sect. 2.1). Thus, we focus on register-specific variation in the following section.

6.2 Register-specific variation

In this section, we continue our analysis based on BERT Large and the propagated, multilabel data by focusing on register-specific variation. To this end, we examine the PR-AUC score in relation to the number of examples in the training data, which is visualized in Fig. 6. These scores have demonstrated a number of important properties associated with the registers.

First, there was expected positive correlation between the PR-AUC scores and number of training examples. Several registers with a large number of training examples were discriminated very accurately, with a PR-AUC score of 90% or greater, which is shown in the upper right corner in Fig. 6. These included such subregisters as Sports report (sr) and News (ne) and the main registers Narrative (NA), Interactive discussion (ID), and Lyrical (LY). Another group of registers were associated with a mid-range PR-AUC score of ~80%. These included such

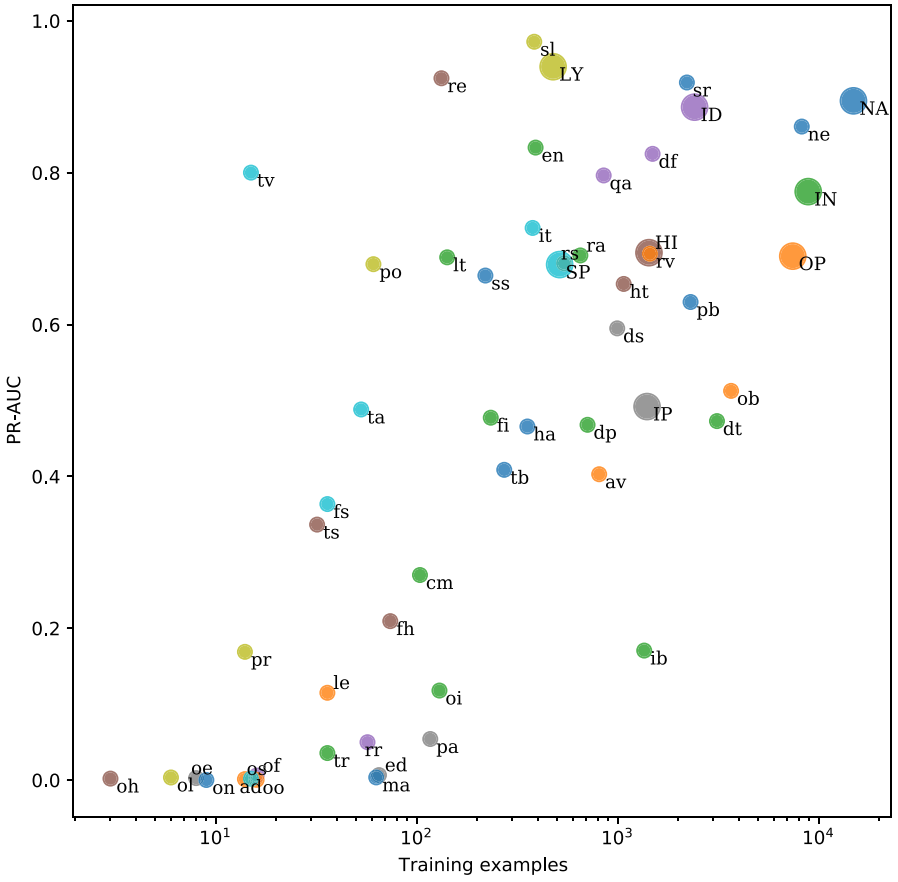

Fig. 6 The relationship between the PR-AUC score and number of training data. The subregisters are color coded according to the main register category (given as large points)

registers as Encyclopedia article (en), Question-Answer forum (qa), Discussion forum (df), and Informational description (IN).

As expected, registers associated with a smaller number of examples were more difficult to discriminate, and these included such registers as Technical report (tr), Reader/viewer response (rr), and Advertisement (ad), which can be seen in the lower left corner in Fig. 6. It is worth pointing out that although this correlation was to be expected, certain registers, nonetheless, deviated from this pattern, for example, Information blog (ib). From a linguistic perspective, this particular register is less likely to display well-structured linguistic characteristics that could be used in their automatic identification (Biber & Egbert, 2018; Biber et al., 2020).

A second pattern visible in Fig. 6 is related to those registers that were still discriminated from other registers but covered a small number of documents in the training data. These included such registers as TV script (tv), Poem (po), and Transcript of video/audio (ta). This suggests that these registers exhibited distinct linguistic characteristics that supported their accurate identification, regardless of the overall number of documents associated with them in the training data.

