Received December 23, 2020, accepted January 27, 2021, date of publication February 3, 2021, date of current version February 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056927

Leveraging Machine Learning to Explain the Nature of Written Genres

MARTA VICENTE, MARÍA MIRÓ MAESTRE, ELENA LLORET, AND ARMANDO SUÁREZ CUETO

Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain

Corresponding author: Marta Vicente (mvicente@dlsi.ua.es)

This work was supported in part by the Ministry of Science and Innovation of Spain for the project "Integer: Intelligent Text Generarion" under Grant RTI2018-094649-B-I00, and in part by the Generalitat Valenciana through project "SIIA: Tecnologias del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible" under Grant PROMETEU/2018/089.

ABSTRACT The analysis of discourse and the study of what characterizes it in terms of communicative objectives is essential to most tasks of Natural Language Processing. Consequently, research on textual genres as expressions of such objectives presents an opportunity to enhance both automatic techniques and resources. To conduct an investigation of this kind, it is necessary to have a good understanding of what defines and distinguishes each textual genre. This research presents a data-driven approach to discover and analyze patterns in several textual genres with the aim of identifying and quantifying the differences between them, considering how language is employed and meaning expressed in each particular case. To identify and analyze patterns within genres, a set of linguistic features is first defined, extracted and computed by using several Natural Language Processing tools. Specifically, the analysis is performed over a corpora of documents—containing news, tales and reviews—gathered from different sources to ensure an heterogeneous representation. Once the feature dataset has been generated, machine learning techniques are used to ascertain how and to what extent each of the features should be present in a document depending on its genre. The results show that the set of features defined is relevant for characterizing the different genres. Furthermore, the findings allow us to perform a qualitative analysis of such features, so that their usefulness and suitability is corroborated. The results of the research can benefit natural language discourse processing tasks, which are useful both for understanding and generating language.

INDEX TERMS Applied computing, communicative objectives, discourse analysis, genre characterization, human language technologies, natural language processing.

I. INTRODUCTION

Among the important aspects and stages that are involved in the process of automatic interpretation and generation of language, the contribution of the communicative goal is a fundamental factor that needs to be considered in research. The process of language understanding and generation is often influenced by the communicative goal pursued [1]. In this manner, an informative text will be written differently from another one whose aim is to persuade, which in turn will be different from a complaining one.

Research focused on these communicative goals is deeply aligned with research that seeks to unravel the nature of textual genres, which represent the framework that shapes discourse so that it can fulfill its communicative goal. Indeed, it is through the textual conventions and patterns underlying each genre that communication and understanding within a community of speakers is possible [2]. The study of such shared patterns in textual genres enriches the automatic processing of text from both the perspective of understanding the text and generating the language.

The concept of genre, according to traditional linguistic theory, involves several overlapping dimensions that include not only the textual nature but also an analysis of the sociological dimension or context of the discourse [3]. Specifically, when it comes to its description, the following three components are considered: i) the typical linguistic characteristics, ii) the situational context and iii) the functional relationship between both, the latter attempting to explain, therefore, the textual choices in relation to the purpose

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

(i.e., communicative goal) for which the text is produced [4].

Natural Language Processing (NLP) is the discipline concerned with the automatic understanding and the production of language. NLP includes the following types of procedures: interpretation, performed for example in information extraction; generation, such as in reporting; and, those procedures that involve both interpretation and generation, as in the case of summarization or automatic translation.

In all these cases, knowledge of the linguistic characteristics that define the genre to which they are ascribed can increase the quality of the results, reinforcing the awareness or manifestation of the discourse's purpose.

The motivation to undertake the present research is grounded in the aforementioned scenario whereby the work of defining textual genres necessitates, in parallel, an inquiry into the relationship between these genres and their respective communicative goals.

Therefore, the main objective of this research is to present a data-driven approach to discover and analyze specific patterns occurring in different textual genres. To achieve this objective, these two research questions need to be addressed: RQ1) *Can a genre be sufficiently distinguished on the basis of a series of given characteristics?*, and RQ2) *Can the presence of each of these features be quantified and evaluated in different genres to identify the specificity of each one?* We consider that a good understanding of these characteristics and patterns with respect to each genre may benefit discourse processing tasks, which would be useful for understanding or generating language.

In order to address this task properly, we have selected corpus from different genres as collections of documents that share communicative goals. In this manner, our corpora is composed of news, tales and user-generated reviews, which are illustrative cases of three communicative goals—inform, entertain and persuade—. These genres differ in terms of purpose, but also share certain traits, all of them being examples of narratives. This aspect can therefore give us the opportunity to characterize both what they have in common and what distinguishes them.

The methodology adopted to respond to the questions posed is based on the use of machine learning (ML) techniques, and more specifically, on the definition of a classification task. The selection of this approach serves a double purpose since, first, one of the factors that determines the effectiveness of the classification is the quality of the features employed—which addresses RQ1—and secondly, certain mechanisms that ML offers, such as feature selection algorithms or the possibility to determine how a feature contributes to each genre classification, allow us to perform a detailed analysis of the presence of certain features—which aligns with RQ2.

Currently, there is a growing trend that places most of the research in NLP in the field of Deep Learning (DL) techniques. DL has become the standard place for almost every NLP task, probably due to the attractive results achieved as well as the constant innovation and development of resources.[1] However, without trying to diminish the relevance of this trend, and despite the evolving sophistication in the type of features used, most of DL research in NLP until now has been based on the use of certain types of characteristics sometimes referred to as ''surface features'', which are easily and rapidly extractable from large quantities of text [5]–[8]. This would be the case, for instance, with BOWs features, one-hot vectors or n-gram approaches; all of them able to procure, as indicated above, very competent results for certain tasks, but unable to provide parameters that could help to properly identify and represent a genre—its structure, its composition—or to supply the guidelines that can help reveal or shape its purpose.

By contrast, this work presents a non-arbitrary collection of characteristics that results from a detailed review of diverse linguistic studies related to discourse and genre, from previous research in the field of NLP and also from the direct examination of the texts concerned. The selection of characteristics is premised on both their processability by means of automatic linguistic tools, and their relevance with regard to the genre identification. The results reveal how the selected group of characteristics provides an insight on the different genres by examining the extent to which these features are present, and shared or not among the genres.

The paper is organized as follows. First, we introduce the related work on linguistic and computational approaches in Section II. Second, how the documents from the corpora were gathered and their description is detailed in Section III. Next, the feature engineering process required to design, extract and compute the set of features is explained in Section IV. The experimentation which helped us to answer the research questions is presented in Section V. Section VI raises further issues for consideration. Finally, the conclusions are summarized in Section VII, where future developments and applications of this work are also suggested.

II. RELATED WORK

Research devoted to genre analysis can be traced in different fields, from purely theoretical or linguistic studies to projects using no more linguistic information than merely the words that comprise the texts. We have referenced relevant works from non-computational areas as well as some that have advanced NLP developments related to our proposal.

As for linguistic approaches, we focus on studies that address narrative typology, given that all three genres considered in this research are examples of narrative. Currently, narrative constitutes one of the most researched textual typologies, a trend perhaps motivated by the successful interaction between linguistics and computers. However, most of the previous research focused its attention exclusively on a generic study of narratives or on the classification of the

[1] Impressive results can be found at nlpprogress.com, which is a repository that helps researchers track progress in NLP, by including both datasets and the state of the art for the most common NLP tasks. (last accessed in February 2021)

textual genres which comprise this typology [4], [9], but papers which delve into several genres individually from this category or even compare them are scarce.

In order to briefly analyze the research results that have been published in this linguistic field, it is worth mentioning the article written by [10] about the analysis of the language of children's literature used by four well-known English writers (including authors such as Potter or Carrol among them). She bases her study on corpus linguistics with the aim of showing the features used by these four authors that make this genre a very distinguished text in comparison to adult literature. We also need to include in this brief review the divergent approach on news by [11], where he also takes corpus linguistics as a basis in order to analyze future-oriented or unreal news using Danish articles with political themes that show a growing speculative intention. Finally, it seems clear that the rise of online linguistic genres offers us a wider spectrum of studies, with the paper published by [12] notable for its innovative approach to the much studied genre of reviews in the hotel domain, which emphasizes the communicative functions of each passage and also the linguistic structures that make them perfect examples of this genre.

Arguably, the scientific research in this area indicates a rather delimited scope within the narrative typology, focusing on generic features of each genre (or subgenre), without delving into further properties of the defined characteristics or comparisons of several genres from the same typology. The lack of research concerning the relationship among different genres belonging to the same textual typology gives computational linguistics the chance to bring to light results that may ease the task of studying each interlinked linguistic level—lexicon, structure, semantics and functions—of such genres more dynamically and thoroughly [13].

From the NLP perspective, genre research is mostly focused on the task of classification, and more specifically, on the achievement of optimal models that allow such classification to be efficient [14]–[16]. This is one of the reasons that explain the aforementioned tendency to employ for the task those ''surface features'' introduced before (bag-of-words, ngrams, word-to-vectors), which happens to be quite easily extractable. Features considered for these types of approaches can be defined without any knowledge of the syntactic or semantic structures that may underlie the discourse; moreover, they can be definable from an representation of the text as a simple set of characters [17]. In contrast to this type of research, we designed our approach to focus on selecting those features that can give a better insight on the common structure and shape that texts belonging to the same genres typically exhibit. In this manner, our approach sets out to define the best set of characteristics in relation to each genre rather than to design the best classifier. Notwithstanding, some interesting work was identified that inspired and helped us to compose the robust superset of features with which we worked, although the work cited did not adopt our approach.

Dealing with a very different set of genres to this study, [5] shows results specifically related to syntactic features as discriminative clues to distinguish among genres. First, a binary classification is performed considering as classes, spoken versus written discourse. In later stages, the classification becomes multiple, but the genres considered include for example *printed*, *non printed*, *public* or *private* as genres to classify.

In the work of [18], the author conducts a study considering several ML algorithms and groups of features, that includes among other features what he terms as ''linguistic expert knowledge set of features'', containing frequency of part-of-speech tags, pronouns and nouns, for example. Some of these features we shared for the present work, but he also considers a type of feature which depends on the most discriminant words. In our case, we decided not to include this type of characteristic because it could lead to a classification based on semantic domains (e.g. discriminating texts related to one sport against another), which is undesirable for our experiment since two texts belonging to two of the genres we consider could share the semantic domain (e.g. a piece of news and a children's tale, both related to a football match). Using these types of features is convenient for Onan's study since the genres he classifies are subgenres of reviews, specifically book reviews and camera reviews.

A different approach is proposed by [19] based on clustering techniques. Related to what they refer as ''*Popular Science*'', several documents are selected from a range of sources (science abstracts, science news, Wikipedia articles,...) and disciplines (medicine, biology, technology,...) and a set of linguistic features is also defined for the purpose of interpreting the clusters obtained considering the type of texts and domains. Moreover, they include within their experiments the possibility of inferring from the clusters communicative objectives and consider as such four purposes: narrate, discuss, describe/explain and summarize scientific or technical information.

A distinction between fiction and non-fiction as genres to classify can be found in [20]. The former genre would include news, reviews or hobbies while the second one, fiction or adventure. In this paper, the authors start from a set of 19 features based on part-of-speech counts, considering also different ratios among them, and look for a successful minimum set using feature selection techniques. As a result, they found that the two groups of texts that they differentiate only require two characteristics to be distinguishable in a classification task, both of which are ratios among word types. This result shows the relevance of considering such features in the composition of the texts, and underscores the importance of taking into account the occurrence of certain linguistic elements in comparison with others. This comparative approach has been incorporated into the design of our features. Nevertheless, the balance between the large variety of texts considered versus the low number of characteristics found to be valuable suggests that the strategy is not sufficient when the objective is not focused on the genre classification task, but on the adequate definition of the genres by features.

The research presented in this paper is inspired by the previously cited authors but develops the scope of the scientific research, both that produced in more linguistic environments and that developed in more computational contexts, covering a wider spectrum of genres and providing a more complete and functionally descriptive set of features. Moreover, as opposed to works that focus on the search for a better classifier to solve tasks in an optimal way, our research does not focus on the design of the best classifier, but on the definition of a set of characteristics and their analysis. By using the ML tools we can assess the quality of the designed set of features in terms of their prevalence in specific textual genres.

