Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora

Amanda Myntti[1], Liina Repo[1], Elian Freyermuth[2], Antti Kanner[1], Veronika Laippala[1], Erik Henriksson[1],
[1]University of Turku, [2]National Graduate School of Engineering of Caen
Correspondence: amanda.a.myntti@utu.fi

Abstract

Web-scale corpora present valuable research opportunities but often lack detailed metadata, making them challenging to use in linguistics and social sciences. This study tackles this problem by exploring automatic methods to classify web corpora into specific categories, focusing on text registers such as *Interactive Discussion* and literary genres such as *Politics and Social Sciences*. We train two machine learning models to classify documents from the large web-crawled OSCAR dataset: a register classifier using the multilingual, manually annotated CORE corpus, and a genre classifier using a dataset based on Kindle US&UK. Fine-tuned from XLM-R Large, the register and genre classifiers achieved F1-scores of 0.74 and 0.70, respectively. Our analysis includes evaluating the distribution of the predicted text classes and examining the intersection of genre-register pairs using topic modelling. The results show expected combinations between certain registers and genres, such as the *Lyrical* register often aligning with the *Literature & Fiction* genre. However, most registers, such as *Interactive Discussion*, are divided across multiple genres, like *Engineering & Transportation* and *Politics & Social Sciences*, depending on the discussion topic. This enriched metadata provides valuable insights and supports new ways of studying digital cultural heritage.

1 Introduction

Automatically collected web-scale corpora, encompassing billions of words, offer significant opportunities for research across a range of disciplines, including computational linguistics, natural language processing, social sciences, and digital humanities. These extensive resources have been, and continue to be, instrumental in advancing large language models, such as the one underpinning ChatGPT. Additionally, these corpora contain vast amounts of text produced in varied contexts and for diverse purposes, serving as repositories for new and evolving forms of digital cultural heritage. Consequently, web-scale corpora not only open new research avenues in the humanities and social sciences but also necessitate multidisciplinary collaboration to ensure their effective use (Laippala et al., 2021b; Välimäki and Aali, 2022).

A notable challenge in utilizing web-scale corpora is the lack of detailed metadata describing their contents. Without such metadata, texts of diverse varieties—such as legal notices, advertisements, news articles, fiction, and song lyrics—are treated equally, despite their distinct linguistic characteristics. This complicates the filtering and selection of data relevant to specific research tasks. Among others, these procedures are often crucial for building accurate language models, where the absence of metadata increases the risk of models learning from biased, toxic, or irrelevant data (e.g. Gehman et al., 2020; Carlini et al., 2021; Dodge et al., 2021; Feng et al., 2023; Bannihatti Kumar et al., 2023; Mallen et al., 2023). Text type metadata has also been shown to enhance the performance of various Natural Language Processing (NLP) applications, including part-of-speech taggers, parsers, and information retrieval systems (e.g. Karlgren and Cutting, 1994; Vidulin, 2007; Giesbrecht, 2009; Santini et al., 2011; Van Der Wees et al., 2018; Argamon, 2019).

To address this challenge, we explore a dual approach to classifying web corpora into specific text categories, focusing on two key approaches: register and genre. Registers, as they are typically applied in corpus linguistics, refer to culturally recognizable text varieties characterized by their communicative situation and functionally related linguistic features (Biber 1988; Egbert and Biber 2019; Biber and Egbert 2023). Genres, in literary studies, are often utilized to examine various forms of literary work, focusing on content, context, and narrative tools (e.g., Goyal and Vuppuluri 2022;

Zhang et al. 2022).

In recent years, text classification and specifically web register identification has taken leaps forward, with web register classifiers achieving nearly human-level performance (Laippala et al. 2023; Kuzman et al. 2023b; Henriksson et al. 2024). These advancements enable us to enhance document metadata substantially. However, when register classification is applied to web-scale corpora, the predicted register classes are still very broad and include a wide range of linguistic variation.

Therefore, in this study, we examine whether combining two approaches, namely registers and genres, can enhance the depth of the available information for a given document. Specifically, we examine the intersections between register and genre labels in a text classification setting and explore how these intersections, along with the resulting new metadata, can allow for novel uses of the corpus in other studies. To achieve this, we apply machine learning to train two text classifiers: one targeting registers and the other focusing on genres. These classifiers are then used to predict classes for one million documents from the widely used web-scale OSCAR dataset (Open Super-large Crawled ALMAnaCH coRpus; Ortiz Suárez et al. 2019; Laippala et al. 2022).

