The Competition on Automatic Classification of Literary Epochs

Irina Rabaev[1,*], Marina Litvak[1], Vladimir Younkin[1], Ricardo Campos[2], Alípio Mário Jorge[3] and Adam Jatowt[4]

[1]*Shamoon College of Engineering, Beer Sheva, Israel*
[2]*University of Beira Interior, INESC TEC, Ci2 - Smart Cities Research Center - Polytechnic Institute of Tomar, Portugal*
[3]*University of Porto, Porto, Portugal*
[3]*University of Innsbruck, Innsbruck, Austria*

Abstract

This paper describes the shared task on Automatic **C**lassification **o**f **Li**terary **E**pochs (CoLiE) held as a part of the 1st International Workshop on Implicit Author Characterization from Texts for Search and Retrieval (IACT'23) held at SIGIR 2023. The competition aimed to enhance the capabilities of large-scale analysis and cross-comparative studies of literary texts by automating their classification into the respective epochs. We believe that the competition contributed to the field of information retrieval by exposing the first large benchmark dataset and the first study's results with various methods applied to this dataset. This paper presents the details of the contest, the dataset used, the evaluation procedure, and an overview of participating methods.

Keywords

Text Classification, Implicit Information Retrieval, Implicit Temporal Context Retrieval

1. Introduction

Automatic epoch classification in the context of literary texts can be viewed as a form of implicit temporal information retrieval. Literature reflects the language styles, grammatical variations, thoughts, emotions, and perspectives of different times. The classification of literary texts into their respective epochs involves extracting implicit temporal information embedded in the language [1], enabling the retrieval of the historical context and characteristics unique to each literary period.

Literature can be classified by movements, genres, or periods. In this competition, we focused on the division of literature into different periods, a.k.a. epochs. According to different academic sources, some epochs are well-defined, while others may overlap [2, 3, 4], which is often a point

In: M. Litvak, I.Rabaev, R. Campos, A. Jorge, A. Jatowt (eds.): Proceedings of the IACT'23 Workshop, Taipei, Taiwan, 27-July-2023

*Corresponding author.
irinar@ac.sce.ac.il (I. Rabaev); marinal@c.sce.ac.il (M. Litvak); vladiyo@c.sce.ac.il (V. Younkin); ricardo.campos@ubi.pt (R. Campos); amjorge@fc.up.pt (A. M. Jorge); adam.jatowt@uibk.ac.at (A. Jatowt)
0000-0002-8542-8342 (I. Rabaev); 0000-0003-3044-3681 (M. Litvak); 0000-0002-8767-8126 (R. Campos); 0000-0002-5475-1382 (A. M. Jorge); 0000-0001-7235-0665 (A. Jatowt)

of contention between scholars. One possible way to categorize literature by epochs from 1700 to our days is as follows:

1. Romanticism (1798-1837) [5]: Romanticism focused on individualism, emphasized emotions over reason, imagination, freedom of form, and the natural world.
2. Victorian Literature (1837-1901) [6]: Named after Queen Victoria's reign, tended to depict daily life, and focused on realism, social reform, and a growing interest in science and technology. Novel became the leading literary genre during this period.
3. Modernism (1900-1945) [7]: Literature during this period often employed blended writing elements, experimentation with form and language, nonlinear plot, and introspection.
4. Postmodernism (1945-2000) [8]: Postmodernism is characterized by self-reflexivity, unreliable narrators, unrealistic and impossible narratives, parody, dark humor, and irony.
5. Our days (from 2000): Contemporary literature reflects technological advances, globalization, questions conventions, and often breaks traditional writing rules.

Every literary epoch is characterized by its voices, themes, and styles. In recognizing and understanding these epochs, we can acquire a more profound insight into the progression of human thought throughout history and the extensive range of human experiences and creativity. This motivated us to conduct the CoLiE task, which is, to the best of our knowledge, the first to be held on the automatic classification of text into five literary epochs. The main goal is to advance the field of implicit temporal information retrieval from a text and to compare the performances of different models and systems on a new dataset.

