LiMe: a Latin Corpus of Late Medieval Criminal Sentences

Alessandra Bassani[1], **Beatrice Del Bo**[2], **Alfio Ferrara**[3], **Marta Mangini**[2], **Sergio Picascia**[3], **Ambra Stefanello**[4]

[1]Università degli Studi di Milano, Department of Italian and Supranational Public Law
[2]Università degli Studi di Milano, Department of Historical Studies
[3]Università degli Studi di Milano, Department of Computer Science
[4]Università degli Studi di Firenze, Department of History, Archaeology, Geography, Fine and Performing Arts

Abstract

The Latin language has received attention from the computational linguistics research community, which has built, over the years, several valuable resources, ranging from detailed annotated corpora to sophisticated tools for linguistic analysis. With the recent advent of large language models, researchers have also started developing models capable of generating vector representations of Latin texts. The performances of such models remain behind the ones for modern languages, given the disparity in available data. In this paper, we present the LiMe dataset, a corpus of 325 documents extracted from a series of medieval manuscripts called *Libri sententiarum potestatis Mediolani*, and thoroughly annotated by experts, in order to be employed for masked language model, as well as supervised natural language processing tasks.

Keywords: latin corpus, medieval case law, natural language processing

1. Introduction

The manuscripts called *Libri sententiarum potestatis Mediolani*, preserved at the *Archivio Storico Civico and Biblioteca Trivulziana in Milan, Cimeli, 146-152*, represent all that remains of the documentation recorded in the late medieval period at the court of justice of the city of Milan. The seven manuscripts of the series cover the activity of the court during the years 1385, 1390-1392, 1397-1398, 1398-1399, 1400-1401, 1427 and 1428-1429, respectively, resulting in the delivery of approximately 3,000 criminal sentences[1] discussed in the presence of the Milanese judges, pronounced by the *podestà*[2] and publicly recorded by the notaries who worked at the court in the *Loggia degli Osii*[3]. Although, as evident, the chronological span of each *Liber* varies considerably according to the length of time each *podestà* was in office, the structure, the material aspect and even the form employed in the drafting of these manuscripts present elements of a certain homogeneity and uniformity. This is due to the fact that the notaries in charge of assisting mayors and judges during trials recorded the sentences according to a pattern that is repeated almost unchanged in all manuscripts.

Each verdict, preceded by the verbal invocation - *In nomine Domini, amen*[4] - is pronounced by the *podestà* in accordance with the seigniorial decrees and statutes of the municipality of Milan. It contains the names of the accused, the narration of the legal proceeding, whether it was an *inquisitio* or an accusation, with the salient phases of the trial and the final pronouncement. In addition to the sentences, whose pattern is formally identical for all defendants, there are also numerous subsequent interventions: e.g. annotations relating to receipts for full or partial payment of penalties or cancellations of sentences.

The *Libri sententiarum potestatis Mediolani* are pivotal sources for law historians, like all Medieval and Early Modern trial outcomes preserved in the European archives: they allow us to measure the distance between the discipline established by *statuta* and *ius comune* and its actual application before the courts of medieval cities (Padoa-Schioppa, 2017). Indeed, the seven *Libri* photograph the complex balance of social and political forces that characterised the city of Milan during the Visconti rule (Gamberini, 2014).

This documentary typology constitutes a source of great importance for historians of medieval law (Storti, 2021; Valsecchi, 2021; Bassani, 2021; Isotton, 2021; Bianchi Riva, 2021; Minnucci, 2021), meanwhile fulfilling the same function for medievalists tout court. It provides inspiration for those who deal with political and institutional history, since it allows one to investigate in practice the dynam-

[1]Throughout the article, the term "sentence" will be used with its meaning of *a punishment that a judge gives to someone who has committed a crime*.

[2]A chief magistrate of a medieval Italian town.

[3]A historical building of Milan, from whose balcony sentences and edicts were proclaimed by the Milanese judges.

[4]*In the name of the Lord, amen.*

LT4HALA 2024@LREC-COLING-2024, pages 41–49

ics of the exercise and management of power, the men, the methods and timing through which justice is administered, including through the selection of judges (Pagnoni, 2021); at the same time, a collection of sentences issued by a city lord provides very useful elements for the study of society and economy, through the analysis and reconstruction of the type of crime, its scene and circumstances, the weapons used, the profiles of the people involved, including their reputation, qualification and profession.