We have selected four registers to illustrate the differences they display in terms of their distinctiveness. For purposes of this quantitative analysis, we generated PR curves showing the relationship between precision and recall across different probability thresholds to probe into the distinctiveness of a particular register. These curves are visualized in Fig. 7. First, Information blog (ib) was difficult to discriminate, as supported by the fairly flat PR curve. This supports the vagueness of the register: the classifier predicted the wrong register label, even for documents associated with a high prediction probability. Second, Lyrical (LY) appears to consist of subregisters associated with linguistically highly distinctive characteristics: almost all the documents were correctly classified. Third, Short story (ss) displayed a weaker discriminability, as indicated by the relatively shallow PR curve in relation to Lyrical. Nonetheless, most documents associated with this register label were likely to be accurately identified. Fourth, Travel blog (tb) displayed a tendency similar to Information blog, here with the distinction that its PR curve was less shallow. In terms of linguistic characteristics, it is likely that Web documents associated with this subregister may tend to contain less variation, giving rise to a better intracategory similarity.

Finally, our analysis offers evidence that the Web registers displayed variation in terms of their ease of automatic identification. In our view, this can be taken as a reflection of the linguistic distinctiveness of a particular register. This relationship is further supported by our analysis of the individual PR curves. The curves do not, however, provide insights into which registers were mixed up by the classifier. To explore this aspect of the results more deeply, we used a heatmap to visualize the relationship between the true register labels and their misclassifications, as provided in Fig. 8.

This visualization shows that the documents with one register label were misclassified relatively rarely, and most misclassifications occurred with hybrids, suggesting that hybridism is associated with less clear linguistic characteristics. In particular, hybrid documents were often misclassified when the model predicted only a single register label and, in the reverse case, when the model predicted hybrid register labels

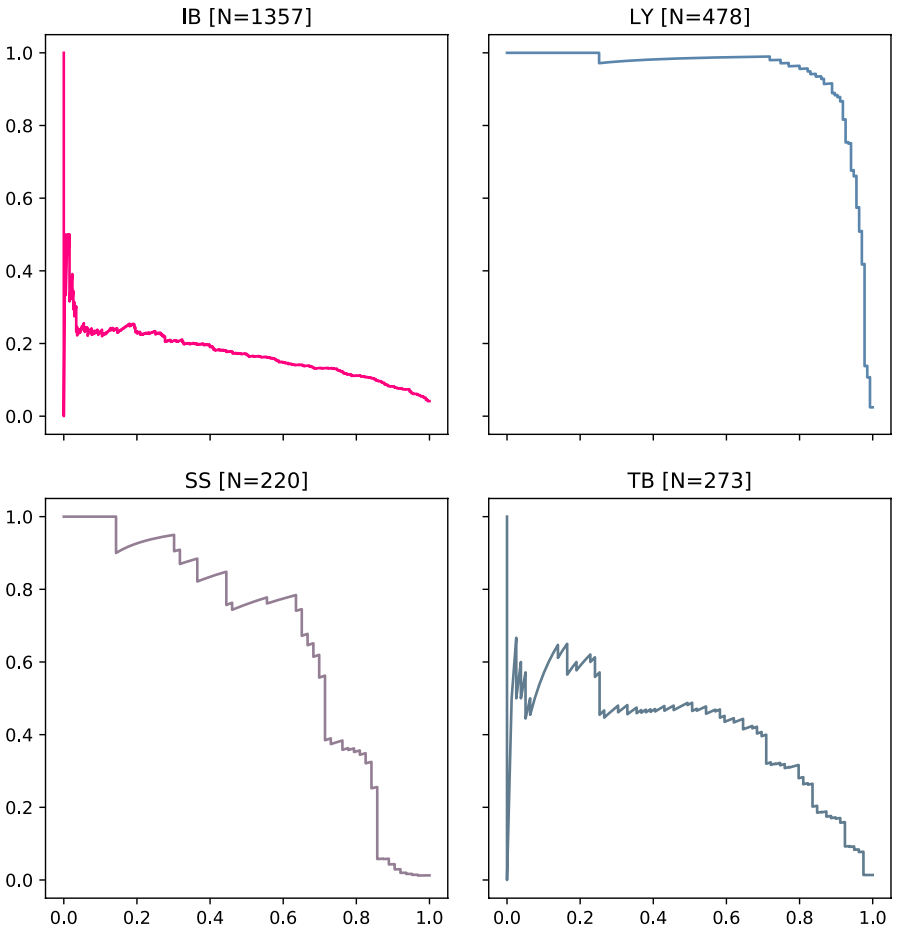

Fig. 7 The relationship between the precision (y-axis) and recall (x-axis) for Information blog (IB), Lyrical (LY), Short story (SS), and Travel blog (TB) in the propagated multilabel data

