sustainability

Article

Semantic Indexing of 19th-Century Greek Literature Using 21st-Century Linguistic Resources

Dimitris Dimitriadis [1,*], Sofia Zapounidou [2] and Grigorios Tsoumakas [1]

[1] School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; greg@csd.auth.gr
[2] Library and Information Centre, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; szapoun@lib.auth.gr
[*] Correspondence: dndimitri@csd.auth.gr

Abstract: Manual classification of works of literature with genre/form concepts is a time-consuming task requiring domain expertise. Building automated systems based on language understanding can help humans to achieve this work faster and more consistently. Towards this direction, we present a case study on automatic classification of Greek literature books of the 19th century. The main challenges in this problem are the limited number of literature books and resources of that age and the quality of the source text. We propose an automated classification system based on the Bidirectional Encoder Representations from Transformers (BERT) model trained on books from the 20th and 21st century. We also dealt with BERT's constraint on the maximum sequence length of the input, leveraging the TextRank algorithm to construct representative sentences or phrases from each book. The results show that BERT trained on recent literature books correctly classifies most of the books of the 19th century despite the disparity between the two collections. Additionally, the TextRank algorithm improves the performance of BERT.

Keywords: semantic indexing; text classification; Greek literature; TextRank; BERT

check for updates

Citation: Dimitriadis, D.; Zapounidou, S.; Tsoumakas, G. Semantic Indexing of 19th-Century Greek Literature Using 21st-Century Linguistic Resources. *Sustainability* **2021**, *13*, 8878. https://doi.org/10.3390/su13168878

Academic Editor: Julie Ernst

Received: 30 June 2021
Accepted: 4 August 2021
Published: 9 August 2021

1. Introduction

The role of cultural heritage in sustainability is mainly perceived in terms of monuments and sites and their role in raising awareness of the landscape or in contributing to the touristic development of the area. Literature can have an important role in this respect, as it helps people understand the cultural context of an era, of a nation, of a monument, of a place, etc., thus enabling them to adopt more inclusive and equitable attitudes and behaviors. From an opposite perspective, the sustainability of cultural heritage through information technology is also very important, as several pieces of cultural content have not yet been fully digitized. This is particularly relevant to works of literature, especially of the past centuries.

Semantic indexing of works of literature using concepts related to their subject, such as genre or form terms, is an important process enabling the work to be searched and retrieved by such concepts. Taking into account that people often exhibit preferences for specific genres [1–3], information about genre is a useful search filter for finding literature. Moreover, without indexing with relevant metadata, works of literature are not easy to discover and eventually use, which puts the maintenance and sustainability of such cultural content at risk. A well-known literature classification scheme is the Genre/Form Terms for Library and Archival Materials (LCGFT) [4] , which is used by most libraries around the world [5]. Manual indexing of works of literature with concepts is a time-consuming task that requires domain expertise. Automated indexing systems based on natural language understanding can help humans do this work faster and more consistently.

This paper documents our approach towards the construction of a system for automatically classifying Greek books of the 19th century with genre/form concepts. Our work is part of the ECARLE research project [6], which concerns the semantic enrichment of 19th-century Greek books with metadata, such as layout information, named entities, relations among named entities, and form/genre terms. ECARLE compiled a dataset of 57 Greek books of the 19th century, classified as essays, prose, poems, and manuals. A distinctive aspect of our work is that Greek books of the 19th century were written mainly in the *katharevousa* variety of the Greek language, a language in between ancient Greek and the modern *demotic* variety of Greek. Working with text in katharevousa is challenging, as state-of-the-art natural language processing resources, such as embeddings and pre-trained neural language models, are based on modern Greek corpora. Most of the past approaches for literature classification have focused on modern Greek [7–9].

Nowadays, state-of-the-art models for text classification employ deep learning and in particular pre-trained language models, which are fine-tuned on training data from the task at hand [10,11]. Following this paradigm, our approach leverages a version of the BERT model pre-trained on modern Greek corpora [12]. To the best of our knowledge, this is the first work to use BERT for classifying Greek literature in general. Most of the past studies put more emphasis on feature engineering [9,13].

One limitation of BERT is that it cannot take large pieces of text as input. To address this issue, we use the TextRank algorithm [14] to extract from each book a representative set of sentences or phrases, which are then passed as input to BERT. This has been recently proposed in [15], but for a different task (information retrieval) and domain (legal documents) than in our case. Other approaches to dealing with long documents include extending BERT [16,17], and selecting parts of the document [10,18].

As the size of the ECARLE dataset was too small for training or fine-tuning models and at the same time evaluating the accuracy of the trained models, we constructed a second annotated dataset of 755 books from the 20th and 21st century by leveraging the content of the *Open Library* (OL) [19], a repository of more than seven thousand Greek digital books distributed freely and legally on the Internet in PDF format. In our empirical study, we use a large part of the OL dataset for fine-tuning BERT and the rest of the OL dataset, as well as the ECARLE dataset, for measuring the accuracy of the fine-tuned model.

In summary, this work makes the following main contributions: (i) a case study on semantic indexing of 19th-century Greek literature using 21st-century state-of-the-art models and linguistic resources, and (ii) two collections of Greek books classified as essays, prose, poems, and manuals, which span three centuries [20].

Our work aims to answer the following research questions:

1. Is it possible to create an accurate learning model for automatically classifying Greek books of the 19th century written in the katharevousa variety of the Greek language, using resources from the 20th and 21st century written in the demotic (modern Greek) variety of the Greek language?
2. Can a representative set of sentences/phrases extracted from each book improve the results of a transformer text classification model compared to using consecutive parts of the book?

To answer the first question, we experimented with two data sets, one from the 19th century (ECARLE) and one from the 20th and 21st century (OL). We used the BERT model, which has previously achieved state-of-the-art results. We hypothesize that BERT will manage to transfer knowledge from modern Greek to katharevousa, as it models language at the sub-word level and the two Greek language varieties share common sub-words. To answer the second question, we experimented with the TextRank extractive summarizer to extract sentences/phrases, hypothesizing that it will manage to distill the necessary information for genre/form classification from each book.

