Exploring Neural Topic Modeling on a Classical Latin Corpus

Ginevra Martinelli[1], Paola Impicciché[2], Elisabetta Fersini[2], Francesco Mambrini[1], Marco Passarotti[1]
[1]Università Cattolica del Sacro Cuore di Milano,
[2]Università degli Studi di Milano Bicocca
Largo Fra Agostino Gemelli, 1, 20123 Milano (Italy),
Viale Sarca 336, 20126 Milano (Italy)
ginevra.martinelli01@icatt.it, p.impicciche@campus.unimib.it,
elisabetta.fersini@unimib.it, francesco.mambrini@unicatt.it, marco.passarotti@unicatt.it

Abstract

The large availability of processable textual resources for Classical Latin has made it possible to study Latin literature through methods and tools that support distant reading. This paper describes a number of experiments carried out to test the possibility of investigating the thematic distribution of the Classical Latin corpus Opera Latina by means of topic modeling. For this purpose, we train, optimize and compare two neural models, Product-of-Experts LDA (ProdLDA) and Embedded Topic Model (ETM), opportunely revised to deal with the textual data from a Classical Latin corpus, to evaluate which one performs better both on the basis of Topic Diversity and Topic Coherence metrics and from a human judgment point of view. Our results show that the topics extracted by neural models are coherent and interpretable and that they are significant from the perspective of a Latin scholar. The source code of the proposed model is available at https://github.com/MIND-Lab/LatinProdLDA.

Keywords: Topic modeling, Classical Latin, Neural Models

1. Introduction

During the last decades, many digital resources and natural language processing tools have been developed for ancient languages. In particular, Latin has drawn much attention due to the large size of its textual tradition and its cultural significance spread across two millenia. Though most languages prove to evolve over the centuries, Latin, after an initial period of stabilization in the first millennium BCE, did not undergo radical developments until Late Antiquity. Moreover, the evolutionary phase of the Latin language known as "Classical Latin" - which covers the period from the end of the 2nd century BCE to the 2nd century CE and is essentially the language of the most famous Latin works that are still studied in schools today - is even more homogeneous. The stability of Classical Latin has significant research implications as it favors cross-text content studies on the most representative centuries of Latin literature.

Since the time of Humanism, these texts have been studied and investigated through an approach based on the so-called "close reading" (Moretti, 2000), which has permitted to penetrate between the folds of words. In contrast, several methods of "distant reading" (Moretti, 2000) have been developed to uncover the latent thematic structure of datasets. In particular, topic modeling has been a popular and successfully applied technique in the Digital Humanities (Jockers, 2013).

Our aim is to test how the latest developments in the area of topic modeling help to investigate the Latin literary heritage and to evaluate the performances of the state-of-the-art models. For this purpose, we use OCTIS (Terragni et al., 2021a) that is a framework that trains, optimizes and compares different models, to study the topic distribution of a Classical Latin dataset, i.e. the "Opera Latina" (Dominique Longree and Margherita Fantoli, 2023) corpus built by the LASLA laboratory. In particular, we provide two main contributions: (i) we extend a set of neural models to deal with Classical Latin corpora and (ii) we perform a quantitative and qualitative evaluation of those models, to highlight their strength and how they improve on statistical-based topic models.

2. Related Work

Topic Models. Topic models are a class of unsupervised machine learning techniques aimed at extracting the underlying themes from a document corpus. Recently, Neural Topic Models (NTMs) have been developed to address the limitations of statistical-based topic models (e.g. LDA (Blei et al., 2003)), which usually assume that words are generated independently of others. Among them, Srivastava and Sutton (2017) proposed Product-of-Experts LDA (ProdLDA), an algorithm based on Variational Autoencoder (VAE) (Kingma and Welling, 2014). ProdLDA replaces the mixture model in LDA with a product of experts, leading to an increase in the number of interpretable topics.

LREC-COLING 2024, pages 6929–6934

Instead, Dieng et al. (2020) proposed Embedded Topic Model (ETM), which incorporates word embeddings so that semantic relatedness between terms is taken into account when extracting topics.

Topic modeling on Classical Latin literature. Regarding ancient languages, a few of them have been explored using traditional statistical-based topic models. Wishart and Prokopidis (2017) employed LDA to extract topics from Hellenistic corpora in Ancient Greek and compared lists of semantically related words. Köntges (2020) "measured philosophy" in the first thousand years of Greek literature through LDA and developed a multilingual topic modeling tool for Greek, Latin, Arabic and other languages. Topic modeling studies have also been conducted on Ancient Chinese (Allen et al., 2017) and Sanskrit literature (Neill, 2019). To the best of our knowledge, there is no work that has applied topic modeling to Classical Latin literature.