in instances where the document was associated with a single register label. This tendency suggests that hybrids might be represented more accurately based on a continuous space of register categories; that is, the categories do not always have discrete boundaries. Finally, No label refers to the documents for which there was no two-way agreement between the coders. These documents do not form a specific register, and their lack of well-defined linguistic characteristics was also evident in the proportion of the misclassifications. In practice, the documents were predicted as belonging to a number of different categories.

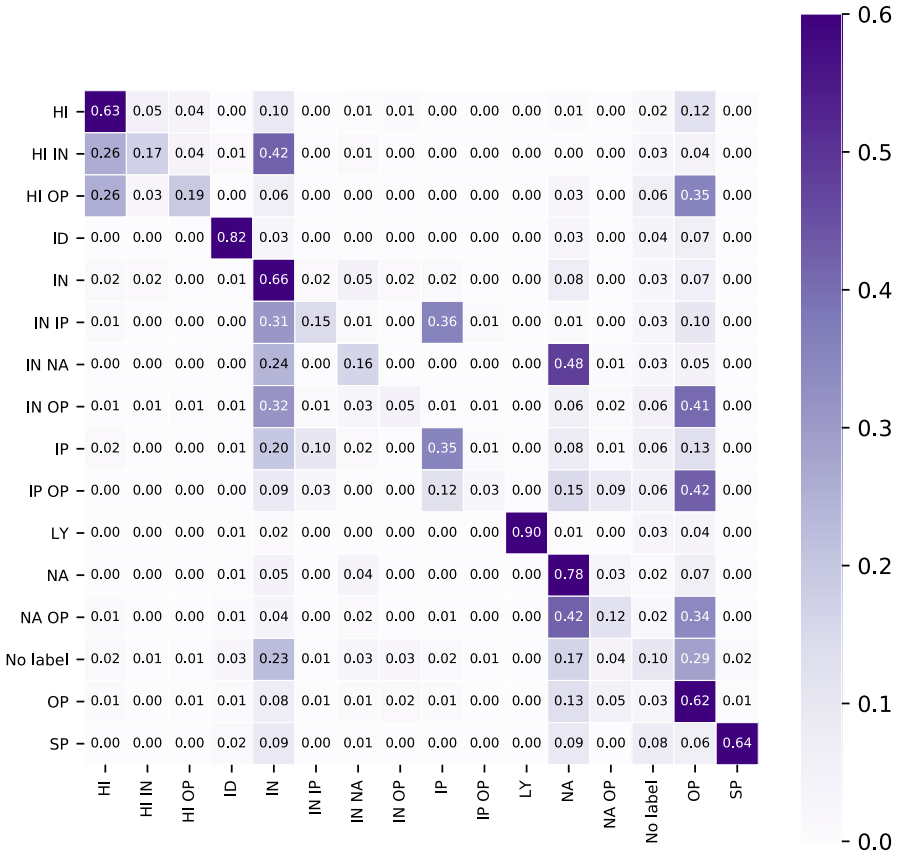

Fig. 8 A heatmap of the most frequently misclassified register labels. Subregisters are reduced to their corresponding main registers. The rows correspond to true labels and columns to predicted labels

7 Document-level analysis and experiments

The BERT models presented in the previous sections were trained on data that only included the beginning of the document. This artificial limitation may affect the classification performance because important information may be excluded from the linguistic representation of a document. Furthermore, this limitation opens up the question of whether specific parts other than the beginning of a given document might serve as a source of additional information and, consequently, affect the model's performance, either decreasing or increasing it.

7.1 Predictive power of the different parts of a document

To investigate how the length of a given document may have affected model performance, the text in a document was divided into blocks consisting of 512 tokens

because this corresponds to the context size of BERT models. The BERT model uses its own tokenizer that is preconfigured based on the frequencies of subword components and a fixed vocabulary size. The tokenizer produces full word tokens for frequent words and breaks less frequent words or word forms into subword tokens. In our training data, on average, each word was broken into 1.7 tokens, and the training set translated into 329.9 words per block on average. To examine the variation of the model performance in relation to the length of the document, we implemented several steps, as described below.