The rest of this paper is structured as follows. Section 2 presents a comprehensive review of related work and the current state of the art in text classification. Section 3 outlines the methodological approach we used. Section 4 describes the evaluation of our approach, including the experimental setup (parameter tuning and datasets preprocessing), the results on both the OL and the ECARLE datasets, and the discussion of the results. Finally, Section 5 concludes this paper and mentions directions for future work.

2. State of the Art and Related Work

We first discuss traditional and state-of-the-art methods for text classification in general. Next, we review related work focused on semantic indexing of literature, contrasting them with the approach we adopted in this work.

2.1. State of the Art

Traditional approaches to text classification typically involve feature engineering and supervised learning. The most common machine learning algorithms used in this task are decision trees, pattern rule-based classifiers, support vector machines (SVMs), neural networks, and Bayesian classifiers [21], while in the feature selection phase, several evaluation metrics have been explored, such as term frequency, information gain, and chi-square [22].

Recent approaches work directly with the text, using complex neural networks for extracting patterns. The authors of [23] use dynamic embedding projection-gated convolutional neural networks, outperforming other approaches on four well-known text classification datasets. A hybrid approach was used by [24] incorporating predictive text embedding and graph convolutional networks for text classification to address the limitations of both methods enabling faster training and improving performance in small labeled training set scenarios. The study in [25] uses a vector representation based on a supervised codebook in document classification in Nepali. These approaches perform well in text classification, but they have not been tested in a large number of benchmarks and natural language processing tasks.

Transformer-based models generally perform well in many natural language processing tasks, including text classification. A recent study shows that BERT and its variations outperform all the previous machine and deep learning techniques, such as SVMs, naive Bayes, convolutional, and recurrent neural networks [11]. The study shows the results of many text classification tasks and benchmarks, such as sentiment analysis, question answering, and topic classification. Towards this direction, we use the BERT model to classify Greek books based on their form/genre concepts, and we build on previous research to overcome the problem of long documents by leveraging the TextRank algorithm to create appropriate inputs for the model.

2.2. Related Work

For genre classification of literary texts, the most common features are based on stylometrics [26], which are extracted by statistical analysis of the texts [27–30]. The authors of [27] used as style markers (i.e., countable linguistic features) the frequency of occurrence of the most frequent words of the entire written English language using the British National Corpus to automatically detect the genre of the given text showing that these frequent words are reliable discriminators of text genre. In the same spirit, we use the TextRank algorithm to extract sentences/phrases considering that the output of the algorithm can be used as a discriminator of genre/forms concepts. Stylometrics, content-based features, and social features were used in [28] for genre classification of German novels. The authors showed that even though topics are orthogonal on genre, the SVMs considering topic-based features achieved the highest accuracy compared to other traditional machine learning algorithms such as k-nearest neighbor, naive Bayes, and multilayer neural networks. Our study shares the same limitations with this work in terms of the amount of data for training the models. However, we also have an extra obstacle related to the language varieties.

Other studies on analyzing literary texts place more emphasis on the representation of each document as a multidimensional vector, which is given as input to a classifier. Towards this direction, Yu [31] experimented with four different text representations (based on absence/presence of a word, frequency of words, normalized frequency of the words, and idf-weighted frequency of the words) in eroticism classification of Dickinson's poems and sentimentalism classification of chapters in early American novels. The authors of [32] propose the encoding of books as binary vectors, where each dimension corresponds to the existence or not of a character 4-gram in the document. In contrast to these approaches and other similar ones, we use the BERT model, which can both represent the given input as a dense vector and categorize it into a genre/form concept. In this way, the model itself encodes the information needed for solving the task, instead of being affected by human choices on document representation.

In the case of very long documents such as books, the extraction of smaller parts of texts is necessary. In this context, Worsham and Kalita [18] propose different methods to select a representative text for training the learning models, such as extracting the first, last, or random 5K words of a book or 2.5K words of each chapter of the book. We similarly sliced the text into parts, but we also proposed the use of TextRank for constructing more sophisticated sets of sentences/phrases for the classifier.

There are also approaches which categorize books by focusing on other parts of the book, such as the book cover [33], as well as approaches which categorize web documents [34]. These approaches are out of the scope of this paper, as the first one considers text with images or only images, while the second one considers web documents that have a very different structure to literature books.

There are few studies on text classification of Greek literature. Most of them show the language independence of their approach in the genre identification task [8,9,35], or test several feature engineering approaches [7,13]. None of the existed studies dealt with the language variety problem, while as far as we know, none of the existed studies have experimented with text classification in Greek literature leveraging transformer-based models.

3. Methodological Approach

This section introduces the approach that we used for constructing a 19th-century Greek literature semantic indexing model, as well as the two datasets that are involved in our study.

3.1. Approach

Our approach is based on a BERT model that was trained on the Greek language [36]. Specifically, it was trained on the Greek part of Wikipedia, the Greek part of European Parliament Proceedings Parallel Corpus, and the Greek part of OSCAR, a cleaned version of Common Crawl. We fine-tune this model on the semantic indexing task of our case, using the modern Greek books of the OL dataset. We then evaluate it on the 19th-century books of the ECARLE dataset.

As BERT cannot accept very long sequences of text as input, we had to find a way to distill representative content from the books that could both fit as input to BERT and contain useful information for discriminating among the four different literature categories. A common way to deal with this problem is to select parts of a long document, e.g., randomly, up to the desired number of tokens. Another way is to use transformer-based models to represent several parts of the document, which are then used by a traditional supervised model as input. Here, we adopt a simpler method that employs TextRank [14], an extractive summarization algorithm, to distill a small set of representative sentences/phrases from each document. These pieces of text are then given as input to the BERT model for the classification task.

Figure 1 illustrates this approach. The long document is given as input to TextRank, which extracts the top N sentences/phrases and passes them to the BERT model. The input sentences/phrases are separated by the [SEP] special token which corresponds to the end of a sentence.