3. Exploring the LASLA Latin Corpus using Neural Topic Models

3.1. Opera Latina

The datasets used for our work are obtained from the Classical Latin corpus "Opera Latina" built by the LASLA laboratory in Liège, which consists of 130 texts, both poetry and prose, composed by 21 authors, for a total of 1,700,000 lemmatized and morphosyntactically analyzed tokens. Since topic models estimate the distribution of topics over a collection of documents, we choose two different units to partition the lemmatized corpus. One of the datasets comes from partitioning the corpus into sentences ("sentence-based collection"),[1] while the other consists of entire literary works or single books ("work-based collection") based on LASLA files. Both sentence-based and work-based batches are then pre-processed to avoid highly frequent words over-influencing the outputs. In particular, both datasets are filtered according to the Part-of-Speech (POS) tags in order to keep only the content words, i.e. nouns, verbs and adjectives. Moreover, we exclude the term *sum* 'to be' from the sentence-based collection and we eliminate *sum* 'to be', *possum* 'to be able', *facio* 'to do', *dico* 'to say', *res* 'thing', and *video* 'to see' from the work-based collection. The main characteristics of the final datasets are reported in Table 1.

3.2. Topic Models

In order to investigate the distribution of topics in the "Opera Latina" corpus, we consider ETM and ProdLDA as candidate neural topic models. While

	Sentences	Works
n. of batches	92214	236
vocabulary size	21841	21836
n. of tokens	986197	943627

Table 1: The datasets

ProdLDA is naturally suitable to deal with any type of corpus, thanks to its ability to encode any document with a Variational Autoencoder, the original ETM needs to be properly extended. In particular, we extend ETM by exploiting a set of Lemma embeddings for the Latin language (Rachele Sprugnoli, Giovanni Moretti and Marco Passarotti, 2020). Lemma vectors are built on the LASLA corpus with Continuous Bag-of-Words (CBOW) architecture. In our case, we use vectors of dimension 100 in word2vec format.[2] For the sake of completeness, we compare the two neural models with LDA as baseline. The final model **Latin Product-of-Expert** (**L-ProdLDA**) is available as extension of OCTIS. To guarantee a fair comparison of the models considered, they are optimized according to a Multi-Objective Bayesian Optimization Strategy (MOBO). In particular, the optimal hyper-parameter configuration of each model is determined by the simultaneous maximization of Topic Coherence, i.e. Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014), and Topic Diversity (TD) (Dieng et al., 2020). Hyper-parameters and their values are reported in Table 2.

Model	Hyper-parameter	Values/[Range]
LDA	Num. of topics	[20, 50]
	α	[0, 2]
	β	[0, 2]
ProdLDA	Number of topics	[20, 50]
	Dropout	[0, 0.60]
	Num. of neurons	50, 100, 200, 300
	Num. of layers	1, 2
	Activation function	softplus, relu, sigmoid
ETM	Num. of topics	[20, 50]
	Dropout	[0, 0.60]
	Hidden size	50, 100, 200, 300
	Activation function	softplus, relu, sigmoid

Table 2: Hyper-parameters and values

The selection of the number of topics and the other hyperparameters is targeted by the multi-

[1] The division into sentences derives from the syntactic annotation provided by the LASLA corpus.

[2] Embeddings available at https://embeddings.lila-erc.eu/samples/download/word2vec/

objective optimization, according to the corpora under consideration. The multi-objective optimization determines such hyper-parameters that simultaneously optimize the coherence of each topic and the diversity of the set of topics, which should lead to interpretable results.

3.3. Evaluation Metrics

The quality of topics is measured using the two target metrics, i.e. NPMI, to measure Topic Coherence and Topic Diversity to measure the diversity of topics. Only for ETM we estimate the Word Embedding-Based Pairwise Similarity (WEPS) (Terragni, Fersini, and Messina, 2021b), a similarity measure that captures the extent to which the words appearing in two different topics tend to be close to each other in the word embedding space. In order to measure WEPS, we made use of the same Latin embeddings used for L-ProdLDA.
In our experiments we compute the above metrics using the top-10 words from each topic. We determine the overall quality of a model's topics by considering both the quantitative metrics and human judgment.