In this article, we present the LiMe dataset, an annotated Latin corpus consisting of 325 judicial documents from the first volume of the *Libri sententiarum potestatis Mediolani*. We illustrate the process undertaken for digitizing the documents and annotating them with detailed information, such as entities and relations, in order to make the manuscript more accessible and valuable to researchers. The paper is structured as follows: Section 2 provides the motivations behind this research; Section 3 outlines relative contributions in the field literature; in Section 4 we define how the data has been extracted and the final structure of the LiMe dataset; Section 5 gives examples of possible statistical and machine learning applications; in Section 6 we discuss the results and the future steps.

2. Motivation

The study of society through the filter of the judicial machine allows a better understanding of the objectives of "political discipline" and the effectiveness of this governing instruments (Campisi, 2019; Luca, 2021). At the same time, the registers of sentences still preserved in the archives of Italian cities of the last centuries of the Middle Ages, constitute a valuable field of research for those who deal with the history of gender in the medieval age (Del Bo, 2021; Dean, 2008). The analysis of such documentation on the basis of the interpretative categories typical of this historiography benefits from the possibility of questioning the source on the characteristics of alleged victims and perpetrators, the type of condemnation/absolution, the granting of pardon (*gratia*), the timing of the execution of the sentence, the type of crime, the weapons used, the place and circumstances of the offence (*delictum*), single or group action, the presence of accomplices or leaders and their gender, the personal/familial condition, the words used to identify and define each person, to mention only a few aspects of the research. Starting from the identification modalities of women and men from the language of sentences, exploiting qualifying attributes, the source offers the possibility of dismantling stereotypes and historiographical clichés.

Despite their undoubted relevance, the *Libri sententiarum potestatis Mediolani* have received little, if any, historiographical attention overall. In fact, they have not been taken into account in wide-ranging studies dedicated to the subject of the documentation issued by medieval Italian judicial bodies (Giorgi et al., 2012; Lett, 2021; Dean, 2007; Vallerani, 2012) and, until very recent years, few scholars have dealt with them specifically (Verga, 1901; Santoro, 1968; Padoa-Schioppa, 1996; Covini, 2012). The first manuscript in the series contains 126 criminal sentences pronounced by the *podestà* of Milan Carlo Zen (1385). This manuscript was recently edited by (Pizzi, 2021) and analysed in (Bassani et al., 2021).

3. Related Work

Despite being a dead language with far less resources with respect to modern languages, Latin has recently received significant attention from the research community, in both the production of annotated datasets and the training of language-specific models.

3.1. Latin Corpora

Several projects are currently dealing with the digitization and annotation of a considerable amount of Latin texts, often coming from different sources, with the purpose of being explored and exploited by history and linguistics scholars. Some of these corpora mainly present detailed syntactic and morphological annotations. It is the case of the five Latin Universal Dependencies[5] treebanks: PROIEL (Haug and Jøhndal, 2008), Perseus (Bamman and Crane, 2011), ITTB (Passarotti, 2019), LLCT (Cecchini et al., 2020), UDante (Flavio et al., 2020). LatinISE (McGillivray and Kilgarriff, 2013) is a Latin corpus for Sketch Engine, gathering documents from different websites; the corpus can be searched through the usage of tokens (13 million those present in the documents), or filtered on metadata, such as the author or the time period of each work. The LIRE (Kaše et al., 2021) dataset is another example of data integration, collecting Latin inscriptions dating back to the Roman Empire from two sources: the Epigraphic Database Heidelberg[6] (EDH) and the Epigraphik Datenbank Clauss-Slaby[7] (EDCS). The Opera Latina corpus (Denooz, 2007), created and maintained by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) includes 154 works from 19 classical Latin authors. The recent LiLa[8] (Passarotti et al., 2020) (Linking Latin) project has the object of building

[5] https://universaldependencies.org/la/
[6] https://edh.ub.uni-heidelberg.de
[7] http://www.manfredclauss.de
[8] https://lila-erc.eu

a common knowledge base, capable of describing several scattered Latin datasets with a unique vocabulary.

There are just a few cases of Latin corpora presenting detailed annotations for a specific task. The dataset presented in (Besnier and Mattingly, 2021) contains proper nouns of people and places in three Medieval languages, Latin included; the dataset can be employed to build named entity recognition (NER) models for low-resource languages. Addressing the task of authorship analysis, MedLatinEpi and MedLatinLit (Corbara et al., 2022) are two datasets consisting of 294 and 30 curated texts, respectively, labelled with the respective author; MedLatinEpi texts are of epistolary nature, while MedLatinLit texts consist of literary comments and treatises about various subjects.

