Weakly Supervised Named Entity Recognition for Historical Texts

Marco Sorbi[1,2], **Laurent Moccozet**[2], **Stephane Marchand-Maillet**[2]

[1]Research Institute for Statistics and Information Science, University of Geneva, Switzerland
[2]Centre Universitaire d'Informatique, University of Geneva, Switzerland
Correspondence: Marco.Sorbi@unige.ch

Abstract

Named Entity Recognition has emerged as a critical task in natural language processing, particularly for extracting meaningful information from unstructured text. Although traditional approaches rely heavily on large annotated datasets, recent advances have explored weak supervision techniques to address the limitations of resource-intensive annotation processes. Historical texts provide unique challenges to this task because of their linguistic peculiarities, and several approaches exist to address texts of this domain in a supervised way, but they involve lengthy manual annotations of the documents of interest by domain experts. To address this issue, this paper explores how recent weakly supervised NER techniques can be adapted to historical texts, analyzing their suitability for this domain. The experiments show that domain-specific architectures can be effectively trained on low-resource corpora with weak supervision over a small set of entity labels. Using only 10% of the annotations, the performance of these architectures remains above 80% of the supervised quality in terms of F1-Score.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing to identify and classify named entities from unstructured text into predefined categories such as people, organizations, and locations, and serves as an important step in the structuring of textual data (Lauriola et al., 2021). General purpose NER methods range from traditional rule-based approaches to modern deep learning architectures, mainly transformer-based (Keraghel et al., 2024). Various domains exist where ad hoc methods have been studied to address domain-specific peculiarities, such as biomedicine (Lauriola et al., 2021) and history (Ehrmann et al., 2023).

Historical documents present distinct challenges compared to modern texts, ranging from an archaic language with non-standardized orthography (Santini et al., 2025), to degraded scripts and complex document layouts (Ehrmann et al., 2020). The lack of large annotated corpora in this low-resource context makes it difficult to design and train specific NER architectures (Novotny et al., 2023) that are required for effective NER in historical documents, with a significant impact on digital humanities and cultural heritage activities.

When manually annotated data are scarce, weakly supervised NER (WS-NER) has emerged to address the lack of training data. One common WS-NER framework is distantly-supervised NER (DS-NER), which takes advantage of knowledge bases or dictionaries to automatically generate labels to use as supervision signals (Fang et al., 2021). However, DS-NER faces significant challenges due to noisy and incomplete labels caused by limited dictionary coverage and imperfect distant annotation (Zhou et al., 2022). Previous works have studied multiple approaches to DS-NER, which we will present in Section 2, which have the potential to progress NER in the historical domain given its noise context and lack of extensive labeled datasets.

This paper aims to study WS-NER in the historical domain, exploiting existing DS-NER techniques for modern texts, to evaluate their effectiveness in the domain, and use them to adapt architectures from supervised Historical NER studies to the WS-NER task[1]. This work is motivated by the feasibility of creating a partial domain-specific knowledge base relative to labeling a complete domain-specific document to train a supervised NER system. For this purpose, we identify two research questions to address:

- **RQ1**: Varying the size of the dictionaries, how do DS-NER techniques that are designed for modern texts perform on historical documents with domain-specific dictionaries? How do

[1] Code is available at `www.github.com/msorbi/hwsner`

Proceedings of SIGHUM 2026, pages 48–65
March 28-29, 2026 ©2026 Association for Computational Linguistics

they compare to domain-specific supervised NER methods in WS-NER and fully supervised NER settings?

- **RQ2**: Can we exploit these DS-NER techniques to adapt domain-specific supervised NER architectures to WS-NER?

The remainder of this paper is structured as follows. In Section 2, we discuss and organize a selection of related works concerning DS-NER and supervised NER in historical documents. Then, Section 3 presents the datasets used in this work to evaluate the models. Section 4 describes the experiments carried out and presents an overview of the metrics used for evaluation in this work. The results are presented in Section 5 and discussed in Section 6, where the research questions are also answered. Finally, Section 7 presents conclusions and suggestions for future work.

2 Related Work

The positioning of this paper is at the intersection of weak supervision and the historical domain of Named Entity Recognition. The presentation of an overview of related work on the two topics is therefore appropriate.