This paper describes the contest details. Section 2 provides an overview of the task and a description of the dataset. Section 3 presents a summary of participating systems, followed by Section 4, which presents results and discussions. Section 5 draws conclusions and proposes future directions.

2. Task Description and Dataset

The task on Automatic Classification of Literary Epochs (CoLiE) aimed at automatic identification of the following literary epoch of a given text from its writing style: (1) Romanticism (1798-1837), (2) Victorian Literature (1837-1901), (3) Modernism (1900-1945), (4) Postmodernism (1945-2000), and (5) Our days (from 2000). In this section, we describe the dataset and the format of the competition.

2.1. Dataset

In this competition, we introduce "BookSCE" — a new large-scale dataset of books, mostly published over the last three centuries. BookSCE is built upon the online book repository Project Gutenberg: Free eBooks, which focuses on literature and other written works. The books in BookSCE were annotated with labels that include the book's meta-data and authors-related information, such as name, residence, age, and publication date. Some labels were automatically extracted from the Project Gutenberg site. When the specific information was not present in the Project Gutenberg database, we tried to automatically retrieve it from other sources, e.g.,

Table 1
The BookSCE split for the CoLiE classification task. The numbers are in the form (#books, #chunks).

Epoch	Train	Val	Test	Total
Romanticism	(242, 20161)	(17, 1158)	(65, 5212)	(324, 26531)
Victorian	(3386, 240365)	(226, 16938)	(905, 63512)	(4517, 320815)
Modernism	(3671, 238088)	(245, 14848)	(977, 61454)	(4893, 314390)
PostModernism	(886, 22139)	(60, 1713)	(236, 6363)	(1182, 30215)
Our days	(537, 25457)	(36, 1600)	(144, 6490)	(717, 33547)
Total	**(8722, 546210)**	**(584, 36257)**	**(2327, 143031)**	**(11633, 725498)**

from the pdf file itself, Wikipedia, and Wikibooks. To verify the automatic annotation, we performed manual label validation on a random dataset sample. Because this competition aimed at automatic epoch classification, we used only a subset of BookSCE with a verified year label converted to the corresponding epoch. The dataset for the CoLiE task consists of around 11K books from literary epochs described at the beginning of Section 2. Each book is split into multiple consequent disjoint 1000-word chunks. Each chunk is provided as a text file. The dataset is divided into training, validation, and testing sets while preserving the epochs ratio in each set. Table 1 summarizes the BookSCE subset compiled for the CoLiE task.

The training and validation sets were released at the beginning of the competition. The test set was released (without labels) a week before the competition's deadline.

The whole dataset with the corresponding ground-truth labels for the train and validation sets can be downloaded from https://www.kaggle.com/competitions/colie/data. Our decision not to publish the ground truth for the test set is primarily due to our plans to organize future editions of the competition. Compilation of a new test set, including its collection and annotation, is very time- and labor-consuming.

2.2. The Competition Format

The competition was hosted on the Kaggle platform https://www.kaggle.com/competitions/colie/ - a popular online platform for data science competitions. Kaggle provides a robust infrastructure for competition management, ensuring a smooth and efficient contest experience for organizers and participants alike. Every Kaggle competition has a public and private leaderboard. Competition hosts split the test dataset into two parts, using one part for the public leaderboard and another part for the private leaderboard, 60% and 40% of the test set, respectively, for this competition. Participants are unaware of which samples are public or private. The public leaderboard is visible to the participants when the competition is alive. The private leaderboard is kept secret until after the competition deadline and is used for determining the final rankings. Therefore, the rankings on the public leaderboard are not necessarily the same as those on the private leaderboard.

The evaluation was based on average accuracy:

$$Acc = \frac{correct\ classifications}{all\ classifications}.$$

In addition, participants were required to provide a short description of their methods together with the confusion matrix for the validation set.

