Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09695-8

SURVEY

Automatic genre identification: a survey

Taja Kuzman[1,2] · **Nikola Ljubešić**[1,3]

Accepted: 18 September 2023

Abstract

Automatic genre identification (AGI) is a text classification task focused on genres, i.e., text categories defined by the author's purpose, common function of the text, and the text's conventional form. Obtaining genre information has been shown to be beneficial for a wide range of disciplines, including linguistics, corpus linguistics, computational linguistics, natural language processing, information retrieval and information security. Consequently, in the past 20 years, numerous researchers have collected genre datasets with the aim to develop an efficient genre classifier. However, their approaches to the definition of genre schemata, data collection and manual annotation vary substantially, resulting in significantly different datasets. As most AGI experiments are dataset-dependent, a sufficient understanding of the differences between the available genre datasets is of great importance for the researchers venturing into this area. In this paper, we present a detailed overview of different approaches to each of the steps of the AGI task, from the definition of the genre concept and the genre schema, to the dataset collection and annotation methods, and, finally, to machine learning strategies. Special focus is dedicated to the description of the most relevant genre schemata and datasets, and details on the availability of all of the datasets are provided. In addition, the paper presents the recent advances in machine learning approaches to automatic genre identification, and concludes with proposing the directions towards developing a stable multilingual genre classifier.

Keywords Text genre · Web genre · Automatic genre identification · Genre schemata · Genre datasets · Survey paper

✉ Taja Kuzman
taja.kuzman@ijs.si

Nikola Ljubešić
nikola.ljubesic@ijs.si

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

[3] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

Published online: 16 November 2023

1 Introduction

Genres are text categories, defined by the author's purpose, common function of the text, and the text's conventional form (Orlikowski & Yates, 1994). They have been studied already in Antiquity, as the Greek philosophers and orators recognized that the manner in which the information is conveyed plays an important role on the receivers' understanding of the message, as well as the efficiency of communication (Kwasnik & Crowston, 2005). As genre involves multiple discourse, language and textual phenomena, it has been studied by researchers from various fields. The term is widely used in linguistics, rhetoric, literary theory, discourse studies, media theory, as well as computational linguistics and information retrieval studies (Chandler, 1997). The importance of providing information on genres included in a text collection was also recognized by the authors of reference corpora, such as the British National Corpus (BNC) (Davies, 2004) and the Corpus of Contemporary American English (COCA) (Davies, 2008).

Since the advent of the World Wide Web, there has been an increased interest in identifying genres in text collections, as the web allowed querying and collecting unprecedented quantities of texts. There has been large interest in genre identification in the area of information retrieval (see Stubbe and Ringlstetter 2007; Eissen and Stein 2004; Vidulin et al. 2007; Roussinov et al. 2001; Finn and Kushmerick 2006; Boese 2005; Priyatam et al. 2013; Stein et al. 2010), as the information on the genre in a query in information retrieval tools would allow users to find the texts that are relevant for them more efficiently (Crowston et al., 2010). Furthermore, collecting texts from the web has allowed a significantly faster and cheaper creation of corpora which are essential for language research and development of advanced language resources and technologies, especially in cases of under-resourced languages. However, in contrast to corpora that were carefully manually selected, web corpora are built in an automated way due to which their composition remains unknown (Baroni et al., 2009).

To obtain the information on the genre of specific texts in large collections of text, automatic genre identification (AGI), a text classification method which categorizes texts into genres, is the only feasible option due to the high data volume. AGI is primarily researched in the field of language technology which is concerned with computational processing of human languages and which encompasses computational linguistics and natural language processing (NLP), combining linguistics and computer science, especially artificial intelligence. As the authors of smaller, manually collected collections of texts are usually aware of the source of texts and their types, automatic genre identification is mostly studied to be applied on large web-based text collections where the source of texts is unknown. That is also why most of genre-annotated datasets surveyed in this paper contain texts from the web. However, since many high-coverage genre schemata, presented in this work, aim to cover all of the genre categories found in texts, the results of web genre identification research can be applied to collections of texts which do not originate from the web as well.

In addition to the benefits of genre identification for information retrieval and corpus creation and curation, it was shown that annotating documents with genre is also beneficial in many other NLP tasks. For instance, in part-of-speech (PoS) tagging of web corpora, Giesbrecht and Evert (2009) showed that using different granularity of PoS sets depending on the genres improves the tagging accuracy. Stewart and Callan (2009) showed that automatic genre identification improves automatic summarization, as the genre-oriented and goal-focused summarization algorithms outperformed other summarization methods. In addition to this, genres were revealed to be beneficial for machine translation (Van der Wees et al., 2018) and zero-shot dependency parsing (Müller-Eberstein et al., 2021). Recently, genre has also been studied in the area of information security, where it can be used for automated genre-aware assessment of credibility of web documents (Agrawal et al., 2019).

Obtaining information on genres can be of great value in a wide range of disciplines, including linguistics, corpus linguistics, computational linguistics, natural language processing, information retrieval and information security. Therefore, automatic genre identification models would be a great resource, as manual annotation of thousands of texts with genre labels is expensive and time-consuming, and, despite the best efforts, it can often result in unreliable annotations. However, although various attempts have been made to provide schemata, genre-annotated datasets and classifiers, there are no reference datasets (Sharoff, 2010) that can be directly used for building an AGI classifier which could be applied to any collection of web documents. In previous work, researchers used various genre definitions, annotation practices and corpora collection methods, each of them having its advantages and disadvantages. Thus, when approaching the automatic genre identification (AGI) task, one must not only ponder the appropriate machine learning strategy, but needs to carefully define each step of the data collection, schema creation and data annotation process.

The objective of this paper is to present a detailed survey of the existing research on automatic genre identification and available genre datasets. The rest of the paper is organised as follows: Sect. 2 presents various terms and definitions used to describe genres, which are the basis of genre schemata, discussed in Sect. 3. Section 4 compares existing genre-annotated datasets given the data collection methods and annotation approaches. The most frequently used datasets are presented in more detail. Section 5 presents an overview of text classification machine learning methods typically used for the automatic genre identification task, the results of various experiments, and the main challenges for automatic genre identification. Finally, Sect. 6 concludes the paper by summarising its findings, and proposing directions for improvements in the area.

2 Genre definition

One of the main challenges of studying genre is that there is no consensus among researchers on the definition of the genre phenomenon, or how to differentiate between genres, what qualifies for a genre class, and, finally, what the optimal genre schema is. Even more, there is no consensus on the term used for this phenomenon.

While many researchers use the term "genre" and describe this task as "automatic genre identification" (AGI) or "web genre identification" (WGI), others avoid this term and prefer derived analogues, such as "registers" (see Egbert et al. 2015 and further research by Laippala et al. 2019; Repo et al. 2021; Ronnqvist et al. 2021), "functional text dimensions" (see Sharoff 2018) or "text types" (see Stubbs 1996). While extensive work has been dedicated to navigate this "terminological maze", as Moessner (2001) put it, especially to distinguish between registers and genres (see Biber and Conrad 2019, Sharoff 2021, Lee 2002A), some researchers use the two terms interchangeably and argue that the distinction between them is "not relevant, and the choice between genre and register comes down to personal preference or tradition" (Egbert et al., 2015). However, in linguistics and especially sociolinguistics the term "register" is used to describe a different aspect of language variation, connected with conversational context. Frequent register categories are "academic language", "non-standard language", "formal language" and so on (see Lukin et al. 2011). According to Lee (2002A), the largest difference between the concepts of genres and registers is that genres describe whole texts, while registers are defined based on internal linguistic patterns, which are independent of text-level structures.

In addition to this, another challenge is how to define genres, an endeavour described by Chandler (1997) as a "theoretical minefield". In general, most definitions describe genre classes based on the socially recognized form of a document and its intended communicative purpose (Kwasnik & Crowston, 2005). In addition to this, some definitions also consider the target audience (Eissen & Stein, 2004), expectations of the reader (Santini, 2006), style (Finn & Kushmerick, 2006; Argamon et al., 1998), or the content of the texts (Rosso, 2008). To identify genres, researchers observe their intrinsic features, i.e., the linguistic and other "look'n'feel" features in the text (text-internal perspective), their extrinsic features, that is the function of the texts (text-external or functional perspective), or both (see Sharoff (2010, 2021) for a detailed discussion on both perspectives).

Moreover, studying genres, especially genres of digital texts, is a "formidable undertaking", to cite Kwasnik and Crowston (2005), due to fundamental difficulties connected with the genre notion itself. Firstly, conventional characteristics of a genre category can change over time and instances can be more or less prototypical in a specific time period (see Santini et al. (2010), Santini (2006)). The evolution of genres was especially notable after the introduction of the World Wide Web, when researchers observed emergence of new genres, unique to the web, and existing genres that evolved to adapt to the new medium (Williams & Crowston, 2000). Although many genres of web documents exist also in traditional texts, genres on the web are less fixed (Kwasnik & Crowston, 2005). Secondly, some web documents can be hybrids, which means that a single text can display characteristics of multiple genres, e.g., an advertorial, a document that has the form of a news article, but the purpose of a promotion of a product (see Sharoff (2021) and Repo et al. (2021)). Thirdly, some web documents might not have any discernible characteristics of a genre class. They can be too short to reveal the purpose of the communication. They can also be of an emerging genre, or a specialized genre with which the reader is not familiar with, e.g., a shipping invoice, used by accountants (see Williams and Crowston (2000)). In addition to this, it is argued that genre can only be realized

in a complete text (Biber & Conrad, 2019), which is not always available after the extraction of the text from a web page, a process that often includes removal of boilerplate content and other imperfect processing methods (see Pomikálek (2011)).

3 Genre schemata

Eissen and Stein (2004) noted that an "inherent problem of web genre classification is that even humans are not able to consistently specify the genre of a given page". Text instances can be more or less prototypical examples of its genre classes, can be hybrids or can lack characteristics of any genre. Thus, it is crucial for the reliability of a genre-annotated dataset that the genre schema supports the annotators so that their decisions are as consistent as possible. As no reference schema exists, most studies use their own, devised in accordance with the aim of the research. Schemata, used in previous work, vary significantly in terms of the number of classes, which range from seven (Santini, 2007; Sharoff, 2010; Lee & Myaeng, 2002b) to more than a hundred (Roussinov et al., 2001) or almost 300 classes (Crowston et al., 2010). They are either hierarchical (Stubbe & Ringlstetter, 2007; Egbert et al., 2015) or not (Asheghi et al., 2016; Sharoff, 2018).

Most existing schemata consist of around 10 to 20 categories, while some of them comprise of more than 50 genre classes (see Crowston et al. (2010); Roussinov et al. (2001); Williams and Crowston (2000); Egbert et al. (2015); Berninger et al. (2008)). Larger granularity usually increases the annotation complexity which results in a lower inter-annotator agreement and a less reliable dataset (see Sharoff (2010)). However, it should be noted that although a smaller set results in broader categories, they should still be limited by functional and form constraints so that the texts belonging to them share certain distinguishing properties, as argued by Asheghi et al. (2016). As in the case of a high granularity, too broad categories can also negatively impact the reliability of the annotated dataset and can be less useful for automatic identification. In addition to this, if the aim of a genre schema is to capture the entire diversity of genres on the web, Sharoff (2021) warned that genres are not fixed and writers are not constrained to following genre norms, thus, one should be aware that there will always exist web documents which do not fit under any genre class. While most studies used a pre-defined closed set of labels, some previous work took this issue into consideration, i.e., Asheghi et al. (2016) and Kuzman et al. (2022b) included a category *Other* in the schema, and Pritsos and Stamatatos (2018) used an open-set classification approach, allowing for a document not to be associated with any genre class.