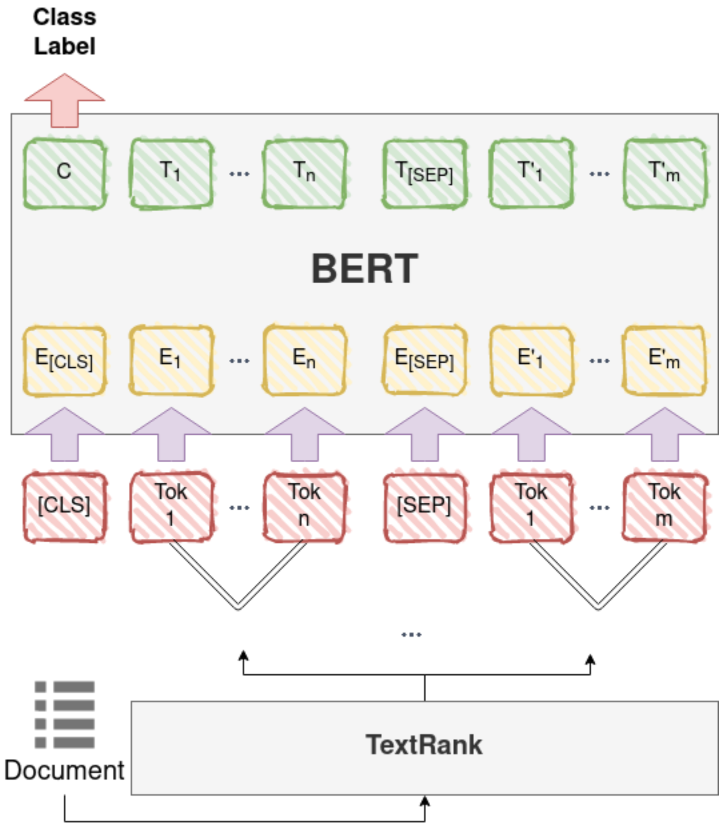

Figure 1. A transformer-based classifier trained on sentences/phrases extracted from TextRank.

3.2. Data

Literature experts selected 107 Greek books [20] from the library of the Aristotle University of Thessaloniki for the purposes of the ECARLE project. The experts classified these books into eight categories: prose, poetry, letters, essays, lexicons, encyclopedias, magazines, and manuals. The books were already in digital form through scanning. However, extracting their text via OCR proved to be very challenging. Despite employing several state-of-the-art tools, from open source libraries and commercial software to training deep learning models on sample transcribed pages of the books [37], the OCR accuracy was far from satisfactory. Some characters were not recognized correctly, e.g., πέχνην instead of τέχνην *(art)*, or were completely missed, e.g., ςϱεϑλόν instead of παϱελϑόν *(past)*, by the OCR. In some cases, the intensity of this phenomenon led to whole words and phrases missing or without meaning. In addition, sometimes the OCR process misinterpreted page headers and footers as normal page text. Due to these difficulties, the extraction was accomplished for only 57 of the books, 29 of which were essays, 15 prose, 11 poems, and 2 manuals.

As the size of the ECARLE dataset was too small for training machine learning models, we constructed a second dataset by leveraging the content of the *Open Library*, a repository of more than seven thousand Greek digital books distributed freely and legally on the Internet in PDF format. The digital books of Open Library are classified into 40 thematic categories. In addition, each book is accompanied by one or more tags and metadata, entered by the platform's administrators. The literature books of this repository are classified into 8 main categories: classic literature, novels-novellas, short stories, poems, essays, plays, children's literature, and comics. Two out of the four categories of interest exist in this categorization: poems and essays. The novels-novellas and short stories categories were jointly considered as members of the prose category. Lacking a category related to manuals, we considered the books having the word manual in their metadata.

Typically, the category of each book is also included in the metadata. This is not the case for all books, however. In addition, we observed that some books that were classified in one of the categories of our interest, had another one in the metadata. For example, the book Ήσουνα κάποτε εδώ (*You were once here*) belongs to the *poetry* class, but includes prose in the metadata. To avoid noisy examples, we decided to keep books that include their category in the metadata and at the same time do not contain another member from our category set in their metadata. Furthermore, we manually removed books that did not contain any readable characters due to the PDF extraction process. The final dataset contains 124 essays, 254 prose, 177 poems, and 200 manuals from the 20th and 21st century. For extracting the plain text from the PDF files of Open Library, we used the Python library PDFMiner [38].

Figure 2 illustrates the workflow of creating the two datasets including the conversion and preprocessing of the original collections. Firstly, we mined the books from the Open Library and the Library of Aristotle University and applied PDF extraction using PDFMiner tool and OCR conversion accordingly (technical details about OCR can be found in [37]). Some books of the Open Library did not have readable characters at all, so we manually removed them. Each dataset passes through the TextRank algorithm using PyTextRank [39] from SpaCy library for creating sentences/phrases for each book. Then, the changed books are tokenized based on the BERT tokenizer provided by the transformers [40] Python library. After the tokenization phase, the OL and ECARLE datasets are ready for the training and testing.

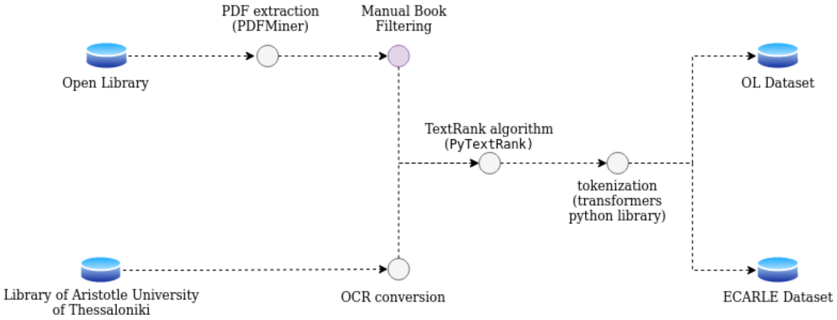

Figure 2. The workflow of creating the OL and ECARLE datasets.