Regarding legal texts, the Justinian's Digest has been digitized and included in a relational database (Ribary, 2020): the texts can be accessed and filtered, querying information about jurists, thematic sections and compositional structure.

3.2. Latin Language Models

In recent years, both non-contextual and contextual embedding models have been exploited for the representation of Latin text. In (Burns et al., 2021) the authors train a word2vec model on a large Latin corpus, achieving state-of-art performances on synonym detection and inter-textual search. Latin BERT (Bamman and Burns, 2020) is a contextual language model for Latin, trained on a large corpus spanning over twenty-two centuries; a fine-tuned version of Latin BERT (Lendvai and Wick, 2022) has been proposed for a word sense disambiguation task.

LatinCy (Burns, 2023) is an entire Latin NLP pipeline built for the Python library spaCy (Honnibal et al., 2020): it consists of several models, capable of performing part-of-speech tagging, dependency parsing, and named entity recognition. Stanza (Qi et al., 2020) is a collection of tools and models for the linguistic analysis of many human languages, including Latin, trained on Universal Dipendencies treebanks. UDPipe (Straka, 2018) is a pipeline for tokenization, tagging, lemmatization and dependency parsing, trainable on CoNLL-U files.

Shared tasks are being proposed in order to foster research in the field of language technologies for Classical languages. The EvaLatin 2022 Evaluation Campaign (Sprugnoli et al., 2022) proposed three tasks relative to lemmatization, part-of-speech tagging, and features identification.

4. Dataset

LiMe[9] (Bassani et al., 2024) is a publicly available Latin corpus consisting not only of criminal sentences, but also of many additional notes gathered from the first manuscript of the *Liber sententiarum potestatis Mediolani* (1385-1429), the oldest known registers of criminal sentences for the city of Milan. The original source, preserved in very good conditions and presenting just three mutilated texts, has been edited and transcribed in the curated edition (Pizzi, 2021). The texts have then been digitized and annotated in the context of the Fight Against Injustice Through Humanities (FAITH) project (Ferrara et al., 2023b), whose main objective is to provide common tools and methodology for the collection, digitization and integration of different historical sources. For each document, named entities, relations between them and events have been manually identified; moreover, the texts have been classified depending on the type of document and, in case of criminal sentences, they have been segmented according to a predefined annotation schema. The result is a collection of 325 documents, made of 87110 tokens, in Latin language. The annotations, performed by a team of experts, have been organized according to a custom schema; an example of the annotations is provided in Section 4.2.

4.1. Data Extraction

The main source of information in the manuscript are the criminal sentences, gathered in dossiers and ordered according to an arbitrary number given from the curator, e.g. *Sentenza I.1* refers to the first (1) judgment from the first (I) dossier. Each dossier is usually opened by a "protocol", i.e., a textual section in which the notary explicitly declares his identity and announces, following a very precise formulary, the name of the judge and *podestà* who presided over the trials. The "eschatocol" is the section closing each dossier, where the notary refers to the group of judgments he has transcribed, citing the witnesses present. Additionally, there are three other types of sources, constituting supplementary information to the judgements: an "addendum" is a document added later to the text of the judgment, indicating further developments happened after the end of the trial; an "insert" is a piece of text, reported within a judgment or addendum, usually certifying orders received from the *podestà*; finally, a "news" is an indirect evidence of an order or document that existed at the time but was not transcribed, useful in justifying decisions made by authority or actions taken by officials.

[9] https://doi.org/10.13130/RD_UNIMI/EN2TFH

The texts of criminal sentences, being them legal texts (thus with a rigid structure and a content pattern based on formulas), present the same sections and reflect a precise and largely stable structure. At the beginning, sometimes there it lies the *significatio*, i.e., the communication of the misdemeanor(s) to the *podestà* by a faithworthy person, the elder of the parish, in charge of the surveillance of a living area; this communication, however, did not always occur, so it is not always found in the text. The following part of the judgment, the *inquisitio*, narrates the events that occurred as they were reconstructed: here, the details regarding each misdemeanor (*misfatto*) are reported, such as the criminal offences, the perpetrator of the violence, the victim and any item involved. The motivational section (*motivazioni*), usually introduced by the words *qua de causa* ("the cause of"), *et predicta* ("and the aforesaid") or *et constat nobis* ("and it is agreed with us"), states the reason why the verdict was reached. Finally, the last part of the sentence consists of the decision (*dispositivo*) of conviction or acquittal and, in the former case, also of the type and amount of punishment; it generally begins with the word *idcirco* ("therefore", "about that"). A summary of the structure of a typical dossier with details on the form of a judgment is depicted in Figure 1.