III. DATA COLLECTION & DESCRIPTION

To ensure a more comprehensive study in relation to defining the different genres, a series of documents have been compiled from a variety of sources. Therefore, while the news items are taken from several editions of the Document Understanding Conferences[2] (DUC), both the stories and the reviews that are included in the respective corpus are extracted from three different collections of texts. All the sources are detailed next.

- **News**. Documents from different editions of DUC [21] were included in the main corpus. Specifically, we compile and process the ones from DUC 2002 and 2004 editions, which include news about different topics, such as natural disasters or politics, among others.
- **Reviews**. For this textual genre, the documents were gathered from the SFU corpus [22], the Multidomain Sentiment Data collection [23], and Opinrank [24]. We selected these three corpora due to the fact that each of them contains reviews about different domains, which brings more variety in the style of the documents, and therefore allows for the consideration of a wide range of phenomena in the genre. The first corpus contains reviews from several type of products (books, movies,...) extracted from Epinions[3] web site. The second also compiles product reviews belonging to different domains, but taken from Amazon.[4] Finally, the third corpus contains user reviews of cars and and hotels collected from Tripadvisor[5] and Edmunds.[6]
- **Tales**. Regarding this genre, we focused on children's tales, using the existing Lobo and Matos corpus of fairy tales [25]. An automatic extraction of stories from certain websites was also performed. In this case, the documents needed were extracted from the website of *Hans Christian Andersen: Fairy Tales and Stories*[7] in addition to those obtained from the *Bedtime stories* site.[8]

Table 1 presents the statistics of the corpora. Although the number of documents per genre is different, the internal composition is very similar and well-balanced among the three. This can be viewed on the right hand side of the table, where the statistics show the number of sentences per document, words per document, and words per sentence.

TABLE 1. Statistical description of the corpora.

	# docs	# sents	# words	sents /doc	words /doc	words /sent
News	981	24,680	520,195	25	530	25
Reviews	1,121	35,267	659,578	31	588	21
Tales	433	12,308	284,219	28	656	25
Total	2,535	72,255	1,463,992			

IV. FEATURE ENGINEERING PROCESS

The objective of this research work is to perform an in-depth analysis of the relation between a certain set of characteristics and a series of textual genres, in order to define and parameterize such a connection so that it may foster an improvement in the different NLP tasks. Therefore, the constitution of a sufficiently solid set of features is a key task. In this section, a detailed explanation of this process is provided.

The definition of the set of features implies, first, determining which linguistic information should be considered relevant to identify the genres (see Section IV-A). In this manner, it will be decided that grammatical classes or the tense of verbs, for instance, could make a difference in the task of distinguishing between reviews and tales. In relation to such linguistic elements, apart from their frequency in a document, it is also important to compute other operations on the elements. Therefore, in Section IV-C, we define these operations performed on the linguistic information to compute, for example, the proportion of sentences of the document that are exclamatory—which can give insights on the structure of the document—or the ratio between two kinds of elements. Those operations represent the final step to characterize the complete set of features that will be computed for each document of the dataset. This means that, after extracting the linguistic information by using a series of NLP tools (Section IV-B), and having computed the correspondent operations, each document of the dataset will be represented as a collection of features quantified to enable an analysis from a genre perspective.

A. LINGUISTIC INFORMATION

In line with previous research and the examination of several documents of each genre, we defined a set of linguistic elements to be considered when performing the feature computation. In this manner, not only grammatical elements were included, but also constituents (e.g. different types

[2]https://duc.nist.gov/
[3]www.epinions.com
[4]https://www.amazon.com/
[5]https://www.tripadvisor.com/
[6]https://www.edmunds.com/
[7]http://hca.gilead.org.il/
[8]https://freestoriesforkids.com/

of subjects) or phrase types (e.g. verb phrases).[9] Besides, other information related to the verbs (e.g. modal verbs, tense and aspect) was found to be interesting. Semantic information was provided through named entities, event or time particles types. Discourse markers, which are key to understanding or establishing the structure of the discourse, are also part of the set of elements considered as well as several sentences types, such as exclamatory or interrogative.

One of the methods selected to obtain features directly related to the text as discourse, namely, a set of coherent sentences, was the use of coreference tools (detailed in the next subsection). Additionally, shallow features (e.g. length of words or length of named entities) were included as elements of interest. In Table 2, all the linguistic information that has been considered to obtain the features for the experiment is defined.

B. TOOLS AND RESOURCES

Gathering the useful characteristics for this investigation begins with the document processing via a series of linguistic analysis tools: CAEVO [26], AllenNLP [27] and Freeling [28].

CAEVO (*CAscading EVent Ordering system*) [26] generates a labeled XML document containing all the annotations, from which we are mainly interested in the *event phenomena*, not just the terms that the tool retrieves as *events*, but their semantic environment. Therefore, we store information about the events and their types, the temporal links between those events, the time expressions and their types.

CAEVO was designed to take into account the TimeML specification [29], whereby an *event* refers to something that *occurs* or *happens*, and can be articulated by different kinds of expressions such as verbs, nominalizations, or adjectives. In this sense, and depending on the context, *full* or *innocent* in *"Their pockets would always be **full**"* or *"You three are **innocent**"* would be events in the same way that more commonly accepted verbs such as *decide* or *call* are. In addition, the tool semantically classifies events into one of seven categories: aspectual; perception; state; reporting; intensional action; intensional state; and, occurrence. A short description and examples from the corpora have been included in the supplementary materials (see Appendix) in Table 8. Apart from event information, time expressions (*timex*) are also annotated and classified in one of these types: date (e.g., *the second of December*); time (e.g., *half past noon*); duration (e.g., *3 days*); and, set (e.g., *every two weeks*). More examples appear in Table 9, in Appendix.

We complete the linguistic analysis with Freeling [28] and AllenNLP-coref [27] taking as a base the work of [30].

[9] Considering grammatical elements and at the same time their correspondent phrases shall not be considered as reporting twice the same element, given that they refer to different linguistic phenomena. Some examples of noun phrases can illustrate such peculiarity. The most simple one could be, for instance, *"the child"*, but also *"the Theory of Relativity"*, *"an impressive device"* or *"the servant working in the front desk"*. All of them would act as noun phrases, showing the convenience of capturing both types of elements, nouns and noun phrases.

TABLE 2. Types of linguistic elements considered for computing the features.

Elements quantified	Description
Grammatical elements	Verbs, nouns, adjectives, adverbs, pronouns, numbers
Grammatical phrases	Verb, noun or proper noun phrases
Verb tenses	Present, past, future
Verb aspects	Infinitive, gerund, participle
Modal verbs	can, may, must, ought, shall, should, will, would
Predicative complement	Presence of a noun phrase completing the meaning of a linking verb
Constituents	Subject and object
Events	Aspectual, Occurrence, Perception, Reporting, State, Intensional action, Intensional state
Time related particles	Timex and time links
Named entities	Locations, persons, organizations and miscellaneous
Corefence features	Number of occurrences, distance among them, etc
Discourse Markers	Discourse markers from [31]
Semantic Phrases	Introductory phrases, exclamatory, interrogative, parentheses
Quotation	Presence of quotation marks
Length of named entities	Counting from 1 to 4 and then +5
Length of words	Counting from 1 to 6 and then +7

The information that we obtain from Freeling relates to multiple linguistic levels of the text. Among others, presence of quotation marks, specific grammatical elements and types of phrases or named entities are obtained by the analysis Freeling performs. Although the tool also performs coreference, we observed that the results from the AllenNLP model where more adequate and complete, apart from being more easily processable. Therefore, regarding coreference, chains of text that refer to the same entity are analyzed from those results to extract the appropriate features.

Additionally, a *lexicon of prototypical discourse markers* [31] is employed to identify discourse markers across the documents that could be subsequently used to provide an argumentative representation of the text.

C. FEATURE COMPUTATION

The linguistic tools provide us with the material to properly build the features of each document in order to generate the dataset that serves as basis of the ML exploration. From the linguistic elements identified by each tool, we proceed to compute a series of calculations. These operations are carried out over the elements previously presented in Table 2.

The different features are extracted from each document, considering a level of granularity sufficient to convey the composition of each sentence. In this manner, the feature calculation allows us not only to quantify characteristics relative to the whole document (e.g. How many events of the type *"occur"* can be found in document 22?), but also other aspects relative to the sentences as components of the text (e.g. What proportion of sentences includes time elements of the type "*Date*"?). Moreover, the feature calculation also

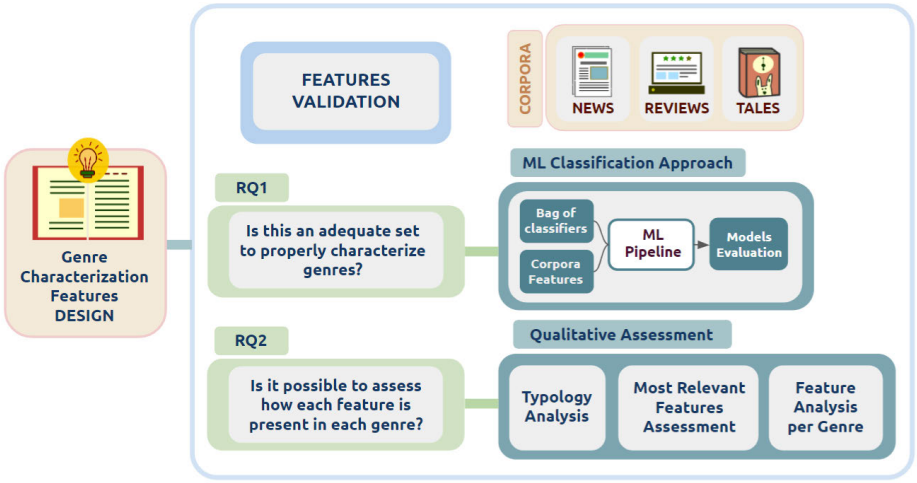

FIGURE 1. Outline of the complete approach, specifying the different strategies for addressing each of the research questions posed.

allows to quantify the composition of the sentence itself (e.g. What is the average number of pronouns per sentence?). Regarding personal pronouns and verb tenses, the predominant element within the document has been also computed, with the value of such features being categorical, unlike the rest, which are of a numerical nature.

In relation to the phenomenon of coreference, AllenNLP provides both the chains of elements that point to the same referent, and their positions. This information is processed to reflect five different measures: the number of coreference chains; the average length of the chains or how many times the element is referenced within the text; the mean spread of the coreference chains as the distance between the first and last reference; the number of chains reporting three or more occurrences; and finally, a measure of the concentration of the references, which points to the relation between the length of the chains and their spread and serves to express if an entity is present along the whole document or only within a small fragment.

A detailed description of the operations is presented in Table 2. Additionally, a complete relation of the whole set of features computed can be consulted in Table 10 (see Appendix).

V. EXPERIMENTS AND EVALUATION

To properly verify that the features we have engineered enable the identification of properties that are specific to each genre (RQ1) and to analyze how those features actually behave within each of the genres (RQ2), we designed a classification task which would help to answer those research questions by leveraging the diverse mechanisms and algorithms that ML provides.[10] Fig. 1 displays a graphical summary of the whole approach.

[10]We use scikit-learn [32] implementations throughout all the experimental stage, and seaborn [33] library to better visualize the results.

In this manner, we devise a workflow which includes a multiclass classification process for which each document would be represented by a linguistic-motivated feature vector as described in Section IV. To tackle RQ1, we hypothesize that the set of features we modeled could be considered valid if it is proved that using them to feed a number of classifiers of a diverse nature results in a good performance for all classifiers. The reasoning behind this is that the effectiveness of a classification model depends on both the algorithm selected and the features that feed the model. Section V-A describes the experimental setup, the decisions made and the results regarding RQ1.

As for the second research question posed, RQ2, firstly, we perform a qualitative analysis of the set of features grounded in the definition of certain typology, and secondly, we define a series of experiments exploiting several feature selection techniques. Additionally, to conduct this part of the research, adopting techniques to discretely analyze the contribution of the different features to each separate genre was of paramount importance. The extensive assessment of the features designed is presented in Section V-B.