4. Results and Evaluation

Table 3 shows the results of the topic modeling experiments. The values of the selected metrics are averaged over 30 runs for each model.

Model	Dataset	NPMI	TD	WEPS
LDA	sentences	-0.06	0.99	-
	works	-0.002	0.07	-
L-ProdLDA	sentences	0.009	0.89	-
	works	-0.07	0.85	-
ETM	sentences	0.002	0.79	0.04
	works	0.02	0.90	0.03

Table 3: Metrics results

It can be observed that L-ProdLDA (and LDA as well) trained on the sentence-based dataset outperforms those trained on the work-based dataset both with Topic Diversity and Topic Coherence measure. L-ProdLDA appears to be the best model according to the quantitative metrics, while LDA fails at generating non redundant topics for work-based input. ETM shows good performance over all the metrics and it is the strongest model in dealing with the work-based collection in terms of NPMI and TD.

With regard to human evaluation, we assess the interpretability of a topic by considering the top-10 words. If these words are semantically related and allow us to assign a label to the topic, we classify it as interpretable. According to the expertise of multiple interpreters familiar with Latin literature, L-ProdLDA trained on the sentence-based collection is the model that predicts the highest number of interpretable topics.[3] Table 4 shows some of the topics predicted through this model.

Label	Top words
War	*miles* 'soldier' *hostis* 'enemy' *dux* 'commander' *eques* 'knight' *signum* 'war banner'
Nature	*mare* 'sea' *terra* 'land' *unda* 'wave' *mons* 'mountain' *ventus* 'wind'
Court	*crimen* 'crime' *causa* 'case' *poena* 'penalty' *scelus* 'criminal deed' *iudex* 'judge'

Table 4: L-ProdLDA topics

Figures 1-4 summarize a few insights of the experiment. Figure 1 illustrates the topic distribution of the works of each author included in the LASLA corpus. In Figure 2 we evaluate the weight of three topics in Vergilius' production. The bubble charts in Figures 3 and 4 display the top words of two topics scaled according to their probability of being part of them (Word Importance) and according to the number of their occurrences in the corpus (Word Count).

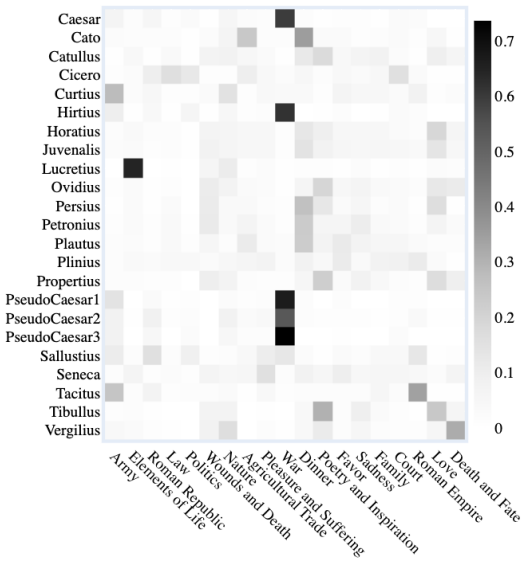

Figure 1: Topic distribution of the LASLA corpus

[3] In particular, LDA predicts 4 interpretable topics out of 47, L-ProdLDA 19 out of 30 and ETM 12 out of 29.

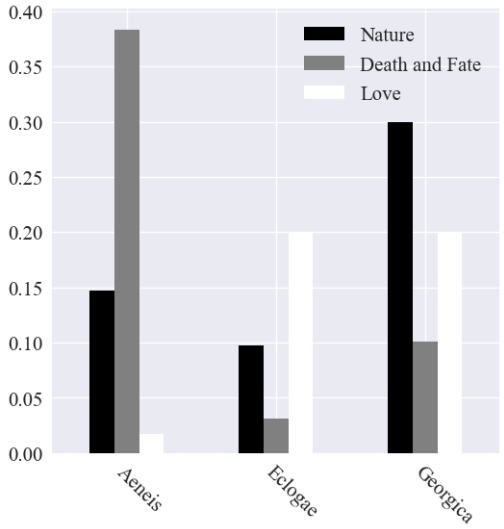

Figure 2: Weight of some topics in Vergilius

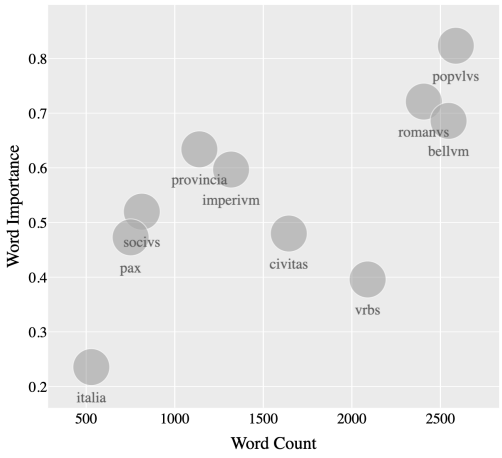

Figure 3: Bubble chart of the topic “Roman Republic”