Figure 3 shows the histogram of the publication year of the books in the two datasets. As we can see, there is a gap from 1900 to 1970 between the two datasets apart from three books which were published in 1917, 1959, and 1963, respectively. All the books in the ECARLE dataset were published before 1900, while most of the books in the OL dataset were published after 2010. One important difference between the books of the two datasets, stemming from the different century that they were written, is the variety of the Greek language that they use. Books in the ECARLE dataset are mainly written in the katharevousa variety, while books in the OL dataset are written in modern Greek (demotic variety).

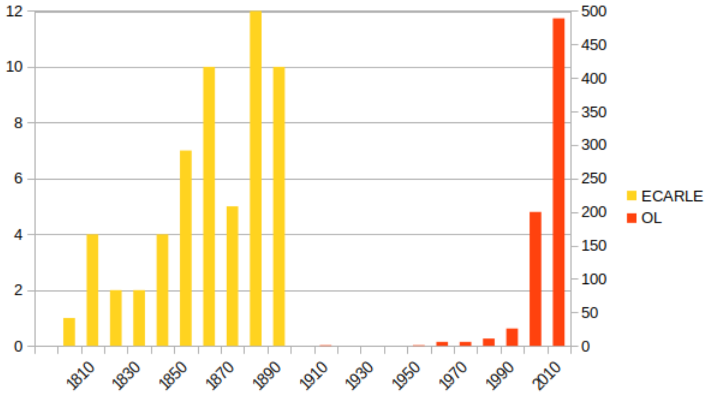

Figure 3. Histogram of the publication year of the books in the OL and ECARLE datasets. The x-axis corresponds to ranges of years, the left y-axis to the number of books in the ECARLE dataset in a specific range, and the right y-axis to the number of books in the OL dataset.

Differences between the two datasets have also been observed in the number of words contained in each book (Figure 4). The books in the OL dataset have approximately three thousand words on average, while the books in the ECARLE dataset more than four thousand. There are also some outlier books with more than seven thousand words. To find the words in the datasets, we used the el_core_news_lg vocabulary of the spaCy [41] Python library and we ignored tokens belonging to PUNKT and SPACE classes since the first one includes all punctuation marks and the second one all space characters.

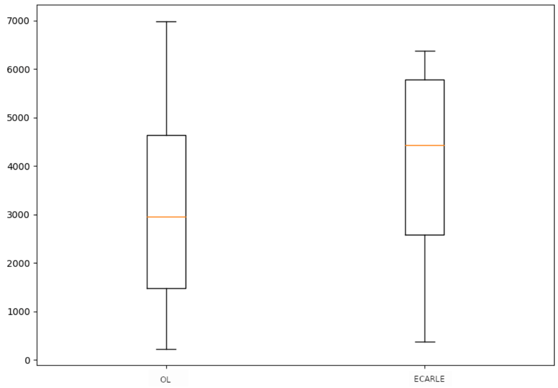

Figure 4. Distribution of the number of words in books of the OL and ECARLE datasets. The outliers are missing for the sake of visualization.

4. Evaluation

This section describes the experimental setup and discusses the results. We first present the datasets that were used for training and testing. Next, we mention the hyperparameter tuning process and present the results on both datasets. Finally, we discuss and explain the results.

4.1. Experimental Setup

We split the OL dataset into a train and a test set, in a way such that the distribution of the classes in the test set is the same as in the ECARLE dataset. This allows for a more informative comparison between the accuracy of the model at in-sample modern Greek text and out-of-sample katharevousa text. As a result, the training set consists of 698 books and the test set 57. Table 1 presents the number of instances per class and the total number of instances for the train and test sets of the OL dataset, as well as for the ECARLE dataset.

Table 1. Number of instances per class and total number of instances for the train and test sets of the OL dataset, as well as for the ECARLE dataset.

Dataset	#Instances	#Instances/Class
OL train set	698	95 essays 166 poems 239 prose 198 manuals
OL test set	57	29 essays 11 poems 15 prose 2 manuals
ECARLE dataset	57	29 essays 11 poems 15 prose 2 manuals

TextRank was used in three different variations: we extracted phrases with a rank score greater than 0.01, as well as the top 5/10 sentences. In addition, we experimented with splitting each book into three equal parts and considering the first 256 tokens of each part, as 256 is the maximum sequence length that our BERT model can accept.

To find the appropriate hyper-parameters for fine-tuning BERT to our classification task, we used stratified 5-fold cross-validation on the train set of OL. We followed the instructions of BERT's creators [42] for the fine-tuning process, experimenting with the following set of parameters: (i) learning rate 2×10^{-5}, 3×10^{-5}, 5×10^{-5}, (ii) batch size 16, 32, and epochs 2, 3, 4. Table 2 shows the selected hyper-parameters, with respect to each different method of input selection for the BERT model, along with the corresponding accuracy of the model.

Table 2. The selected hyper-parameters for fine-tuning BERT with respect to each method for selecting the input to the BERT model.

Method	Batch Size	lr	#Epochs	*Acc.*
first part	16	5×10^{-5}	4	0.8525
second part	16	3×10^{-5}	4	0.9012
third part	32	2×10^{-5}	3	0.8693
TextRank phrases	16	2×10^{-5}	4	0.9039
TextRank 5 sentences	32	2×10^{-5}	2	0.8725
TextRank 10 sentences	16	3×10^{-5}	4	0.8926

To further enhance our assumption about the effectiveness of BERT to generalize beyond the training set, we also experimented with the most common traditional machine learning algorithms that have achieved great performance in text classification. Particularly, we experimented with the support vector machines (SVMs), naive Bayes (NB), and logistic regression (LG). To give appropriate inputs to the classifiers, we used count vectorization converting the training/test sets of text documents to a matrix of token counts. As the vocabulary, we used the top 60,000 tokens ordered by term frequency across the training set. Since the entire document can be represented using such method, we did not experiment with different input methods. We used stratified 5-fold cross validation to select the models with the highest accuracy considering a set of parameters for each one. For SVMs, we experimented with the regularization parameter (C) with values (0.1, 1, 10, 100), the kernel coefficient for radial basis function (rbf), polynomical and sigmoid kernels (gamma) (1, 0.1, 0.01, 0.001) and the kernel type to be used in the algorithm (kernel) (rbf, polynomial, sigmoid, linear). The degree of the polynomial kernel function was set fixed to 3 and tolerance for stopping criterion to 1 × 10^{-3}. To support multiclass classification, we used the one-against-one scheme which is used as multiclass strategy and performs better than other schemes such as one-against-all [43]. For NB, we experimented with the additive smoothing parameter (alpha) with values (0.5, 1, 2). Finally, for LG, we experimented with different solvers (newton-cg, lbfgs, sag, and saga), the norm used in penalization (L1, L2), and the inverse of regularization strength (C) (0.1, 0.5, 1.0). The tolerance for stopping criteria was set fixed to 1 × 10^{-4}. To implement the infrastructure for the traditional machine learning algorithms, we used the Scikit-learn Python library [44].