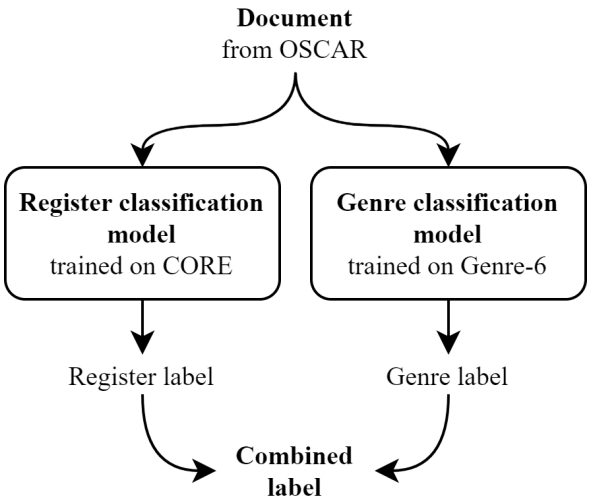

Figure 1: Workflow of our process.

To assess the quality of the new metadata–combined label of register and genre classifiers–we evaluate two conditions. First, we examine the overlap between the two labelling schemes, as cross-labelling has little value if the classifiers produce similar results. Ideally, each register class should map onto multiple genre categories, adding previously unattainable information to each document. Second, we evaluate whether these cross-labellings are meaningful by using topic modelling to extract topic words for each register-genre combination.

2 Data

We use three different datasets for our experiments, namely the Corpus of Online Registers of English (CORE), the Genre-6 literary genre corpus, and the Open Super-large Crawled ALMAnaCH coRpus (OSCAR), each for a different task.

2.1 Register Data

The register classifier is trained using the CORE corpus[1] (Egbert et al. 2015; Laippala et al. 2023), which consists of manually register-annotated English web texts. The corpus contains nearly 50,000 documents and covers the full range of English web registers. The annotation process involved four individual annotators independently assigning each document a main register label and, when possible, a subregister label for a more detailed description. In cases of annotator disagreement, a document could be assigned multiple register labels. The annotation process resulted in a hierarchical multilabel register scheme with eight main register categories with broad, functional labels such as *Narrative*, *Informational description* and *Opinion*, and tens of more detailed subcategories such as *News report*, *Research article* and *Review*. Following Laippala et al. (2022), we slightly modify this hierarchy by mapping some subregisters together to enhance classifier performance. All the main registers and subregister categories of CORE used in this study are shown in Table 1, and the mapping from original CORE labels to our modified labels can be found in Appendix A.

2.2 Genre Data

For the genre classifier, we train using the Genre-6 dataset[2], which is derived from Kindle UK & US. Genre-6 comprises over 20,000 short stories and novels, with genre categories assigned by the authors. The genre labels are selected from the available categories on Kindle UK & US, resulting

[1]Specific version available at https://github.com/TurkuNLP/CORE-corpus.

[2]The original dataset is available at https://huggingface.co/datasets/marianna13/the-eye, and the cleaned version used in training is available on our Huggingface page https://huggingface.co/datasets/TurkuNLP/genre-6.

Main register	Sub register	Support
How-to/Instruction (HI)		**2047**
	Recipe (re)	189
	(no subregister)	1858
Interactive discussion (ID)	–	**3449**
Informational description (IN)		**13047**
	Description of a thing or a person (dtp)	5444
	Encyclopedia article (en)	556
	FAQ about information (fi)	337
	Legal terms and conditions (lt)	202
	Research article (ra)	936
	(no subregister)	5572
Informational persuasion (IP)		**2011**
	Description with intent to sell (ds)	1422
	Editorial (ed)	94
	(no subregister)	495
Lyrical (LY)	–	**680**
Narrative (NA)		**21534**
	News report (ne)	11785
	Narrative blog (nb)	3620
	Sports report (sr)	3164
	(no subregister)	2965
Opinion (OP)		**10754**
	Advice(av)	1161
	Opinion blog (ob)	5242
	Reviews (rv)	2065
	Religious blogs/sermons (rs)	776
	(no subregister)	1510
Spoken (SP)		**736**
	Interview (it)	537
	(no subregister)	199
Total		**54258**

Table 1: Main and subregister categories of CORE used in this study. Original scheme in Egbert et al. (2015) and mapping to this scheme in Appendix A.

Genre	N
Cookbooks, Food & Wine (Cook)	370
Engineering & Transportation (Engn)	1688
Literature & Fiction (Lit)	4969
Medicine & Health Sciences (Med)	763
Politics & Social Sciences (Pol)	2134
Science & Math (Sci)	1474
None (in prediction only)	–
Total	11398

Table 2: Chosen genre labels of Genre-6 dataset.

in a multilabel genre annotation scheme with categories such as *Children's Books*, *Science & Math*, and *Action & Adventure*. Initially, we performed minor preprocessing steps to improve data quality, such as excluding categories with minimal support.

Because some of the genre classes in the Genre-6 dataset are overlapping, for this experiment we further chose a subset of genres that maximizes the performance in two ways: Firstly, the chosen genres need to be present in our target corpus, OSCAR. As OSCAR is a web corpus, genres most suitable for our task include common topics in online sources. Secondly, to try to maximize classifier performance, we chose categories by training the classifier with different candidate subsets and evaluating the classifiers' performance. Table 2 presents the final genre categories used in this study. Lastly, we included a "None" category for uncertain classifications and to address those that fall outside our chosen category scheme, acknowledging that our categories do not fully represent the entire web and some common internet topics, such as religion, are not well covered by this set of labels.

