Classification of Literary Documents Using the Choquet Integral

Charles Planque, Marie-Jeanne Lesot, Grégory Smits

Classification of Literary Documents Using the Choquet Integral

Charles Planque[1], Marie-Jeanne Lesot[2], and Gregory Smits[3]

[1]École Nationale des Chartes – PSL ;
charles.planque@chartes.psl.eu
[2]LIP6 ; marie-jeanne.lesot@LIP6.fr
[3]IMT Atlantique ; gregory.smits@imt-atlantique.fr

Keywords: classification, coalitions, metadata, literary documents classification

Introduction: Motivation and Objective
In literature-oriented Digital Humanities (DH), classifying documents into empirically observed categories can help researchers better understand those concepts they are studying. In machine learning terms, this task corresponds to a classification problem. We address this task with the following characteristics: 1) the considered features are the texts metadata, e.g. author, title and publication date, in addition to the text content; 2) the features are considered both individually and through their coalition, i.e when several features can reinforce each other to influence the prediction. 3) the considered concepts correspond to the author's gender or to more complex, possibly imprecisely defined, categories, e.g. whether the texts are canonical;

Proposed Approach
We propose to use the AI method called CHOCOLATE [4] that offers the following properties: (i) capture of imprecise class definitions (as is for instance the case of the text canonicity) in the formalism of fuzzy sets, (ii) taking into account feature coalitions (iii) use of tiny sets of class examples. To do so, it relies on the aggregation operator called Choquet integral (see [4] for a formal definition, omitted in this short paper). More precisely, CHOCOLATE determines a membership degree for each class aggregating two main factors: the relevance of each text feature and that of the coalitions of these features. It assigns weights to the relevance of coalitions and individual features.

Related Work
Several DH papers provide broad overviews on the classification of literary cor-

pora, but few of them take metadata into account. Specifically, in [5] the performances of many combinations of text classifiers are studied, giving an overview of embedding techniques, but without incorporating metadata. Moreover, it is not possible to easily determine which feature or coalition of features in the document leads to its class assignment. Metadata are exploited in [3] but the method can be seen as a black-box model, due to the neural network applied at the end of the classification pipeline. As a result, it does not support the DH goal of using classification to generate new insights into literary corpora. Along the same lines, the MATCH framework [6] uses a transformer model at the end of its classification pipeline. Moreover, these techniques do not consider the use of attribute coalitions. Another benefit of Chocolate is that it only needs to be fed a few data which are examples representing the positive class, and it has no training phase. Therefore, it is computationally lighter and focuses on a qualitative choice of data instead of scaling data quality over quantity. This kind of approach corresponds to methods traditionally used in humanities works.

Experimental Protocol and Results

The experiments apply the proposed approach to predict the author's gender and the text canonicity (see [2]). Both tests are conducted on the ANRChapitres dataset[1], an XML-formatted anthology of 2000 French books published between 1811 and 1994. Each file contains the text and metadata. The experimental protocol is the following: textual data are transformed into probabilities by TF-IDF+Logistic regression. CHOCOLATE is then to aggregate these probabilities, returning a membership degree for each document. Scores are calculated from the relevance of both singular features and coalitions for each document, based on a continuous similarity measure to best use prediction values. Ultimately, a threshold separates the two classes according to membership degrees.

Those results are achieved with a positive class of size 1700 or 1400 for each task. A treatment with more traditional machine learning techniques would have needed a lot more data.

In both experiments, Chocolate outperforms the baselines considered, which are simpler. For canon prediction, it achieves **0.92** accuracy / **0.86** F1, compared to 0.71 / 0.33 for TF-IDF+Linear SVM and 0.75 / 0.76 for TF-IDF+Logistic regression. For gender prediction, Chocolate reaches **0.98** accuracy / **0.98** F1, again exceeding standard TF-IDF and regression baselines.

Conclusion and Future works

This study presents an innovative methodology using the Choquet integral to classify literary documents, participating in increasing diversity of techniques in DH. Additionally, it takes into account feature coalitions that can generate interesting insights into their interactions, incorporating text metadata. Through this study, CHOCOLATE demonstrates that advanced aggregation methods can enrich DH research by creating accurate classification possibilities for literary

documents.

Future work will explore its application to more diverse corpora with different sets of meta-data and other concepts to classify.
Additionally, Chocolate allows for insights into its classification choices. It would then be possible to exploit them in order to get explanations of the results given, in an XAI perspective. Thus, since the factors leading to the algorithm's choices can be extracted and interpreted, they could highlight relevant attribute coalitions in the class, and detections of previously unnoticed biases. Those could be, for example, that texts considered as canons are mostly written by men, even more so before the XXth century. Detecting biases and patterns, even the more obvious ones, could help make sure the prediction model detects what it is aiming at and potentially lead to questioning previous definitions of concepts, reshaping our idea of concepts such as canonicity or gender.

Acknowledgments: this work has been funded by the École Nationale des Chartes – PSL.

References

[1] ANRChapitres. ANRChapitres/2000romans19e20e: Corpus Chapitres, December 2022.

[2] Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature. *Journal of Cultural Analytics*, 8(3), October 2023.

[3] Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. Enriching BERT with Knowledge Graph Embeddings for Document Classification. *Proc. GermEval 2019 Workshop*, 2019. Version Number: 1.

[4] Grégory Smits, Ronald R. Yager, Marie-Jeanne Lesot, and Olivier Pivert. Concept Membership Modeling Using a Choquet Integral. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 1237, pages 359–372. Springer International Publishing, Cham, 2020.

[5] Oleg Sobchuk and Artjoms Šeļa. Computational thematics: comparing algorithms for clustering the genres of literary fiction. *Humanities and Social Sciences Communications*, (1):438, March 2024.

[6] Yu Zhang, Zhihong Shen, Dong Yuxiao, Kuansan Wang, and Jiawei Han. Match: Metadata-aware text classification in a large hierarchy, 02 2021.