Table 3 summarizes the results of the classifiers along with the selected parameters and the mean accuracy over the five folds during validation.

Table 3. The best hyper-parameters for each machine learning algorithm based on the accuracy.

Classifier	Parameters	*Acc.*
SVMs	C = 0.1 , kernel = linear, gamma = 1	0.8738
LG	C = 1, solver = lbfgs, penalty = L2	0.9154
NB	alpha = 1	0.8982

To evaluate the performance of the learning models we used the following measures:

1. **Accuracy** counts the correct predictions over the total number of examples.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, FN correspond to the true positives, true negatives, false positives, and false negatives, respectively.

2. **Kappa coefficient** [45] indicates how much better a trained classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. The smaller the value, the more likely it is that the classifier would randomly classify the instances.

$$K = \frac{c * s - \sum_k^C p_k * t_k}{s^2 - \sum_k^C p_k * t_k} \tag{2}$$

where c is the total number of instances correctly predicted, C the total number of classes, s the total number of instances, p_k the number of times that k was predicted, and t_k the number of times that k truly occurs.

3. **F1 score** is the harmonic mean of the precision and recall for a class.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

where:

$$precision = \frac{TP}{TP + FP} \tag{4}$$

and

$$recall = \frac{TP}{TP + FN} \tag{5}$$

4. **Weighted average F1 score** estimates the weighted average of the harmonic means of the precision and recall for all classes.

$$WAF1 = \frac{\sum_{i=1}^C n_i * F1_i}{\sum_{i=1}^C n_i} \tag{6}$$

where n_i is the number of instances of the class i, C the total number of classes, and the $F1_i$ is the F1 score of the class i.

Although accuracy is an appropriate measure for balanced datasets, in our case it can be misleading, since the dataset is significantly imbalanced. F1 score and Kappa coefficient can give us better insights on the outcomes.

4.2. Results

Firstly, we present results on the OL test set (Table 4) and the ECARLE dataset (Table 5) in terms of accuracy and kappa coefficient (K), based on the hyper-parameters selected earlier. As expected, we notice that the results in the ECARLE dataset are worse than those in the OL test set. The models trained on the first part of the book or TextRank phrases have high performance in OL test set. In the ECARLE dataset, the model trained on five sentences extracted from TextRank has the best accuracy (68.42% acc.) The models trained on second/third parts of the books have the worst performance in OL test set with 75.44% and 71.93%, respectively, while in the ECARLE dataset, the models trained on second part of the book or TextRank phrases have equally the worst performance with 59.65% accuracy. Models trained on 10 sentences extracted from TextRank have high performance on OL test set with 85.96% accuracy. However, the performance is lower in the ECARLE dataset.

Table 4. Results on OL test set. Bold indicates the best performance.

Method	***Acc.***	**K**
first part	**0.8772**	0.8100
second part	0.7544	0.6580
third part	0.7193	0.6010
TextRank phrases	**0.8772**	**0.8140**
TextRank 5 sentences	0.8246	0.7410
TextRank 10 sentences	0.8596	0.7930

Table 5. Results on ECARLE dataset. Bold indicates the best performance.

Method	***Acc.***	**K**
first part	0.6140	0.3160
second part	0.5965	0.3590
third part	0.6316	0.4410
TextRank phrases	0.5965	0.4120
TextRank 5 sentences	**0.6842**	**0.5090**
TextRank 10 sentences	0.6140	0.3580

Regarding the K score in OL test set, there is greater agreement between the raters (actual and predicted values) for the model trained on TextRank phrases (0.8140) in OL dataset, while in the ECARLE dataset, the model trained on 5 sentences extracted from TextRank has K equal to 0.5090.

Table 6 presents the results on OL test set and ECARLE dataset considering the F1 score. The models trained on TextRank phrases/sentences have the highest weighted average F1 score on the OL test set (88.32%) and on the ECARLE one (67.75%). The difference in the performance on the two datasets of the model trained on the TextRank 5 sentences, is the second-lowest (15.04%). The model trained on the first part of the book has a high weighted average F1 score on the OL test set. However, its performance in the ECARLE dataset is low (55.19%). Finally, the model trained on the third part has the lowest difference between the results on the two datasets (10.13%), but this depends on the lowest performance on the OL test set (74.81%).

Table 6. Results based on F1 score per class and weighted average F1 score (WAF1) on OL test set and ECARLE dataset with and without TextRank. Bold indicates the highest weighted average F1 score either on OL dataset or ECARLE one.

		F1 Score/Class				
Dataset	**Method**	**Essays**	**Prose**	**Poems**	**Manuals**	**WAF1**
OL	first part	0.89	0.88	0.86	0.80	0.8784
	second part	0.77	0.88	0.84	0.29	0.7956
	third part	0.79	0.78	0.67	0.33	0.7481
	TextRank Phrases	0.91	0.84	0.85	1.00	**0.8832**
	TextRank 5 sentences	0.86	0.80	0.81	0.67	0.8279
	TextRank 10 sentences	0.88	0.90	0.88	0.50	0.8719
ECARLE	first part	0.72	0.21	0.67	0.00	0.5519
	second part	0.73	0.36	0.55	0.00	0.5723
	third part	0.72	0.56	0.63	0.33	0.6468
	TextRank Phrases	0.68	0.69	0.42	0.00	0.6086
	TextRank 5 sentences	0.75	0.67	0.62	0.00	**0.6775**
	TextRank 10 sentences	0.71	0.30	0.67	0.00	0.4993

We further present the corresponding confusion matrices to show the predictions of the models over the true labels of the books with and without TextRank. Table 7 shows the confusion matrices of the OL test set. In all cases the models predict correctly the manuals. Firstly, we observe that only the models trained on an input constructed by TextRank correctly predicts all the poems. The model trained on TextRank phrases is the only one that did not misclassify other books as manuals while the models trained on the second/third parts of the books misclassified 10 and 7 books, respectively, as manuals. The model trained on the first part of the books is biased towards the essay class while it is the only one that successfully classified 25/29 books as essays. The models trained either on the second or on the third parts of the books are biased towards manual and prose classes which justifies the low performance of the models based on the accuracy.