The input to the classifier is a 1000-word chunk of a book in text format, and its output is a single value (epoch). The submission file must contain two columns: one represents the file name (chunk ID) in the test set, and the second is its epoch's label. For convenience, the participants were provided with the "sample_submission.csv" file as an example of the submission format.

3. Participating Teams

Seven teams enthusiastically participated in the contest, six of which agreed to share their identities and briefly overview their methodology. Below we present a summary of the participating methods. Readers who are interested in more details should contact the representatives of the teams.

WebSty. Submitted by Tomasz Walkowiak, CLARIN-PL, Wroclaw University of Science and Technology, Poland.

Each text was vectorized by the TF-IDF weights scaled to z-scores. The method used 5,000 of the most common training set words from the texts for this process.For classification, a multilayer perceptron (MLP) was employed. The network consisted of 5,000 input neurons, two hidden layers (with 1,000 and 500 neurons, respectively), and an output layer (with 5 neurons). The ReLU was used as the activation function in the hidden layers, while SoftMax was applied in the final layer. The dataset includes information about the book identifier for each text. It is in the first column of provided data. This means that texts from the same book can be selected. As an entire book consists of a sequence of texts belonging to the same literary period, the team decided to improve recognition efficiency by leveraging this information [9]. To achieve this, they adopted a sequence classification method proposed in [10], which utilizes logits from the neural classifier trained for classifying individual texts. The logit is the raw output of the final layer before applying the SoftMax activation function to convert it into probabilities. The logits are calculated by combining the weighted sum of the outputs from the last hidden layer with biases. The sequential classification of texts (x_i) from the same book employs the summing of logits ($\sum_i f_c(x_i)$) and is defined as follows:

$$argmax_c \sum_i f_c(x_i)$$

The selected class is assigned to all texts x_i from the same book.

Back to the ... Past. Submitted by Pietro Maldini, an independent participant. From each file provided, stop words and punctuation were filtered out, and some portion of the first words were taken. This dataset with reduced dimensions was used to train a Deep Neural Network using Keras. At first, the documents were vectorized, after that, they were fed to an Embedding Layer to get a representation for each word. This representation was passed through a Bidirecional GRU layer, then through a Dropout layer, a Dense layer, another Dropout

layer, and a final Output layer. The network was trained using AdamW optimizer with a *SparseCategoricalCrossentropy* loss function. The model predicted a Literary Epoch for each fragment of a book. The predictions for each fragment of the book were combined and used to predict the label of the book.

Behrooz Qiassi. Submitted by Behrooz Qiassi, an independent participant. This method uses feature extraction followed by classification. The TF-IDF vectorizer was employed for feature extraction and the Logistic Regression model was used as the classifier.

AMXingu. Submitted by Daniel Quintão de Moraes, Giuseppe Vicente Batista, and Gustavo Pádua Beato, Instituto Tecnológico de Aeronáutica - ITA. The model consists of a three-step pipeline as follows: (1) TF-IDF with sublinear term-frequency[11]; (2) TruncatedSVD (Singular Values Decomposition) with 128 components, which is a sparse version of SVD also known as Latent Semantic Analysis [11, 12]; and (3) an XGBoost classifier with 0.05 learning rate [13]. Words with a relative maximum document frequency above 0.7 and with absolute minimum document frequency below 2 were excluded from the vocabulary in order to avoid stop words and unimportant words, respectively. TruncatedSVD contained 128 components.

Although the dataset (as well as the expected submission format) had been originally split into chunks, the participants concatenated all book chunks belonging to the same book before classification. Accordingly, they made validation and test sets predictions per book and replicated it for all the book's chunks before test submission. The team motivated this step by the fact that a book belongs to a single literary epoch, although some models may benefit from chunk splitting (e.g., deep learning methods with limited input dimension).