2.1 Distantly-supervised Named Entity Recognition

DS-NER is a WS-NER framework that relies on the availability of dictionaries or knowledge bases instead of labeled sequences as training supervision signal. It avoids the need for large-scale human annotation, but introduces noise due to incomplete and inaccurate annotations (Zhang et al., 2021). Early research sees the development of AutoNER (Shang et al., 2018), which introduced a Tie-or-Break tagging scheme with a Fuzzy-LSTM-CRF architecture designed to be robust to noisy distant supervision. We divide subsequent work into two main branches: Positive-Unlabeled Learning (PUL) and Self-Supervised Learning (SSL).

Positive-Unlabeled Learning It formulates DS-NER as training on positively labeled data, created using the knowledge base, and unlabeled data. Contributions using this approach include:

- AdaPU (Peng et al., 2019), which firstly introduced PUL in DS-NER, designing a training algorithm that can unbiasedly estimate the task loss as if there were fully labeled data.
- Conf-MPU (Zhou et al., 2022), which extended PUL in DS-NER to a multi-class setting and introduced confidence-based risk estimation.
- CuPUL (Li et al., 2025), which adds curriculum learning to PUL to stabilize training and weaken the impact of noise.

Self-Supervised Learning Used as a training stage, it allows reducing the impact of noisy labels to works like:

- BOND (Liang et al., 2020), which first adapts Pre-trained Language Models to noisy distant labels, and then applies teacher-student self-training for refinement.
- SCDL (Zhang et al., 2021), which trains two teacher-student networks to jointly denoise labels.
- CENSOR (Si et al., 2024), which introduces uncertainty-aware teacher learning, to reduce reliance on miscalibrated high-confidence labels, and student-student label sharing, to mitigate error propagation.

Other approaches There are other recent works using different approaches, for example:

- SANTA (Si et al., 2023), addressing inaccurate and incomplete annotations separately, with a memory-smoothed focal loss and a noise-tolerant loss.
- MProto (Wu et al., 2023), employing a prototype network to capture intra-class variance, and formulating token-prototype assignment as an optimal transport problem.

2.2 Historical Named Entity Recognition

In the domain of historical documents, NER becomes substantially more difficult, as it faces multiple challenges due to the linguistic characteristics of these texts, which the survey by Ehrmann et al. (2023) identifies in language dynamics, noisy input, and lack of resources. Analyzing methodologies from rule-based to deep learning, the survey emphasizes the need for models robust to multilingualism, spelling variation, and sparse data. When dealing with digitized documents, non-standard layout and Optical Character Recognition (OCR) errors add a layer of noise that can significantly degrade performance (Kettunen et al., 2017).

We can distinguish the application of different techniques for NER in this domain.

Rule-based Early research mainly employs rule-based systems tailored to documents of interest, using lexical heuristics, gazetteers, and token-level rules, in order to extract people and places from British parliamentary records (Grover et al., 2008), for example. This approach is transparent, but typically struggles to face the challenge of integrating variability in the rules when increasing the corpus size.

Machine learning It enables the use of more flexible feature-based techniques that commonly rely on Hidden Markov Models, Support Vector Machines, Conditional Random Fields (CRF), and decision trees (Yadav and Bethard, 2018). In historical context, Torres Aguilar et al. (2016) applied a CRF on latin charters, obtaining good results based on a large number of features. These methods reduce the effect of variability, but they require a large number of hand designed features.

Deep learning These systems, often relying on sequence labeling and transformer-based models such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) for automated feature learning, and CRF for classification, have further reduced the effect of variability. Among deep learning works in the historical domain,

- Boros et al. (2020) propose a hierarchical transformer stack for historical German and French datasets, specifically to mitigate OCR noise in digitized historical texts,

- Blouin et al. (2021) investigate transfer learning from modern to historical domains, studying annotation effort, domain mismatch, and pre-training data selection for historical NER,

- Torres Aguilar (2022) creates a human-labeled dataset of multilingual medieval charters, and addresses multilingual NER by combining stacked embeddings with BERT-based models fine-tuned on the dataset,

- Novotny et al. (2023) create a small dataset with sentences containing known entities from late medieval European texts, and propose a bootstrapping annotation pipeline to build larger copora of the texts.

However, these methods rely on heavily annotated datasets to learn features, which are rarely available.

Few-shot learning It exploits Large Lanugage Models (LLMs) through prompting in order to label the historical texts without relying on annotated corpora. Among these works in the historical domain,

- Hiltmann et al. (2025) prompt LLMs with historical context, outperforming established frameworks on historical texts,

- Zhang and Colavizza (2025) use few-shot learning of LLMs on the HIPE dataset (Ehrmann et al., 2020).