A. FEATURE ASSESSMENT BY MODEL ESTIMATION

To determine if a genre can be sufficiently distinguished considering the set of features designed (RQ1), the strategy that we have adopted implies the evaluation of a collection of models. This procedure has several stages, since we want each model to be tuned with the adequate parameters, and to be evaluated accordingly.

In general terms, ML refers to the process by which an estimator is trained on a dataset with the objective of providing a model able to infer or predict certain knowledge from new data. In our case, the estimator is a classifier and the classification task involves predicting a genre out of three, i.e., news, tales, and reviews. Typically, the learning process starts with

TABLE 3. Performance of classification algorithms with default parameters. Scores for precision, recall, F-score and accuracy, ranked by the latter.

Model	Precision	Recall	F-score	Accuracy
Linear SVC	0.9956	0.9955	0.9955	0.9955
Random Forest	0.9941	0.9941	0.9941	0.9941
Linear Discriminant Analysis	0.9941	0.9940	0.9940	0.9941
SGD Classifier	0.9941	0.9940	0.9940	0.9941
Ridge	0.9941	0.9940	0.9940	0.9941
Multi Perceptron Classifier	0.9927	0.9926	0.9926	0.9926
SVC	0.9926	0.9926	0.9926	0.9926
Decision Tree Classifier	0.9869	0.9866	0.9866	0.9866
Bagging Classifier	0.9855	0.9851	0.9851	0.9851
AdaBoost Classifier	0.9824	0.9822	0.9822	0.9822
K-Neighbors Classifier	0.9778	0.9777	0.9777	0.9777
Bernoulli NB	0.9614	0.9614	0.9613	0.9614
NuSVC	0.9582	0.9569	0.9570	0.9569
Extra Tree Classifier	0.9514	0.9509	0.9508	0.9510
Quadratic Discriminant Analysis	0.9355	0.9288	0.9275	0.9287
Gaussian NB	0.9055	0.8948	0.8903	0.8945
Dummy	0.3358	0.3358	0.3358	0.3358

the division of the dataset into two subsets, one used for the algorithm in order to learn the appropriate parameters and validate the training (i.e., the training dataset) and the other engaged in testing its generalization capacity (i.e., the test dataset). Two types of parameters are associated with each classifier so that certain parameters are learned during the training process while others need to be provided to the algorithm prior to the start of the process—hyperparameters—. Their optimal value depends on each task and can be settled after a fitting procedure which is known as hyperparameter-tuning. In this regard, although a model can be used directly with default parameters, a more reliable and stable performance is expected after performing the fine-tuning [34].

A successful model is one that can generalize to an independent dataset of new examples that the model has not seen during training. This capacity determines its predictive power. To assess the goodness of such a model, different measures can be taken into account, such as accuracy, F-score or AUC. However, whatever this measure is, specific techniques are required to handle the stochastic nature of ML algorithms, which explains why running the algorithm several times is going to produce different results. A widely accepted technique for dealing with this issue is cross-validation (CV) in any of the different modalities. Cross-validation also prevents overfitting, which is what happens to the model when what it learns is too close to the training data, causing a deterioration in generalization capacity. Cross-validation consists of training and evaluating the model repeatedly over different train/validation folds on the same data.

In the current experiments, accuracy is provided to evaluate the model. In order to compute a correct measure for the model, data imbalance and the multiclass scenario needs to be considered. As Table 1 shows, not all the classes are equally represented within the dataset used, so we apply oversampling to increase the volume of the under-represented classes [35]. The class with higher examples is reviews, so new tales and news examples are produced to balance the dataset by sampling with replacement. Furthermore, a macro-average strategy [36] has been adopted along the experimentation such that the final score results from the average of three independent accuracies, thereby giving equal weight to the classification of each label without favoring any particular class.

The first step of the experimental stage related to feature assessment by model estimation was to gather a bag of candidates. We selected 16 models by their popularity in text classification tasks. Scikit-learn provides several options to compute simple baselines, and we included a Dummy classifier which applied a stratified strategy that generates predictions respecting the class distribution in the training data. A preliminary screening was carried out over this initial bag to select a smaller set of models on which to perform the hyperparameter search. For this first stage, the examples were split into a typical train/test setup with 80% of the examples for training and the rest for test, but with both subsets containing equally balanced classes.

Table 3 shows the results for the maximal set of models we have chosen to perform the classification task, trained over the total set of features. Precision, recall and F-score are presented, together with the accuracy, which is the measure considered to apply the cutoff threshold. We observed that all results are quite high, but we still needed to make a deeper evaluation of the models to ensure that such results are stable in repeated experiments, also considering the most appropriate hyperparameters. Therefore, we worked next with those models that in this stage presented a performance accuracy above 0.99, resulting in 7 models that were thereafter tuned.

The best seven models (BMs) are fine-tuned and their performance measured through a more consistent method. Grid search of proper parameters is conducted considering

a 5-fold CV. Specifically, we adopt here a nested CV strategy, that implies double training the model to prevent overly-optimistic scores. In this manner, each model is first trained with CV to search the hyperparameters space (inner training), and then, once the parameters are selected, re-trained using again the CV strategy to provide the precise score for its performance (outer training). Results are included in Table 4, which reports accuracy scores for both stages of the training phase as well as scores for the test set. A slight decrease in performance can be observed between the two training phases, as expected, while high results are reported regarding the test evaluation. This was the performance over data that has not been seen by the model before.

TABLE 4. Mean accuracy for the best models after hyperparameter tuning (nested cross-validation) for the training and test sets.

Best Classifiers	Training Accuracy		Test Accuracy
	Inner score *grid search*	Outer score *cross-validation*	
Random Forest	0.9918	0.9907	0.9941
Ridge	0.9888	0.9888	0.9941
Linear Discriminant Analysis	0.9888	0.9885	0.9941
Multi Perceptron Classifier	0.9903	0.9874	0.9941
SGD Classifier	0.9911	0.9870	0.9941
Linear SVC	0.9818	0.9781	0.9851
SVC	0.9766	0.9770	0.9822

The results for all models, macro-averaged as before, indicate their high predictive power. Rather than comparing the performance of different models or discussing the best ones, our objective was to demonstrate that having optimally adjusted a series of different classifiers, regardless of the type considered, our features effectively help to generate noteworthy models. From that, we can conclude that those features convey relevant traits that relate to the genre to which the texts belong.

In order to answer the RQ1, it was stated that for a set of features to be considered sufficiently descriptive, feeding them into different fine-tuned models would yield an adequate evaluation of such models as classifiers. The results show an outstanding performance of the models, thereby demonstrating that the modeled collection of features is sound enough to answer RQ1.

B. QUALITATIVE ASSESSMENT OF THE FEATURES

After having confirmed that the collection of features enables an adequate distinction of the three genres and once a series of models were obtained, i.e. BMs, able to properly identify a document's genre from its linguistic features, we focused on RQ2: *Can the presence of each of these features be quantified and evaluated in different genres to identify the specificity of each one?* An adequate answer demands a detailed analysis of the features in relation to the different genres, delving into how these features are present in each of them.

To conduct a comprehensive analysis of the collections of features, several paths were followed. First, a feature typology was elaborated and used to categorize the individual features accordingly. The models' performance was assessed when fed by the resulting feature sets. Section V-B1 describes the typology and the model assessments. Afterwards, Section V-B2 explains how feature selection techniques were leveraged to detect and rank the subset of features found to be most relevant for the genre classification task. To make the ML algorithms more explainable, some methods have emerged that attempt to shed light on how different features affect the classification of each individual example. The third and final part of our study exploits one of these strategies to review each genre separately. Therefore, Section V-B3 provides a detailed analysis of the features prominent for each specific genre, discussing also how such features are present within the other genres.

1) FEATURE ASSESSMENT BY PROPOSED TYPOLOGY

To better analyze the behavior of the features in relation to the genres, a typology of features was defined so as to classify them considering three different criteria: their calculation complexity, their semantic load, and whether they are an expression of the discourse structure. According to the first criterion, i.e. considering complexity in the calculation, a feature could be considered as *basic* or *derivative* with the former one requiring only the quantification of its presence while the second one implying additional operations to obtain, for instance, ratios, averages or proportions.

Considering the second criterion, i.e. *semanticness*, each feature was categorized according to its semantic load. Following this, features representing the length of a word, for example, would be considered as *non-semantic* in opposition to features related to named entities or coreference traits, labeled as *semantic*.

A third criterion is selected that enables a distinction between *structural* and *non-structural* features. Consequently, characteristics that indicate, for example, the proportion of sentences in the document that include certain elements would be considered *structural*, as opposed to characteristics that mention a quantification of certain elements in the entire text, which would be tagged as *non-structural*.

Table 5 summarizes these criteria, including some examples, while Table 10 in Appendix, indicates the precise label for each of the 153 features regarding the three aspects considered, namely complexity of calculation, *semanticness,* and relation to the discourse structure.

Following such criteria, 13 subsets of features were created and studied to better understand their importance as generators of meaning related to each genre. We have first considered each group by itself (basic features, derived features, ...) and then different combinations such as *Basic+semantic* or *Semantic+Structural*. The complete list of feature sets is presented in Table 6.

To understand the importance of each type of feature, the BMs were retrained at this time feeding them with the different subsets just described. The process was repeated under the same conditions detailed in Section V-A and thus, the models were evaluated using CVs as before, over the

TABLE 5. Description of criteria applied to define sets of features, with examples.

Criterion	Name and Description	Example
Complexity in the calculation	Basic Only counts required	Number of adjectives per document
	Derived measure: Distribution, Proportion, Averaged measures	Proportion of quoted sentences per document
Semanticness	With semantic weight	Number of named entities in the document
	No semantic weight	Number of words in the document with length N
Related to the discourse structure	Structural	Proportion of sentences with at least one proper noun
	Non-structural	Number of events in the document

TABLE 6. Groups of features classified by type, including their description.

Type	Description
Primary groups	Basic Derived Non-semantic Semantic Structural Non-structural
Combined groups *Features within each group belonging to both categories*	Basic + Semantic Basic + Non-semantic Derived + Semantic Structural + semantic Structural + Non-semantic Non-structural + semantic Derived + Sentence-related

test set. The accuracies resulting from such testing are shown in Fig. 2, where each subset is tagged, including the number of features that comprise it.

The figure includes the BMs and the feature groups from the typology and their respective accuracies derived from feeding each model with each set of features. Several key findings can be summarized. Although the results in general are quite good, with practically all sets of features yielding accuracies greater than 0.9, an exception can be observed when using the feature set that combines structural characteristics that are non-semantic. In this case, the results show values below 0.9, which may be due to the fact that the number of features included is very small (19 features). However, when observing the performance of the systems with the derived structural features, although the number of characteristics is slightly lower (17 features), the accuracies improve, which would indicate that this small set is especially suitable for distinguishing the genres, producing better models than some combinations that include more characteristics. Finally, we noted that most of the combinations that explicitly discard non-semantic features score worse than combinations that include them or include both, semantic and non-semantic features. Features groups of this type mainly occupy the lower part of the table, with a variable number of features.

At this point in the research, we could not discern whether these types of features are prominent in any specific genre. However, we could conclude, in light of these figures, that a thorough characterization of the genres was achieved, and this is indicated by the feature groups at the top of the figure, which are more comprehensive than those at the bottom as they include a mixed collection of features. Thus, a good characterization of genres would need to include features of different types.

2) MOST RELEVANT FEATURES EXTRACTION

In statistics and machine learning, *feature selection* refers to the process by which a set of features is selected with the aim of improving the performance of a defined model. It is usually applied to reduce overfitting by removing redundant features. The selection can be performed manually, defining certain criterion to conduct the screening, or automatically applying ML strategies. In what follows, we explain how feature selection strategies have been applied in this research to determine and analyze which features are most valuable for distinguishing the three genres studied.

When performing feature selection, the employed method usually belongs to one of these types: filter, wrapped or embedded methods, the objective of all them being the removal of non-relevant features to optimize the classifier performance [37]. Filter methods apply statistical tests over the data to extract the features with a strongest relationship with the label assigned. They are independent from the use of any classifier, as opposed to the other methods, which make use of estimators to perform their selection. In this manner, neither *wrapper* or *embedded methods* are model agnostic. These tend to provide better subsets of features, despite involving a greater processing burden. *Wrapper* approaches, such as as Recursive Feature Elimination (RFE) [38], run a recursive procedure to remove the irrelevant features and require a base estimator for which the resultant set of features is optimal. The *embedded* strategy actually embeds the selection within the learning process and provides an *importance score* that helps to build the ranking of the features.