Table 8 shows the confusion matrices for the ECARLE dataset. Models trained either on the 5 or 10 sentences extracted from TextRank correctly predict 9/11 poems while the model trained on the third part of the book has the worst performance in this class (6/11). The model trained on the first part classifies 26/29 essays correctly, but misclassifies 13/15 prose books as essays. Only the model trained on TextRank phrases classifies 11/15 prose books correctly, while it is the one with the worst performance in the essay class predicting 16/29 books. Furthermore, the model trained on the third part of the book predicts 1/2 manuals.

Table 7. Confusion matrices of OL test set. Rows correspond to predictive values and columns to the actual ones for the four categories (essays (E), prose (Pr), poems (P), and manuals (M)).

	First Part				**Second Part**				**Third Part**				**TextRank Phr.**				**TextRank 5 Sent.**				**TextRank 10 Sent.**			
	E	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**
E	25	1	1	0	18	0	0	0	19	0	0	0	24	0	0	0	22	0	0	0	23	0	0	0
Pr	2	14	1	0	1	15	3	0	2	14	5	0	3	13	0	0	3	12	0	0	1	13	0	0
P	1	0	9	0	0	0	8	0	1	0	6	0	2	2	11	0	2	3	11	0	1	2	11	0
M	1	0	0	2	10	0	0	2	7	1	0	2	0	0	0	2	2	0	0	2	4	0	0	2

Table 8. Confusion matrices of ECARLE dataset. Rows correspond to predictive values and columns to the actual ones for the four categories (essays (E), prose (Pr), poems (P), and manuals (M)).

	First Part				**Second Part**				**Third Part**				**TextRank Phr.**				**TextRank 5 Sent.**				**TextRank 10 Sent.**			
	E	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**	**E**	**Pr**	**P**	**M**
E	26	13	3	1	22	7	1	1	19	4	1	0	16	2	0	0	21	5	0	1	23	11	1	1
Pr	0	2	1	1	0	4	2	1	6	10	4	1	1	11	4	1	0	9	2	1	0	3	1	1
P	3	0	7	0	6	4	8	0	2	0	6	0	12	2	7	1	8	1	9	0	6	1	9	0
M	0	0	0	0	1	0	0	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Finally, we present the results with traditional machine learning (ML) algorithms for the OL test set (Table 9) and the ECARLE one (Table 10) . As we expected, the algorithms have very good performance in the OL test set, since they have achieved great performance in a variety of text classification tasks before. The LG algorithm outperforms all models with 89.47% accuracy and WAF1 89.84%. However, in the ECARLE dataset, the results are significantly worse. The NB algorithm has the worst performance (19.30% acc.), while the best algorithm LG has 52.63% accuracy.

Table 9. Results on OL test set using ML.

Model	***Acc***	**WAF1**
SVMs	0.8596	0.8810
LG	0.8947	0.8984
NB	0.8070	0.8158

Table 10. Results on ECARLE dataset using ML.

Model	***Acc***	**WAF1**
SVMs	0.2807	0.2082
LG	0.5263	0.4439
NB	0.1930	0.0624

4.3. Discussion

The results indicate that we can build an efficient classifier for Greek books of the 19th century using resources from the 20th and 21st century since the BERT model classified most of the ECARLE books correctly. We observe that the model on both the OL test and ECARLE dataset has equally high performance. Furthermore, our assumption about the effectiveness of the BERT model has been confirmed since the alternatives, the traditional machine learning algorithms, had the worst performance in the ECARLE dataset.

All learning models had high performance during the hyperparameter tuning process with stratified 5-fold validation and OL training set. We expected good performance of the traditional machine learning algorithms since they have achieved high accuracy in text classification generally and also because all books are from the same collection and produced by the same extraction method. We also expected good performance of the BERT model since (1) it has achieved state-of-the-art results in many natural language processing tasks; (2) all books are from the same collection produced by the same extraction method and are written in modern Greek; (3) the BERT model has also been pre-trained on modern Greek texts; and (4) BERT is highly adaptable in downstream tasks.

The TextRank algorithm improves the results of the BERT model. The model fine-tuned on TextRank phrases has the highest weighted average F1 score (88.93%), the highest Kappa Coefficient score (81.40%), and the equal highest overall accuracy (87.72%). An explanation for the performance of TextRank is that BERT discovered more patterns in the phrases than among the sentences. These are not syntactic and semantic patterns but maybe simple statistics such as word and punctuation frequencies. Considering also the fact that most previous approaches achieve high performance using stylometric features including word and punctuation frequencies [27–30], the results can be justified. The performance of the models fine-tuned on consecutive parts of texts is affected by the category distribution in the test set. Although the models trained on the second/third part of the text have high accuracy during hyperparameter tuning (90.12% and 86.93%, respectively), they have the lowest performance in the test set (75.44% and 71.93%, respectively).

We expected the results on the ECARLE dataset for a multitude of reasons. First, we were unable to find a plethora of 19th-century books, and the training of the learning models happened using modern Greek books of the 20th century and the 21st century. Thus, we expected some failures due to fundamental differences between the language used in the 19th-century and 20th-century Greek texts. These failures are more apparent when we use traditional machine learning algorithms. Descriptive statistics showed that there is a chronological gap between the two datasets, as well as a difference between the distribution of the words. The differences which occurred from the language itself related to the so-called *Greek language question* between the high variety of katharevousa and the low variety of demotic. This conflict regarding the dominance of one variety over the other took place in the late 19th and the 20th century.