The thematic structure predicted in this experiment generally aligns with the expectations of a Latinist. To illustrate, in Figure 1, we can observe that Caesar, renowned for his works such as the *Commentarii De Bello Gallico* and *De Bello Civili*, which document his military campaigns, predominantly engages with the topic we labeled “War”. Conversely, for Lucretius, the author of *De Rerum Natura*, a didactic poem explicating the universe through Epicurean physics, “Elements of Life” emerges as the predominant theme.

Even these seemingly straightforward observations, which might be anticipated by a scholar of Latin literature, are nonetheless intriguing as they allow for an evaluation of the significance of each theme for the considered authors. In the cases mentioned, for instance, it is evident how Caesar and Lucretius predominantly focus on specific themes (albeit broad ones), which almost exclusively constitute the subjects of their works. This

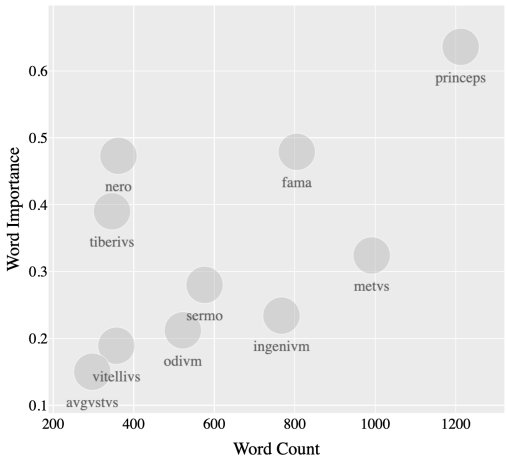

Figure 4: Bubble chart of the topic “Roman Empire”

specificity thus characterizes Caesar’s historiography (and that of the other authors of the so-called *corpus Caesarianum*) and distinguishes his thematic choices from those of other historians, such as Sallustius, who opts to cover a much wider range of topics.

However, the type of visualization provided by a heatmap also proves useful in uncovering unpredictable patterns that can inspire further literary investigations. For instance, we can observe the pervasive nature of a theme like “Wounds and Death”, which seems to be a constant focus for Latin authors. In fact, this theme encompasses not only the metaphorical wounds of romantic separation for an elegiac author like Propertius but also the wounds received and inflicted by epic heroes narrated by Vergilius, the wounds of Senecan tragic characters, and even, in a way, the processes of atomic disintegration leading to death in Lucretius’ *De Rerum Natura*.

A more nuanced analysis can also be conducted by delving into the works of individual authors and examining the trends of specific themes therein (Figure 2). In the case of Vergilius, a close examination reveals distinct thematic focal points across his three major works. In the epic poem *Aeneid*, which narrates the vicissitudes of the hero Aeneas, “Death and Fate” emerges as fundamental motif. In contrast, in the pastoral poetry of the *Eclogae*, where young shepherds lament their unrequited loves amidst idyllic landscapes, “Love” reigns supreme as the most recurrent theme. Meanwhile, in the *Georgica*, which expounds upon agriculture, “Nature” takes center stage, drawing parallels between human existence and the natural world. Of particular intrigue is the notable presence of “Love” within this work, traditionally perceived as a didactic poem focused on the agricultural world. This departure from convention signals the author’s intent to transcend the confines of genre and engage with a

broader audience.

Furthermore, it is possible to focus on individual themes to delve deeply into their nature and significance within Latin literature. In this regard, it becomes apparent that even minor distinctions between similar subject areas hold significance (Figure 3 and Figure 4). For instance, while "Roman Republic" and "Roman Empire" may fall under the broad category of "Government of Rome", they are identified by the tool as distinct topics with nuanced differences. This distinction underscores the complex dynamics within Roman governance. On one hand, "Roman Republic" is characterized by a horizontal power structure, epitomized by concepts such as *populus* 'population', *civitas* 'citizenship' and *Italia* 'Italy'. Conversely, "Roman Empire" introduces a vertical hierarchy, symbolized by figures like *Augustus*, *Nero*, *princeps* 'prince' and *ingenium* 'intelligence'. These subtle delineations shed light on the intricacies of Roman political and social systems, providing valuable insights into the evolution of governance and power dynamics throughout history.