2.3 The Labelled Target Corpus

We use the OSCAR corpus (Ortiz Suárez et al., 2019) for our analysis of the register–genre intersection. The OSCAR corpus was created by extracting and filtering text from Common Crawl[3]. It is a multilingual corpus comprising billions of words across 166 languages. For our study, we specifically use the pre-classified version, Register OSCAR[4] (Laippala et al., 2022), which has un-

[3] https://commoncrawl.org

[4] https://huggingface.co/datasets/TurkuNLP/register_oscar

dergone further processing, including additional boilerplate text removal, resulting in higher quality than the original OSCAR corpus. We do not use the existing register labels as they only cover the main register level labels, but re-predict the dataset with our register classifier to also get access to the supplementary information of the sublabels. However, in some of our analyses, such as topic modelling, we present results at the level of the main labels for simplicity; in these cases, the sublabels have been aggregated into the main labels according to the label hierarchy shown in Table 1. We label a sample of 1 million documents from OSCAR in this study.

3 Experimental Setup

Figure 1 illustrates the workflow of the classification process. We use two classifiers to predict labels for each document from the OSCAR corpus, and the resulting labels are then combined into a single, combined label.

3.1 Classifier training

We approach classification using a multilabel setting, which previous research on register identification has shown to produce significantly higher scores compared to a single-label, multi-class approach (e.g. Egbert et al. 2015; Madjarov et al. 2019; Sharoff 2021; Laippala et al. 2023). The data is split into training, development, and test sets using stratified sampling, with proportions of 80%, 10%, and 10% for the Genre-6 corpus, and 70%, 10%, and 20% for the CORE corpus. We limit each document to the first 512 tokens, based on (1) the maximum token limit of the models used and (2) the findings of Laippala et al. (2023), which indicate that the best performance in register classification is achieved using the initial 512-token chunks of text.

The register classifier is trained using the CORE corpus, whereas the genre classifier is trained using the Genre-6 corpus, as mentioned in the previous section. Both classifiers are implemented by fine-tuning XLM-RoBERTa-Large (XLM-R; Conneau et al. 2020) for a multilabel classification task using the Huggingface Transformers library. We selected XLM-R due to its efficiency and strong performance in earlier studies of register classification (e.g., Repo et al. 2021). Both models use a multilabel setup and Binary Cross-Entropy Loss, and the prediction threshold is optimized for the F1-score. We also experimented with Focal Loss (Lin et al., 2018), but ultimately chose Binary Cross-Entropy Loss, as it produced higher-quality predictions in manual evaluation, despite Focal Loss yielding slightly better F1-scores.

3.2 Topic modelling

For our topic modelling experiments, we utilized the Latent Dirichlet Allocation (LDA) algorithm implemented in the `gensim` library[5]. We extracted topics separately from each register–genre intersection class. We also experimented with transformer based BERTopic (Grootendorst, 2022). Using BERTopic, we were able to extract more detailed topics; however, the initial experiments showed that the support of each intersection class influenced the quality of the results, whereas this effect was diminished with the LDA model. Additionally, we preferred the simpler presentation of the LDA model. We used the following parameters: 30 passes, 1 topic, 10 best words per topic. For the largest combination classes, where support is in the hundreds of thousands, we randomly select a subset of 10000 documents for the analysis. We lemmatize, and remove punctuation and English stop words using the `nltk` library[6]. We also experimented with extracting multiple topics for each register–genre combination, which revealed the structure of some intersection classes better. However, for simplicity, we present one topic per class. Additionally, in this part of the experiment, we focus on the main hierarchy level of the register labels to maintain clarity in the presentation.

4 Results

4.1 Classifier Evaluation

Our results show that the register classifier is able to reach an F1-score of 0.74, whereas the genre classifier's F1-score is 0.70. The class-specific performance of each model is detailed in Tables 3 and 4. Although these scores vary considerably, ranging from 0.45 for *Science & Math* to 0.89 for *Lyrical*, they are consistent with previous register identification results (Egbert et al., 2015), indicating that the predictions are reasonably reliable.