We conducted our study to elucidate what are the most relevant features by considering a total of 16 alternative

FIGURE 2. Cross-validated accuracy for the 13 features groups.

methods.[11] First, Chi-square, mutual information and ANOVA were employed as filter methods. Next, five common estimators were selected to perform RFE as wrapper methods, namely Decision Tree, Random Forest, Linear Regression, Perceptron and Support Vector Classifier, and we also perform a permutation technique usually referred to as *permutation importance* or *Mean Decrease Accuracy* over the Random Forest classifier. Finally, among the embedded models, which intrinsically attribute an importance score to each feature, we selected for this research two widely known methods, Lasso (*Least absolute shrinkage and selection operator*) and XGBoost (*Extreme Gradient Boosting*). Additionally, given that five of the BMs evaluated in the Section V-A belong to this *embedded* category, thereby providing feature importance, we have also included them within this step. All these methods, along with their specific types, are listed in Table 7.

All the methods adopted present benefits and drawbacks, either related to their efficiency, their accuracy results or, in some cases, their *ad-hoc* nature. To overcome their weaknesses and harness their strengths, we propose a strategy that aggregates their results as an heterogeneous ensemble,

[11] While it is not within the scope of this paper to discuss the specifics of such methods, detailed descriptions can be found in [37] and [39]

TABLE 7. Types of feature selection strategies applied over the general feature set, including the specific method or estimator. * indicates that the extraction of features proceeds from the BMs (see Section V-A).

Type	Method
Filter Selection	Chi square Mutual information ANOVA
	Base Estimator
Wrapper Selection	Decission Tree Random forest Linear Regression Perceptron Support Vector Classification Permutation Importance (RF)
	Model
Embedded Selection	Lasso XGBoost Linear Discriminant Analysis* Linear Support Vector* Random Forest* Ridge* Stochastic Gradient Descent*

drawing inspiration from [40], whose work introduces the concept of *ensemble feature selection* taking as basis the principle underlying the *ensemble learning* idea. This idea states that the results produced by an appropriate combination

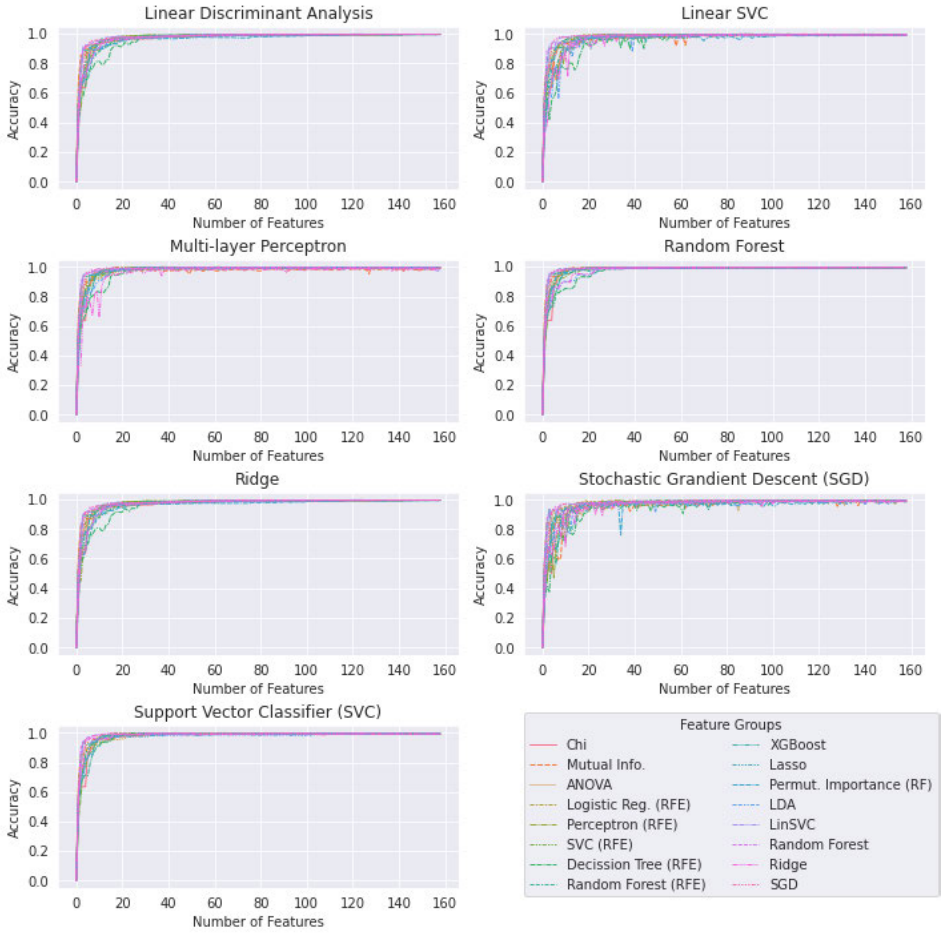

FIGURE 3. Accuracy for the different models over an incremental set of features, ranked by their importance computed from the different ML strategies. The set of features is gradually incremented from 1 to 153 features.

of models are better than any singular model. Therefore, the combination of good enough feature selection methods—each of them different, and hence, *heterogeneous*—should lead to the completion of the best set of features, *best* as enabler of the effective distinction between genres.

Creating the ensemble of features from the different methods requires of three steps: i) determining the optimal number of features per method, ii) creating the final set (*ML Selected*) gathering those top features and iii) recalculating their importance score to establish a new ranking within the *ML Selected* group.

- *Which is the optimal number of features for each method?* In order to address the first task and identify the appropriate number of features in accordance with each method, a specific algorithm was implemented. It assumes that for a given feature selection method *fsm*, each feature f has received an importance score that allows to build an ordered set F as $\{f_0, .., f_{n-1}\}$. This set contains the total number of features n in descending order on the basis of that rating. Therefore, given a model m and an ordered F, the prediction accuracy can be computed n times for the model and a subset of F, starting with the most relevant feature f_0 and including one more feature in each iteration. Subsequently, a set of accuracies $A_m = \{a_0{}^{\mathrm{m}}, .., a_{\mathrm{n}}{}^{\mathrm{m}}\}$ is calculated with $a_{\mathrm{j}}{}^{\mathrm{m}}$ being the accuracy of the model m computed with the j first features of F. Since features are included by importance, accuracy growth is expected to be the fastest for the set of features.
We compute such a process for the 16 feature selection methods combined with the BMs resulting from Section V-A. Fig. 3 shows accuracy measures for the combinations mentioned. Notably, the accuracy increased significantly when the first features of the set, ranked by importance, were input. Thereafter, at a certain point accuracy stabilized.
For a given feature method, seven A sets have been computed and it is possible to calculate a new set of mean accuracies $A_{\mathrm{fsm}} = \{a_{m0}, .., a_{mn}\}$ related to such *fsm* method with a_{mj} as the *m*ean of the a_j accuracies calculated for each model in *BM*. Equation 1 reflects this part of the process:

$$A_{\mathrm{fsm}}[j] = a_{mj} = \frac{\sum_{i \in BM} a_{\mathrm{j}}{}^{\mathrm{i}}}{\mid BM \mid} \tag{1}$$

In order to extract the optimal number of relevant features from a ranking F determined by a given *fsm*, we need to find out how many features are needed for the accuracy mean to stabilize. In this way, being w_k a window over A_{fsm} that contains 5 consecutive elements $\{a_{k-4}, ..a_k\}$, we will consider that the number of features s from which the accuracy is stabilized is such that the standard deviation *std* of w_s is lower than a certain threshold. The *std* provides a measure of the spread of a set of values with respect to their mean. Therefore, we can expect that when the accuracy stabilizes, the differences between consecutive values of accuracy become very small, resulting in a very low *std* value, thus indicating that there is hardly any variation. Fig. 4 shows the values for the sequence of standard variations related to each *fsm*.

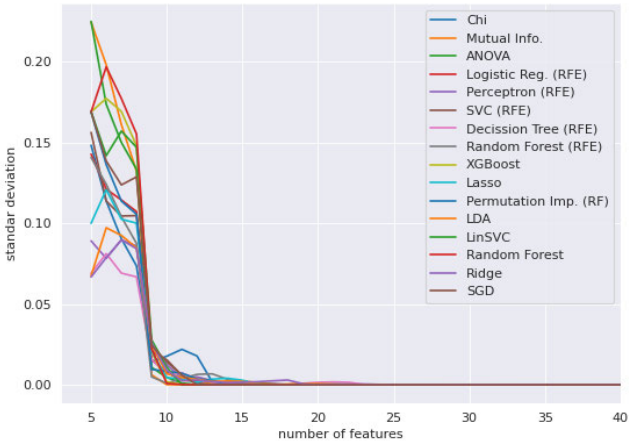

FIGURE 4. Standard deviation computed for consecutive accuracy values.

To ensure that all interesting features would be included in the final set, we empirically established 0.003 as the value for the threshold. Next, we computed the cut-off value s for all the *fsm*. In this case, s refers to the number of features to be extracted, therefore we retrieved s features from each F selecting first those with the highest importance score. The quantity of features for each of the 16 resultant groups F_{sg} ranges between 16 and 34 features each.

- *What are the features that comprise the ML Selected group?*
  The *ML Selected* group was then created from the union of such F_{sg} groups, and aggregated 95 different features. Before this procedure, each *fsm* provided an ordered F. Next, we designed an algorithm to rank the features within the *ML Selected* group.
- *How is the ranking determined among the ML Selected features?*
  Two scores were assigned to each feature belonging to the *ML Selected* group. First, a *votes* score v that indicates the number of F_{sg} groups in which the feature appears and a second one, *importance mean im*, which results from averaging the importance scores computed for this feature by the *fsm* that generated the F_{sg} to which it belongs. With this information, an *importance coefficient ic* was calculated for each selected feature as the product of such values, i.e. $ic = v * im$.

The *ML Selected* group is examined from a dual perspective. First, we analyzed it considering the typology defined in Section V-B1. An examination of the group configuration was carried out considering different stages according to the number of features included, so that the composition of the 10 most relevant characteristics could be reviewed, next the composition of the 30 most relevant ones, next the 60 most relevant, and finally the whole set of 95 features. The results are shown in Fig. 5.

The configuration of the top 10 features is mostly balanced between *basic/derived* features and *non-semantic/semantic* ones. However, although in the first stage 6 out of 10 features are *derived*, this ratio changes, and the difference between both types, *basic/derived*, keeps increasing with the number of features considered. Nonetheless, in the final group of 95, almost all the *derived* features have been included (30 out of 34), indicating the relevance of this feature type. Regarding the semanticness, although *non-semantic* features are more abundant in every stage, a balance is perceived and the difference between them never exceeds ten features, which does not occur with the other types. As for the discourse-related type—features related to the structure and composition of the text—, from the very beginning, the *structural* features appear minimal, notwithstanding that half of the total *structural* features are considered relevant and are finally included in the *ML Selected* group. In line with the results of the general analysis performed in Section V-B1, all types of features need to be included to obtain a fully representative set that is capable of conveying the peculiarity of different genres.

Taking into account the resultant feature ranking computed by the designed algorithm, a second analysis of the thirty most important features of the *ML Selected* group was conducted. Fig. 6 includes these features displaying their importance coefficient. A first glance at the figure shows that those features at the top of the list have a much higher coefficient of importance than the rest. It is to be expected that these characteristics are predominant in one of the genres, as their importance derives from the capabilities they provide the models for correctly discriminating and classifying documents. A connection could be suggested *a priori* between the features and the genres to which they are related. For instance, parenthesis may be more common in reviews, and maybe proper nouns in news or quotation marks in dialogues within tales. Nevertheless, this type of analysis would be limited to our interpretations, underscoring the importance of a more objective and deeper approach.