Most of the schemata were developed following the top-down approach, which is based on theoretical principles, researchers' knowledge and understanding of the domain, and categorization from previous research. In regard to this approach, Rehm et al. (2008) proposed that the researchers should take advantage of accumulated expert knowledge, to avoid choosing the genre set based solely on how they perceive the classes, because the end users without a linguistic education might understand genre categories differently than genre researchers. By deriving the schema from category sets proposed by various research groups, this approach leads

to an improved set of categories that have already been proven to be applicable to the variety of language in actual document collections. An alternative approach is the bottom-up approach, taken up by some researchers (Crowston et al., 2010; Dewe et al., 1998; Eissen & Stein, 2004; Egbert et al., 2015), which is based on surveys of web users. This approach revealed to be problematic as the respondents were not consistent, they had difficulties with naming genres or used general or unspecific terms (see Crowston et al. (2010)).

The main factor defining the composition of genre schemata in previous works was the aim with which the genre dataset was constructed. Some researchers aimed to create a schema which would cover all possible genre variation found in (mostly web) texts. Such datasets allow for creation of general-purpose classifiers which would be able to classify any text into some genre class. In contrast, other researchers were not interested in covering all of the diversity found in text collections. They rather focused on a smaller set of specific genres that they deemed to be the most useful for search engine users who would find texts more efficiently based on genre criteria. In addition, some other schemata and corresponding genre datasets were developed with the aim to study characteristics of specific genres, such as poetry and prose, or with the aim of improving manual annotation. Thus, based on the aim of genre research, we divided the genre schemata into the following groups:

1. High-coverage schemata (complete genre taxonomy) which aim to cover the entire genre composition of the web. These studies mostly used larger sets of specific classes (see Stubbe and Ringlstetter (2007); Sharoff (2018); Egbert et al. (2015); Laippala et al. (2019); Kuzman et al. (2022b); Suchomel (2020)), a smaller set of broader categories (see Sharoff (2010)) or a combination of the two (see Santini (2010)).
2. Information retrieval schemata, created with the aim to provide genre identification inside the search engines. Thus, they consist of genre categories that are deemed to be useful for the search engine users (see Eissen and Stein (2004); Vidulin et al. (2007); Roussinov et al. (2001); Dewe et al. (1998); Lim et al. (2005); Santini (2007)).
3. Schemata, developed with other aims: limited sets of categories used in studies which analyse one or a few genre classes, e.g., a smaller set of genres of interest (Boese (2005); Lee and Myaeng (2004); Asheghi et al. (2016)), academic genres (Rehm, 2002), e-shop genres (Levering et al., 2008), poetry and prose genres (Shavrina, 2019), news articles and reviews (Finn & Kushmerick, 2006), and home pages (Kennedy & Shepherd, 2005).

The most relevant schemata are presented in more details in Sect. 4.3, together with the datasets for which they were used.

To explore which categories appear most frequently in the schemata, we analysed 16 of the previously mentioned schemata, i.e., schemata from the works of Egbert et al. (2015), Asheghi et al. (2016), Stubbe and Ringlstetter (2007), Eissen and Stein (2004), Santini (2007), Vidulin et al. (2007), Sharoff (2010), Laippala et al. (2019), Santini (2010), Sharoff (2018), Dewe et al. (1998), Lim et al.

Table 1 Most frequent genre labels from 16 genre schemata

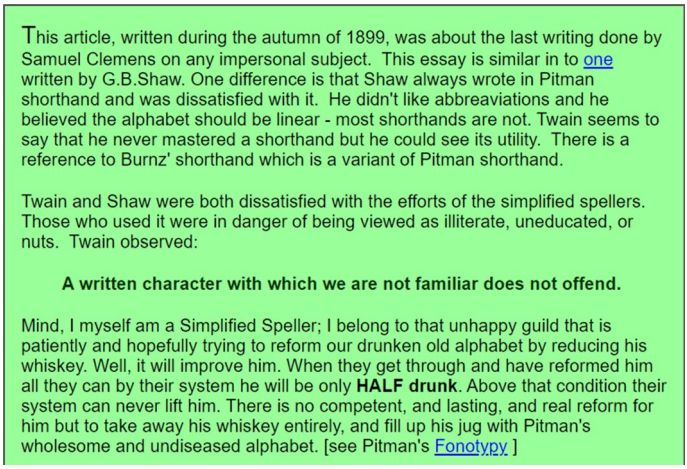
Genre label	Occurrences in schemata
FAQ	9
Instruction	7
News	6
Legal	6
Personal home page	6
Review	6
Information	5
Discussion	5
Research article	5
Interview	5
Lyrical	5
Poem	4
Recipe	4
E-shop	4
Editorial	4
Promotion	3
Link Collection	3
Journalistic	3
Blog	3
Prose	3

Labels with very similar names, e.g., "FAQ" and "FAQs" or "news" and "news/reporting", were counted as the same label name

(2005), Boese (2005), Lee and Myaeng (2002b), Suchomel (2020), Kuzman et al. (2022b). Together, the schemata have 280 genre labels, out of which 214 have unique names. After merging genre labels that have very similar names, such as "FAQ" and "FAQs" or "news" and "news/reporting", there are 177 unique genre labels, out of which 129 appear only in one schema and 28 only in two schemata. Table 1 shows the 20 labels that are used in at least 3 schemata. Out of these labels, the most frequent categories are *FAQ* and *Instruction*, appearing in 7 or more schemata, and *News*, *Legal*, *Personal Home Page* and *Review*, appearing in 6 schemata.

Schemata also vary in terms of the simplicity of names, used for genre labels. As shown in Table 1, the most frequent categories are named with common words with the aim of making the labels understandable to the end users, e.g. *Instruction*, *News* or *Interview*. In contrast, some researchers opted for more abstract label names, such as *content delivery* (Vidulin et al., 2007), *resources* (Stubbe & Ringlstetter, 2007), *recreation* (Sharoff, 2010), *commpuff* (Sharoff, 2018).

While Table 1 shows which category names are most frequently used in genre schemata, we should note that this is a harsh generalization of the existing categories, primarily for purposes of obtaining an overview. The reality in the



This article, written during the autumn of 1899, was about the last writing done by Samuel Clemens on any impersonal subject. This essay is similar in to one written by G.B.Shaw. One difference is that Shaw always wrote in Pitman shorthand and was dissatisfied with it. He didn't like abbreaviations and he believed the alphabet should be linear - most shorthands are not. Twain seems to say that he never mastered a shorthand but he could see its utility. There is a reference to Burnz' shorthand which is a variant of Pitman shorthand.

Twain and Shaw were both dissatisfied with the efforts of the simplified spellers. Those who used it were in danger of being viewed as illiterate, uneducated, or nuts. Twain observed:

A written character with which we are not familiar does not offend.

Mind, I myself am a Simplified Speller; I belong to that unhappy guild that is patiently and hopefully trying to reform our drunken old alphabet by reducing his whiskey. Well, it will improve him. When they get through and have reformed him all they can by their system he will be only **HALF drunk**. Above that condition their system can never lift him. There is no competent, and lasting, and real reform for him but to take away his whiskey entirely, and fill up his jug with Pitman's wholesome and undiseased alphabet. [see Pitman's Fonotypy]

Fig. 1 An example of a text from the category *Article* in the Genre-KI-04 (Eissen & Stein, 2004) dataset

underlying datasets is much more complicated. Labels that have same or similar names do not necessarily cover similar texts, as this depends on the researchers' and annotators' understanding of the labels. Moreover, the texts inside similarly named classes in different datasets could differ significantly also based on the approaches, with which the datasets were collected, as described in the following section.

The case of journalistic genres illustrates well the differences between the schemata and shows why it is infeasible to merge any datasets, even if their labels have similar names. For instance, the Genre-KI-04 dataset (Eissen & Stein, 2004) includes a label *Article*, for which one might assume that it represents news articles. However, in the documentation accompanying the dataset, the label is defined as "Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters". The label thus includes a large variety of texts, including opinionated texts, such as the instance in Fig. 1. In the 20-Genre Collection (MGC) (Vidulin et al., 2007), a genre that could be connected to news articles based on the name is *Journalistic*. Its definition further confirms this: "conveys mostly objective information on current events" (Vidulin et al., 2007). However, on their website,[1] authors included the following types of texts under the *Journalistic* genre: news, reportages, editorials, interviews and reviews, indicating that the class is quite broad and includes objective as well as subjective texts. The great inner variation between the texts, included in this genre, can be seen in Fig. 2. In the CORE dataset (Egbert et al., 2015), the news article genre is divided into two categories – *News Report/Blog* and *Sports Report* – based on the topic of articles. In addition,

[1] The website can be accessed through the Internet Archive Wayback Machine: https://web.archive.org/web/20120513113203/http://dis.ijs.si/mitjal/genre/list_of_genres.htm

Fig. 2 Two instances, labelled with the category *Journalistic* in the 20-Genre Collection (MGC) dataset (Vidulin et al., 2007)

A message to all the Olympic moaners... BELT UP! It's a crisis. It's a complete disaster. It's not just a shambles out there, people; it's an Olympic omnishambles. The Games are six days away and nothing works, absolutely everything is broken and the only solution is to cancel the event and send all those arriving at Heathrow straight back home on the next available flight. Why? Because according to reports, any fool venturing into London will die of carbon monoxide poisoning as they sit in month-long traffic jams. Or drown in their body sweat on overcrowded Tube trains. Or sink into oblivion trying to negotiate the mud flats otherwise known as the Olympic Park. The mobile phone networks will fail, the internet will collapse into a black hole in cyberspace, pickpockets will steal everything, including your kidneys, and the entire country will end up bankrupt. It's a nightmare - and all because of the 2012 Games. You have been warned! Stop moaning! Great

Ash tree disease found for first time in Wales Forestry Commission Wales has confirmed the discovery of a serious disease of ash trees for the first time in Wales. Chalara dieback of ash has been found in a small, privately-owned young broadleaved woodland in Carmarthenshire. The trees, which were planted in 2009, are still quite small and measures are being put in place to minimise the risk of the disease spreading to the wider environment. A containment notice will be issued shortly for the site and FC Wales will work with the woodland owner to deal with the infected trees. The discovery follows a rapid survey of the whole of Wales over the past four days to check the condition of the country's ash trees as plant health authorities across the United Kingdom stepped up their efforts to tackle the disease, caused by the fungus Chalara fraxinea. FC Wales and Welsh

Fig. 3 Two instances, labelled with the category *News Report/Blog* in the CORE dataset (Egbert et al., 2015)

the schema also includes a class, named *Magazine Article*. The *News Report/Blog* genre is defined in the annotation guidelines[2] as a "news report written by journalists and published by news outlets" with the purpose to report on recent events. Despite a clear description, instances in this class vary from objective reports on recent events to opinionated blog-like posts, as can be seen in Fig. 3.