First, we divided the text in a given document into 512-token blocks. The blocks were padded if they contained less than 512 tokens. Then, these documents were partitioned into groups based on the number of blocks contained in them (1, 2, 3, 4, 5-9, and 10+). The partitions of the documents covering 1 and 2 blocks were evenly sized and accounted for about 60% of the data. In contrast, 83% of the documents consisted of 1-4 blocks, and 6% contained 10+ blocks. The relationship between the F1-scores (BERT Base) and these partitions is visualized in Fig. 9 (left panel). We observed a clear negative correlation between the F1-score and block size; that is, longer documents were classified less accurately than shorter ones. This might indicate that the relevant characteristics associated with these registers might be excluded from the data simply because of the limitations pertaining to the length of the document. On the other hand, the decrease in model performance with longer documents might also be related to their relatively low frequency in the data. Indeed, there was a positive relationship between the F1-score and sample size, as visualized in Fig. 9 (right panel).

Next, we examine the relationship between the positions of the blocks in a document and the performance of BERT, as measured by F1-score. In each document, in increments of 20%, the blocks were extracted based on their relative position: starting from 0% and ending at 100%. BERT Base was fine-tuned and tested on the blocks in four different conditions to tease apart the potential influence of the position of the block in model performance: (1) trained and tested on the same blocks,

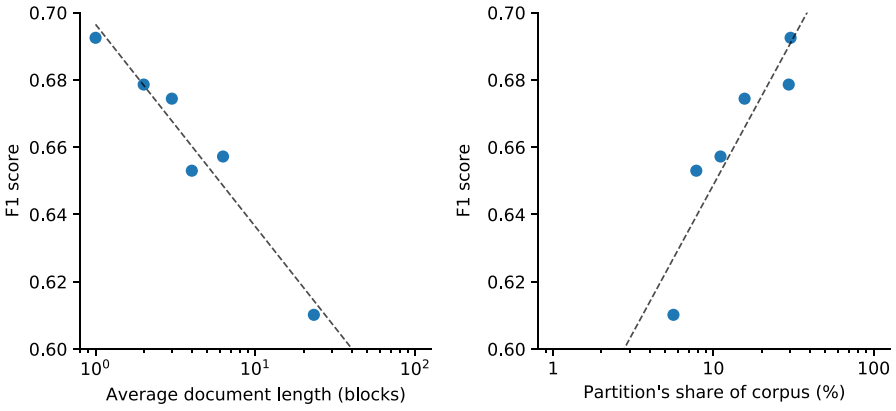

Fig. 9 The relationship between the F1-score and length of the document as measured in blocks per partition (in the left panel) and the proportion of the blocks in the corpus (in the right panel)

(2) trained on different blocks and tested on the first block (0%), (3) trained on the first block (0%) and tested on the different blocks, and (4) trained on all blocks of each training set document and tested on different blocks.

The relationship between the F1-score and position of the block is visualized in Fig. 10.

The results clearly demonstrate that when a model was trained on the beginning of the document, it always provided a better classification performance, regardless of the position of the block in the testing data (golden line). Interestingly, even a model trained on all of the blocks displayed a considerably worse performance (red line) than the point of reference condition. Thus, the results offer strong evidence that the beginning of a document likely contains the linguistic characteristics that serve as the most informative in terms of its register category, here as reflected by the F1-score. Furthermore, it is likely that the remaining blocks in a given document are likely to contain linguistic characteristics that are less informative about the register category of a given document. This is strongly supported by the surprising drop of almost 10% when the model was trained on the beginning and tested across the different positions (green line).

7.2 Building document-level predictions from blocks using an LSTM

In the previous section, we have shown that an increase in the length of a document came with a decrease in predictive power. Although the beginning of the document was shown to offer an increase in predictive accuracy, it is plausible that combining the predictions made at the different parts of a document can enhance the predictive accuracy of the model. To this end, we explore whether information from the different blocks can complement each other, hence leading to a better classification performance when all the blocks are taken into consideration using an LSTM.