The high variety of katharevousa retained the ancient Greek synthetic character and was established as the official national language of the Greek state in 1827. Regarding literary texts, katharevousa was mostly used in official documents, essays, prose, while the dominant language variety in poetry was the demotic [46]. In the late 19th century, many poems and scholars started defending the use of the demotic variety, establishing the movement of *demoticism*. The demotic variety was the one spoken by Greeks of the time. Concerning the katharevousa, the demotic is not a synthetic language but an analytic one. In practice, this means that the demotic uses more words and phrases to express a meaning in contrast with the katharevousa. In the 20th century, the *Greek language question* took on political dimensions; the conservatism supported the use of katharevousa, while the communists supported the use of the demotic. The use of demotic in written texts expanded in the 20th century and became dominant in the 1960s. In 1976, demotic was finally recognized as the official language of the Greek state.

The above statements justify the performance of TextRank phrases in predicting the prose books with high accuracy (11/15). The input of the BERT model based on the TextRank phrases is a concatenation of non-consecutive small pieces of texts that mitigates the differences between the demotic and katharevousa. Indeed, there are not more words and phrases in demotic books to express a meaning in contrast with the books in katharevousa. Although the prose books are written in katharevousa in the 19th century and in Demotic in the 20th century, this difference did not influence the performance of the model. On the other hand, the model trained on the first part of the book predicts 14/15 books as prose books in OL test set, while it classified only 2/15 books in the ECARLE dataset.

The performance of the models is also justified for the poems. Many books have been misclassified as poems. Sentences extracted from TextRank can classify 9/11 poems correctly. Poems have already been written in *Demotic* since the 19th century. Thus, books in the OL dataset have similar way of writing with the books of ECARLE dataset.

An observation is that the models are biased towards essays in ECARLE dataset except the model trained on TextRank phrases. This is an evidence that the OL dataset can be used for training models that can be used for predicting books from an earlier century. The high performance of the models in OL test set for the essay class is equally high for the ECARLE dataset despite the fact that we have a small number of essays during training in contrast to the number of books of other classes (e.g., prose).

The corrupted sets due to OCR conversion and PDF extraction seem not to affect the performance of the learning models. The noisy and missing sentences did not significantly affect the performance of the BERT model, neither did the differences between the Greek demotic and katharevousa. Although the test sets are small enough to provide a full explanation of the performance of the BERT model and TextRank algorithm, we observe that despite the limitations and obstacles presented in the paper, the models are capable of classifying an important set of books correctly.

BERT is known to learn complex features, such as syntactic patterns and semantic dependencies [47]. Considering that previous studies have shown that stylometrics play a key role in genre identification, we believe that BERT manages to learn such types of features during fine-tuning.

5. Conclusions and Future Work

This paper addressed the problem of constructing a model for classifying Greek literature of the 19th century by genre/form concepts, under the limitations of a small collection of works of literature and the low quality of the source text. To address these challenges, we compiled a collection of modern Greek books and employed the state-of-the-art BERT model in conjunction with the TextRank algorithm for extracting significant sentences/phrases from each book. We posed two research questions and experimented with state-of-the-art algorithms for answering them. We found that recent books written in the modern Greek language helped us train efficient models correctly classifying most of the literature books in our target collection which were written in katharevousa. The assumption that the BERT model can efficiently build such a classifier has been confirmed considering that traditional machine learning algorithms had the worst performance in katharevousa. In addition, we found that using TextRank leads to better results compared to consecutive text parts extracted from the start, middle, or end of each book.

In future work, we aim to extend our collections of literature books to conduct more data-intensive experiments. Furthermore, we aim to experiment with several extractive and abstractive summarizers to confirm that a set of representative sentences/phrases can carry enough information for training an efficient classifier in this task. Finally, more experiments with traditional machine learning and deep learning models will give us a better perspective about the efficiency of the transformer-based models in this task and domain of interest.

Author Contributions: Conceptualization, D.D. and G.T.; methodology, D.D.; software, D.D.; validation, D.D.; formal analysis, D.D.; investigation, D.D.; resources, D.D., S.Z. and G.T.; data curation, D.D.; writing—original draft preparation, D.D., S.Z. and G.T.; writing—review and editing, D.D., S.Z. and G.T.; visualization, D.D.; supervision, G.T.; project administration, G.T.; funding acquisition, G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-05580).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work was supported by computational time granted from the National Infrastructures for Research and Technology S.A. (GRNET S.A.) in the National HPC facility—ARIS—under project ID pa181002-NEBULA.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kraaykamp, G. Literary socialization and reading preferences. Effects of parents, the library, and the school. *Poetics* **2003**, *31*, 235–257. [CrossRef]
2. Tveit, Å.K. Reading habits and library use among young adults. *New Rev. Child. Lit. Librariansh.* **2012**, *18*, 85–104. [CrossRef]
3. Schutte, N.S.; Malouff, J.M. University student reading preferences in relation to the big five personality dimensions. *Read. Psychol.* **2004**, *25*, 273–295. [CrossRef]
4. Library of Congress Genre/Form Terms. LC Linked Data Service: Authorities and Vocabularies | Library of Congress. Available online: https://id.loc.gov/authorities/genreForms.html (accessed on 6 August 2021).