The examples in Figures show that a detailed exploration of texts inside the collections would be necessary to be able to compare the genre schemata thoroughly, which is outside of the scope of the paper. What is more, more detailed analyses of genre classes and the compatibility of datasets are hampered by unavailability of the datasets that use the analysed schemata.

[2] https://turkunlp.org/register-annotation-docs/register/NA-ne.html

4 Genre-annotated datasets

This section discusses the process of constructing datasets that are manually annotated for genre. They are the basis for training machine learning models for automatic genre identification. Some genre researchers used traditional general-purpose corpora that have genre information, such as Corpus of Contemporary American English (COCA) (Davies, 2008) (see Ströbel et al. (2018)), British National Corpus (BNC) (Lee, 2002A) (see Sharoff (2010)) and Brown Corpus (Kučera & Francis, 1967) (see Karlgren and Cutting (1994)). However, these datasets do not represent well the challenges for automatic genre identification imposed by less curated web text collections, where there is much wider variation between styles and quality of texts (Sharoff, 2010). That is why in the great majority of studies, the genre researchers decided to construct their own datasets based on web data. The process of building genre-annotated datasets regularly starts with a selection of data to be manually annotated, as described in Sect. 4.1. Next, manual annotation for genre is performed, as discussed in Sect. 4.2. We wrap up this section with an overview of various types of manually annotated datasets in Sect. 4.3.

4.1 Data Selection

Most datasets that are manually annotated for genre consist of texts originating from the web. The texts are either taken from existing web corpora, or new document collections are constructed in a manual or an automated manner.

As genre-annotated datasets were created with different aims, and financial and technical constraints, they largely differ in terms of size, collection methods and document format. In many studies, the datasets were limited to document formats which are easier to obtain from the web and to process with automated methods, such as HTML. This influences the genre distribution in the final dataset, as by not including PDF files or Word documents, certain genres for which the preferred format is PDF, such as research papers and catalogues, are likely under-represented (Santini et al., 2010).

Based on the collection method, web corpora can be categorized into designed and crawled corpora (Kilgarriff, 2012). In designed corpora, texts are collected based on certain criteria, defined by the researchers. Although this collection method facilitates the annotation and machine classification, Sharoff (2010) argues that designed corpora are not representative of the actual genre distribution on the web. They are thus less useful for training a genre classifier that would be applied on a new text collection. In contrast, crawled corpora represent a "(more or less faithful) snapshot of the web" (Asheghi et al., 2016). Examples of designed corpora are those by Stubbe and Ringlstetter (2007), Santini (2007), Boese (2005), Berninger et al. (2008), while Laippala et al. (2020), Kuzman et al. (2022b) and Sharoff (2018) used samples of crawled corpora. Rehm et al. (2008) proposed using both methods, which was done by Asheghi et al. (2016). They first collected a designed corpus consisting of prototypical instances to test their method of manual annotation, and then extended their research to a random collection of web texts to analyse the coverage of the genre

schema and the reliability of annotation in a more realistic setting. Another commonly used method for data collection is retrieving top hits of automated queries made to a search engine (see Egbert et al. (2015); Sharoff (2010); Lim et al. (2005)). However, one should be aware that the final composition of such collection might be impacted by the choice of keywords in the queries and by the searching and ranking algorithms used by the search engines which favour more popular pages (see Santini et al. (2010); Asheghi et al. (2016)). This can result in a bias towards certain genres, such as e-shops. Furthermore, this method allows collecting only documents that can be found on the searchable web (Egbert et al., 2015).

Being aware of the impact of the collection method on the resulting dataset, Asheghi et al. (2016) and Lim et al. (2005) devoted attention to assuring that the texts from the crawled corpus originate from a wide variety of sources, to avoid biases towards certain topics, authors or sources, thereby guaranteeing generalizability of the resulting datasets. Furthermore, Asheghi et al. (2016) and Kuzman et al. (2022b) assured temporal diversity of sources, by selecting sources published at several time points, as some genres, such as *News*, have strong temporal connection to topics.

4.2 Annotation practices

In addition to the corpus collection methods that the datasets are based on, one of the most crucial factors for producing a high-quality dataset that can be used for machine learning experiments is reliable manual annotation. The annotation can be performed by a smaller group of experts, such as in the case of Santini (2006); Vidulin et al. (2007); Eissen and Stein (2004). However, as the task is very time-consuming, annotation campaigns in previous research mostly resulted in relatively small datasets, consisting of less than 2,000 texts. To achieve faster and cheaper manual annotation, some researchers leveraged the advent of crowd-sourcing. This is how two of the biggest English genre-annotated corpora were created, the Leeds Web Genre Corpus (Asheghi et al., 2016) and the CORE corpus (Egbert et al., 2015). The decision to opt for crowd-sourcing was based on the previous findings (see Snow et al. (2008)) which compared expert and crowd-sourcing annotation on various tasks and concluded that if four non-experts annotate the same instance, the final annotations are often comparable to the expert-annotation quality. In addition to this, Asheghi et al. (2016) argue that this is the most appropriate annotation method, because it avoids potential bias from the experts who both developed the genre schema and annotated the dataset.

Furthermore, annotation strategies vary based on how researchers dealt with hybridity of web texts. As mentioned in the Introduction, classifying web texts in genre categories can be challenging, because some texts, known as hybrid texts, can have characteristics of multiple genres. The hybridity of texts might be intentional, which is the case of an advertorial – a promotion of a product written in a form of a news article to appear more trustworthy to the readers. It might also occur due to authors' lack of expertise or willingness to follow traditionally accepted genre characteristics (Sharoff, 2021). In contrast to non-digital texts which are often reviewed

before being printed and published, a much larger number of texts on the web does not undergo any editorial control which would compel the authors to follow genre norms.

In addition to hybrid texts where characteristics of multiple genres are intertwined, there exist multi-class documents, which consist of a selection of functionally-separated texts of different genres that were published on the same web page and thus belong to a single document (Stubbe & Ringlstetter, 2007). An example of a multi-class document is a document from a web page consisting of a news article, followed by the readers' comments. A special and the most-challenging case of multi-class documents are documents where text in another genre is embedded in another text (see Williams and Crowston (2000)), such as a blog post that includes an entire letter. To assure a reliable annotation, challenges with hybrid texts and multi-class documents need to be addressed in the data selection and annotation procedure. Some previous studies tackled the issue with hybrid texts by allowing multi-label annotation, i.e., assigning multiple labels to a single text (Vidulin et al., 2007; Laippala et al., 2019; Sharoff, 2018; Suchomel, 2020; Egbert et al., 2015; Kuzman et al., 2022b).

The reliability of the annotation procedure is commonly evaluated based on the inter-annotator agreement, i.e., agreement of two or more annotators on the label of the same instance. That is why most of the datasets were annotated by at least 2 annotators. However, their reliability cannot be directly compared, because different methods were used to calculate the agreement: percentage agreement (in Asheghi et al. (2016), Roussinov et al. (2001), Williams and Crowston (2000), Berninger et al. (2008)), Fleiss' kappa (Fleiss, 1971) (in Asheghi et al. (2016), Egbert et al. (2015)), Cohen's kappa coefficient (Cohen, 1960) (in Vidulin et al. (2007)), Krippendorff's alpha (Krippendorff, 2018) (in Vidulin et al. (2007); Sharoff (2018); Suchomel (2020); Kuzman et al. (2022b)), Jaccard similarity coefficient (in Suchomel (2020)), accuracy (in Suchomel (2020)) and F1 score (in Repo et al. (2021)).

Despite the best efforts in creating optimal genre schemata, with annotation guidelines and surveys that provide guidance to the annotators, previous studies revealed that achieving high inter-annotator agreement is very challenging. For instance, Suchomel (2020) concluded that a high agreement in this task is "hardly possible". This is also the case for two of the datasets that have been recently most frequently used, the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the Functional Text Dimensions (FTD) dataset (Sharoff, 2018). For CORE (Egbert et al., 2015), an analysis of the inter-annotator agreement revealed that there was no majority agreement, i.e., agreement between at least three of four annotators, on one-third of texts on the level of 8 main categories, and on half of the texts on the level of 53 subcategories. When Sharoff (2018) applied the CORE schema and annotation process on another dataset, the inter-annotator agreement reached a nominal Krippendorff's alpha of 0.66 and 0.53 for main categories and subcategories respectively, which is on the verge and below the acceptable threshold of 0.67, defined by Krippendorff (2018). For the FTD dataset, the inter-annotator agreement initially revealed encouraging results, reaching Krippendorff's alpha above 0.76 (Sharoff, 2018). However, when the annotation was extended to another

dataset in subsequent research by Suchomel (2020), the minimal acceptable Krippendorff's alpha was not reached despite using a smaller set of labels. Recently, based on the findings from the previously discussed work, Kuzman et al. (2022b) considered assuring a reliable annotation at every step of their research. They devised their genre schema based on the categories from previous schemata so that the categories would be widely used and the label names would be recognizable to the annotators. The authors also opted for expert annotators instead of crowd-sourcing, and devised a decision tree survey and detailed guidelines with examples which led the annotators through their decision process. The results of the annotation campaign showed that this endeavour resulted in an inter-annotator agreement up to the nominal Krippendorff's alpha of 0.71.

To avoid time-consuming and costly manual annotation, Santini (2006) proposed "annotation by objective sources". This means that she identified genre-specific web domains, such as e-shops, and considered all texts from the domain as automatically belonging to the same genre without any further manual annotation. Automatic annotation was also applied to texts that contained the name of the genre in the title. This approach is based on the idea that if a text originates from a web domain which specializes in publishing texts of one genre, this means that web users collectively consider the text to belong to this genre. This approach was recently taken up by Lepekhin and Sharoff (2021) as well, who named it "natural annotation".

If researchers opted to use an existing genre schema for performing automatic genre identification experiments, they most often chose those from the few openly available datasets. For instance, two earlier schemata, Genre-KI-04 (Eissen & Stein, 2004) and 7-Web-Genre-Collection (Santini, 2006), were often used for evaluation of automatic genre identification algorithms (Jebari, 2021; Pritsos & Stamatatos, 2018; Kanaris & Stamatatos, 2007, 2009; Mason et al., 2009). However, today, most AGI experiments, as well as new genre datasets, are based on the schemata of the Corpus of Online Registers of English (CORE) (Egbert et al., 2015), the Functional Text Dimensions (FTD) approach (Sharoff, 2018), or the GINCO corpus (Kuzman et al., 2022b). The datasets and their schemata are described in more details in the following section.

4.3 Comparison of datasets

An overview of genre-annotated datasets with information on their size, language, number of classes and hierarchy depth, annotation approach and availability is shown in Table 2. The most relevant datasets, divided into high-coverage datasets, information retrieval datasets and other datasets, are described in more details in the following subsections.