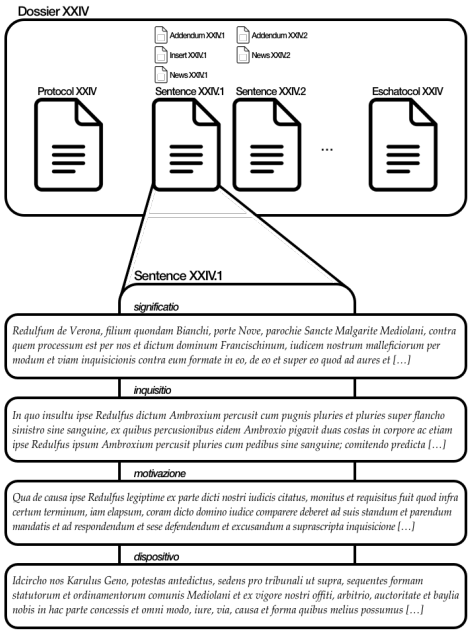

Figure 1: Example of the structure of a dossier and details on judgment's sections.

The text of each source, strictly written in Latin language, has been thoroughly studied by experts, combining the findings extracted from the text with their domain knowledge in order to provide accurate and detailed annotations about people, places and items. For each person involved in the facts, demographic and social information have been identified: name and nicknames, biological gender, social class (*dominus*), profession, place of origin or residency, possible relationships with relatives, and roles played in the events. For instance, we know that *Laurentius de Roncho*, also referred to as *Beleius* and son of *Belollus*, was murdered in March 1385 by *Iohanollus de Raude*, also known as *Barachinus*.

Knowledge about places is important to understand where crimes were being committed and the geographical origin of the criminals: places inside the city regard the *parochiae* (parishes) and *portae* (gates), that were used to divide the territory of Milan; places outside the city are used for both towns under the jurisdiction of Milan, and for cities inside or outside of Italy; finally, generic places are used to indicate where a misdemeanor has taken place, e.g., a public street or a private house. The murderer of *Laurentius de Roncho* took place in a public street near its residency, in *Parochia Sancti Babile foris*, *Porta Horientalis*.

Within the narrative of a criminal event, it is possible to read about items used within an assault or that had been stolen by pickpockets, along with the indication of the body parts struck or striking. Additionally, for stolen artifacts, it is also specified their value, expressed in the currency of the time. For example, *Iohanollus de Raude* struck *Laurentius de Roncho* dead in the occipital bone (*in capite de retro*) with a tuck (*stocho*), an ancient type of longsword.

4.2. Annotation Structure

The annotation activity has been performed by a team of domain experts, that defined and mutually agreed on the custom guidelines followed throughout the entire process. The resulting dataset consists of a collection of 325 documents, of which most comprise the Latin text, the document type, named entities, events, relations, and text segmentation labels.

The documents are classified according to the six document types identified at the beginning of the previous section; the counts of documents for each type is reported in Table 1.

Type	Count
Sentences	127
Addendum	71
News	48
Protocol	30
Insert	26
Eschatocol	23

Table 1: The list of document types ordered by number of occurrences.

Objects under the "news" type, given the fact that they are orders or information from non transcribed documents, do not have any text; thus, knowledge about "news" can be indirectly acquired from the text of another object they refer to, usually an "addendum". However, this knowledge is still reported in the "news" object in order to keep it logically distinct from the others.

In each document, there are eight types of named entity recognised: "PERSON" (e.g. *Laurentius de Roncho*), "PLACE" (*Parochia Sancti Babile foris*), "DATE" (*01/03/1385-31/03/1385*), "ITEM" (*stocho*), "ANIMAL" (*equum brunum*, brown horse), "MEASURE" (*valoris*, value), "UNITY OF MEASURE" (*librarum imperialum*, imperial pounds), "QUANTITY" (*viginti quinque*, twenty-five). For some of them, further sub-types have been defined, such as "GIVEN NAME" and "NICKNAME" for "PERSON", or "CITY" and "CHURCH" for "PLACE". The counts of named entities types and subtypes is reported in Table 2; since the same named entity can occur in multiple documents, the counts refer to the unique occurrences in the entire dataset.