5. Conclusions and future work

In this paper, we tested how neural topic models can be used to detect the thematic distribution of a large portion of Classical Latin literature. Our results show that (i) neural topic models ProdLDA and ETM, trained on the datasets considered, provide a promising starting point to investigate the distribution of topics in a Classical Latin corpus; (ii) that the outputs obtained are meaningful from the point of view of a scholar of Latin.

In particular, claim (ii) allows us to infer that the traditional way of studying ancient literature can be fruitfully complemented by that based on distant reading, made possible by the development of topic modeling techniques and by the large availability of processable textual resources. In fact, while through close reading it may be possible to state that an author addresses a specific topic, the systematic analysis derived from a topic modeling experiment on an extensive collection of Latin literature enables to discern quantitatively the extent to which a considered topic is explored in a work or in a group of works. Topic modeling makes it possible to examine the degree to which other authors within the corpus discuss the same topic and it provides a basis for comparing the thematic distributions of different authors. Furthermore, it empowers the identification of unforeseen patterns and trends that may prompt additional inquiries. In other words, it provides a holistic perspective on thematic exploration across a spectrum of works and authors. This methodological advancement leads to the possibility of acquiring new knowledge, since Latin literature can thus be considered not only in its separate constituents, but as an articulated whole. Interdisciplinarity is one further added value for achieving these goals. The development of the tools and models requires, on the one hand, competence in statistics and computational knowledge; but, on the other, the interpretation of the outputs carried out by a Latinist is essential for the qualitative evaluation of the results.

Our next step is to extend the dataset diachronically and synchronically in order to broaden the coverage of Classical literature and the temporal span that can be analyzed with regard to the thematic distribution, but the goal is above all to impact literary research in ancient languages so that this field can progress more and more in depth thanks to the use of advanced tools and large sets of empirical evidence provided by language resources. An additional forward step regards the applicability of the proposed frameword to the analysis of smaller corpora. In order to use the L-ProdLDA on a small-scale corpora, two main strategies can be adopted. The first one is to train the model on large corpora and adopt a transfer learning approach to adapt the model parameters to the smaller one. The second approach is to learn parameters and hyper-parameters on large corpora that are similar in terms of distribution shape and adopt the learned parameters on the smaller corpora. However, to the best of our knowledge, the research related to transfer topic models is still unexplored.

Acknowledgments

The work of Elisabetta Fersini has been partially funded by MUR under the grant *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca, Italy.

6. Bibliographical References

Colin Allen, Hongliang Luo, Jaimie Murdock, Jianghuai Pu, Xiaohong Wang, Yanjie Zhai, and Kun Zhao. 2017. Topic modeling the Hàn diăn Ancient Classics. *Journal of Cultural Analytics*, 2(1).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3:993–1022.

Adji Bousso Dieng, Francisco J.R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguistics*, 8:439–453.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *The International Conference on Learning Representations (ICLR)*, Banff.

Thomas Köntges. 2020. Measuring philosophy in the first thousand years of Greek literature. *Digital Classics Online*, 6:2.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 530–539.

Franco Moretti. 2000. Conjectures on world literature. *New left review*, 2(1):54–68.

Tyler Neill. 2019. LDA topic modeling for pramāṇa texts: A case study in Sanskrit NLP corpus building. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 52–67, IIT Kharagpur, India. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference of Representation Learning*.

S. Terragni, E. Fersini, B. Galuzzi, P. Tropeano, and A. Candelieri. 2021a. OCTIS: Comparing and optimizing topic models is simple! In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*, pages 263–270. Association for Computational Linguistics (ACL).

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems*, pages 33–45, Cham. Springer International Publishing.

Ryder Wishart and Prokopis Prokopidis. 2017. Topic modelling experiments on Hellenistic corpora. In *Proceedings of the Workshop on Corpora in the Digital Humanities*.

7. Language Resource References

Dominique Longree and Margherita Fantoli. 2023. *LASLAfiles_Latin_BPNFormat_SharedwithDTA_2019*. ULiège Open Data Repository.

Rachele Sprugnoli, Giovanni Moretti and Marco Passarotti. 2020. *Latin embeddings*. Zenodo.