As shown in Table 2, less than half of the datasets are available online for use in further experiments. As none of the datasets, except the GINCO dataset (Kuzman et al., 2022b), is published in a repository, they are hard to find and can become unavailable over time. For instance, previously most often used datasets, e.g., the 7-Web Genre Collection (Santini, 2007), the Hierarchical genre collection (Stubbe & Ringlstetter, 2007), KRYS I (Berninger et al., 2008) and others, used to be

Table 2 Overview of the genre-annotated datasets, compared in terms of size (number of texts), language (ISO codes), number of classes in the genre schema and multi-label approach

Dataset	Size (texts), Language	Classes	Multi-label
Dewe et al. (1998)	1358 (EN)	2/11	No
Stamatatos et al. (2000)	250 (EL)	10	No
Roussinov et al. (2001)	1076 (EN)	116	No
Lee and Myaeng (2002b)	7828 (KO), 7615 (EN)	7	Yes
Genre-KI-04 (Eissen & Stein, 2004)[a]	1239 (EN)	8	No
Lim et al. (2005)	1328 (KO)	2/16	No
Boese (2005)	343 (EN)	10	No
Freund et al. (2006)	800 (EN)	16	No
7-Web Genre Collection (Santini, 2007)	1400 (EN)	7	No
20-Genre Collection (MGC) (Vidulin et al., 2007)[b]	1539 (EN)	20	Yes
Hierarchical genre collection (HGC) (Stubbe & Ringlstetter, 2007)	1280 (EN)	7/32	No
KRYS I (Berninger et al., 2008)	6300 (EN)	10/70	No
Maeda and Hayashi (2009)	870 (JA)	7	No
SANTINIS-ML (Santini, 2010)	2480 (EN)	15	Yes
Syracuse dataset (Crowston et al., 2010)[c]	3027 (EN)	292	No
I-EN-Sample, I-RU-Sample (Sharoff, 2010)	250 (EN), 250 (RU)	7	No
CORE (Egbert et al., 2015)[d]	48,420 (EN)	8/47	Yes
Leeds Web Genre Corpus (Asheghi et al., 2016)	4964 (EN)	20	No
Functional Text Dimensions (FTD) (Sharoff, 2018)[e]	1562 (EN), 1,930 (RU)	18	Yes
FinCORE (Laippala et al., 2019)[f]	2237 (FI)	8/22[g]	Yes
FreCORE, SweCORE (Laippala et al., 2020)[h]	688 (FR), 1085 (SV)	8/22	Yes
Suchomel (2020)	1974 (EN)	9	Yes
LiveJournal dataset (Lepekhin & Sharoff, 2022)	13,629 (RU)	10	No
GINCO (Kuzman et al., 2022b)[i]	1002 (SL)	24	Yes
FinCORE (extended) (Skantsi & Laippala, 2023)[j]	10,754 (FI)	9/30	Yes

If the schema is hierarchical, the number of classes includes the main classes and sub-classes, e.g., 7/32 stands for 7 main classes and 32 sub-classes

[a]The Genre-KI-04 dataset is available here: https://webis.de/data/genre-ki-04.html.

[b]The 20-Genre Collection can be accessed through the Internet Archive Wayback Machine: https://web.archive.org/web/20120512021524/http://dis.ijs.si/mitjal/genre/

[c]Details on the dataset are based on the description in Rezapour Asheghi (2015).

[d]The CORE corpus is available here: https://github.com/TurkuNLP/CORE-corpus and can be queried here: https://www.english-corpora.org/core/.

[e]The FTD dataset is available here: https://github.com/ssharoff/genre-keras.

[f]The FinCORE dataset is available here: https://github.com/TurkuNLP/FinCORE.

[g]Based on the annotation guidelines published at https://turkunlp.org/register-annotation-docs.

[h]The FreCORE and SweCORE datasets are available here: https://github.com/TurkuNLP/Multilingual-register-corpora.

[i]The GINCO dataset is available here: http://hdl.handle.net/11356/1467.

[j]The new extended FinCORE dataset is available here: https://github.com/TurkuNLP/FinCORE_full

Main Category	Journalism	Literature	Information	Documentation	Dictionary	Communcation [*sic*]	Nothing
Sub-category	Commentary	Poem	Science Report	Law	Person	Mail, Talk	Nothing
	Review	Prose	Explanation	Official Report	Catalog	Forum, Guestbook	
	Portrait	Drama	Receipt	Protocol	Ressources	Blog	
	Marginal Note		FAQ		Timeline	Form	
	Interview		Lexicon, Word List				
	News		Bilingual Dictionary				
	Feature Story		Presentation				
	Reportage		Statistics				
			Code				

Fig. 4 Hierarchical schema, proposed by Stubbe and Ringlstetter (2007) and used in The hierarchical genre collection (HGC)

available at a WebGenreWiki website which was last edited in 2012 and ceased to be available in 2015.[3] Table 2 includes non-English datasets which are in general rare. Although the genre is mainly studied for English, there exist some research groups who work towards contributing datasets in other languages: Sharoff and colleagues from the University of Leeds, UK, with interest in Russian (Sharoff, 2010, 2018, 2021; Lepekhin & Sharoff, 2022) and Arabic (Bulygin & Sharoff, 2018), and the TurkuNLP Group from the University of Turku, Finland, with interest in extending the CORE schema to other languages to leverage the promising advances in cross-lingual learning, providing Finnish (Laippala et al., 2019; Skantsi & Laippala, 2023), Swedish and French (Laippala et al., 2020; Repo et al., 2021) datasets, and smaller evaluation datasets in 8 additional languages (Laippala et al., 2022b). In addition to this, there exist genre-annotated datasets for Korean (Lim et al., 2005; Lee & Myaeng, 2002b), Japanese (Maeda & Hayashi, 2009), Greek (Stamatatos et al., 2000) and Slovene (Kuzman et al., 2022b).

4.3.1 High-coverage datasets

In this section, we present the most relevant high-coverage datasets, i.e., datasets created with the aim of capturing all of the genre variation that can be found in (mostly web) texts. This is reflected in genre schemata, which consist either of a large number of specific categories or a smaller number of broader genre classes.

The hierarchical genre collection (HGC) (Stubbe & Ringlstetter, 2007) is one of the first datasets annotated with a hierarchical schema. The schema was devised following a combination of a top-down and bottom-up approach. It was based on the genre set proposed by Dewe et al. (1998) and improved based on the feedback from a user study. The schema, shown in Fig. 4, consists of 7 main categories and 32 subcategories. The dataset was collected manually, which means that it is a designed

[3] The website as it was in 2015 can be viewed via Internet Archive Wayback Machine at https://web.archive.org/web/20150615095927/http://www.webgenrewiki.org/index.php5/Main_Page.

Main Category	Informational Description/ Explanation	Opinion	Narrative	Informational Persuasion	Interactive Discussion	How-To/ Instructional	Lyrical	Spoken
Sub-category	Course materials	Advertisement	Historical article	Description with intent to sell	Discussion forum	FAQ about how-to	Song lyrics	Formal speech
	Description of a person	Advice	Magazine article	Editorial	Reader/viewer responses	How-to	Poem	Interview
	Description of a thing	Letter to editor	News report/blog	Persuasive article or essay	Question/answer forum	Technical support	Prayer	TV/movie script
	Encyclopedia article	Opinion blog	Travel blog	Other	Other forum	Recipe	Other	Transcript of video/audio
	FAQ about information	Reviews	Personal blog			Other		Other
	Information blog	Religious blogs/sermons	Sports report					
	Legal terms and conditions	Other opinion	Short story					
	Technical report		Other narrative					
	Research article							
	Other information							

Fig. 5 The CORE schema as it is used in the CORE dataset (Egbert et al., 2015). The figure was created based on the information that was provided to us by the dataset authors

corpus, and consists of 1,280 texts that are prototypical for the assigned genre classes.

The *KRYS I* (Berninger et al., 2008) dataset consists of over 6,000 texts that are exclusively in the PDF format. The authors used their own hierarchical schema of 10 main categories and 70 subcategories, such as *Poetry Book, Menu, News Report, Contract, Essay, Sheet Music*. The dataset was collected manually by the students who collected instances of a genre that was assigned to them. Later, the instances were reclassified independently by two annotators.

The *SANTINIS-ML* dataset (Santini, 2010) is a multi-label extension of the 7-Web Genre Collection (Santini, 2007). Here, the authors introduced a double-layered genre schema with the aim of being able to cover all of the diversity of texts found on the web without having a very high granularity. The schema consists of four functional genres (*Descriptive-Narrative, Explicatory-Informational, Argumentative-Persuasive* and *Instructional*) and 11 specific genre categories – the labels from the 7-Web Genre Collection and 4 BBC web genres (*Editorials, DIY Mini-Guides, Short Biographies, Feature Articles*). The dataset consists of 2,480 texts, annotated by the author of the research.

The *I-EN-Sample* and *I-RU-Sample* (Sharoff, 2010) datasets consist of 250 texts, annotated with the Functional Genre Classification schema. The schema consists of 7 macro-genres: *Discussion, Information, Instruction, Propaganda, Recreation, Regulation*, and *Reporting*. They are defined solely based on the function or the purpose of the document and the 7 of them should be able to cover all the texts on the web. The novelty of this approach lies in the fact that the schema aims for a high coverage with a low granularity by using broad functional categories. The dataset was collected based on the results of random queries made to a search engine.

The *CORE* dataset (Egbert et al., 2015) is the result of one of the most extensive works on the genre schema and annotation process. The CORE schema was constructed following the bottom-up approach, where the end web users were asked to provide situational characteristics of web texts that would be relevant for them. The

Category Group	Objective Informative	Subjective Reporting	Opinion	Promotion	Dialogue	Literature	Formatted Text	Other
Category	News/Reporting	Opinionated News	Opinion/ Argumentation	Promotion	Interview	Script/Drama	FAQ	Other
	Announcement		Review	Promotion of a Product	Forum	Lyrical	List of Summaries /Excerpts	
	Information/ Explanation			Promotion of Services	Correspondence	Prose		
	Research Article			Invitation				
	Instruction							
	Recipe							
	Call							
	Legal/Regulation							

Fig. 6 The GINCO schema

initial schema was then improved based on a series of 10 pilot studies. The final schema as it is used in the dataset which is available online consists of 8 higher-level categories that cover the situational characteristics or functions of texts and 47 specific subcategories. The schema is shown in Fig. 5. A slightly modified schema with a smaller set of categories was used for annotation of Finnish (Laippala et al., 2019; Skantsi & Laippala, 2023), French and Swedish (Laippala et al., 2020; Repo et al., 2021) datasets, for which the detailed guidelines are available.[4] For the purposes of evaluation of a genre classifier model, 8 more datasets were recently annotated with CORE labels: an Indonesian dataset of around 1,000 texts, and 7 smaller evaluation datasets of 92 to 334 texts for Arabic, Catalan, Chinese, Hindi, Portuguese, Spanish and Urdu (see Laippala et al. (2022b)).