Type	Sub-Type	Count
PERSON	GIVEN NAME	721
	NICKNAME	30
	NAME VARIANT	7
PLACE	CHURCH	75
	GENERIC	37
	CITY	18
DATE	DAY	105
	RANGE	42
ITEM	GENERIC	45
	BONE	25
QUANTITY	GENERIC	38
UNIT OF MEASURE	GENERIC	10
MEASURE	GENERIC	7
ANIMAL	GENERIC	3

Table 2: The list of named entity types and subtypes ordered by the number of unique occurrences.

Events are the most complex structure in the dataset; each of them is characterised of a type, usually of a subtype, and one or more arguments. There are 5 types of events: "TRIAL STAGE", "TRIAL INTEGRATION", "ESCHATOCOL", "OFFENCES", and "DEATH". A type of event may have one or multiple subtypes, for a total of 37 event subtypes: for example, an event of type "OFFENCES" may be, among others, of subtype "INSULT", "MURDER" or "THEFT". Depending on its type and subtype, an event has a different set of attributes, each of them having a role and an entity playing that role: in a "THEFT" event, we expect to have a time and place of the event, a victim, a thief, and the object or quantity of money stolen.

Sentence I.1
SEGMENT 2 - start: *1293*, end: *1990*, type: *inquisitio*

In quo quidem insultu predictus Iohanollus dictus Barachinus cum stocho uno evaginato, quem suis tenebat manibus, percussit et vulneravit suprascriptum Laurentium in capite de retro una percussione cum sanguinis effuxione, ex qua percussione dictus Laurentius mortuus fuit et est, dictum homicidium idem Iohanollus dictus Barachinus suis propriis manibus comitendum; et predicta omnia et singula commissa et perpetrata fuerunt per suprascriptum Iohanollum dictum Barachinum superius inquisitum de anno presenti curente MCCCLXXXV et mense marcii proximi praetiriti dicti anni, in strata publica sita in suprascriptis porta Horientalis et parochia Sancti Babille foris, choerentiae in inquisitione.

PERSON ITEM PLACE DATE

Named Entities

[P27] *Iohanollus de Raude*, PERSON, GIVEN NAME
[P27] *Barachinus*, PERSON, NICKNAME
[P30] *Laurentius de Roncho*, PERSON, GIVEN NAME
[I57] *Stocho*, ITEM
[I127] *Capite de retro*, ITEM, BONE
[L19] *Strata publica*, PLACE
[L15] *Porta Horientalis*, PLACE
[L17] *Parochia Sancti Babile foris*, PLACE, CHURCH
[D37] *1/3/1385-31/3/1385*, DATE, RANGE

Events

[E131], OFFENCES, MURDER
- *date*: D37
- *place*: P19
- *victim*: P30
- *murderer*: P27
- *weapon*: I57
- *bodyPartHit*: I127

[E122], DEATH
- *date*: D37
- *place*: P19
- *deceased*: P30

Relations

P27, *hasBiologicalGender*, Male
P30, *livesIn*, L15
P30, *livesIn*, L17
P30, *hasBiologicalGender*, Male
L17, *isLocatedIn*, L15
E122, *subsequentTo*, E131

Figure 2: Example of the annotations of a segment taken from *Sentence I.1*.

Relations between entities are defined by a triple of the form ("ENTITY1", "PREDICATE", "ENTITY2"), where "ENTITY1" is one of the named entities or events, "PREDICATE" defines the type of relation, and "ENTITY2" can be a named entity (or event) or a group. For instance, *Laurentius de Roncho isSonOf Belollus* or *Laurentius de Roncho*

hasBiologicalGender Male. In the dataset there are 37 unique predicates, which define 3397 unique relations.

Finally, for documents of type "sentences", the text has been divided into segments, each of them classified with a label that specifies the section in which they appear, according to the annotation schema defined in the previous section: *significatio*, *inquisitio*, *motivazioni*, *dispositivo*. The segments are outlined by a starting and ending index, enclosing a specific span of text.

An example of all the annotations that can be found in a text is portrayed in Figure 2: this shows the amount of details that can be extracted even from a very short piece of text, like the one presented.