As set out previously, the intention of this work is to leverage the *ML Selected* group of features so that they can be exploited as more than a mere ingredient for an ML setup, and

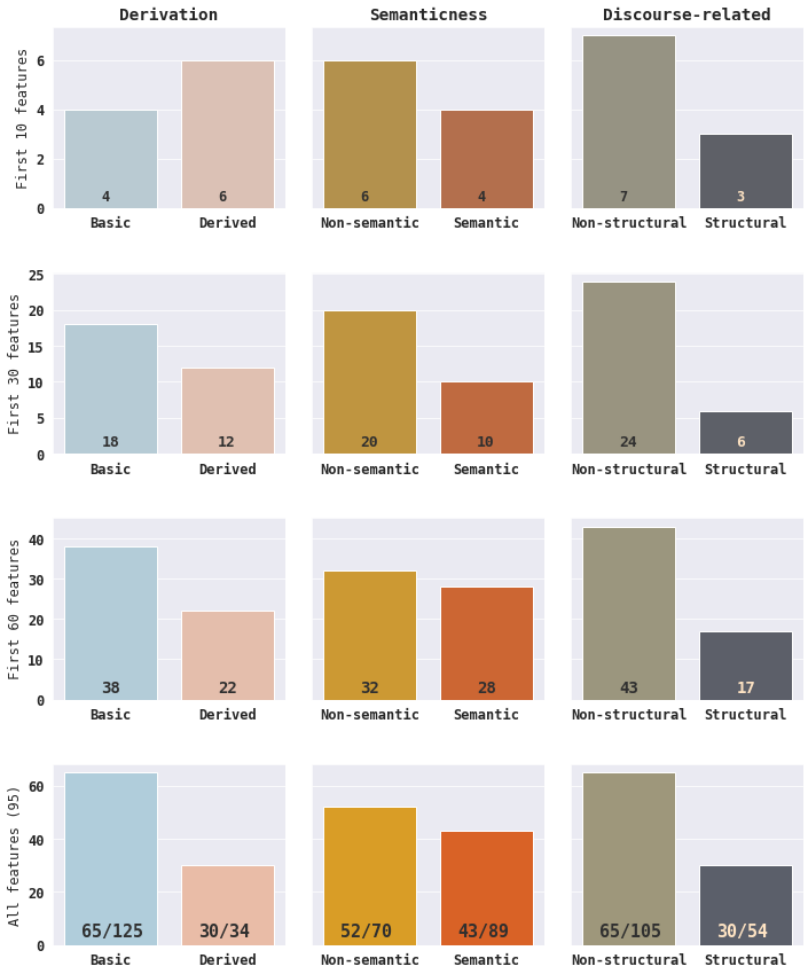

FIGURE 5. Analysis of the *ML Selected* type composition. Each row corresponds to the inner composition of the first 10, 30, 60 and 95 most relevant *ML Selected* features, considering the triple criteria introduced in Section V-B1.

can provide a description of the feature set and specific and well-founded conclusions about each genre. For this purpose, a further analysis is performed considering the contribution of the relevant features for each concrete genre, once again harnessing ML strategies. Details and results are described in the next section.

3) FEATURE ANALYSIS PER GENRE

Artificial Intelligence and ML currently represent two powerful tools for progress in multiple areas of science and everyday life. In recent times, a necessity to understand how their complex algorithms reach conclusions has encouraged a shift towards developing techniques able to provide such insights, namely *Explainable ML* or *Augmented ML* [41], [42]. The creation of tools that can explain how features condition a model's decisions is one of the objectives of *Explainable ML*, which is why we have adopted such perspective in this study.

Among the several libraries devoted to improving model interpretability, we selected ELI5[12] (*Explain like I'm five*) a Python tool which explains the weights given to each feature and predictions made by scikit-learn models. ELI5 was used then to evaluate the *ML Selected* group of features against one of the models in the BMs set, in this case, the SGD model.

Fig. 7 displays the results provided by ELI5. Each column corresponds to a genre and the weights therein show how much a feature has contributed to the classification of the genre. Their absolute values indicate the relevance of the feature, either in a positive or a negative way. A thorough study of such information was conducted also considering previous theoretical knowledge regarding the nature and composition of the genres involved. Subsequently, each genre has been discussed separately.

- **News**
  According to the results ELI5 has elaborated for this first genre, the characteristics which help most to identify any text as a piece of news are both proper nouns and the named entities mentioned in news. This is reflected in Fig. 7 with the high and positive weight reported for the frequency of proper nouns per sentence in news documents (*NNPXsent*), as well as with the total number

[12]https://eli5.readthedocs.io/en/latest/

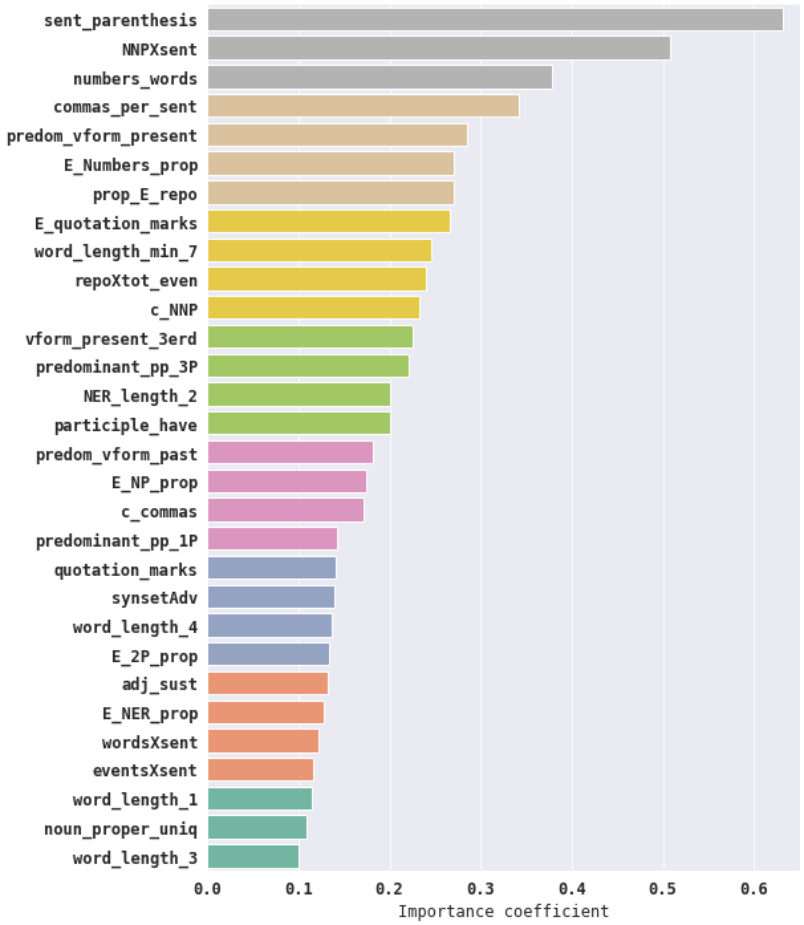

FIGURE 6. The 30 most relevant features of the 95 selected by the different ML strategies, after sorting them by their *importance coefficient*, which results from a combination of the *votes*—i.e. number of methods that selected the feature—, and the averaged importance score also retrieved from such methods.

of proper nouns per document (*c_NNP*). Moreover, this is also confirmed by the high weight assigned to named entities made up of two words (*NER_length_2*), which include proper nouns, organizations, locations and other types.

The high frequency that prevails in these features coincides with Van Dijk's linguistic theory [43] where he stated that news prototypical structure followed a 'top-to-bottom' pattern. Considering this pattern, we observed that the most relevant information is introduced at the beginning of the text, and when all the important elements have been mentioned, then the content goes back again to pieces of information already introduced, in order to add further information of less importance, such as comments, secondary details, etc. Therefore, proper nouns are a fundamental feature of this genre, as they lead the introduction and development of the information by means of their repetition so as to add new excerpts of the events, always recognizing what or whom we are referring to at any moment. This structure is directly related to another feature which also favors the identification of the news genre, and that is the number of sentences containing time expressions (*E_c_timexes*), because if we allude to the same events along the piece of news, we need to make use of several time expressions in order to locate the events in the temporal context and thus arrange the timeline of those actions.

In the same vein, the results shows that two other features help to identify this genre. First, is the number of sentences containing quotation marks (*E_quotation_marks*). Quotation helps to distinguishes between the narrator level and the statements of the people involved in the event by using the first person with direct speech. According to Van Dijk's, this type of quote-based narrative is one of the linguistic mechanism journalist use in order to provide a plausibly disinterested piece of news [43]. Second, we also find the total number of words which contain at least 7 letters (*word_length_min_7*), which shows this genre's preference for the use of terminology from each of the linguistic fields related to the facts, by means of

News - Top Features		**Reviews** - Top Features		**Tales** - Top Features	
Weight	**Feature**	**Weight**	**Feature**	**Weight**	**Feature**
+372.350	NNPXsent	+854.513	sent_parenthesis	+448.416	commas_per_sent
+308.949	word_length_min_7	+231.894	chain_spread_mean	+292.598	c_commas
+273.913	E_quotation_marks	+225.278	synsetAdv	+244.194	pers_pronouns_3P
+224.283	E_c_timexes	+221.333	vform_present_3erd	+241.900	word_length_1
+184.150	NER_length_2	+204.424	DiscM	+211.567	eventsXsent
+183.168	B-MISC	+162.876	VPXsent	+207.849	sent_exclamative
+178.338	c_NNP	+150.824	numbers_words	+181.588	mean_chain_len
+131.159	prop_E_event	+142.560	sent_interrogative	+181.472	occuXtot_even
... 36 more positive ...		+141.914	predicative_complement	+175.812	sent_interrogative_prop
... 41 more negative ...		*... 33 more positive ...*		+150.434	prop_E_perc
-122.831	NER_length_3	*... 38 more negative ...*		*... 40 more positive ...*	
-124.440	vform_present_3erd	-124.594	SIMULTANEOUS	*... 35 more negative ...*	
-132.748	predominant_pers_pron_2P	-147.764	E_REPORTING	-143.941	numbers_words
-135.061	VPXsent	-155.203	quotation_marks	-146.797	sent_interrogative
-152.456	E_2P_prop	-173.945	E_c_timexes	-163.130	B-ORG
-176.657	DiscM	-180.134	NNPXsent	-190.228	NER_length_2
-180.518	mean_chain_len	-184.959	entity_concentration_mean	-213.265	c_NNP
-191.314	adj_nouns	-226.366	participle_have	-252.347	vform_gerund
-202.978	word_length_4	-228.316	c_commas	-254.757	B-MISC
-217.819	sent_exclamative	-235.702	pers_pronouns_3P	-327.484	word_length_min_7
-297.071	synsetAdv	-350.693	commas_per_sent	-365.343	NNPXsent
-618.319	sent_parenthesis	-464.806	E_quotation_marks	-424.998	sent_parenthesis

FIGURE 7. **Feature contribution analysis performed by ELI5 for the classification of different textual genres (SGD model evaluated).**

the creation of nominalizations [43], mainly due to the formal register of news.

In the same way that the profuse presence of certain features in the documents helps to better classify the texts as news, the scarcity of others also affects the identification of the genre. This shortage is expressed by the low numerical value that such features present, which appears in the ELI5 summary indicated by the negative weights.

According to the results, one of the features whose small value seems to help to the classification of a text as news is the number of words with 4 letters that are included along the documents (*word_length_4*). This might occur, as we have previously mentioned, because longer words benefit the identification of terminology linked to the formal register of news and can convey more precise meaning, whereas 4-letter words may belong to any type of register, being more frequent in general speech than inside the terminology used in specialized fields.

Likewise, another two features to consider are the proportion of adjectives to the number of nouns (*adj_nouns*) and the total number of adverbs (*synsetAdv*) used in the documents, which appear to be more related to the reviews and they will be analyzed and better explained within that context. Similarly, the feature that accounts for the number of sentences in parenthesis in each text (*sent_parenthesis*) becomes the feature with the least weight when classifying texts as news because, as will be shown hereafter, a high frequency of this feature, i.e., a high number of parenthesis within the document, helps the model to classify and define a document as a review.

The overall trend of news texts is to make a greater use of nominalizations and noun phrases above every other grammatical category, since they convey a higher informational density [44]. This notion consistently reflects the importance of longer words in these texts, as aforementioned and also explains the low ratio of verbal phrases per sentence (*VPXsent*).