The CORE corpus consists of web texts that were extracted from the "General" part of the Corpus of Global Web-based English (GloWbE) (Davies & Fuchs, 2015). The GloWbE corpus was collected via Google searches with high frequency English 3-grams as the queries (Davies & Fuchs, 2015). The texts were then annotated via crowd-sourcing with 908 participants, where each text was annotated by 4 annotators. Despite leading the annotators through the decision process with a decision-tree survey, the inter-annotator agreement was shown to be low (see Sect. 4.2 for more details). The instances where there was no majority agreement on one label are annotated with multiple labels. The size of the dataset, as reported by the authors, is 53,000 texts, however, the dataset that is available online contains 48,420 texts. The dataset was used in further studies which provide a detailed analysis of linguistic characteristics of the CORE genres (Biber & Egbert, 2018), analyse the role of lexical and grammatical features in the performance and stability of genre classifiers (Laippala et al., 2021), or experiment with automatic genre classification (Ronnqvist et al., 2021; Repo et al., 2021).

The Slovene *Genre Identification Corpus (GINCO)* (Kuzman et al., 2022b) consists of 1,002 texts and uses a schema that is based on the subcategory level of the CORE schema, but also on other high-coverage schemata. In contrast to the CORE

[4] The guidelines for the annotation with the CORE schema are available here: https://turkunlp.org/register-annotation-docs.

	Principal Categories			
Category Group	**information**	**discussion**	**narration**	**promotion**
Categories	instruct	argum	fictive	compuff
	hardnews	scitech	personal	ideopuff
	legal	eval		appell
	info			

Additional Categories
emotive
flippant
informal
specialist
dialogue
poetic

Fig. 7 The FTD schema

schema, it is not hierarchical and it has a lower granularity of categories to improve the inter-annotator agreement. The schema, shown in Fig. 6, consists of 23 specific genre categories and the category *Other*, dedicated to texts which do not fall into any other label. To alleviate the annotation, the genre categories were presented in category groups. The dataset was annotated by 2 expert annotators. They followed a multi-labelling approach where texts can be annotated with up to three genre labels. In this setting, the primary label is deemed to be the one that is most prevalent and the one that is used for single-label text classification, whereas the secondary and tertiary labels provide additional information on the fuzziness of the text, and the three levels can be used for multi-label classification. Extensive guidelines are available.[5] The dataset is a crawled corpus, as it is randomly sampled from two Slovenian web corpora from different time periods – the slWaC 2.0 corpus (Erjavec & Ljubešić, 2014) from 2014 and the MaCoCu-sl 1.0 corpus from a 2021 crawl (Bañón et al., 2022b). To be representative of the conditions on the web, it also includes noise, that is, a subset of "not suitable" texts, such as non-textual documents and automatically generated documents, multi-genre documents and other web-specific challenges.

The *Functional Text Dimensions* (FTD) dataset (Sharoff, 2018) in English and Russian follows another approach, introducing 18 Functional Text Dimensions (FTDs). As stated by the authors, the novelty of this approach is that "[u]nlike the traditional approach to designing genre lists or hierarchies in which texts need to be members of the respective classes, the FTD approach to detecting the genre of annotated texts is based on their similarity to prototypes, which is measured as distance in the FTD space" (Sharoff, 2018). The schema is a set of broad functional categories instead of specific labels, so that a text can be described through a combination of multiple labels, which could be regarded as parameters. By providing a possibility of assigning multiple labels to a text, the schema also deals with hybrid texts. In this approach, annotators define the presence of each dimension on a scale from 0 to 2, representing the distance of the text from the prototypical texts based on presence or absence of text features, typical for the genre class. As shown in Fig. 7, the schema consists of 12 principal categories, which means that a text needs to be given a non-zero score in at least one of them, and 6 additional categories, which describe additional variation between texts. While the principal categories

[5] The guidelines for annotation with the GINCO schema are available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

differentiate texts based on their function, additional categories provide information on other textual characteristics, for instance, on the stylistic differences, such as is the case of category *informal*, used for texts, written in informal language.

The FTD dataset consists of texts from multiple sources: the Pentaglossal corpus (5 g) (Forsyth & Sharoff, 2014), which consists of Russian and English texts with known origin (fiction, corporate communication, political debates, TED talks, UN reports, etc.), and a random selection of texts from the ukWac web corpus (Baroni et al., 2009) and the Russian GICR web corpus (Piperski et al., 2013). The FTD schema was extended to Arabic (Bulygin & Sharoff, 2018), a smaller set of FTDs was used for annotation experiments of additional English datasets (Suchomel, 2020), and a modified schema, consisting of 10 FTD labels, was used for annotation of a large Russian corpus, the *LiveJournal* dataset (Lepekhin & Sharoff, 2022). Guidelines for annotation are available.[6]

4.3.2 Information retrieval datasets

Information retrieval datasets were created with the aim to provide genre identification inside the search engine, so that the users would be able to find the texts more efficiently based on genre criteria. Thus, their schemata consist of categories that are deemed to be useful for the search engine users.

The *Genre-KI-04* (Eissen & Stein, 2004) dataset consists of 1,239 HTML files, annotated with a schema of 8 categories, devised based on a survey on genre usefulness (bottom-up approach). The schema consists of the following labels: *Help, Article, Discussion, Shop, Portrayal (non-priv), Portrayal (priv), Link Collection* and *Download*. Recently, the schema was used by Agrawal et al. (2019) for assessing the credibility of web pages.

The *7-Web Genre Collection* (Santini, 2007), in some works also referenced as SANTINIS, consists of 1,400 texts. It is annotated with 7 genre labels that are exclusive to the web: *Blog, e-Shop, FAQs, Online Newspaper Front Page, List, Personal Home Page*, and *Search Page*. The dataset, consisting of 200 instances of each genre, was manually collected by the author of the research, following the criteria of "annotation by objective sources".

The *20-Genre Collection* (MGC) (Vidulin et al., 2007) consists of 1,539 texts. They are classified into 20 genres: *Blog, Childrens', Commercial/Promotional, Community, Content Delivery, Entertainment, Error message, FAQ, Gateway, Index, Informative, Journalistic, Official, Personal, Poetry, Pornographic, Prose fiction, Scientific, Shopping*, and *User Input*. As the aim of the research was to construct a genre classifier which would identify genres within a search engine, the schema includes categories which would be of interest to the search engine users or which the users would want to filter out, e.g., the *Error message*. Based on the collection method, the corpus has both characteristics of a designed as well as a random corpus: first, instances were retrieved based on the searches made to the Google search

[6] Guidelines for annotation with the FTD schema are available here https://github.com/ssharoff/genre-keras/blob/master/annot-v4.md.

engine using popular keywords. Secondly, the collection was enlarged by adding random web pages, and finally, instances of less frequent genre categories were manually collected to obtain a balanced corpus. The collection was one of the firsts that opted for the multi-labelling approach. Recently, the schema was transformed into a hierarchical set of categories, which was shown to improve the automatic genre prediction (Madjarov et al., 2019).

4.3.3 Other datasets

The *Leeds Web Genre Corpus* (Asheghi et al., 2016) is one of the largest English genre datasets, consisting of almost 5,000 texts. The dataset consists of two subcorpora, created with the aim of evaluating the proposed method for manual annotation: a balanced, manually collected subcorpus, and a random subcorpus, collected based on the automated queries made to a search engine. The balanced subcorpus was used to test the reliability of the manual annotation on a more controlled set of texts, and then the annotation campaign was extended to the random subcorpus to test the method in realistic conditions. The authors devised their own schema, constructed following the top-down approach, starting from the Genre-KI-04 schema (Eissen & Stein, 2004). The schema consists of 20 specific genres, i.e., *Personal Homepage, Company/Business Homepage, Educational Organizational Homepage, Personal Blog/Diary, Online Shops, Instruction/How to, Recipe, News Article, Editorial, Conversational Forum, Biography, Frequently Asked Questions, Review, Interview, Story, Dictionary/Thesaurus Entries, Link Lists or Directories of Links, Song Lyrics, Quotes* and *Encyclopedic Articles*. The categories were shown to be able to cover 74% of the texts in the Leeds Corpus, while the remaining were annotated with the category *Other*. Despite the fact that the introduction of the category *Other* allows classification of any web text with the Leeds schema, the authors state that their main focus was not on "completeness of the genre inventory but on genre annotation methodology". Since the authors note that the aim behind the construction of the dataset and the corresponding schema is not to provide a high-coverage schema, but to analyse genre annotation, we do not include the Leeds Web Genre Corpus under high-coverage collections, but rather describe it separately. The dataset was annotated via crowd-sourcing where each text was annotated by 5 participants. The annotation campaign was shown to be successful and a high inter-annotator agreement was reported, with the percentage agreement of 88% and Fleiss' kappa (Fleiss, 1971) of 0.87 on the balanced subcorpus, and the percentage agreement of 78% and Fleiss' kappa of 0.71 on the random subcorpus.

5 Automatic genre identification

Genre researchers used various technologies for automatic genre identification, mostly depending on the state-of-the-art of machine learning algorithms at the time of research. In the past, support vector machines (SVMs) were most frequently used (Rezapour Asheghi, 2015; Laippala et al., 2017, 2021; Sharoff et al., 2010; Pritsos & Stamatatos, 2018; Petrenz & Webber, 2011), as it was shown that they are very

suitable for text categorization (Joachims, 1998). Other methods, previously used for genre classification, are the discriminant analysis (Feldman et al., 2009; Biber & Egbert, 2015) which is the earliest method that was applied to this task (Karlgren & Cutting, 1994), decision tree classifiers, more specifically the C4.5 algorithm (Finn & Kushmerick, 2006; Dewdney et al., 2001), Naive Bayes algorithm (Feldman et al., 2009; Priyatam et al., 2013) and graph-based models using hyper-link connections between web pages (Asheghi et al., 2014; Zhu et al., 2011). In recent studies, SVMs were used to obtain an insight into which feature sets are most relevant for genre identification (Laippala et al., 2021; Sharoff et al., 2010; Pritsos & Stamatatos, 2018; Asheghi et al., 2014). As noted by Laippala et al. (2021), it is very valuable to include linguistic analyses into the research to "not only achieve high performance but also to guarantee the validity of the results". For these further analyses, linear support vector machines were shown to be an appropriate choice.