5. Applications

In this section we provide examples of some possible use cases for the LiMe dataset, starting from simple exploratory analysis, that can be useful for medievalist researchers, to more elaborate Natural Language Processing (NLP) tasks.

5.1. Exploratory Analysis

The detail of annotation in the LiMe dataset allow for a methodological and technical study about social, demographic, judicial and economical aspects of the city of Milan in the XIV century. Extracting all the events of type "OFFENCES", and grouping them by subtype, it is possible to have an overview of the nature of crimes at the time. As shown in Figure 3, beside some usual types of crime, such as insults, murders and thefts, there are some kinds of particular crimes, typical of that period, such as *decapilatio*, the act of pulling someone's hair, and *descapuzatio*, which consists in stealing a wool hat.

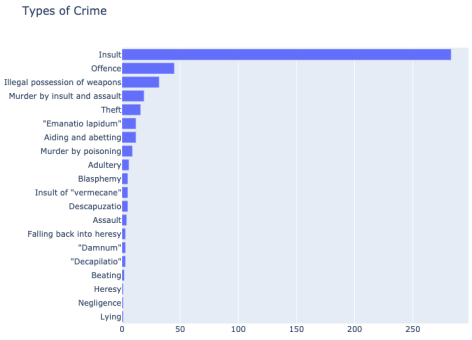

Figure 3: Number of criminal offences by type.

There are also some kinds of condemnation typical of the time, like flogging or corporal punishment (Figure 4).

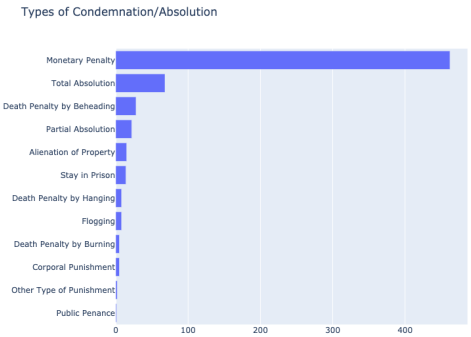

Figure 4: Number of condemnation/absolution by type.

It is also interesting to notice the difference in gender distribution of victims and criminals: despite them being mainly men in both cases, the percentage of females is almost triplicated when it comes to victims (Figure 5).

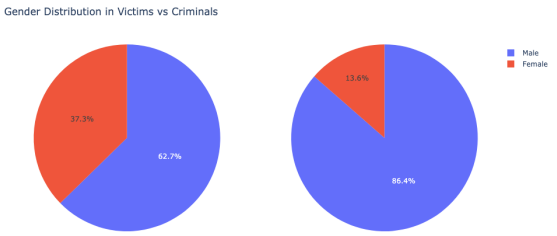

Figure 5: Distribution of males and females in victims (left) and criminals (right).

5.2. NLP Tasks

Given the peculiarity of the dataset, we believe that LiMe can be employed for many machine learning tasks involving the usage of NLP techniques. Here we provide two examples of traditional problems: document classification and text segmentation.

5.2.1. Document Classification

A document classification task regards the process of automatically assigning predefined labels to documents based on their content. For this reason, we decided to employ the 276 documents having a text, leaving out the "news" documents and ending up with five possible labels: "addendum", "eschatocol", "insert", "protocol", "sentence". We employ Latin BERT (Bamman and Burns, 2020), a contextual language model trained on a large corpus in Latin language, and fine-tune it on the training set (221 documents) for this specific classification task. The model achieves a weighted F1 score of 0.96 on the

test set (55 documents), performing remarkably well on every class (Figure 6).

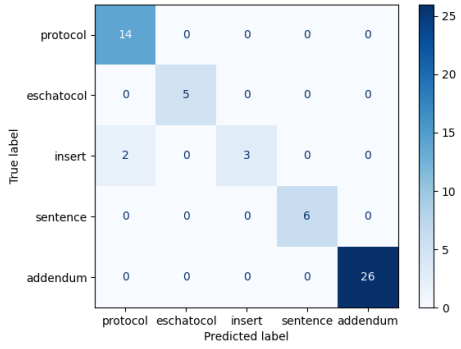

Figure 6: Confusion matrix for the document classification task.