The variability in identification performance can be attributed to how well the registers and genres are defined linguistically, which affects how ac-

[5] `https://radimrehurek.com/gensim/models/ldamodel.html`

[6] `https://www.nltk.org/`

Label	F1-Score	Support
LY	0.8949	135
SP	0.7032	146
ID	0.8475	686
NA	0.8405	4264
HI	0.6788	411
IN	0.7176	2596
OP	0.6854	2129
IP	0.5591	402
it	0.7045	104
ne	0.8120	2359
sr	0.8942	635
nb	0.6745	722
re	0.8116	37
en	0.8079	108
ra	0.6686	189
dtp	0.5271	1090
fi	0.5000	69
lt	0.5763	40
rv	0.7040	411
ob	0.5591	1051
rs	0.7278	157
av	0.5119	236
ds	0.6427	280
ed	0.0000	19
μ (micro)	0.74	18276

Table 3: Results of our trained register classifier. The threshold for classification is set at 0.4, optimized wrt. F1-score.

curately they can be classified (Biber and Egbert, 2018; Biber et al., 2020; Laippala et al., 2021a). For example, *Lyrical* texts, which mostly consist of song lyrics and poems, have distinct characteristics that make them easier to classify accurately. In contrast, texts within the *Advice* subregister vary widely and can be mistaken for other opinionated registers, such as *Opinion Blog*, leading to lower identification scores. A similar pattern is observed in the genre-specific performance, where some classes, like *Science & Math* and *Medicine & Health Sciences*, contain texts that are hard to distinguish clearly between these categories.

4.2 Register and Genre Intersection

Figure 2 illustrates the intersections between the registers and genres. The figure confirms that no register and genre categories fully overlap, demonstrating that cross-labelling with our setup achieves the intended outcome: it refines the classification and enriches the information for each document.

Label	F1	N
Cookbooks, Food & Wine	0.59	35
Engineering & Transportation	0.65	172
Literature & Fiction	0.81	535
Medicine & Health Sciences	0.61	72
Politics & Social Sciences	0.53	194
Science & Math	0.45	144
μ (micro)	0.70	1152

Table 4: Results of our trained genre classifier. The threshold for classification is set at 0.3, optimized wrt. F1-score.

To evaluate the increase in information quantitatively, we calculate mutual information (MI) between the register and genre labels. Mutual information measures the information one label provides about the other and is calculated from the joint probability distribution of the genre and register labels. We use the `scikit-learn`[7] library to calculate this value, with multilables separated for this step and main-subregister combinations treated as separate classes. Although values of mutual information are not comparable, values close to zero indicate low levels of dependency between the variables, which is ideal in our case. We also calculate the increase of information using Shannon's entropy H as $\sum_{x \in \mathcal{X}} -P(x) \log_2 P(x)$, which measures the informational value of the variable, with P standing for the marginal probability function is separate label cases and joint probability function in the combined label case. The results are presented in Table 5. These values show that the genre label cannot be inferred from the register label and that the information content increases with the combined labelling compared to using each label scheme separately. Specifically, the additional information contributed by the genre labels to the register labels is $H(\text{genre}|\text{register}) = H(\text{register,genre}) - H(\text{register}) = 2.073$ bits.

MI	H(register)	H(genre)	H(register, genre)
0.109	3.370	2.229	5.443

Table 5: Mutual information (MI) between register and genre and the entropy H of register labels (main and subregisters), genre labels, and the combined labelling.

From Figure 2, expected combinations between certain registers and genres can be seen. For in-