Genre researchers experimented with various feature sets, e.g., bag-of-words, consisting of lexical features (words, word or character n-grams) and/or grammatical features (part-of-speech tags) (Sharoff, 2021; Laippala et al., 2021), punctuation, text statistics (Finn & Kushmerick, 2006), visual features of HTML web pages, such as HTML tags and images (Lim et al., 2005; Levering et al., 2008; Maeda & Hayashi, 2009), and URLs of web documents (Abramson & Aha, 2012; Jebari, 2014; Priyatam et al., 2013). However, results revealed that the discriminative features vary across studies and datasets, as well as across genre classes (see Laippala et al. (2021)). Nevertheless, multiple studies showed that lexical features describe genres better than grammatical ones, which was based on the comparison of lexical features and part-of-speech tags (Pritsos & Stamatatos, 2018; Laippala et al., 2021; Sharoff et al., 2010). Thus, when recent studies performed machine learning experiments with symbolic classifiers, that is, classifiers which are explainable and non-neural, they most commonly used lexical features – character and/or word n-grams (Pritsos & Stamatatos, 2018; Bulygin & Sharoff, 2018; Lepekhin & Sharoff, 2022) or a combination of lexical and grammatical features (Asheghi et al., 2014; Laippala et al., 2021). The latter was shown to provide better results than training on lexical information alone and to assure more stable models, that is, models which are capable of generalizing beyond the training data (see Laippala et al. (2021)). Recently, Kuzman and Ljubešić (2022) analysed the impact of using yet another type of grammatical features, namely syntactic dependencies, and showed that this textual representation is able to capture genre information better than lexical and part-of-speech features. In addition to this, when trained on syntactic dependencies, the classification model learns the structure of the sentences instead of word meanings, which quite likely assures that the prediction is not topic-dependent. This assumption, however, has not been proven yet.

Recent developments in NLP shifted the focus to neural networks due to significant improvements on all fronts. Researchers experimented with deep as well as shallow neural networks and found that for this task, the linear fastText (Joulin et al., 2017) model, which is a shallow neural model, is comparable to convolutional neural networks (CNN) with pre-trained word embeddings, while being computationally more efficient (see Laippala et al. (2019)). Furthermore, it was shown that the fastText model outperforms traditional methods, i.e., SVMs,

decision trees, random forest classifiers, logistic regression classifiers and the Naive Bayes classifier at this task (Kuzman & Ljubešić, 2022).

Recently developed Transformer-based pre-trained language models led to a significant breakthrough in this field. Repo et al. (2021) and Kuzman et al. (2022b) showed that these deep learning models can achieve strong performance on small amount of training data, and, what is even more, they can reach around 30 points higher micro and macro F1 scores than the previous best model, fastText (see Kuzman et al. (2022b)). One likely reason for such drastic improvements obtained through the Transformer models in comparison to previous methods could be the fact that Transformer text representations incorporate information on syntax as well, which was shown to be very informative for genre classification (Kuzman and Ljubešić (2022)). Moreover, the models are capable of good levels of cross-lingual transfer. This was demonstrated when zero-shot cross-lingual experiments were performed by learning on a large English CORE dataset and testing on smaller datasets in Finnish, French, and Swedish (Repo et al., 2021). The cross-lingual experiments resulted in F1 scores between 0.61 and 0.69, while the results of training and testing the classifier on the same datasets in a monolingual setting range between 0.73 and 0.83 F1 (see Table 4). This was further researched by Ronnqvist et al. (2021) who showed that by combining all the four CORE datasets into a multilingual dataset, the performance is further improved, reaching micro F1 scores between 0.72 and 0.84. The multilingual classifier also achieved better zero-shot performance, as the classifier trained on three of the datasets and tested on the fourth reached micro F1 scores between 0.63 and 0.80. The highest scores were achieved with the multilingual XLM-RoBERTa model (Conneau et al., 2020) which outperformed other multilingual pre-trained Transformer language models in multiple genre studies (Ronnqvist et al., 2021; Repo et al., 2021; Kuzman & Pollak, 2022a). Furthermore, the XLM-RoBERTa model was revealed to be comparable (Kuzman et al., 2022b) or better (Repo et al., 2021) than monolingual models. This is in contrast to other NLP tasks, where the monolingual models were shown to outperform the multilingual models (Ulčar et al., 2021).

In addition to choosing the optimal features and machine learning algorithms to achieve good classification performance, genre researchers need to tackle additional genre-specific challenges, such as handling noise, i.e., texts which do not fit under any of the categories, and dealing with hybrid texts by performing multi-label or span-based classification. Handling noise was researched by Pritsos and Stamatatos (2018), and regarding the hybrid texts, Santini (2006) proposed a new method of multi-label classification: classifying texts by applying several classification models on them, trained on different genre datasets with different genre schemata. Recently, multi-label classification was also taken up by Madjarov et al. (2019), who performed multi-label classification and hierarchical multi-label classification using Predictive Clustering Trees (PCTs), Sharoff (2021), who used a Bi-directional Long Short-Term Memory (BiLSTM) classifier (Yogatama et al., 2017), by Ronnqvist et al. (2021), who used multilingual Transformer models for cross-lingual multi-label classification experiments, and Laippala et al. (2022a) who used a monolingual Transformer model for multi-label classification experiments.

Table 3 Accuracy obtained on the best-performing feature (for each of the datasets), reported in the experiments with the SVM (Sharoff et al., 2010)

Dataset	Accuracy
7-Web Genre Collection (Santini, 2007)	97.14
Genre-KI-04 (Eissen & Stein, 2004)	85.81
Hierarchical genre collection (Stubbe & Ringlstetter, 2007)	65.72
KRYS I (Berninger et al., 2008)	62.02
I-EN-Sample (Sharoff, 2010)	60.40
20-Genre Collection (Vidulin et al., 2007)	56.45

Another challenge for genre identification is assuring stability of the trained models. If the genre-annotated dataset is not general enough, the models could learn to classify texts based on other characteristics that are common to the texts of one class, such as the topic, instead of genre characteristics, and would not be able to generalize beyond the dataset (Sharoff et al., 2010). Petrenz and Webber (2011) assessed the stability of various machine learning methods by analysing their performance across changes in topic-genre distribution and advised that "stability should join accuracy as a criterion for assessing any new developments in genre classification". This was further explored by Laippala et al. (2021) who analysed which features provide the most stable SVM models. Since the advent of deep neural models, exploration of their stability has been tightly connected with research on their explainability, as the models act as black boxes. Thus, to explore which genres are more affected by topical biases, Lepekhin and Sharoff (2021) performed adversarial attacks on Transformer models, and Ronnqvist et al. (2022) introduced an Integrated Gradients input attribution method to achieve stable explanations of genre predictions based on keyword lists. The issue of topical shift which impacts the performance of the models was recently addressed by Lepekhin and Sharoff (2022) who performed experiments on two genre-annotated datasets with different genre distribution. The results showed that training the Transformer models on multiple datasets improves the performance. In addition to that, they experimented with ensembles of a monolingual Russian RuBERT (Kuratov & Arkhipov, 2019) model, multilingual XLM-RoBERTa (Conneau et al., 2020) and the Logistic Regression classifier and showed that ensembles of pre-trained models mostly outperform single Transformer models.

As genre-annotated datasets were created with different end goals and as no reference genre-annotated dataset exists, previous machine learning experiments focusing on automatic genre identification are "self-contained, and corpus-dependent" (Rehm et al., 2008). In addition to using different datasets, genre classes, classification algorithms and feature sets, researchers reported performance of the models with different metrics, making comparison between the experiments impossible. The last comparison of genre-annotated datasets based on AGI experiments was published by Sharoff et al. (2010) who compared most of the relevant genre datasets that existed at the time of research: the Hierarchical genre collection (Stubbe & Ringlstetter, 2007), the I-EN-Sample (Sharoff, 2010), Genre-KI-04 (Eissen & Stein,

Table 4 Overview of the reported results from the most relevant in-dataset experiments on recently developed genre datasets

Research	Dataset	Number of labels	Model	Best score (per dataset)
Asheghi et al. (2014)	Leeds corpus (Asheghi et al., 2016)	15 (subset of Leeds labels)	Semi-supervised multi-class min-cut graph-based algorithm (Ganchev & Pereira, 2007)	Accuracy: 0.90
Sharoff (2021)	FTD (English), FTD (Russian) (Sharoff, 2018)	10 (subset of FTD labels)	BiLSTM (Yogatama et al., 2017)	Precision: 0.77, 0.75
Repo et al. (2021)	FinCORE, FreCORE, SweCORE, CORE (Egbert et al., 2015)	7 (subset of CORE main labels)	XLM-RoBERTa large (Transformer model) (Conneau et al., 2020)	F1: 0.73, 0.77, 0.83, 0.76
Kuzman et al. (2022b)	GINCO	12 (merged GINCO labels)	SloBERTa (Transformer model) (Ulčar & Robnik-Šikonja, 2021)	Micro F1: 0.70, Macro F1: 0.67
Laippala et al. (2022a)	CORE (Egbert et al., 2015)	56 (CORE main and sub-categories; multilabel approach)	BERT large (Transformer model) (Kenton & Toutanova, 2019)	Micro F1: 0.68

2004), the KRYS I (Berninger et al., 2008) dataset, the 20-Genre Collection (Vidulin et al., 2007) and the 7-Web Genre Collection (Santini, 2007) dataset. They used a linear SVM model, and the part-of-speech n-grams, character and word n-grams as the features. The results that were obtained on the best-performing feature set, shown in Table 3, show very high results obtained by the classifiers trained on the 7-Web Genre Collection or the Genre-KI-04 dataset. While they performed with accuracy over 85, the results for classifiers trained on other datasets ranged between 56 to 66 in terms of accuracy. The worst results were obtained on the 20-Genre Collection. Despite the high results of some of the classifiers, the authors noted that "these impressive results might not be transferable to the wider web" due to the fact that the collections are not comparable between each other even on similar categories, that they are lacking in representativeness and that the annotations either cannot be tested for reliability or were revealed to not be sufficiently reliable. This was confirmed via cross-dataset experiments, that is, by testing the classifiers on all other compared datasets. There, the scores for accuracy were mostly below 40.

The results from the machine learning experiments on the genre datasets that were developed since then are reported in Table 4. The reported results give the reader a feel of the current state-of-the-art performance of genre classifiers on the datasets, however, as the experiments are corpus- and technology-dependent, the results are hardly comparable. Furthermore, as can be seen from Table 4, most of the experiments used only a subset of labels from their high-coverage schemata, and thus the results do not show the trained classifiers' capacity of identifying genres in the entire dataset. Thus, to be able to compare the suitability of genre datasets for automatic genre identification, a study is needed which would include all of the datasets.

First step towards comparing recent genre datasets was made by Kuzman and Pollak (2022a) who explored the comparability of the Slovene GINCO dataset (Kuzman et al., 2022b) and the English CORE dataset (Egbert et al., 2015). They mapped the original GINCO categories and CORE subcategories to a joint schema and performed cross-dataset experiments. As the datasets are in different languages, the experiments were cross-lingual at the same time. The experiments showed that the datasets are comparable enough to allow cross-dataset and cross-lingual transfer, reaching micro and macro F1 scores between 0.60–0.64 and 0.52–0.66. For comparison, the in-dataset results ranged between 0.78–0.81 in micro F1 and 0.73–0.84 in macro F1 for GINCO, and reached 0.77 micro F1 and 0.72 macro F1 for CORE. The results are comparable to the cross-lingual experiments, performed by Repo et al. (2021) on the English, Swedish, French and Finnish datasets, annotated with the CORE schema. Both studies also revealed that despite the fact that the CORE dataset has up to 40-times more instances than other compared datasets, the difference in size does not result in significantly better results. This might indicate that the pre-training of Transformer models gives such informative representations that Transformer models reach high results already after a few thousand instances observed, and adding tens of thousands of instances to the training does not significantly improve the results. This hypothesis is supported by recent experiments on the CORE dataset, which showed that the learning curve of a Transformer language model started flattening already after using 30% of training data (around 10,000

texts) (Laippala et al., 2022a). In addition, the study revealed that the beginning of the document is the most informative part for genre prediction, which means that the Transformer models are potent enough to only need a short part of a text to recognize a genre.