Sorbonne University. Submitted by Iglika Nikolova-Stoupak, Kyoto University, Gaël Lejeune, Sorbonne University, and Eva Lacroix, Sorbonne University. The team used a sample of 50,000 entries (while keeping the balance between the 5 labels) as train data and the whole validation set as validation data. The pipeline of the best system consists of the following: (1) Cleaning of the textual data (including removal of capitalization and symbols except common punctuation); (2) Application of the TF-IDF vectorizer from python's *sklearn* library on the textual data with the following settings: *char_wb* analyser with *n-gram* range (5,6); and (3) Training a Logistic Regression model (with the following settings: penalty "l2", C "1", solver "lbfgs").

Debajyoti Mazumder. Submitted by Debajyoti Mazumder, the Department of Data Science and Engineering, Indian Institute of Science Education and Research Bhopal, India. The pre-trained RoBERTa model have been used from huggingface[1]. Pooler output from the pretrained model is taken and a linear layer is stacked on top of it for classification purpose. Only the last layer of RoBERTa-base[14] was trained and the rest layers were frozen. The maximum sequence length 500 was chosen. The learning rate of 2e-4 was chosen with weighted cross entropy loss and AdamW[15] optimizer for mitigating the imbalance in this large dataset. The class weights are given according to the distribution of classes. A stepwise learning rate scheduler

[1] https://huggingface.co/roberta-base

Table 2
Literary epochs classification results.

Team	Accuracy
WebSty	0.79367
Back to the Past	0.77684
Behrooz Qiassi	0.76629
AMXingu[2]	0.76258
Sorbonne University	0.71518
Debajyoti Mazumder	0.65998
The baseline	0.56615

with gamma=0.95 was used, and the model was allowed to run on an early stop strategy with patience=3.

4. Results and Discussion

We received a total of 71 submissions from seven different teams. Each team chose the two best submissions that counted toward the final rank. For comparison purposes, we implemented a very simple baseline–logistic regression applied on normalized count vectors, which achieved an accuracy of 0.566. A summary of the overall rankings of the submitted methods is provided in Table 2. Table 3 shows the confusion matrixes on the validation set. The classification accuracy ranges between 65 and 79 percent. The best results were achieved by the 'WebSty' (first place) and 'Back to the ... Past' (second place) teams. As can be observed, text classification approaches using traditional classifiers and shallow neural networks above classic text representations (such as TF-IDF) outperformed classification with pretrained language model (such as RoBERTa). This outcome is very interesting, given that language models are reported to outperform traditional approaches in most IR tasks. Moreover, it can be observed from Table 3(f) that the method applied RoBERTa obtained the best results for the 'Our days' category, which can be explained by the fact that the RoBERTa was pretrained on modern texts. Also, the representation of documents with TF-IDF vectors seems to be a better option than counting word appearances, as the results of a baseline and two other systems that used the same classifier (Logistic Regression) demonstrate. An additional observation from all confusion matrices is that underrepresented categories have much more misclassifications than the 'Modernism' and 'Victorian' categories, which constitute the majority of the dataset. Also, we noticed that all classifiers did not distinguish very well between adjacent categories. This may be explained by the way literature has evolved, where the changeover between the writing styles representing different epochs was gradual and took place over a long period of time.

[2]After the end of the competition, the team noted that their submission used misspelled labels which degraded their score. The team decided to rerun the model with the correct labels and reported that their scores went up by 3%.

Table 3
Confusion matrices on the validation set. The classes are: (1) Romanticism, (2) Victorian, (3) Modernism, (4) PostModernism, and (5) Our days. Rows: true labels; columns: predicted labels.