5.2.2. Text Segmentation

A text segmentation task consists in dividing a given text into meaningful and coherent segments based on an underline annotation schema. The documents interested by this task are the "sentences" that, together, are made of more than one thousand textual segments. Each of them has a section associated to it, according to the following schema: "significatio", "inquisitio", "motivazioni", "dispositivo". In order to solve the task, we employ Rewired Conditional Random Fields (Ferrara et al., 2023a), a recent approach developed for the textual segmentation of Italian judgments, capable of working in a few-shot scenario, which is ideal given the low number of available observations. We train the above model on the segments of one hundred "sentences": the model achieves a weighted F1 score of 0.84 on the remaining 20% of the dataset left out for evaluation purposes (Figure 7).

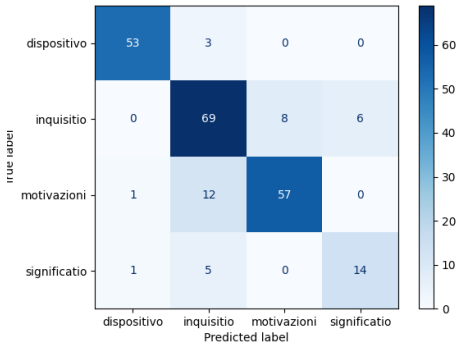

Figure 7: Confusion matrix for the text segmentation task.

6. Conclusion

The *Libri sententiarum potestatis Mediolani* are a valuable resource not only for scholars studying medieval law, but also for historians and linguists. The LiMe dataset proves how the digitisation and annotation of these kinds of sources allow for a methodological and technical analysis of the data, thanks to the usage of statistical and machine learning tools. In the future, we expect to: exploit the current dataset for more complex tasks, such as named entity recognition or event extraction; increase the number of annotated documents, with information coming from subsequent volumes of the *Libri*, which are currently being examined by experts; extend the current annotations with features at syntactical and morphological levels.

7. References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. *arXiv e-prints*, page arXiv:2009.10053.

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alessandra Bassani. 2021. Le assoluzioni nel Liber comunis potestatis Mediolani: riflessioni sull'ipotesi di una giustizia giusta. *Notariorum Itinera*, 7:177–204.

Alessandra Bassani, Marta Calleri, and Marta L. Mangini. 2021. Liber sententiarum potestatis mediolani (1385): Storia, diritto, diplomatica e quadri comparativi. *Notariorum Itinera*, 7.

Alessandra Bassani, Beatrice Del Bo, Alfio Ferrara, Marta Mangini, Sergio Picascia, and Ambra Stefanello. 2024. LiMe - Liber sententiarum potestatis Mediolani.

Clément Besnier and William Mattingly. 2021. Named-entity dataset for medieval latin, middle high german and old norse. *Journal of Open Humanities Data*, 7(0):23.

Raffaella Bianchi Riva. 2021. Iniuria e insultus tra diritto e politica. Le offese alle magistrature comunali nella legislazione statutaria e nella prassi giudiziaria in età viscontea. *Notariorum Itinera*, 7:239–264.

Patrick J. Burns. 2023. Latincy: Synthetic trained pipelines for latin nlp.

Patrick J. Burns, James A. Brofos, Kyle Li, Pramit Chaudhuri, and Joseph P. Dexter. 2021. Profiling of intertextuality in latin literature using word embeddings. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4900–4907.

Luca Campisi. 2019. Prassi giudiziaria a vercelli nel xiv secolo. *Studi di storia medioevale e di diplomatica - Nuova Serie*, (2):131–150.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020. A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, Marseille, France. European Language Resources Association.

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. 2022. Medlatinepi and medlatinlit: Two datasets for the computational authorship analysis of medieval latin texts. *Journal on Computing and Cultural Heritage*, 15(3):1–15.

Nadia Covini. 2012. Assenza o abbondanza? la documentazione giudiziaria lombarda nei fondi notarili e nelle carte ducali (stato di milano, xiv-xv secolo). *La documentazione degli organi giudiziari*, pages 483–499.

Trevor Dean. 2007. *Crime and Justice in Late Medieval Italy*. Cambridge University Press.

Trevor Dean. 2008. Theft and gender in late medieval bologna. *Gender & History*, 20(2):399–415.

Beatrice Del Bo. 2021. Tutte le donne (del registro) del podestà fra cliché e novità. *Notariorum Itinera*, 7:83–106.

Joseph Denooz. 2007. Opera latina: le nouveau site internet du lasla. *Journal of Latin Linguistics*, 9(3):21–34.