Another theory could be that the CORE dataset is less suitable for the automatic genre identification, since the reported low inter-annotation agreement (Egbert et al., 2015; Sharoff, 2018) could mean that there is more noise in the data. However, as performance of genre classifiers was often improved by training on multiple datasets (Lepekhin & Sharoff, 2022; Ronnqvist et al., 2021), the next step could be to experiment with training a classifier on the combination of GINCO and CORE, since they were shown to be comparable. The described research explored comparability to some extent, but it should be noted that since the labels vary significantly in terms of granularity, the joint schema did not cover all of the CORE and GINCO labels. In addition to this, the experiments were single-label. Thus, texts belonging to discarded labels, and CORE texts annotated with more than one class were not used in the experiments. This means that the comparison was performed on reduced datasets.

Classification experiments based on a joint schema seem hardly possible without any interventions to the data due to many differences between most relevant high-coverage datasets: the CORE (Egbert et al., 2015), FTD (Sharoff, 2018) and GINCO (Kuzman et al., 2022b) datasets. The FTD dataset is annotated using the dimensional and not categorical approach, meaning that the texts are annotated with multiple labels. Similarly, the CORE corpus is annotated with two levels of categories, main categories and subcategories, where some texts belong to multiple main categories, multiple subcategories or to no subcategory at all. This further complicates mapping to a joint schema, and requires multi-label experiments. However, an analysis of the real usefulness of the datasets and their schemata for classification applied to new data would be very beneficial to the community. To this end, an indirect comparison via downstream evaluation could be performed, either on samples of various web corpora, or on whole web corpora. Each of the datasets could be used to train a separate classifier, using the original schema. The classifiers would be applied to the sample, or the whole corpus. The sample could be used for estimating the accuracy and usefulness of the predictions directly by human annotators, while the whole annotated corpus could be used in various extrinsic evaluations, and evaluated on the target task.

Despite the challenges in automatic genre identification that still exist, the recent technology breakthrough has allowed creation of Transformer-based genre classifiers which can already be applied to large new datasets in numerous languages. Recently, Laippala et al. (2022b) published the Register Oscar dataset, consisting of 351 million documents in 14 languages to which genre labels were automatically assigned. The dataset was created by applying the multilingual genre classifier (Ronnqvist et al., 2021) on the OSCAR datasets (Suárez et al., 2019). The classifier is trained on the CORE datasets and uses the 8 main CORE labels. The zero-shot classification was evaluated based on annotated datasets in eight new languages (Arabic, Catalan, Chinese, Hindi, Indonesian, Portuguese, Spanish and Urdu). The results for the new languages were promising, ranging between 0.58 and 0.82 F1 scores.

Similarly, the genre classifier based on the GINCO dataset (Kuzman et al., 2022b) is planned to be applied to massive monolingual and parallel web corpora, produced for 10 under-resourced European languages in the scope of the MaCoCu project (Bañón et al., 2022a). The MaCoCu corpora, automatically annotated with genre labels, are reported to be released in 2023.[7] As opposed to genre datasets, presented in Sect. 4, which were manually annotated with genre classes and used as the basis for training AGI classifiers, the Register Oscar and MaCoCu datasets are automatically annotated with genre classes and are ones of the first results of applying genre classifiers on unseen collections of texts.

6 Conclusion

This survey presents an exhaustive overview of previous research on automatic genre identification. Thereby, it uncovers numerous challenges of the task and the approaches to dealing with them, further explained in the following paragraphs:

1. Lack of consensus on the main concepts, connected with the automatic genre identification, leading to a proliferation of competing genre schemata and approaches.
2. Existence of hybrid texts and multi-text documents in web datasets, related to the limited control over the content creation and collection of web texts, which pose challenges for manual annotation and automatic classification.
3. Lack of comparability between existing genre datasets, aggravated by the fact that they are hard to find or have ceased to be available.
4. Challenges, connected with assuring stability of genre classifiers and their good performance also in out-of-domain scenarios, so that the classifiers are not dataset-dependent.

The first challenge is that there exists no consensus on the term that should be used for this phenomenon, on the definition of the concept itself, and on the set of labels to be included into the schema to assure high coverage, reliable annotation and good classification results. As researchers define genres differently and perform research on them with different aims, they use very different genre schemata. We analysed 16 of them and showed that out of 177 labels only 48 labels are used in more than one schema.

Secondly, the reliability of genre datasets and consequently also the results of automatic genre identification can be negatively impacted by the fundamental difficulties connected with the genre notion itself. During annotation and machine learning experiments, researchers need to deal with hybrid texts, i.e., texts which have features of multiple genres, texts without any discernible purpose or features, and documents that consist of multiple texts, such as a letter, posted as a part of a blog. On the web, writers are not constrained to following genre norms, thus, the texts

[7] https://macocu.eu/

vary significantly in regard of their genre prototypicality. These difficulties contributed to rather low inter-annotator agreement in some of the annotation campaigns, resulting in genre datasets with questionable reliability.

Thirdly, the last review of existing genre datasets, published in 2010 (see Sharoff et al. (2010)), exposed the lack of comparability, representativeness and reliability of existing genre datasets, and highlighted a need for further research on AGI. They called for efforts to develop a large reference corpus, collected from a diverse range of sources, and with a genre schema which would allow for reliable annotation. At that time, genre datasets almost exclusively contained only English documents, thus, the authors also urged for a creation of datasets in other languages, which would enable cross-lingual automatic genre identification. Our paper is the first work that surveys the research on automatic genre identification that was done in the last 12 years, describing that research in the context of the older work as well.

In this period, the technological progress delivered crawling tools which enable fast collection of thousands, or even millions of texts from the web, assuring representativeness of the datasets. Secondly, the researchers benefited from the advent of crowd-sourcing platforms, which allow quick and relatively inexpensive annotation of tens of thousands of texts. And finally, recently developed deep neural machine learning technologies relying on self-supervision, such as the Transformer language models, were shown to be able to classify genres drastically better than any previous technology, even when trained on only a thousand of texts. Furthermore, the models were revealed to achieve good results on cross-lingual genre classification which opens the doors to empowering a multitude of languages with the benefits of automatic genre identification.

While early genre datasets mostly focused on a small set of specific labels, recently, numerous datasets were developed with schemata which aim to cover the entire genre composition of the web, which is the first step to providing a general dataset with labels that could be applied to any corpus. Such schemata are the FTD schema (Sharoff, 2018), based on which English, Russian and Arabic genre datasets were created, and the CORE schema (Egbert et al., 2015) which served as a basis for English, Finnish, French, Swedish, Slovene and other datasets. These datasets addressed the lack of representativeness, comparability and reliability that was observed for earlier genre collections.

While recently many new genre-annotated datasets emerged, previous datasets ceased to be available. This paper is the first to provide an extensive list of all available genre datasets with information on where to access them. Despite a recent positive trend of publishing genre datasets, the majority of datasets were hard to find. They are not published in a certified repository which would assure reliable long-term archiving. This should be addressed by the community to assure that the datasets are available even 10 years after the research was published. In addition to this, to facilitate further advances in this field, researchers should strive to also publish the code used in their AGI experiments, to make the research reproducible and open.

Recent emergence of multiple high-coverage datasets in various languages has been an important step towards developing a genre classifier which could be applied to any dataset in multiple languages. Recent research (Ronnqvist et al., 2021; Lepekhin & Sharoff, 2022) achieved very promising results by training Transformer

models on multiple datasets joined together. Thus, the key to a general cross-lingual genre identifier could lay in merging comparable high-coverage datasets. By training a model on multiple datasets, it would be less prone to topical and other biases, which would assure its stability. To be able to merge the high-coverage datasets, i.e., the CORE dataset (Egbert et al., 2015), the FTD dataset (Sharoff, 2018) and the GINCO dataset (Kuzman et al., 2022b), a cross-dataset analysis is necessary. To avoid interfering with the datasets and schemata, for each of the datasets, a separate Transformer model could be trained, using its original schema. Then each classifier could be applied to all other high-coverage datasets. This would result in each text from each of the datasets being labelled three times: with a GINCO category, a CORE category and an FTD category. The quantitative analysis of the results would reveal comparability of the genre schemata. Moreover, it could be extended to an extrinsic evaluation to analyse the usability of assigned genre labels for the users. Based on this, two promising directions are possible: 1) merging the datasets into one based on a joint schema to obtain a more stable and potent genre classifier; or 2) instead of creating just one model and applying it on a dataset, texts could be automatically annotated with predictions of all three classifiers, the GINCO-based, FTD-based and CORE-based classifier, allowing the corpora users to choose the labels that are most relevant for their use case.

Development of general and stable genre classifiers is a first step to providing reliable automatic genre identification. However, as observed by previous work, the field needs more research that would address additional challenges related to this task, i.e., performing multi-label classification on hybrid texts, handling noise, and separating multiple texts in a multi-text document. Only then we will have classifiers that will be truly useful in realistic conditions on the web.

Author contributions All authors contributed to the study conception and design. Literature search, data collection and analysis were performed by TK. The first draft of the manuscript was written by TK and NL, who also critically revised the work. All authors read and approved the final manuscript.

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

Availability of data and materials The authors confirm that all datasets analysed during this study are referenced in this published article and the links to their location are provided.

Code availability Not applicable.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Abramson, M., & Aha, D.W. (2012). What's in a URL? Genre Classification from URLs. Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Agrawal, S., Sanagavarapu, L.M., & Reddy, Y.R. (2019). FACT-Fine grained assessment of web page CredibiliTy. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 1088–1097.

Argamon, S., Koppel, M., & Avneri, G. (1998). Routing documents according to style. In: First International Workshop on Innovative Information Systems, pp. 85–92.

Asheghi, N.R., Markert, K., & Sharoff, S. (2014). Semi-supervised graph-based genre classification for web pages. In: Proceedings of TextGraphs-9: The Workshop on Graph-Based Methods for Natural Language Processing, pp. 39–47.

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

Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., & Suchomel, V. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 301–302.

Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., & Zaragoza, J. (2022). Slovene web corpus MaCoCu-sl 1.0. (Slovenian language resource repository CLARIN. SI)

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.

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

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press.

Biber, D., & Egbert, J. (2015). 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*(1), 3–36.

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

Boese, E.S. (2005). Stereotyping the web: Genre classification of web documents (Unpublished doctoral dissertation). Citeseer.

Bulygin, M., & Sharoff, S. (2018). Using machine translation for automatic genre classification in Arabic. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 153–162.

Chandler, D. (1997). An introduction to genre theory.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V.,Wenzek, G., Guzmán, F., & 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.