	1	2	3	4	5
1	461	526	160	11	0
2	222	14386	1980	283	67
3	62	1164	13451	122	49
4	0	459	720	511	23
5	64	342	578	337	279

(a) TWebSty

	1	2	3	4	5
1	13	966	178	1	0
2	53	12888	3993	4	0
3	2	4332	10505	9	0
4	1	647	1023	42	0
5	0	620	961	19	0

(b) Back to the ... Past

	1	2	3	4	5
1	206	777	174	1	0
2	285	13799	2728	78	48
3	7	3181	11311	88	261
4	20	716	733	221	23
5	37	451	719	147	246

(c) Behrooz Qiassi

	1	2	3	4	5
1	455	570	122	11	0
2	182	14663	1866	227	0
3	0	1891	12875	82	0
4	0	550	660	494	0
5	0	638	571	215	176

(d) AMXingu

	1	2	3	4	5
1	147	826	181	4	0
2	252	13855	2797	23	11
3	13	3046	11586	33	170
4	19	665	831	190	8
5	22	600	869	60	49

(e) Sorbonne University

	1	2	3	4	5
1	575	402	126	30	25
2	1372	10860	2890	844	972
3	130	2085	10510	899	1224
4	59	368	463	626	197
5	57	221	345	368	609

(f) Debajyoti Mazumder

	1	2	3	4	5
1	694	203	142	57	62
2	107	10162	6541	74	54
3	228	1006	7443	597	2574
4	98	278	647	685	5
5	56	470	215	43	816

(g) The baseline

5. Conclusion

This competition provided an opportunity to compare and evaluate the effectiveness of different classification methods to accurately categorize texts into their respective literary epochs. By evaluating the performance of various algorithms and techniques, researchers can determine which methods are most effective in achieving accurate and reliable results. Several interesting observations, some more expected and some less, were reported in this study. In the future, we intend to organize new editions of this competition with new tasks related to literary classification and implicit information retrieval.

Acknowledgments

M. Litvak and I. Rabaev were supported by the Internal SCE grant - Excellence Research Track B, no. EX/06-B-Y22/T1/D3/Yr1. Ricardo Campos and Alípio Jorge were financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC. The authors would like to thank Milana Michaeli for her assistance in manually checking the BookSCE labels.

References

[1] R. Campos, G. Dias, A. M. Jorge, C. Nunes, Identifying top relevant dates for implicit time sensitive queries, Information Retrieval Journal 20 (2017) 363–398. doi:10.1007/

s10791-017-9302-1.
[2] M. Abrams, The Norton Anthology of English Literature (Vol. Package 1: Volumes A, B, C) by, WW Norton & Company, 2012.
[3] I. M. Milne, Literary movements for students: Presenting analysis, context, and criticism on literary movements, Gale, 2009.
[4] Wikipedia, List of literary movements, https://en.wikipedia.org/wiki/List_of_literary_movements#CITEREFMilne2009, 2023. [Online; accessed 03.08.2023].
[5] The Cambridge Companion to British Romanticism, Cambridge Companions to Literature, Cambridge University Press, 1993. doi:10.1017/CCOL0521333555.
[6] The Cambridge Companion to the Victorian Novel, Cambridge Companions to Literature, Cambridge University Press, 2000. doi:10.1017/CCOL0521641500.
[7] The Cambridge Companion to Modernism, Cambridge Companions to Literature, Cambridge University Press, 1999. doi:10.1017/CCOL0521495164.
[8] The Cambridge Companion to Postmodernism, Cambridge Companions to Literature, Cambridge University Press, 2004. doi:10.1017/CCOL0521640520.
[9] T. Walkowiak, Author attribution of literary texts in polish by the sequence averaging, in: L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J. M. Zurada (Eds.), Artificial Intelligence and Soft Computing, Springer International Publishing, Cham, 2023, pp. 367–376.
[10] T. Walkowiak, Authorship attribution of literary texts using named entity masking and maxlogit-based sequence classification for varying text lengths, in: Artificial Intelligence and Soft Computing, Springer, 2023, pp. –. ICAISSC 2023.
[11] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, UK, 2008. URL: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.
[12] F. Pedregosa, G. Varoquaux, 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, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[13] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794. URL: http://doi.acm.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.
[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019).
[15] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam, CoRR abs/1711.05101 (2018).