- **Reviews**

Reviews share personal experiences of consuming products or services. As for online reviews, the author tends to be non-professional and addresses their audience as peers [45]. This communicative action entails a high degree of personal involvement compared to other genres such as news, which is typically objective. The feature that was found to be most useful to help the model to identify the text as a review is the presence of sentences in parenthesis within a text (*sent_parenthesis*). While a high value of the feature is indicative of a review, the converse—a low value—will enable the model to classify the text as an alternative genre. Indeed, the result of this feature was negative in news and tales, but nonetheless proved to be a determining feature for the classification of these text genres. Hence, the use of parenthesis in reviews is a informal means by which the author can add brief snippets of information about the product (e.g. *'(800RMB / 8 hours / overtime ok)'*), address peers directly and informally (e.g. *'(Did I mention that this film is a disaster?)'*), or express emotion or mental states (e.g. *'(What the heck?!)'*). This finding is consistent with Vasquez's analysis of the resources deployed by users in review texts [46]. Moreover, question marks have been identified as a relevant rhetorical element for transmitting emotion [47], [48]. This is

indeed consistent with ELI5 findings regarding interrogative sentences (*sent_interrogative*), which appear to be among the most relevant features for classifying reviews. The results show a positive weight, i.e. a high numerical value both for the feature predicative complements (*predicative_complement*) and for the average number of verb phrases per sentence in this genre (*VPXsent*).

This may be due to the very nature of the reviews and the motivation of the reviewer, who voluntarily writes these comments to share experiences and help others, either by criticizing or praising the product. In any case, the user normally tries to include as much information as possible, both in the evaluation part (*'the laptop is incredibly light'*), and in the product description (*'the room was not insulated'*). In this context, an extensive use of the predicative complements seems adequate, and if the user has a lot of information to include, their density will be high. Moreover, another characteristic with a high position in the table for reviews is the number of adverbs in each document of such genre (*synsetAdv*). Adverbs have the capacity to include physical and temporal contexts [49], so as to complete the description of each product, its functioning, the circumstances in which it is used, the environment or other matters. Another way to enrich the description of the product relies on the use of figures to precisely detail certain characteristics (e.g. distance or time to the city center, start-up speed of a system, number of product features…). Consistently, ELI5 detects as a positive distinguishing feature the high value of the ratio between numbers and words (*numbers_words*).

Next, we highlight those features which, due to their low presence, help the model to identify documents as reviews. The results indicate that documents showing a low number of quotation marks, either considering the amount per document (*quotation_marks*) or the number of sentences containing quotation marks (*E_quotation_marks*), tend to be classified as reviews, in contrast to news, where this feature has a high value. Conversely, this feature does not appear in the relevant list. At times, tales can include dialogue comprised of quoted sentences but sometimes this type of dialog is not present and, thus, no quotations are used. Therefore, the value of the feature varies from one document to another, and this variation could be the reason why, for tales, quotation marks related features do not assist the classification.

Absence of reporting verbs appears to be a remarkable feature in the results, conveyed by the low value of the feature that indicates the detection of events of this nature (*E_REPORTING*). This finding is consistent with the lack of quotation marks mentioned before. Moreover, this genre's communicative purpose is ''rating a product or service'' to help the Internet community to check those opinions [50]. Hence, other types of verbs are prominent, particularly those which help to rate or describe users' perception with respect to the product at issue.

The reviews relate personal experiences told from the author's point of view with the aim of expressing his opinion regarding a certain object. Thus, although the third person is used to describe the product under consideration, the number of personal pronouns in third person per document (*pers_pronouns_3P*) is quite low, and appears as negative, in contrast to what happens with the children's tales, which will be dealt with next.

Lastly, the number of commas per sentence (*commas_per_sent*) and the number of commas per document (*c_commas*) also affects the classification, with a negative weight for both. Despite the length of some reviews, most of them typically use a kind of hybrid colloquial text that mixes oral and written text. This is reflected by the omission of apostrophes and punctuation marks [51]. Thus, users prefer to shorten sentences or either separate different events by means of other linguistic tools such as discourse markers, which moreover help to develop evaluative arguments, essential for this genre. The positive weight assigned to this feature confirms the relevant role of this element for reviews (*DiscM*).

- **Children's tales**

  In order to analyze the features which contribute the most to the identification of Children's tales, we need to take into consider that as children are the target audience, excerpts of information need to be simplified so that children can assimilate the events as they unfold. Hence, children's tales should show 'condensed style and brevity' [52]. This notions help us to analyze the results from the ELI5 algorithm, as the features with the highest weight in the table for tales are both the number of commas per sentence (*commas_per_sent*) and the total number of commas per document (*c_commas*). These grammatical elements help the division of the sentences, facilitating the progressive introduction information.

  Similarly, another feature to consider for the classification of this genre are the events that are introduced chronologically for the development of the story, which is considered suitable for children as especially very young children may have problems reconstructing the actual flow of events unless they are rendered chronologically' [53]. Thus, we can confirm the relevance of events in children's tales with the positive weights for both the number of events per sentence (*eventsXsent*) and the prevalence of *occurrence* events among other types of events (*occuXtot_even*)—see Table 8 in Appendix for event types definitions and examples.

  Lastly, on the analysis of findings related to features relevant due to their positive weight, when it comes to children's tales both the proportion of interrogative sentences in the corpus (*sent_interrogative_prop*) and the number of exclamatory sentences in the documents

(*sent_exclamative*) posses greater weight compared to other features. Arguably, the use of interrogative and exclamatory sentences are intended to emphasize emotion in the salient parts of the tale, considering that 'their main rhetorical purpose is first to entertain' [52].

Moving next to analyzing features with negative weights, the one with the highest negative weight is the number of sentences with parenthesis in the documents (*sent_parenthesis*). This is likely due to the fact that parenthesis are mainly used, as the previously mentioned literature affirms, to include even more information or to clarify something already mentioned, whereas in children's tales there is a clear preference for simplicity and concise language.

Furthermore, another feature to highlight for its limited presence in tales is the ratio of proper nouns per sentence (*NNPXsent*), directly connected also to the number of proper nouns within a document (*c_NNP*). This low presence in many children's tales is likely driven by the need for text simplification so that children can follow the story line without confusing them and without diverting attention from the events taking place. This characteristic contrasts with other types of texts, such as news, which need a clear identification of the elements, people and locations, involved in unfolding and real events. Therefore, children's tales show a preference for 'general words and concrete nouns belonging to children's environment' [54], so narratives are created with general terms rather than specific entities so as to not lose the child's attention. Accordingly, characters of the stories may be referenced by common names as *a forest wolf*, *the wonderful musician* or *a shoemaker*, whereas places are mentioned similarly, e.g. *a city*, *a kingdom*, etc. Not surprisingly, other features that convey information on the presence of named entities (*NER_length_2, B-ORG, B-MISC*) receive negative weights, implying that for a document to be qualified as a children's tale, the values of that characteristic have to be low.

Finally, comparing again this genre with news, another feature with a negative weight is the number of words with at least seven letters (*word_length_min_7*). This is an unsurprising result, given that children's tales include frequently used, simple vocabulary and quite informal language of a suitable level for their reading age to challenge and develop their linguistic and vocabulary skills [52].

This section presented a detailed study that analyzed each genre separately, together with the subset of features that were selected as most discriminating in Section V-B2, i.e. the *ML Selected* group. The findings indicate that the automatic processing discovered a set of features relevant for each genre, and the results are aligned with previous work that alludes to the functionality associated with each feature and its role within the genre involved. This step lays the foundation for including such features in NLP strategies and applications, and concludes the qualitative examination conducted to address RQ2, which questioned the possibility of characterizing each of the genres by using the set of features that we designed.

VI. FURTHER ISSUES FOR CONSIDERATION

Our work was developed under the acknowledged assumption that genres are defined by communicative purposes and genre types have a critical influence on the text's shape. Given this, genre-specific linguistic features are likely to be identifiable. The present study proves that it is possible to research, identify, analyze and explain these specific features for three genres, reviews, news and children's tales.

Despite the positive performance of our approach, and the validation of our initial research questions, there are some other issues that require development. First, although a genre is typically related to a communicative objective, and this is considered the dominant type, it is possible to detect several *stages* or *moves* within the text linked to different purposes [55]–[57]. For example, within a review, the main purpose of the author may be to evaluate, and the secondary one to describe the product or narrate the plot [45]. In the same way, a children's tale, which narrates to entertain, can also have a moral intention as happens in fables. The present work has not investigated a genre's granularity in terms of its communicative goal. This represents a merit worthy future project since the more precise the categorization of the communicative objective, the more versatile the features become.

A second concern is related to the assumption of the errors related to the different linguistic analysis tools. A posterior analysis of the tools' output indicated some mistakes when performing tasks such as disambiguation or coreference resolution. Therefore, sometimes the value attributed to a certain feature may not exactly reflect the properties of the text. It would be the case, for example, when a word that appears in capital letters is mislabeled as a named entity. There is no definitive solution for this issue. It is inherent to NLP tools due to the ambiguity and richness of the language. Tools evolve, algorithms and methods become more accurate. However, a constant evaluation of the tools' output needs to be performed not only to facilitate their tuning but to understand and measure the mistakes and their impact, and thus, adjust accordingly.

A third element that we have not addressed, yet interesting for this work, concerns the heterogeneity of the corpus and the form of its texts. Notwithstanding that heterogeneity is preferred because it may increase the classifier predictive power, and thus the robustness of the approach, a careful analysis of the texts brings to light important insights. In this manner, one facet of language whose effects are worthy to study is the temporal evolution of language, evidenced for instance by comparing more recently published children's tales versus the stories of Andersen or Grimm Brothers.

A different phenomenon is produced within the Reviews genre. Language in computer-mediated communication, also named *electronic discourse* [58] or simply *internet language* [59], has been increasingly recognized as a new

TABLE 8. Categories of events and examples.

Event type	Examples
Aspectual	Action that focuses on event history: initiation, culmination, continuation, ... *He lifted her into the carriage , but her feet **continued** to dance* *one of the boldest amongst the boys **began** to sing a song*
Perception	Action that involves physical perception *The girl can only **see** a small piece of field* *It was delightful to **hear** Great-grandfather tell about all this*
State	Circumstances in which something obtains true *They were **dark** as the blackness of night* *So he came home again and was **sad***
Reporting	Action to declare, narrate or inform about an event *"Tweet, tweet" **sang** the bird, as he flew out into the green woods* *"Thank you", **said** the woman*
Intensional Action	An intensional action introduces an event argument *It was very natural that he should **wish** to know* *I **knew** that there would be an awful storm*
Intensional State	Events referencing alternative worlds *I once **wanted** to fit out a ship to sail round the world* *It could **become** the pride of the garden and the joy of the family*
Occurrence	The other kind of events describing something that happens in the world *Then sometimes the public **applauded** too much* *One night a fierce tempest **broke** loose*

TABLE 9. Categories of time expressions and examples from corpora.