Crowston, K., Kwaśnik, B., & Rubleske, J. (2010). *Problems in the use-centered development of a taxonomy of web genres. Genres on the Web* (pp. 69–84). Springer.
Davies, M. (2004). British National Corpus (from Oxford University Press). Available online at https://www.english-corpora.org/ bnc/
Davies, M. (2008). The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/ coca/
Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide, 36*(1), 1–28.
Dewdney, N., Van Ess-Dykema, C., & MacMillan, R. (2001). The form is the substance: Classification of genres in text. In: Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management.
Dewe, J., Karlgren, J., & Bretan, I. (1998). Assembling a balanced corpus from the internet. In: Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998), pp. 100–108.
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*(9), 1817–1831.
Erjavec, T., & Ljubešić, N. (2014). The slwac 2.0 corpus of the slovene web. T. Erjavec, J. Žganec Gros (ur.). *Jezikovne tehnologije zbornik, 17*, 50–55.
Feldman, S., Marin, M.A., Ostendorf, M., & Gupta, M.R. (2009). Part-of-speech histograms for genre classification of text. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4781–4784.
Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology, 57*(11), 1506–1518.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378.
Forsyth, R. S., & Sharoff, S. (2014). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing, 29*(1), 6–22.
Freund, L., Clarke, C.L., & Toms, E.G. (2006). Towards genre classification for IR in the workplace. In: Proceedings of the 1st International Conference on Information Interaction in Context, pp. 30–36.
Ganchev, K., & Pereira, F. (2007). Transductive structured classification through constrained min-cuts. In: Proceedings of the Second Workshop on Textgraphs: Graph-Based Algorithms for Natural Language Processing, pp. 37–44.
Giesbrecht, E., & Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: Proceedings of the Fifth Web as Corpus Workshop, pp. 27–35.
Jebari, C. (2014). A pure URL-based genre classification of web pages. In: 2014 25th International Workshop on Database and Expert Systems Applications, pp. 233–237.
Jebari, C. (2021). Enhancing the identification of web genres by combining internal and external structures. *Pattern Recognition Letters, 146*, 83–89.
Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning* (pp. 137–142). Springer.
Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the Fifteen Conference of the European Chapter of the Association for Computational Linguistics, 2*, 427–431.
Kanaris, I., & Stamatatos, E. (2007). Webpage genre identification using variable-length character n-grams. *IEEE International Conference on Tools with Artificial Intelligence, 2*, 3–10.
Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management, 45*(5), 499–512.
Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In: Proceedings of the 15th International Conference on Computational Linguistics.
Kennedy, A., & Shepherd, M. (2005). Automatic identification of home pages on the web. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 99c–99c.
Kenton, J.D.M.-W.C., & Toutanova, L.K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacl-hlt, pp. 4171–4186.
Kilgarriff, A. (2012). Getting to know your corpus. In: International Conference on Text, Speech and Dialogue, pp. 3–15.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.
Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.
Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 333–339.
Kuzman, T., & Ljubešić, N. (2022). Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In D. Mladenić & M. Grobelnik (Eds.), *Odkrivanje znanja in podatkovna skladišča - SiKDD: 10.* Institut Jožef Stefan.
Kuzman, T., Rupnik, P., & Ljubešić, N. (2022). The GINCO training dataset for web genre identification of documents out in the wild. *Proceedings of the language resources and evaluation conference* (pp. 1584–1594). European Language Resources Association.
Kuzman, T V. N., & Pollak, S. (2022). Assessing comparability of genre datasets via cross-lingual and cross-dataset experiments. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: Zbornik konference* (pp. 100–107). Institute of Contemporary History.
Kwaśnik, B. H., & Crowston, K. (2005). *Introduction to the special issue: Genres of digital documents*. Information Technology & People.
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.
Laippala, V., Luotolahti, J., Kyröläinen, A.-J., Salakoski, T., & Ginter, F. (2017). Creating register subcorpora for the Finnish Internet Parsebank. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, pp. 152–161.
Laippala, V., Rönnqvist, S., Hellström, S., Luotolahti, J., Repo, L., Salmela, A., & Pyysalo, S. (2020). From web crawl to clean register-annotated corpora. In: Proceedings of the 12th Web as Corpus Workshop, pp. 14–22.
Laippala, V., Salmela, A., Rönnqvist, S., Aji, A.F., Chang, L.-H., Dhifallah, A., & Skantsi, V. (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), pp. 215–221.
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, 5*, 1–32.
Laippala, V., Rönnqvist, S., Oinonen, M., Kyröläinen, A.-J., Salmela, A., Biber, D., & Pyysalo, S. (2022). Register identification from the unrestricted open web using the corpus of online registers of English. *Language Resources and Evaluation, 1*, 1–35.
Lee, Y.-B., & Myaeng, S.H. (2002). Text genre classification with genrerevealing and subject-revealing features. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 145–150.
Lee, Y.-B., & Myaeng, S.H. (2004). Automatic identification of text genres and their roles in subject-based categorization. In: 37th Annual Hawaii International Conference on System Sciences.
Lee, D. (2002). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Teaching and learning by doing corpus analysis* (pp. 245–292). Brill Rodopi.
Lepekhin, M., & Sharoff, S. (2021). Experiments with adversarial attacks on text genres. arXiv preprint arXiv:2107.02246
Lepekhin, M., & Sharoff, S. (2022). Estimating confidence of predictions of individual classifiers and their ensembles for the genre classification task. *Proceedings of the language resources and evaluation conference* (pp. 5974–5982). European Language Resources Association.
Levering, R., Cutler, M., & Yu, L. (2008). Using visual features for fine-grained genre classification of web pages. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), pp. 131–131.
Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management, 41*(5), 1263–1276.
Lukin, A., Moore, A.R., Herke, M., Wegener, R., & Wu, C. (2011). Halliday's model of register revisited and explored.
Madjarov, G., Vidulin, V., Dimitrovski, I., & Kocev, D. (2019). Web genre classification with methods for structured output prediction. *Information Sciences, 503*, 551–573.

Maeda, A., & Hayashi, Y. (2009). Automatic genre classification of Web documents using discriminant analysis for feature selection. In: 2009 Second International Conference on the Applications of Digital Information and Web Technologies, pp. 405–410.

Mason, J.E., Shepherd, M., & Duffy, J. (2009). An n-gram based approach to automatically identifying web page genre. In: 2009 42nd Hawaii International Conference on System Sciences, pp. 1–10.

Moessner, L. (2001). Genre, text type, style, register: A terminological maze? *European Journal of English Studies, 5*(2), 131–138.

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2021). Genre as Weak Supervision for Cross-lingual Dependency Parsing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4786–4802.

Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly, 5*, 541–574.

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

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., & Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In: Proceedings of 8th Web as Corpus Workshop (WAC-8), pp. 24–29.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora (Unpublished doctoral dissertation)*. Masaryk university Faculty of informatics.

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

Priyatam, P. N., Iyengar, S., Perumal, K., & Varma, V. (2013). Don't use a lot when little will do: Genre identification using URLs. *Research in Computing Science, 70*, 233–243.

Rehm, G. (2002). Towards automatic Web genre identification: a corpus-based approach in the domain of academia by example of the Academic's Personal Homepage. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, pp. 1143–1152.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., & Vidulin, V. (2008). *Towards a reference corpus of web genres for the evaluation of genre identification systems*. Lrec.

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

Rezapour Asheghi, N. (2015). *Human annotation and automatic detection of web genres (Unpublished doctoral dissertation)*. University of Leeds.

Rönnqvist, S., Kyröläinen, A.-J., Myntti, A., Ginter, F., & Laippala, V. (2022). Explaining Classes through Stable Word Attributions. Findings of the association for computational linguistics: Acl 2022, pp. 1063–1074.

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.

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

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., & Liu, X. (2001). Genre based navigation on the web. In: Proceedings of the 34th annual Hawaii international conference on system sciences, p. 10.

Santini, S.M. (2006). Common criteria for genre classification: Annotation and granularity. In: Workshop on Text-based Information Retrieval (TIR-06). Conjunction with ECAI 2006, Riva del Garda, 2006.

Santini, M. (2007). *Automatic identification of genre in web pages (Unpublished doctoral dissertation)*. University of Brighton.

Santini, M. (2010). *Cross-testing a genre classification model for the web. Genres on the Web* (pp. 87–128). Springer.

Santini, M., Mehler, A., & Sharoff, S. (2010). *Riding the rough waves of genre on the web. Genres on the Web* (pp. 3–30). Springer.

Sharoff, S. (2021). Genre annotation for the web: text-external and textinternal perspectives. Register studies.

Sharoff, S. (2010). *In the garden and in the jungle genres on the web* (pp. 149–166). Springer.

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

Sharoff, S., Wu, Z., & Markert, K. (2010). *The Web Library of Babel: Evaluating genre collections*. Lrec.
Shavrina, T. (2019). Genre classification problem: In pursuit of systematics on a big webcorpus. *Proceedings of Third Workshop Computing, 4*, 70–83.
Skantsi, V., & Laippala, V. (2023). Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics, 1*, 1–31.
Snow, R., O'connor, B., Jurafsky, D., & Ng, A.Y. (2008). Cheap and fast– but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254–263.
Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471–495.
Stein, B., Eissen, S. M. Z., & Lipka, N. (2010). *Web genre analysis: Use cases, retrieval models, and implementation issues Genres on the Web* (pp. 167–189). Springer.
Stewart, J. G., & Callan, J. (2009). *Genre oriented summarization (Unpublished doctoral dissertation)*. Language Technologies Institute, School of Computer ScienceCarnegie Mellon University.
Ströbel, M., Kerz, E., Wiechmann, D., & Qiao, Y. (2018). Text genre classification based on linguistic complexity contours using a recurrent neural network. MRC@ IJCAI, pp. 56–63.
Stubbe, A., & Ringlstetter, C. (2007). Recognizing genres. Towards a reference corpus of web genres: Proceedings.
Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Blackwell Oxford.
Suárez, P.J.O., Sagot, B., & Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In: 7th Workshop on the Challenges in the Management of Large Corpora (cmlc-7).
Suchomel, V. (2020). Genre Annotation of Web Corpora: Scheme and Issues. In: Proceedings of the Future Technologies Conference, pp. 738–754.
Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model.
Ulčar, M., Žagar, A., Armendariz, C.S., Repar, A., Pollak, S., Purver, M., & Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. arXiv preprint arXiv:2107.10614 .
Van der Wees, M., Bisazza, A., & Monz, C. (2018). Evaluation of machine translation performance across multiple genres and languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
Vidulin, V., Luštrek, M., & Gams, M. (2007). Using genres to improve search engines. In: 1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing, pp. 45–51.
Williams, M., & Crowston, Kevin. (2000). Reproduced and emergent genres of communication on the World WideWeb. *Information Society, 16*(3), 201–215.
Yogatama, D., Dyer, C., Ling, W., & Blunsom, P. (2017). Generative and discriminative text classification with recurrent neural networks. In: Thirty-fourth International Conference on Machine Learning (ICML 2017).
Zhu, J., Zhou, X., & Fung, G. (2011). Enhance web pages genre identification using neighboring pages. In: International Conference on Web Information Systems Engineering, pp. 282–289.
Zu Eissen, S.M., & Stein, B. (2004). Genre classification of web pages. In: Annual Conference on Artificial Intelligence, pp. 256–269.