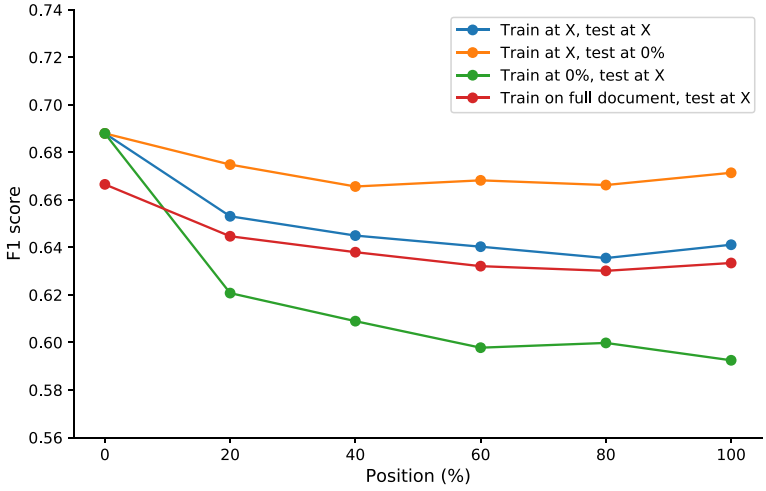

Fig. 10 The relationship between the F1-score and relative position of the block in the document given as percentage

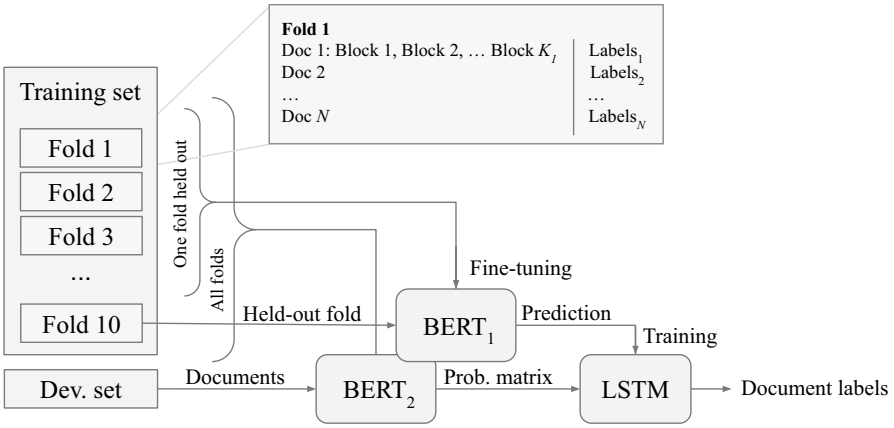

Fig. 11 The setup for applying BERT across blocks and aggregating predictions with an LSTM network. A BERT model (1) is fine-tuned in a cross-validation setup, and its predictions (label probabilities) are used in the training of the LSTM. A separate BERT model (2) is fine-tuned on all folds, and its predictions are run through the same LSTM

We used a 10-fold cross-validation to fine-tune BERT and provide training and development datasets for the LSTM model. The setup is visualized in Fig. 11, in which the tenth fold was held out. The BERT model was fine-tuned on each of the K_i block from document i, belonging to the set of nine folds, and then, the model was applied to each block of a document in the held-out fold. Repeating this procedure with each held-out fold yielded a matrix of probabilities associated with the register labels for each document in the training data. The matrices for each of the 10 folds were concatenated to train the LSTM model.

The LSTM model achieved an F1-score of 66.56% on the development set. Thus, it cannot compete with the BERT models trained only the first block, as presented in Sect. 6. These results indicate that the text blocks following the first 512 tokens in Web documents might not contain additional information that can be used to further enhance the classification performance of BERT. We return to this point in the general discussion section below.

8 General discussion and conclusions

In the current study, we have explored the automatic identification of Web registers using the CORE corpus in combination with deep learning. Although a substantial body of research has been dedicated to examining and modeling registers associated with Web documents, the novelty of the current study stems from our four contributions.

As the first contribution, we release the full CORE corpus under an open license. CORE is the first English-language Web register corpus that aims to represent the unrestricted range of Web registers attested to in the open Internet, here consisting of nearly 50,000 Web documents. Additionally, the hierarchical schema used in the

encoding of the registers in CORE provides a rich source of information for linguistically motivated studies, as well as for NLP. The corpus offers a unique opportunity to analyze and build models for gaining a better understanding of the linguistic variation associated with Web documents.