Timex type	Examples
Date	*Opposition supporters alleged the incident **Wednesday** night was politically motivated.* *Mahathir, who had heart bypass surgery in **1989**, had groomed Anwar, 51, as his successor.* *Yigit bought the bank at a public auction in **August** for dlrs 600 million.* *I love Ford's New Edge design concept, which traces it's roots back directly to the Ford GT-90 concept car of the **early nineties***
Time	*On **Friday evening**, about 1,500 mostly young people, came to the Goteborg Cathedral to try to assuage their grief and bewilderment at a memorial service.* *The vote was announced at **1:24 p.m.** EST, leaving the 42nd president to face trial in the Senate on whether he should be removed from office.* *Thus, **overnight**, Nico became the richest and most respected boobuan in the jungle.* *I was pretty disappointed with movie in general after watching Bad Boys II on **Wednesday night**.*
Duration	*The decline in exports will trim the growth of the American economy, and it was one factor in the Federal Reserve's decision to lower interest rates two times in **recent weeks**.* *For the first time in **decades**, Congress and the White House negotiated tax and spending legislation this year with the budget in surplus.* *And so his life went on, sadly and filled with loneliness, for many **centuries**.* *Ras raps on this track, telling the entire story for **nearly 8 minutes** straight.* *After I loaded all of my software (which took me **about 2 hours**) I started playing around with it.*
Set	*Annual inflation stood at 1 percent in October and 1.6 percent in November 1997, according to Eurostat's **monthly** report.* *We ought to hammer Gingrich **every day**.* *Albert's father was an extremely important and busy man who worked so many hours that he often had to work whole **weekends**.*

linguistic variety. Spontaneity, linguistic economy or the need of new devices to express emotions through written text, are several factors that affect the shape of language in the digital discourse [60]. User-generated content on reviews falls in this category, in which language usage tends to be informal and unconventional. This represent a challenge for the NLP field. Thus, although it has been demonstrated that the features designed in this work can appropriately identify the reviews, an improvement would be expected if the linguistic tools were able to adequately interpret the new word-formations developed for effectively communicate in space of virtual interactions. Some examples of this irregular yet flexible use of language would be the presence of abbreviations (e.g. *'lol'*, *'dunno'*, *'b4'* instead *'before'*) or

TABLE 10. Description of the features designed. It has been included the definition of each feature, together with its classification signature according to the criteria defined in Table 5, with *Comp.* standing for *Complexity in the calculation*, *Sem.* standing for *Semanticity* and *Disc.* referring to whether the feature is related with the structure of the discourse or not

Feat#	Feature ID	Der	Sem	Disc	Description
1	**c_events**	B	S	NST	Events in the document
2	**c_timexes**	B	S	NST	Timex particles in the document
3-9	**X-type of events**	B	S	NST	X type of event in the document
10-13	**X-type of times***	B	S	NST	Timex of type DATE in the document
14	**c_nsubj**	B	NS	NST	Nominal subject in the document
15	**c_csubj**	B	NS	NST	Clausular subject in the document
16	**c_xsubj**	B	NS	NST	Other type of subject in the document
17	**c_dobj**	B	NS	NST	Direct object in the document
18	**c_NP**	B	S	NST	Noun phrases in the document
19	**c_NNP**	B	S	NST	Propoer noun phrases in the document
20	**c_VP**	B	NS	NST	Verb phrases in the document
21	**E_c_events**	B	S	ST	Sentences containing at least one event
22	**E_c_timexes**	B	S	ST	Sentences containing at least one timex
23-29	**E_X-type of events***	B	S	ST	Sentences containing at least one event of type X
30-33	**E_X-type of timex***	B	S	ST	Sentences containing at least one timex of type X
34	**E_c_nsubj**	B	NS	ST	Sentences containing at least one nominal subject
35	**E_c_csubj**	B	NS	ST	Sentences containing at least one clausular subject
36	**E_c_xsubj**	B	NS	ST	Sentences containing at least one subject of any other type
37	**E_c_dobj**	B	NS	ST	Sentences containing at least one direct object
38	**E_c_NP**	B	S	ST	Sentences containing at least one noun phrase
39	**E_c_NNP**	B	S	ST	Sentences containing at least one proper noun phrase
40	**E_c_VP**	B	NS	ST	Sentences containing at least one verb phrase
41	**c_tlinks**	B	S	NST	Time links in the document
42-58	**X-type of time links**	B	S	NST	X type of time link in the document
59	**wordsXsent**	D	NS	NST	Average words per sentence
60	**eventsXsent**	D	S	NST	Average events per sentence
61	**timexXsent**	D	S	NST	Average timex per sentence
62	**NPXsent**	D	S	NST	Average noun phrases per sentence
63	**VPXsent**	D	NS	NST	Average verb phrases per sentence
64	**NNPXsent**	D	S	NST	Average proper noun phrases per sentence
65	**occuXtot_even**	D	S	NST	Distribution of occurrence events
66	**percXtot_even**	D	S	NST	Distribution of perception events
67	**repoXtot_even**	D	S	NST	Distribution of reporting events
68	**prop_E_event**	D	S	ST	Proportion of sentences with at least one event
69	**prop_E_timex**	D	S	ST	Proportion of sentences with at least one timex
70-72	**prop_E_X type of events**	D	S	ST	Proportion of sentences with at least one event of type X (occurrence, perception,reporting)
73	**chain_amount**	B	S	ST	Elements with coreference chain related
74	**mean_chain_len**	D	S	ST	Average length of the chains (how many times an entity is referenced)
75	**chain_spread_mean**	D	S	ST	Average spread of the coreference chains
76	**maximal_len_chains amount**	B	S	ST	Entities referenced three or more times along the text
77	**entity_concentration mean**	D	S	ST	Average of concetration, indicating the relation between the length and spread of the chains
78	**sent_exclamative**	B	S	ST	Exclamative sentences in the document
79	**sent_exclamative_prop**	D	S	ST	Proportion of exclamative sentences in the document
80	**sent_interrogative**	B	S	ST	Interrogative sentences in the document
81	**sent_interrogative_prop**	D	S	ST	Proportion of interrogative sentences in the document
82	**sent_parenthesis**	B	S	ST	Segments in parenthesis in the document
83	**predomi-nant_pers_pron**	B	NS	NST	Predominant personal pronoun in the document
84	**pers_pronouns_1P**	B	NS	NST	First-person pronouns in the document
85	**pers_pronouns_2P**	B	NS	NST	Second-person pronouns in the document
86	**pers_pronouns_3P**	B	NS	NST	Third-person pronouns in the document
87	**pers_pronouns_IT**	B	NS	NST	Pronoun “it’ ‘ in the document
88	**It_pronoun_per_sent**	D	NS	NST	Average pronoun “it’ ‘ per sentence
89	**E_1P**	B	NS	ST	Sentences containing at least one first-person pronoun
90	**E_1P_prop**	D	NS	ST	Proportion of sentences with at least one first-person pronoun
91	**E_2P**	B	NS	ST	Sentences containing at least one second-person pronoun
92	**E_2P_prop**	D	NS	ST	Proportion of sentences with at least one second-person pronoun
93	**NE**	B	S	NST	Named entities in the document
94	**E_NER**	B	S	ST	Sentences containing at least one named entity
95	**E_NER_prop**	D	S	ST	Proportion of sentences with at least one named entity
96	**NER_NP**	D	S	NST	Proportion of Named Entities vs Proper Nouns
97	**nouns**	B	NS	NST	Nouns in the document
98	**noun_proper**	B	S	NST	Proper nouns in the document
99	**noun_proper_uniq**	B	S	NST	Different proper nouns in the document
100	**NP_nouns**	D	S	NST	Proportion of Proper Nouns vs nouns
101	**E_NP**	B	S	ST	Sentences containing at least one proper noun
102	**E_NP_prop**	D	S	ST	Proportion of sentences with at least one proper noun
103-106	**X-type of NE**	B	S	NST	Named entities in the document
107-111	**NER_length_N**	B	S	NST	Named entities in the document with length N [1,2,3,4,5+]
112	**numbers**	B	NS	NST	Number expressions in the document
113	**E_Numbers**	B	NS	ST	Sentences containing at least one number expression

Continued on next page

B:Basic,D:Derived - S:Semantic,NS:Non-semantic - ST:Structural, NST:Non-structural

TABLE 10. *(Continued.)* **Description of the features designed. It has been included the definition of each feature, together with its classification signature according to the criteria defined in Table 5, with *Comp.* standing for *Complexity in the calculation*, *Sem.* standing for *Semanticity* and *Disc.* referring to whether the feature is related with the structure of the discourse or not**

Feat#	Feature ID	Der	Sem	Disc	Description
114	**E_Numbers_prop**	D	NS	ST	Proportion of sentences with at least one number expression
115	**numbers_words**	D	NS	NST	Proportion of numbers vs words
116	**synsetAdj**	B	NS	NST	Adjectives in the document
117	**E_adj**	B	NS	ST	Sentences containing at least one adjective
118	**E_adj_prop**	D	NS	ST	Proportion of sentences with at least one adjective
119	**adj_sent**	D	NS	NST	Average adjectives per sentence
120	**adj_nouns**	D	NS	NST	Proportion of adjectives vs to nouns
121	**adj_words**	D	NS	NST	Proportion of adjectives vs to words
122	**vform_future**	B	NS	NST	Future-tense verbs in the document
123	**vform_past**	B	NS	NST	Past-tense in the document
124	**vform_present**	B	NS	NST	Present-tense in the document
125	**vform_present_3erd**	B	NS	NST	Third person present-tense verbs in the document
126	**predominant_time**	B	NS	NST	Predominant verbal tense in the document
127	**vform_gerund**	B	NS	NST	Verbs in gerund form in the document
128	**vform_infinitive**	B	NS	NST	Verbs in infinitive form in the document
129	**vform_participle**	B	NS	NST	Verbs in participle form in the document
130	**verb_modal**	B	NS	NST	Modal verbs in the document
131	**participle_be**	B	NS	NST	Participles preceded by the modal to be
132	**participle_have**	B	NS	NST	Participles preceded by the modal to have
133	**predicative complem.**	B	NS	NST	Predicative complement in the document
134	**quotation_marks**	B	NS	NST	Quotation marks in the document
135	**E_quotation_marks**	B	NS	NST	Sentences containing at least one pair of quotation mark
136	**synsetAdv**	B	NS	NST	Adverbs in the document
137	**wh-adverb**	B	NS	NST	Wh-pronouns in the document (what,who)
138	**wh-pronoun**	B	NS	NST	Wh-adverbs in the document (where,why, whence,...)
139	**when-adverb**	B	NS	NST	When adverbs in the document
140	**c_commas**	B	NS	NST	Commas in the document
141	**commas_per_sent**	D	NS	NST	Average commas per sentence
142	**DiscM**	B	S	ST	Discourse markers in the document
143	**intr_ph**	B	NS	ST	Introductory phrases in the document
144	**intr_with_PN_NER**	B	NS	ST	Introductory phrases containing a proper noun or a named entity in the document
145	**intr_with_adverb**	B	NS	ST	Introductory phrases containing an adverb in the document
146	**intr_ph_prop**	D	NS	ST	Proportion of sentences with introductory phrase
147-153	**word_length_N**	B	NS	NST	Words in the document with length N [1,2,3,4,5,6,7+]

***Extension of types of events, timex, time links and named entities**

Events: Aspectual, Intensional Action, Intensional State, Occurrence, Perception, Reporting, State

Timex: Date, Time, Duration, Set

Time Links: Before, After, Ibefore, Iafter, Includes, Is included, Begins, Begun by, Ended by, Ends, Simultaneous, None, Vague, Unknown, Overlap, Before or overlap, Overlap or after

Named Entities: B-LOC, B-MISC, B-ORG, B-PER

B:Basic,D:Derived - S:Semantic,NS:Non-semantic - ST:Structural, NST:Non-structural

the employment of non-standard spellings and interjections (e.g. *'Noooooo!!!','Arghhh!!'*) together with textual emoticons (e.g. *'xD', ':-)'*) to express emotions.

VII. CONCLUSION AND FUTURE WORK

The analysis conducted in this paper was performed to broaden the comprehension of the genres as expressions of communicative objectives so that findings may be beneficial to the NLP field. Our efforts were focused on the construction of a set of linguistically inspired features that needed to be expressive enough to articulate what is specific to a given genre, as opposed to others. This robust set of features could be employed in NLP disciplines, both those that involve tasks which need to understand language to be successful, as well as those which aim to create text. Therefore, two research questions were raised whose answers would confirm whether the designed set of features could contain enough information to distinguish between several genres and also express the peculiarities of each one separately.

Based on the extant literature and observation, we built the set of features with a significant linguistic load, a total of 153 features containing lexical, grammatical, semantic information beyond the sentence level. A classification framework was set up to verify the feature suitability in terms of our main purpose and, subsequently, a corpus composed of news, stories and reviews in English was assembled specifically for the project. Next, we performed a twofold assessment of the features.

First, we fine-tuned a series of models to find out whether the set of features would succeed regardless of the classifier. Then, we analyzed the models performance in terms of accuracy, and all of them delivered outstanding results. Second, we conducted a quality assessment of the features at different levels. A typology was defined to ascertain which type of features were more influential for the classification. The findings suggest that a group of heterogeneous features provides a more accurate representation of the genre's peculiarities, presumably because a particular genre's distinctive nature is not determined by a unique linguistic level. Furthermore,

we leveraged feature selection techniques to analyze which features were more discriminant among the 153 designed. We ended up with a set of 95 features, not surprisingly comprised of all the types of features proposed previously. Lastly, we performed a specific analysis for each of the three genres, reviewing the behavior of the features that most contributed to distinguish each one.