[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif

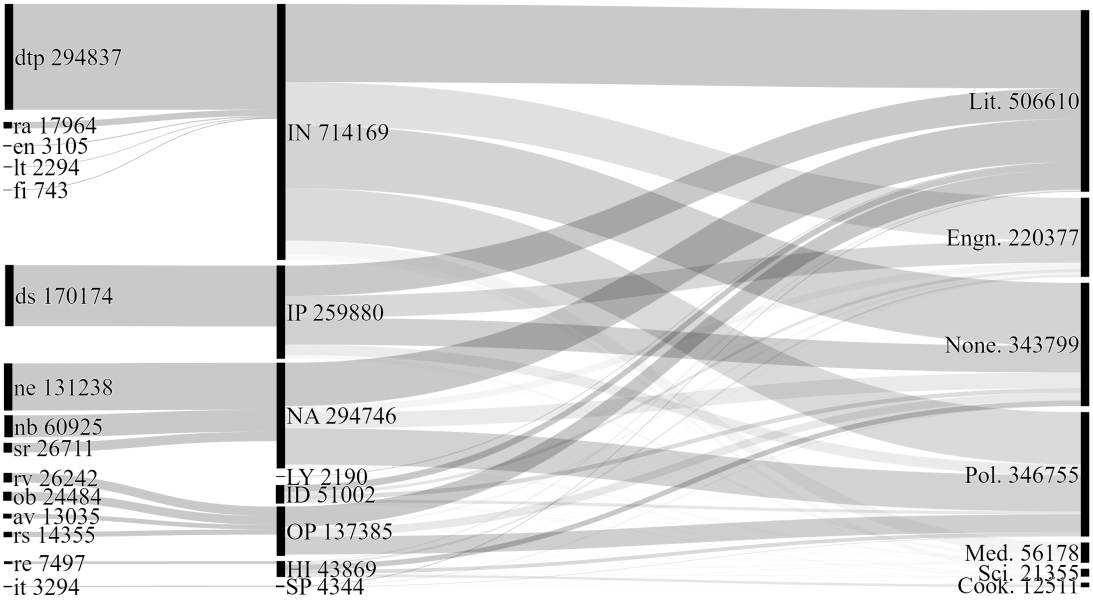

Figure 2: Intersection of registers and genres in the OSCAR corpus. Register sublabels on the left, main level labels on the centre and genres on the right. See Table 1 for register abbreviations and Table 2 for genre abbreviations. The thickness of the connection reflects the relative frequency of co-occurrence, with opacity adjusted to make register-class-wise frequencies more visible. Numbers indicate the number of documents in OSCAR labelled as each class, with multilabels separated. A small number of instances with incorrect register label hierarchy removed.

stance, as could be anticipated, the *Lyrical* register and the *Literature & Fiction* genre co-occur very often, although *Lyrical* also intersects with other genres – such as *Politics & Social Sciences*. Manual evaluation shows this particular intersection class contains religious poetry and lyrical texts with social commentary aspects. An example of this can be seen as the fourth example in Table 7.

Similarly, the *Spoken* register seems to be mostly associated with the *Politics & Social Sciences* and *Literature & Fiction* genres. These combinations convincingly suggest documents with spoken elements, such as conversations or interviews. In *Literature & Fiction*, these might include dialogues between characters, and in *Politics & Social Sciences*, they could be interviews or political speeches. It is also noteworthy that the *Literature & Fiction* class is very large, both in terms of support and the content used for classifier training, and thus covers a variety of texts, including low-quality content. As stated previously, our genre selection criteria were influenced by both the predicted contents of the target documents and the measured classifier performance, which in this case resulted in the selection of the large and sometimes vague class of *Literature & Fiction*.

All other registers are divided into multiple genres—in particular, the largest registers cover documents across all of them. The intersections are logical; for instance, the largest intersections of *Informational persuasion* are with *Engineering & Transportation* and *Literature & Fiction*, suggesting that these are persuasive documents such as editorials or company websites, discussing topics relevant to these genres, like advertisements for technological devices. Similarly, in addition to these two genres, the largest register class *Informational description* intersects mostly with *Politics & Social Sciences* and *None*, which facilitates the identification of different types of informative/descriptive documents, which were previously only marked by their register characteristics. *Interactive Discussion* intersects with all genres, with the strongest connections to *Engineering & Transportation* and *Politics & Social Sciences*, depending on the discussion topic. Finally, the *Narrative* register also intersects with all these genres. This register primarily includes news articles and narrative blogs, suggesting that these subregisters discuss topics related to these genres. Previously, identifying such combinations required manual searches within the documents, but with the addition of genre labels, these intersections can now be easily categorized, filtered, or selected from the corpus.

From the perspective of the genres, similar results can be seen. For example, as expected, the genre *Medicine & Health Sciences* mostly intersects with *Informative description*, however, we were able to extract conversations and ads with health care related topics with the intersection of registers *Interactive discussion* and *Informational persuation*.

Finally, the *None* category for uncertain genre predictions contains documents from all the register classes, most notably from *Informational description* and *Informational persuasion*. This is expected due to the class size imbalance, however, these classes are over-represented in the *None* genre category compared to the *Narrative* register class, which despite its large size rarely intersects with the *None* genre. The reasons behind this lie in the variety of topics discussed in the respective registers. The registers frequently intersecting with the *None* genre seem to cover topics not included in the genres, or the documents within the register simply do not feature a well-defined genre at all. For instance, in Table 7, a text combining *Interactive discussion* with *None* a disproved message on a forum, which does not fit any of the specified genres. This shows that the *None* class is linguistically motivated and provides meaningful intersections.

4.3 Contents of the intersection classes

We use topic modelling to gain insights into each register–genre intersection class, and the results are presented in Table 6. Some register–genre combination classes produced topic keywords that reflect features of both the register and the genre involved. For instance, most keywords extracted from combinations with the register *How-to Instructions* include characteristic verbs such as "make" and "use", which are frequently used in instructive texts. Similarly, the *Interactive discussion* register produced topic words containing personal pronouns like "I('m)" and "us", which are typical of discussions. In the *Informational persuasion* register, words like "product", "help", and "business" suggest that the texts feature ads and other commerce-related documents. The *Informational description* register is associated with verbs in the past tense, which is characteristic of texts such as encyclopedic entries.