Alfio Ferrara, Sergio Picascia, and Davide Riva. 2023a. Few-shot legal text segmentation via rewiring conditional random fields: A preliminary study. In *Advances in Conceptual Modeling*, pages 141–150, Cham. Springer Nature Switzerland.

Alfio Ferrara, Sergio Picascia, Elisabetta Rocchetti, Gaia Varese, et al. 2023b. The FAITH project: integrated tools and methodologies for digital humanities. *Proceedings of the Statistics and Data Science Conference*, pages 323–327.

Cecchini Flavio, Rachele Sprugnoli, Moretti Giovanni, Passarotti Marco, et al. 2020. Udante: First steps towards the universal dependencies treebank of dante's latin works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, pages 99–105. Accademia University Press.

Andrea Gamberini. 2014. *A Companion to Late Medieval and Early Modern Milan: The Distinctive Features of an Italian State*, volume 7. Brill.

Andrea Giorgi, Stefano Moscadelli, and Carla Zarrilli. 2012. *La documentazione degli organi giudiziari nell'Italia tardo-medievale e moderna: atti del convegno di studi, Siena, Archivio di Stato, 15-17 settembre 2008*. Number v. 1 in Pubblicazioni degli Archivi di Stato. Saggi. Ministero per i beni e le attività culturali, direzione generale per gli archivi.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Roberto Isotton. 2021. La repressione dei reati di furto e rapina nel Liber sententiarum potestatis Mediolani del 1385: acquisizioni e questioni aperte. *Notariorum Itinera*, 7:205–238.

Vojtěch Kaše, Petra Heřmánková, and Adéla Sobotková. 2021. Classifying latin inscriptions of the roman empire: A machine-learning approach. In *Proceedings of the Conference on Computational Humanities Research 2021CEUR-WS*, volume 2989, pages 123–135.

Piroska Lendvai and Claudia Wick. 2022. Finetuning Latin BERT for word sense disambiguation on the thesaurus linguae latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 37–41, Taipei, Taiwan. Association for Computational Linguistics.

Didier Lett. 2021. *I Registri Della Giustizia Penale Nell'Italia Dei Secoli XII-XV*. Ecole française de Rome.

Campisi Luca. 2021. *L'impatto sociale. I protagonisti delle pratiche giudiziarie a Vercelli fra XIV e XV secolo*. Phd thesis, Università degli Studi di Milano.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of latin. *New methods in historical corpus linguistics*, 1(3):247–257.

Giovanni Minnucci. 2021. Intorno al Liber sententiarum potestatis Mediolani e ad altre fonti giudiziarie. Alcune note conclusive. *Notariorum Itinera*, 7:373–380.

Antonio Padoa-Schioppa. 1996. *La giustizia milanese nella prima età viscontea (1277-1300)*. Giuffrè.

Antonio Padoa-Schioppa. 2017. *A History of Law in Europe: From the Early Middle Ages to the Twentieth Century*. Cambridge University Press.

Fabrizio Pagnoni. 2021. Selezione e circolazione dei giudici ai malefici nel dominio visconteo fra Tre e Quattrocento. *Notariorum Itinera*, 7:61–81.

Marco Passarotti. 2019. *The Project of the Index Thomisticus Treebank*, pages 299–320. De Gruyter Saur, Berlin, Boston.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Pier Francesco Pizzi. 2021. *Liber sententiarum potestatis Mediolani (1385), Edizione critica*. Società Ligure di Storia Patria.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Marton Ribary. 2020. A relational database of roman law based on justinian's digest. *Journal of Open Humanities Data*, 6(1):5.

C. Santoro. 1968. *Gli offici del comune di Milano e del dominio visconteo sforzesco (1216-1515).* 1. collana: Monografie, ricerche ausilierie, opere strumentali. A. Guiffrè.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Claudia Storti. 2021. 1385: un anno tra politica e giustizia a milano. *Notariorum Itinera*, 7:7–31.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

M. Vallerani. 2012. *Medieval Public Justice*. Studies in Medieval & Early Mo. Catholic University of America Press.

Chiara Valsecchi. 2021. «Per viam inquisicionis». Note sul processo criminale a Milano in un'età di transizione. *Notariorum Itinera*, 7:127–176.

E. Verga. 1901. *Le sentenze criminali dei podestà milanesi 1385-1429: appunti per la storia della giustizia punitiva in Milano*. P. Confalonieri.