The results highlight how our features are consistent with previous theoretical studies, and thereby endorse our strategy as a valuable framework to boost and enhance research that investigates the links between linguistic dimensions and discourse functions. Our contribution here is twofold: we provide a solid feature set that can be adjusted to any other genre; and also a methodology that enable the inclusion and evaluation of new features if needed.

Nonetheless, there is still plenty of room for improvement. A thorough examination of the features not selected for the *ML Selected* group, for example, would be advisable. The approach presented could be extended not only to more genres, but to different languages, assuming the existence of the required linguistic tools. Furthermore, a step beyond regarding events and their relation with agents and time would aid understanding of narrative and argumentation, maybe by including graphs in the process. These would be beneficial for tasks like question answering, textual entailment or story generation. From here, emerges the next challenge; to include these features within a generation framework and thus, investigate how to create text according to a specific communicative objective. The findings achieved in this work pave the way for what comes next.

APPENDIX

As additional information, we provide some examples of the type of events and the type of time expressions the tool CAEVO provides. Events appear in Table 8 whereas time expressions in Table 9. Besides, Table 10 includes a complete description of all the designed features, together with their classification according to the typology defined in Section V-B1.

REFERENCES

[1] R. Thomason, J. Hobbs, and J. Moore, "Communicative goals," in *Proc. Eur. Conf. Artif. Intell. Workshop Gaps Bridges, New Directions Planning Natural Language Gener.*, 1996, pp. 1–7.

[2] J. Swales, *Genre Analysis: English in Academic and Research Settings*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[3] L. Flowerdew, "Corpus-based discourse analysis," in *The Routledge Handbook of Discourse Analysis*, J. P. Gee and M. Handford, Eds. Evanston, IL, USA: Routledge, 2012, ch. 13, pp. 174–187.

[4] D. Biber and S. Conrad, *Register, Genre, Style*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[5] M. Wan, A. C. Fang, and C.-R. Huang, "The discriminativeness of internal syntactic representations in automatic genre classification," *J. Quant. Linguistics*, pp. 1–34, Sep. 2019, doi: 10.1080/09296174.2019.1663655.

[6] S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers, and A. Löser, "SECTOR: A neural model for coherent topic segmentation and classification," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 169–184, Nov. 2019.

[7] X. Lin, W. Jian, J. He, T. Wang, and W. Chu, "Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 41–52.

[8] R. Kruiper, J. Vincent, J. Chen-Burger, M. Desmulliez, and I. Konstas, "In Layman's terms: Semi-open relation extraction from scientific texts," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1489–1500.

[9] V. Propp, *Morphology of the folktale. Indiana University Research Center in Anthropology*, vol. 10. Dallas, TX, USA: Folklore, Linguistics, Publication, 1958.

[10] E. Laily Zen, "Corpus-driven analysis on the language of Children's literature," in *Proc. 1st Int. Conf. Recent Innov.*, 2018, pp. 17–22.

[11] K. R. Hansen, "News from the future: A corpus linguistic analysis of future-oriented, unreal and counterfactual news discourse," *Discourse Commun.*, vol. 10, no. 2, pp. 115–136, Apr. 2016.

[12] A. Thumvichit, C. Gampper, and J. van de Weijer, "Composing responses to negative hotel reviews: A genre analysis," *Cogent Arts Humanities*, vol. 6, no. 1, Jan. 2019, Art. no. 1629154.

[13] K.-I. Mavridou, A. Friedrich, M. P. Sørensen, A. Palmer, and M. Pinkal, "Linking discourse modes and situation entity types in a cross-linguistic corpus study," in *Proc. 1st Workshop Linking Comput. Models Lexical, Sentential Discourse-level Semantics*, 2015, pp. 12–21.

[14] D. Croce, G. Castellucci, and R. Basili, "GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2114–2119. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.191

[15] H. Chen and Y. Ji, "Learning variational word masks to improve the interpretability of neural text classifiers," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4236–4251.

[16] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in," in *Proc. Natural Legal Lang. Process. Workshop*, 2019, pp. 78–87. [Online]. Available: https://www.aclweb.org/anthology/W19-2209

[17] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[18] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 28–47, Feb. 2018.

[19] A. Lieungnapar, R. W. Todd, and W. Trakulkasemsuk, "Genre induction from a linguistic approach," *Indonesian J. Appl. Linguistics*, vol. 6, no. 2, pp. 319–329, 2017.

[20] M. R. Qureshi, S. Ranjan, R. Rajkumar, and K. Shah, "A simple approach to classify fictional and non-fictional genres," in *Proc. 2nd Workshop Storytelling*, 2019, pp. 81–89.

[21] P. Over, H. Dang, and D. Harman, "DUC in context," *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1506–1520, Nov. 2007.

[22] M. Taboada, C. Anthony, and K. D. Voll, "Methods for creating semantic orientation dictionaries," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2006, pp. 427–432.

[23] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, 2007, pp. 440–447.

[24] K. Ganesan and C. Zhai, "Opinion-based entity ranking," *Inf. Retr.*, vol. 15, no. 2, pp. 116–150, Apr. 2012.

[25] P. V. Lobo and D. M. De Matos, "Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, pp. 1472–1475.

[26] T. Cassidy, B. McDowell, N. Chambers, and S. Bethard, "An annotation framework for dense event ordering," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Short Papers*, vol. 2, Jun. 2014, pp. 501–506.

[27] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A deep semantic natural language processing platform," 2018, *arXiv:1803.07640*. [Online]. Available: http://arxiv.org/abs/1803.07640

[28] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proc. 8th Int. Conf. Lang. Resour. Eval.*, Istanbul, Turkey: European Language Resources Association, 2012, pp. 2473–2479.

[29] J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev, "TimeML: Robust specification of event and temporal expressions in text," *Directions Question Answering*, vol. 3, pp. 28–34, Jan. 2003.

[30] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–10.

[31] L. A. Alemany, "Representing discourse for automatic text summarization via shallow NLP techinques," Ph.D. dissertation, Universitat de Barcelona, Barcelona, Spain, 2005.
[32] F. Pedregosa, G. V. A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[33] M. Waskom and the Seaborn Development Team. (2020). *Seaborn*. [Online]. Available: https://doi.org/10.5281/zenodo.592845
[34] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-sklearn: Automatic hyperparameter configuration for scikit-learn," in *Proc. 13th Python Sci. Conf.*, 2014, p. 50.
[35] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," *Procedia Comput. Sci.*, vol. 159, pp. 736–745, 2019.
[36] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.
[37] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
[39] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, p. 94, 2016.
[40] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: Homogeneous and heterogeneous approaches," *Knowl.-Based Syst.*, vol. 118, pp. 124–139, Feb. 2017.
[41] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, 2016, pp. 1135–1144.
[42] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
[43] T. A. V. Dijk, *News as Discourse*. Trenton, NJ, USA: Lawrence Erlbaum Associates, 1988, pp. 43–44.
[44] D. Biber, *Dimensions Register Variation: A Cross-Linguistic Comparison*. Cambridge, U.K.: Cambridge Univ. Press, 1995.
[45] M. Taboada, "Stages in an online review genre," *Text Talk - Interdiscipl. J. Lang., Discourse Commun. Stud.*, vol. 31, no. 2, pp. 247–269, Jan. 2011.
[46] C. Vásquez, "Narrativity and involvement in online consumer reviews: The case of Tripadvisor," *Narrative Inquiry*, vol. 22, no. 1, pp. 105–201, 2012.
[47] S. Banerjee, A. Y. K. Chua, and J.-J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proc. 9th Int. Conf. Ubiquitous Inf. Manage. Commun.*, Jan. 2015, pp. 1–7.
[48] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.
[49] D. Biber, *Variation Across Speech Writing*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[50] C. Vásquez, "Intertextuality and interdiscursivity in online consumer reviews," in *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age*. Evanston, IL, USA: Routledge, 2014, pp. 66–80.
[51] I. Pollach, "Electronic word-of-mouth: A genre approach to consumer communities," *Int. J. Web-Based Communities*, vol. 4, no. 3, pp. 284–301, 2008.
[52] A. Guijarro and J. Moya, "Thematic and topical structuring in three subgenres. A contrastive study," *Miscelánea, A J. English Amer. Studies*, vol. 27, no. 27, pp. 131–154, 2003.
[53] M. Nikolajeva, "Beyond the grammar of story, or how can Children's literature criticism benefit from narrative theory?" *Children's Literature Assoc. Quart.*, vol. 28, no. 1, pp. 5–16, 2003.
[54] J. I. A. Hernández and A. J. M. Guijarro, *Narración Infantil y Discurso: Estudio Lingüístico de Cuentos en Castellano e Inglés*. Cuenca, Spain: Ediciones de la Universidad de Castilla-La Mancha, 2001.
[55] A. Trosborg, *Text Typology and Translation*. Amsterdam, The Netherlands: John Benjamins, 1997.
[56] C. Smith, "Discourse modes: Aspectual entities and tense interpretation," *Cahiers de Grammaire*, vol. 26, pp. 183–206, 2001.
[57] W. Labov and J. Waletzky, "Narrative analysis. Essays on the verbal and visual arts," in *Proc. Spring Meeting Amer. Ethnolog. Soc.*, Seattle, WA, USA: Univ. Washington Press Seattle, 1967, pp. 11–44.
[58] M. A. Locher, "Electronic discourse," in *Pragmatics Discourse*. Berlin, Germany: Walter de Gruyter GmbH & Co KG, 2014, pp. 555–581.
[59] L. Squires, "Enregistering Internet language," *Lang. Soc.*, vol. 39, no. 4, pp. 457–492, Sep. 2010.
[60] A. O. AbuSa'aleek, "Internet linguistics: A linguistic analysis of electronic discourse as a new variety of language," *Int. J. English Linguistics*, vol. 5, no. 1, p. 135, Jan. 2015.

MARTA VICENTE received the bachelor's and master's degrees in computer science from the University of Alicante, Spain, in 2015, where she is currently pursuing the Ph.D. degree with the Language Processing and Information Systems research group.

She is the author of several communications and scientific articles on the field of natural language generation, and language comprehension and summarization. Her research interests include discourse processing, including both its understanding and generation, with particular emphasis on semantics and pragmatics.

MARÍA MIRÓ MAESTRE received the bachelor's degree in translation and interpreting and the master's degree in institutional translation from the Universidad de Alicante, Spain, in 2017 and 2019, respectively, where she is currently pursuing the Ph.D. degree in computational linguistics with the Language Processing and Information Systems Research Group. In 2020, she joined the Department of Software and Computing Systems, University of Alicante. Her research interests include pragmatics, with particular emphasis on communicative intentions, and how they can be processed automatically.

ELENA LLORET received the Ph.D. degree in text summarization from the University of Alicante, Spain, in 2011.

She is currently a Lecturer with the University of Alicante. She is the author of over 60 scientific publications in international peer-reviewed conferences and refereed journals. She has served on the program committee for several international conferences, such as ACL, EACL, RANLP, and COLING. Her research interests include natural language processing (more specifically text summarization), and natural language generation. She is also a member of the Spanish Society for Natural Language Processing, and has participated in a number of national and EU-funded projects, among which the current and latest are: Canonical Representation and transformations of texts applied to the Human Language Technologies (TIN2015-65100-R) and SAM—Dynamic Social & Media Content Syndication for 2nd Screen. She has also been collaborating with international groups at the Universities of Wolverhampton, Sheffield, and Edinburgh—all in UK—, and the Lorraine Research Laboratory in Computer Science and its Applications in France.

ARMANDO SUÁREZ CUETO received the Ph.D. degree in computer science from the University of Alicante, Spain, in 2004.

Since 2008, he has been working as a Tenured University Lecturer with the Department of Languages and Computer Systems, University of Alicante. He is currently the Deputy Head of the Department. His research interests include word sense disambiguation, automatic language generation, and some other topics of natural language processing. Participating in several high impact publications, he co-chaired Doctoral Theses, contributed to development registered in the Intellectual Property Registry as a software registry, and participated in the coordination and organization of different R & D & I projects and activities. He has also been a member of the Language Processing and Information Systems research group, University of Alicante, since 1992.