From the perspective of the genre classes, similarities are shared over register class boundaries. For instance, "food", a characteristic word for the *Cookbooks, Food & Wine* genre, appears in all but one of the combination classes. Interestingly, the intersection of *Cookbooks, Food & Wine* and *Interactive discussion* contains more documents seemingly about kitchen decor, not only about food. In the *Medicine & Health Sciences* genre, themes present throughout are patients and helping, while in *Science & Math*, recurrent topic words include

Reg.	Genre	N	Topic keywords
HI	Cook.	6838	make, add, recipe, minutes, use
	Engn.	5875	use, one, need, make, new
	Lit.	3815	like, also, love, new, would
	Med.	877	skin, may, help, make, like
	Pol.	10219	use, also, make, time, may, new
	Sci.	975	time, water, may, make, need
ID	Cook.	256	rustic, home, plans, house, would
	Engn.	8759	would, get, new, time, need, work
	Lit.	18947	would, i'm, know, good, think
	Med.	1278	would, time, also, good, i'm
	Pol.	8300	time, people, need, us, know
	Sci.	1001	one, would, know, new, good
IN	Cook.	2228	food, used, make, made, many
	Engn.	99490	new, time, used, us, system, data
	Lit.	190734	dating, free, pdf, online, first
	Med.	35252	patients, health, care, treatment, new
	Pol.	134903	people, use, get, us, information
	Sci.	12814	one, two, water, used, species
IP	Cook.	1396	food, like, product, used, new
	Engn.	41796	new, us, service, quality, system
	Lit.	62779	free, get, great, book, home
	Med.	4352	skin, health, help, body, new
	Pol.	21401	new, business, us, help, people
	Sci.	1112	fishing, new, plants, water, use
LY	Lit.	1908	download, love, would, know, i'm
NA	Cook.	427	make, food, good, would, made
	Engn.	17251	said, would, time, first, us
	Lit.	113824	new, would, first, back, people
	Med.	5983	new, health, said, people, may
	Pol.	93575	said, would, people, us, first
	Sci.	3602	new, said, would, water, two
OP	Cook.	705	food, make, good, get, great,
	Engn.	5652	new, time, car, make, well
	Lit.	52402	would, also, us, people, first
	Med.	2697	get, people, would, may, help
	Pol.	45086	people, would, time, us, new
	Sci.	274	one, people, new, us, many
SP	Engn.	147	would, think, work, time. people
	Lit.	2060	think, people, would, really, know
	Pol.	1370	people, think, know, us, going

Table 6: Selected 5 of 10 top topic words extracted from the intersection classes. "None" class and classes with less than 100 documents omitted. See Table 1 for register abbreviations and Table 2 for genre abbreviations.

Register	Genre	Truncated document
IN, ra	Engineering & Transportation	The management of existing road infrastructures is a multidisciplinary activity that involves structural engineering, material science, management, economics and ecology. The objective is to achieve maximum availability of road links at minimum societal costs. Recently, tools (Bridge Management Systems, BMSs) have been developed to help decision makers to determine the optimal management strategies within available resources.
NA, ne, IN, dtp	Politics & Social Sciences	Second-year medical student Walter Humann is the winner of the 2015 Danny Jones History of the Health Sciences Student Essay Competition. A panel of three judges rated the six entries on the quality of writing, the comprehension of the issues, the clarity of discussion, and the applicability of the topic to the history of health care. His essay Medical Progress in the West: A Historical Perspective addresses the three eras of history in which medical developments struggled against societal norms.
OP	Cookbooks, Food & Wine	White chocolate isn't really chocolate at all. While it contains the cocoa butter of true chocolate, it lacks cocoa solids, the element responsible for milk and dark chocolate's characteristic brown color and nutty roasted flavor. Other pale confections labeled simply "white" chips or bars (these boast less than the 20 percent cocoa butter required to earn the designation "white chocolate") are just as common in the baking aisle of the supermarket.
LY	Politics & Social Sciences	I am obnoxious to each carping tongue Who says my hand a needle better fits. A Poet's Pen all scorn I should thus wrong, For such despite they cast on female wits. If what I do prove well, it won't advance, They'll say it's stol'n, or else it was by chance.
ID	None	I posted a question to the forum an hour ago and received an email saying the moderators had approved the content. Checking a few minutes ago I see a notice that the message has now be disproved. How do I contact a moderator to find out what is going on?

Table 7: Examples of our classification results. Texts truncated and original spelling retained. See register abbreviations in Table 1.

"water" and "plants".

Some drawbacks can also be seen from these topic keywords. The keywords for the genre classes *Literature & Fiction* and *Politics & Social Sciences* offer few distinguishing words, apart from references to people. Manual evaluation confirms this observation; as previously noted, the *Literature & Fiction* class, in particular, contains a large variety of texts, and thus extracting a single topic from this class does not result in coherent keywords. In the case of *Lyrical* and *Spoken* registers, the small support affects the results of the topic modelling. However, for the *Lyrical* register, keywords like "love" and "I'm" are characteristic of song lyrics and poetry; the keyword "download" likely appears due to boilerplate text commonly found on song lyrics web pages.