As the second contribution, we have applied several of the latest, state-of-the-art deep learning methods to model the CORE registers. These included BERT, fastText, and a CNN with fastText vectors. The best results, a 68% F1-score, were achieved with BERT Large. Considering the difficulty of the task, this can be seen as a competitive result. Importantly, the fine-tuned BERT model can be applied to a range of different Web-based language resources that are compiled without restricting the kinds of documents included in them. We can envision that this type of fine-tuned BERT can be applied to massive language resources that commonly do not include any information pertaining to register; hence, this can enhance their utility in linguistics and NLP. An example of such a dataset is OSCAR, a huge multilingual collection of Web data, aimed at training language models (Ortiz Suárez et al., 2020).

As the third contribution, we have systematically examined how to best model registers in a space where not all Web documents necessarily belong to a single category (Biber & Egbert, 2018; Biber et al., 2020; Sharoff, 2018). We treated these instances as hybrids, that is, documents combining the characteristics of different registers. To this end, we experimented with four different experimental settings. The first one consisted of the traditional multiclass setting, where a given document is always associated with one register label. The second one consisted of a multilabel setting, where a given document may be associated simultaneously with several labels. Finally, we created two additional versions of the data, where the register labels were either propagated or nonpropagated. The benefit of the propagated version of the data is that it incorporates further information the register categories represented in CORE. The results demonstrate that the multilabel setting was the most beneficial for CORE and that the model performance was improved by introducing the propagated register labels as part of the modeling framework (Madjarov et al., 2019). Although this increase in performance could be simply related to the larger amount of training data, it is likely not to be the entire story because we analyzed the performance of the model in relation to the sample size and demonstrated that overall, the size of the corpus was sufficient for this task. Thus, it is probable that the propagated register labels positively impacted model performance, at least in this task. Future studies are required to explore the possible motivations behind this behavior.

We have also shown that the registers exhibited different profiles in their discriminability (Biber et al., 2020). Certain registers, such as Lyrical and TV scripts, were accurately captured, regardless of the limited number of documents available for training. An opposite profile emerged with Web registers such as Information blog, that is, low discriminability in conjunction with high number of documents. This was evident when analyzing the confusion matrix featuring the misclassifications and further supported by the precision-recall curves associated with a set of selected Web registers. These findings reflect the variation and nature of language use on the unrestricted Web. Registers tend to follow a continuum from highly distinctive with

clear situational and linguistic characteristics to less crisp category boundaries and vague situational and linguistic characteristics. Indeed, this opens up the questions of the mapping of the situational and linguistics characteristics to specific registers. While in the prototypical case there appeared to be a well-defined one-to-one mapping between them, the category boundaries became blurred once we moved toward hybrids with multiple labels. However, future studies are required to tackle this complicated issue in more detail. For this specific purpose, CORE and the analytical framework presented in the current study can be used to guide this endeavor in the future.

As the fourth contribution, we have examined how the different parts of the document may have influenced model performance. Our analysis demonstrated that the beginning of a document provided the most predictive power and that, additionally, longer documents were more difficult to identify. Even combining the individual predictions based on the different parts of the document with an LSTM did not improve model performance. It is plausible that the importance of the beginning of a document in register identification is related to annotation practices. Specifically, it is logically plausible that the annotator focused only on the beginning in their determination of the register of a given document. Although we cannot formally verify the schema given the number of coders in CORE, the acceptance of this assertion would imply that the annotation would be less than optimal in CORE. However, this is likely to not to be the case. For example, Laippala et al. (2021) reported an F1-score of 74.5% when working with well-defined, nonhybrid CORE registers and linear support vector machines trained on bag-of-word document representation-by definition, these representations would cover the full document. Therefore, a linguistically grounded explanation seems more plausible-the beginning of the document tends to contain the most salient linguistic characteristics of its register. At the same time, the beginning of the document was defined relative arbitrarily in the current study. It is expected certain registers might display different degrees of sensitivity in how the concept of the beginning of a document is quantified.

The beginning of a document sets the stage for the purposes of effective communication. In an ideal situation, the author attempts to convey information as effectively as possible (Gibson et al., 2019; Jaeger & Tily, 2011). Thus, important information w.r.t. the register would likely be placed at the beginning of the document. Similarly, the conventionalized linguistic patterns associated with specific registers are likely to be found in the beginning of the document. Importantly, similar to the situational and linguistic characteristics of a document, the degree of conventionalization is expected to vary depending on the specific register (Biber & Conrad, 2009; Görlach, 2004; Swales, 1990). From this perspective, it is not surprising that the top three subregisters in terms of discriminability were Song lyrics, Sports report, and Recipe. They tended to exhibit well-established conventions and have been upheld by institutional practice. In the same token, the reverse holds for such subregisters as Information blog. Although blogs have certainly emerged as a register of their own right, Information blog might not be not sufficiently well formed or distinctive-at least in CORE-to allow for its reliable automatic identification. However, future studies are required to fully disentangle the possible different contributions to this phenomenon.