5. Bitter, C.; Tosaka, Y. Genre/Form Access in Library Catalogs: A Survey on the Current State of LCGFT Usage. *Libr. Resour. Tech. Serv.* **2020**, *64*, 44. [CrossRef]
6. Ecarle Project. Exploitation of Cultural Assets with Computer-Assisted Recognition, Labeling and Meta-Data Enrichment. Available online: https://ecarle.web.auth.gr/en/ (accessed on 6 August 2021).
7. Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. Automatic text categorization in terms of genre and author. *Comput. Linguist.* **2000**, *26*, 471–495. [CrossRef]
8. Peng, F.; Schuurmans, D.; Wang, S. Augmenting naive bayes classifiers with statistical language models. *Inf. Retr.* **2004**, *7*, 317–345. [CrossRef]
9. Zhang, D.; Lee, W.S. Extracting key-substring-group features for text classification. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006 ; pp. 474–483.
10. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*; Springer: Cham, Switzerland, 2019; pp. 194–206.
11. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning–based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* **2021**, *54*, 1–40. [CrossRef]
12. Koutsikakis, J.; Chalkidis, I.; Malakasiotis, P.; Androutsopoulos, I. Greek-bert: The greeks visiting sesame street. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020; pp. 110–117.
13. Gianitsos, E.; Bolt, T.; Chaudhuri, P.; Dexter, J. Stylometric Classification of Ancient Greek Literary Texts by Genre. In Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Minneapolis, MN, USA, 7 June 2019; pp. 52–60.
14. Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
15. Shao, Y.; Mao, J.; Liu, Y.; Ma, W.; Satoh, K.; Zhang, M.; Ma, S. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, Yokohama, Japan, 11–17 July 2020; pp. 3501–3507.
16. Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Hierarchical Transformers for Long Document Classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 838–844.
17. Zhang, R.; Wei, Z.; Shi, Y.; Chen, Y. {BERT}-{AL}: {BERT} for Arbitrarily Long Document Understanding. 2020. Available online: https://openreview.net/forum?id=SklnVAEFDB (accessed on 6 August 2021)
18. Worsham, J.; Kalita, J. Genre identification and the compositional effect of genre in literature. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018 ; pp. 1963–1973.
19. Ανοικτή Βιβλιοθήκη • Ελεύθερα ψηφιακά βιβλία // Ελληνικά δωρεάν e-Books. Available online: https://www.openbook.gr/ (accessed on 6 August 2021).
20. GitHub. Dndimitri/Semantic-Indexing-of-GREEK-literature: This Repository Accompanies Our Ongoing Work on Classifying 19th Century Greek Literature Using 21st CenturyLinguistic Resources. Available online: https://github.com/dndimitri/semantic-indexing-of-greek-literature (accessed on 6 August 2021)
21. Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In *Mining Text Data*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 163–222.
22. Khan, A.; Baharudin, B.; Lee, L.H.; Khan, K. A review of machine learning algorithms for text-documents classification. *J. Adv. Inf. Technol.* **2010**, *1*, 4–20.
23. Tan, Z.; Chen, J.; Kang, Q.; Zhou, M.; Abusorrah, A.; Sedraoui, K. Dynamic embedding projection-gated convolutional neural networks for text classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–10. [CrossRef]
24. Ragesh, R.; Sellamanickam, S.; Iyer, A.; Bairi, R.; Lingam, V. Hetegcn: Heterogeneous graph convolutional networks for text classification. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, online, 8–12 March 2021; pp. 860–868.
25. Sitaula, C.; Basnet, A.; Aryal, S. Vector representation based on a supervised codebook for Nepali documents classification. *PeerJ Comput. Sci.* **2021**, *7*, e412. [CrossRef] [PubMed]
26. Lagutina, K.; Lagutina, N.; Boychuk, E.; Vorontsova, I.; Shliakhtina, E.; Belyaeva, O.; Paramonov, I.; Demidov, P. A survey on stylometric text features. In Proceedings of the 2019 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 5–8 November 2019; pp. 184–195.
27. Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. Text genre detection using common word frequencies. In Proceedings of the 18th Conference on Computational Linguistics, Saarbrücken, Germany, 31 July–4 August 2000 ; Volume 2, pp. 808–814.
28. Hettinger, L.; Becker, M.; Reger, I.; Jannidis, F.; Hotho, A. Genre classification on German novels. In Proceedings of the 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), Valencia, Spain, 1–4 September 2015; pp. 249–253.
29. Biber, D. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* **1986**, *62*, 384–414. [CrossRef]
30. Douglas, D. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Comput. Humanit.* **1992**, *26*, 331–345. [CrossRef]
31. Yu, B. An evaluation of text classification methods for literary study. *Lit. Linguist. Comput.* **2008**, *23*, 327–343. [CrossRef]

32. Wu, Z.; Markert, K.; Sharoff, S. Fine-grained genre classification using structural learning algorithms. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 749–759.
33. Biradar, G.R.; Raagini, J.; Varier, A.; Sudhir, M. Classification of Book Genres using Book Cover and Title. In Proceedings of the 2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), Visakhapatnam, India, 29–30 June 2019; pp. 72–723.
34. Lee, Y.B.; Myaeng, S.H. Text genre classification with genre-revealing and subject-revealing features. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 11–15 August 2002; pp. 145–150.
35. Peng, F.; Schuurmans, D.; Wang, S. Language and task independent text categorization with simple language models. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, AB, Canada, 27 May–1 June 2003; Volume 1, pp. 110–117.
36. GitHub. Nlpaueb/Greek-Bert: A Greek Edition of BERT Pre-Trained Language Model. Available online: https://github.com/\nlpaueb/greek-bert (accessed on 6 August 2021).
37. Tzogka, C.; Koidaki, F.; Doropoulos, S.; Papastergiou, I.; Agrafiotis, E.; Tiktopoulou, K.; Vologiannidis, S. OCR Workflow: Facing Printed Texts of Ancient, Medieval and Modern Greek Literature. In Proceedings of the Conference on Digital Curation Technologies (Qurator 2021), Berlin, Germany, 8–12 February 2021
38. PDFMiner. Available online: https://pypi.org/project/pdfminer/ (accessed on 6 August 2021).
39. PyTextRank. Available online: https://pypi.org/project/pytextrank/ (accessed on 6 August 2021).
40. Transformers. Transformers 4.7.0 Documentation. Available online: https://huggingface.co/transformers/ (accessed on 6 August 2021).
41. Honnibal, M.; Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Appear* **2017**, *7*, 411–420.
42. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
43. Hsu, C.W.; Lin, C.J. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425. [PubMed]
44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
45. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, 159–174. [CrossRef]
46. Politēs, L. *A History of Modern Greek Literature*; Clarendon Press: Oxford, UK, 1973.
47. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What does bert look at? an analysis of bert's attention. *arXiv* **2019**, arXiv:1906.04341.