Table 7 presents example documents associated with the register–genre intersections. The first example, labelled as a *Informational description – Research article* and *Engineering & Transportation* reflects the extracted topics, including topic keywords like "system". In the second example, labelled as *Informational description – Description of a thing or a person*, *Narrative – News report* for register and *Politics & Social Sciences* for genre, clearly contains characteristics of a news article and describes a person. At first glance, the text could be labelled as *Medicine & Health Sciences*, however, the document actually covers history and struggles against societal norms. In the third example, the vocabulary clearly reflects the class *Cookbooks, Food & Wine*, but the tone is correctly identified as opinionated, as the text contains emotionally charged adjectives, which are also seen in the topic words extracted for the register class *Opinion*. This example also justifies our decision to include the *Cookbooks, Food & Wine* genre, despite the CORE register scheme already containing a *Recipe* subregister, as it allows us to capture a broader range of food-related documents.

5 Conclusion

In this study, we experimented with labelling a large internet corpus using two classifiers and evaluated the new metadata produced by the intersection of two classification schemes. We trained the classifiers on available register and genre datasets in a multilabel classification setting and reached reliable results. We then analyzed the distributions of the intersection classes and extracted topic keywords from them. Our evaluation, based on quantitative analysis using topic modelling and close reading, demonstrated that the predicted genre and register labels provide meaningful auxiliary information, facilitating new ways to use the corpus.

This is particularly valuable for digital humanities and cultural heritage studies, as it allows for

richer contextualization and more nuanced analysis of historical documents, literary texts, and other cultural artefacts. Enhanced metadata can also support the preservation and accessibility of digital archives, ensuring that documents are more easily discoverable and interpretable.

In the future, we will aim to improve model performance by experimenting with different model architectures and refining the chosen classes of the genre classifier. Specifically, we recognize that the current *Literature & Fiction* category is too broad and plan to subdivide it into more specific genres. We are also interested in exploring different data augmentation techniques (e.g. label cleaning tools[8]), particularly for our genre corpus, which has shown apparent label issues during manual review.

6 Limitations

Our work is conducted entirely in English. While previous studies (e.g. Repo et al., 2021; Rönnqvist et al., 2021) suggest that the effectiveness of register classification using the CORE scheme may transcend language barriers, the same may not apply to our genre classification system. Additionally, we based our genre classification training on methods typically used for registers, which may not perfectly align with genre distinctions. Our selection of genre categories relied on the support of the classes, partly due to the limited number of instances in the corpus. As previously mentioned, we recognize the bias towards more technical genres, as the selected genre categories contain both engineering and science related topics but lack coverage of other common internet subjects. Another possible approach to text classification would have been to use recent large language models such as ChatGPT[9]. Kuzman et al. (2023a) compared the performance of an XLM-R-based model to GPT-3.5 and GPT-4 (OpenAI, 2023) in register (genre in their terminology) classification. In their experiments, GPT-4 and XLM-R-Large performed similarly on out-of-domain English testset. This indicates that using GPT-like models for this task holds substantial potential.

Acknowledgements

We wish to acknowledge FIN-CLARIAH (Common Language Resources and Technology Infrastructure), and CSC – IT Center for Science for computational resources. This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Academy of Finland under grant numbers 358720 and 331297. We thank LAION[10] and Ontocord.AI[11] for collaboration on the genre dataset.

[8] https://github.com/cleanlab/cleanlab

[9] https://openai.com/chatgpt/

References

Shlomo Engelson Argamon. 2019. Computational register analysis and synthesis. *ArXiv*, abs/1901.02543.

Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2023. Privacy adhering machine un-learning in NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 268–277, Nusa Dua, Bali. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.

Douglas Biber and Jesse Egbert. 2023. What is a register?: Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, 5.

Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

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

Jesse Dodge, Ana Marasovic, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Conference on Empirical Methods in Natural Language Processing*.

[10] https://laion.ai/

[11] https://www.ontocord.ai/

Jesse Egbert and Doug Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Eugenie Giesbrecht. 2009. In search of semantic compositionality in vector spaces. In *Conceptual Structures: Leveraging Semantic Technologies*, pages 173–184, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anshaj Goyal and Prem Vuppuluri. 2022. *Statistical and Deep Learning Approaches for Literary Genre Classification*, pages 297–305. Springer Singapore, Singapore.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Erik Henriksson, Amanda Myntti, Anni Eskelinen, Selcen Erten-Johansson, Saara Hellström, and Veronika Laippala. 2024. Untangling the unrestricted web: Automatic identification of multilingual registers.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023a. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2023b. Get to know your parallel data: Performing English variety and genre classification over MaCoCu corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103, Dubrovnik, Croatia. Association for Computational Linguistics.

Veronika Laippala, Jesse Egbert, Douglas Biber, et al. 2021a. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation*, 55:757–788.