There are certain limitations to keep in mind when working with the data from CORE. First, it is worth noting that our solution to represent Web registers in a multi-label and propagated space does not necessarily do justice for hybrid documents that would likely be best modeled in a truly continuous space. This is certainly an interesting avenue to pursue in future research. Second, it can be argued that CORE provides only a snapshot of the Internet defined by time and space. Web registers certainly evolve over time, as reflected by changes in their relative frequencies. However, as we have used pre-trained models, it is unlikely that the vocabulary of the data used in fine-tuning goes stale. Furthermore, in terms of space, CORE is certainly a reflection of its compilation process and is likely biased towards the searchable Web as provided by Google's search algorithm. In this respect, a different compilation process, for example based on Common Crawl, would likely provide a different representation of Web registers. That being said, one of the motivations of CORE was to bring forth the Web registers that are visible to readers, typical end-users of the Web. Given this goal, CORE is likely to provide an adequate and accurate representation of Web registers. When working with textual data, it is important to remember that spaces are also defined by language, i.e., English. Additionally, a language comes with the cultural practices that reflect the registers molded through its use. In this regard, further studies are needed to develop resources for Web registers in languages other than English. Indeed, recent studies suggest that cross-lingual modeling of Web registers is achievable (Repo et al., 2021; Rönnqvist et al., 2021), but it is unclear to what extent cultural differences affect the representation of Web registers and even the very existence of specific registers. Therefore, in the future, extending the research of Web registers to a widely multilingual setting would be greatly beneficial.

Appendix 1

This appendix provides the hyperparameters used to obtain the results presented in Sect. 4:

1. *BERT Large* Batch size of 7 with a sequence length of 512 and 5 epochs, with the optimal learning rate varying between datasets. Propagated data: learning rate 7e-6. Nonpropagated data: 1e-5. Multiclass data: 4e-5.
2. *BERT Base* Batch size of 6 with a sequence length of 512 and 5 epochs. Propagated and nonpropagated datasets: learning rate 3e-5. Multiclass data: 4e-5.
3. *fastText* Propagated data: learning rate 0.1, 80 training epochs, ngram length of 3, and context window of 5. Nonpropagated data: learning rate 0.05, 60 training epochs, maximum ngram length of 3, and a context window of 5. Multiclass: learning rate 0.1, 80 training epochs, maximum ngram length of 4, and context window of 5.
4. *CNN* Propagated data: learning rate 0.001, minimum threshold 0.4, kernel size 2, and batch size 32. Nonpropagated data, learning rate 0.001, minimum threshold 0.4, kernel size 1, and batch size 32. Multiclass data: learning rate 0.001, kernel size 1, and batch size 32.

Funding Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital. Funding was provided by Academy of Finland (Grant No. 331297), Emil Aaltosen säätiö, National Science Foundation (Grant No. 1147581).

References

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, (pp. 1638–1649). Association for Computational Linguistics.

Asheghi, N., Sharoff, S., & Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation, 50*(3), 603–641.

Asheghi, R.N., Markert, K., & Sharoff, S. (2014). Semi-supervised graph-based genre classification for web pages. In *Proceedings of TextGraphs-9*, (pp. 39–47).

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio, & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015*.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation, 43*(3), 209–226.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 238–247).

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv:2004.05150.

Berninger, V. F., Kim, Y., & Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. In *BCS-IRSG Workshop on Corpus Profiling*, (pp. 1–10).

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

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

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

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

Biber, D., & Egbert, J. (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.

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge University Press.

Biber, D., Egbert, J., & Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory, 16*(3), 581–616.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5,* 135–146.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Pestimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, (pp. 451–466).

Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467–480.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 8440–8451), Online. Association for Computational Linguistics.

Degaetano-Ortlieb, S., & Teich, E. (2022). Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory, 1*(18), 175–207.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology, 66,* 1817–1831.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis* (pp. 1–32). Oxford: Blackwell. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952–1959, London: Longman (1968).

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407.

Görlach, M. (2002). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In A. Fisher, T. Lutz & P. Schneider (Eds.), *Text types and corpora: Studies in honour of udo fries* (pp. 17–27). Gunter Narr Verlag.