Veronika Laippala, Aki-Juhani Kyröläinen, Jenna Kanerva, and Filip Ginter. 2021b. Dependency profiles in the large-scale analysis of discourse connectives. *Corpus Linguistics and Linguistic Theory*, 17(1):143–175.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

Gjorgji Madjarov, Vedrana Vidulin, Ivica Dimitrovski, and Dragi Kocev. 2019. Web genre classification with methods for structured output prediction. *Information Sciences*, 503:551–573.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 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, pages 183–191, Online. Association for Computational Linguistics.

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

Marina Santini, Alexander Mehler, and Serge Sharoff. 2011. Riding the Rough Waves of Genre on the Web. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 3–30. Springer Netherlands, Dordrecht.

Serge Sharoff. 2021. Genre annotation for the Web: Text-external and text-internal perspectives. *Register Studies*, 3(1):1–32.

Marlies Van Der Wees, Arianna Bisazza, and Christof Monz. 2018. *Evaluation of machine translation performance across multiple genres and languages*, pages 3822–3827. European Language Resources Association (ELRA). International Conference on Language Resources and Evaluation, LREC ; Conference date: 07-05-2018 Through 12-05-2018.

Vedrana Vidulin. 2007. Training the genre classifier for automatic classification of web pages. *Journal of Computing and Information Technology*, 15.

Reima Välimäki and Heta Aali. 2022. *The Ancient Finnish Kings and their Swedish Archenemy: Nationalism, Conspiracy Theories, and Alt-Right Memes in Finnish Online Medievalism*, pages 55–78. Studies in Medievalism XXXI. Boydell and Brewer, Boydell and Brewer.

Jinbin Zhang, Yann Ciarán Ryan, Iiro Rastas, Filip Ginter, Mikko Tolonen, and Rohit Babbar. 2022. Detecting sequential genre change in eighteenth-century texts. In *Proceedings of the Computational Humanities Research Conference 2022*, CEUR Workshop Proceedings, pages 243–255, Germany. CEUR-WS.org.

A CORE label scheme modification

The mapping from CORE original label scheme to the one used in this study is presented in Table 8.

Original CORE		Simplified scheme
Register	Subregister	
Narrative (NA)	News report/blog (ne)	–
	Sports report (sr)	–
	Personal/diary blog (pb)	Narrative blog (nb)
	Historical article (ha)	Narrative (NA)
	Travel blog (tb)	Narrative blog (nb)
	Short story (ss)	Narrative (NA)
	Novel	Narrative (NA)
	Biographical story/history	Narrative (NA)
	Magazine article (ma)	Narrative (NA)
	Obituary	Narrative (NA)
	Memoir	Narrative (NA)
	Other narrative (on)	Narrative (NA)
Opinion (OP)	Opinion blog (ob)	–
	Review (rv)	–
	Religious blog/sermon (rs)	–
	Advice (av)	–
	Letter to the editor (le)	Opinion (OP)
	Self-help	Opinion (OP)
	Advertisement (ad)	Opinion (OP)
	Other opinion (oo)	Opinion (OP)
Informational Description (IN)	Description of a thing (dt)	Description of a thing or a person (dtp)
	Informational blog (ib)	Informational Description (IN)
	Description of a person (dp)	Description of a thing or a person (dtp)
	Research article (ra)	–
	Abstract	Informational Description (IN)
	FAQ about information (fi)	–
	Legal terms and conditions (lt)	–
	Course materials (cm)	Informational Description (IN)
	Encyclopedia article (en)	–
	Technical report (tr)	Informational Description (IN)
	Other informational (oi)	Informational Description (IN)
Interactive Discussion (ID)	Discussion forum (df)	Interactive Discussion (ID)
	Question/answer forum (qa)	Interactive Discussion (ID)
	Reader/viewer responses (rr)	Interactive Discussion (ID)
	Other interactive discussion (of)	Interactive Discussion (ID)
How-to Instructional (HI)	How-to (ht)	How-to Instructional (HI)
	Recipe (re)	–
	Instructions	How-to Instructional (HI)
	FAQ about how-to (fh)	How-to Instructional (HI)
	Technical support (ts)	How-to Instructional (HI)
	Other how-to/instructional (oh)	How-to Instructional (HI)
Informative Persuasion (IP)	Description with intent to sell (ds)	–
	Persuasive article or essay (pa)	Informative Persuasion (IP)
	Editorial (ed)	–
	Other informational persuasion (oe)	Informative Persuasion (IP)
Lyrical (LY)	Song lyrics (sl)	Lyrical (LY)
	Poem (po)	Lyrical (LY)
	Prayer (pr)	Lyrical (LY)
	Other lyrical (ol)	Lyrical (LY)
Spoken (SP)	Interview (it)	–
	Transcript of video/audio (ta)	Spoken (SP)
	Formal speech (fs)	Spoken (SP)
	TV/movie script (tv)	Spoken (SP)
	Other spoken (os)	Spoken (SP)

Table 8: Mapping from original CORE scheme to the scheme used in this study. Dashes indicate the subregister was preserved identically. Subregisters without abbreviations have zero support in the specific version of CORE that we use.