Görlach, M. (2004). *Text types and the history of English*. De Gruyter Mouton.

Halliday, M. (1985). Register variation. In M. Halliday & R. Hasan (Eds.), *Language, context and text: Aspects of language in a social-semiotic perspective* (pp. 29–41). Oxford University Press.

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 187–196, Turku, Finland. Linköping University Electronic Press.

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (pp. 328–339). Association for Computational Linguistics.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(3), 323–335.

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, (pp. 427–431).

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

Kwaśnik, B., Chun, Y., Crowston, K., D'Ignazio, J., & Rubleske, J. (2006). *Challenges in creating a taxonomy of genres of digital documents* (p. 225). Knowledge Organization for a Global Learning Society.

Kyröläinen, A.-J., & Kuperman, V. (2021). Predictors of literacy in adulthood: Evidence from 33 countries. *PLoS ONE, 16*(3), e0243763.

Laippala, V., Egbert, J., Biber, D., & Kyröläinen, A.-J. (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, 55*(3), 757–788.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., & Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 292–297, Turku, Finland. Linköping University Electronic Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Lee, D. (2002). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. In B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 245–292). Brill.

Levy, O. & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (pp. 302–308).

Madjarov, G., Vidulin, V., Dimitrovski, I., & Kocev, D. (2019). Web genre classification with methods for structured output prediction. *Information Sciences, 503,* 551–573.

Maharjan, S., Montes, M., onzález, F. A., & Solorio, T. (2018). A genre-aware attention model to improve the likability prediction of books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (pp. 3381–3391). Association for Computational Linguistics.

Martin, S., Liermann, J., & Ney, H. (1998). Algorithms for bigram and trigram word clustering. *Speech Communication, 24*(1), 19–37.

Meyer zu Eissen, S., & Stein, B. (2004). Genre classification of web pages. In S. Biundo, T. Frühwirth, & G. Palm (Eds.), *KI 2004: Advances in artificial intelligence* (pp. 256–269). Springer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, (pp. 746–751).

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70*(2), 151–167.

Mishra, S. (2019). 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-European languages. In *FIRE*.

Ortiz Suárez, P. J., Romary, L., & Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 1703–1714), Online. Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 2227–2237). Association for Computational Linguistics.

Petrenz, P., & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics, 37*(2), 385–393.

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

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Biber, D., Egbert, J., Pyysalo, S., & Laippala, V. (2021). Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, (pp. 183–191), Online. Association for Computational Linguistics.

Rönnqvist, S., Skantsi, V., Oinonen, M., & Laippala, V. (2021). Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rosso, M. A. (2008). User-based identification of Web genres. *Journal of the American Society for Information Science and Technology, 59*(7), 1053–1072.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Santini, M. (2007). Characterizing genres of web pages: Genre hybridism and individualization. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, (pp. 71–71).

Santini, M. (2008). Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing & Management, 44*(2), 702–737.

Santini, M. (2011). Cross-testing a genre classification model for the web. In A. Mehler, S. Sharoff, & M. Santini (Eds.), *Genres on the Web: Computational Models and Empirical Studies* (pp. 87–128). De Gruyter.

Santini, M., Mehler, A., & Sharoff, S. (2011a). *Riding the Rough Waves of Genre on the Web*, pp. 3–30.

Santini, M., Mehler, A., & Sharoff, S. (2011b). Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff, & M. Santini (Eds.), *Genres on the Web: Computational Models and Empirical Studies* (pp. 3–30). Springer.

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

Sharoff, S., Wu, Z., & Markert, K. (2010). The web library of babel: Evaluating genre collections. In *Proceedings of LREC)*.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, (pp. 384–394).

van der Wees, M., Bisazza, A., & Monz, C. (2015). Translation model adaptation using genre-revealing text features. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, (pp. 132–141). Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Vidulin, V., Lustrek, M., & Gams, M. (2009). Multi-label approaches to web genre identification. *JLCL, 24,* 97–114.

Webber, B. (2009). Genre distinctions for discourse in the Penn treebank. In *Proceedings of ACL-IJCNLP*, (pp. 674–682).

Worsham, J., & Kalita, J. (2018). Genre identification and the compositional effect of genre in literature. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1963–1973, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xiao, Y., & Cho, K. (2016). Efficient character-level document classification by combining convolution and recurrent layers. arXiv preprint arXiv:1602.00367.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.