This work focuses on deep learning systems to reduce the amount of annotations required to train them. Specifically, we study the applicability of weak supervision techniques to train deep learning models on the historical NER, studying the impact of the amount of annotated data on the quality of results. Even if we can consider the approach of Novotny et al. (2023) to belong to the WS-NER paradigm, to the knowledge of the authors, there is no extensive study of WS-NER or DS-NER in the historical domain.

3 Dataset

Our experiments are based on the dataset introduced by Torres Aguilar (2022), consisting of a human-labeled medieval NER dataset whose ancient French charters come from:

- Corpus de la Bourgogne du Moyen Âge ("Corpus of Burgundy in the Middle Ages", *CBMA*) - Cartulary of the city of Arbois (Magnani, 2020): a municipal cartulary commissioned in 1384, containing public issues such as military services and war costs, taxes and customs, or lawsuits in court.

- Diplomata Belgica ("Belgian Diplomatic Sources", *CDBE* – de Hemptinne, Thérèse and Deploige, Jeroen and Kupper, Jean-Louis and Prevenier, Walter, 2015): a database published by the Belgian Royal Historical Commission, containing French charters dated the 13th century, containing legal actions concerning individuals, corporations, and private affairs.

- HOME History of Medieval Europe (*HOME*) - Alcar (Stutzmann et al., 2021): it contains cartularies dated between the 12th and 14th centuries, reporting donations, exchanges, and other legal acts.

	CBMA	CDBE	HOME
Train	38658	235643	114640
Validation	896	18510	18510
Test	8133	56081	18554

Table 1: Number of tokens per each dataset split.

	CBMA	CDBE	HOME
PERS	652	4118	925
LOC	347	2662	922

Table 2: Number of distinct training entities for each type and source, before sampling.

The dataset contains annotations for entities in the Person (PERS) and Location (LOC) classes, and the sizes for each source are reported in Table 1.

Separately for each source, to adapt the dataset to a WS-NER setting, we extracted all the tagged entities from the train split, and randomly sampled the entities to make the dictionaries for each type. Then, we used the sampled dictionaries to create the noisy labels for the train splits of the mentioned sources via string matching. Table 2 shows the dimensions of each dictionary. The validation and test splits remained unchanged.

4 Experimental Setup

To address RQ1, we tested DS-NER techniques belonging to the Positive-Unlabeled and Self-Supervised Learning frameworks, plus the two specific works mentioned in Section 2 and the Stacked Embeddings supervised architecture (Torres Aguilar, 2022), and compared their performances on historical French texts in two settings:

- Fully supervised NER, using the original human-labeled train splits, to analyze the best supervision condition, and

- WS-NER, varying the sampling size for the dictionaries according to Section 3, to see its impact on the performances. We used sampling sizes from 20% to 100%, increasing by 20%, that are commonly used in previous works (Zhou et al., 2022; Wu et al., 2023), and 10% to test with increased noise conditions,

training every method on the train split of each setting and sample size independently. To better summarize the results, we will refer with "small dictionaries" to the 10% and 20% samples, with "medium dictionaries" to the 40% and 60% samples, and with "large dictionaries" to the 80% and 100% samples.

Then, to address RQ2, we used the best performing DS-NER technique to train

1. XLM-RoBERTa base (Conneau et al., 2020), a large multilingual masked language model, and

2. Stacked Embeddings (Torres Aguilar, 2022), a domain-specific architecture for supervised NER in medieval texts,

on the WS-NER task, using the same settings as in the first part, and checked whether DS-NER can generalize their performance from the supervised task.

Evaluation We now discuss the evaluation metrics used to assess NER performance in historical documents. According to the works presented in Section 2, the typical evaluation metric for NER is the F1-score, usually presented alongside precision and recall to show their impact. Novotny et al. (2023) use the $F_{0.25}$-score, which significantly gives more weight to precision than to recall, but otherwise the two metrics are averaged in a balanced way, that is, using the F1-score.

In WS-NER, training supervision is based on incomplete labels, which are characterized by high precision but low recall (Zhou et al., 2022). This effect is amplified in historical documents by orthographic variations that complicate entity identification. As DS-NER training algorithms aim to improve the recall at the expense of precision, we continue to use the common F1-score as a balanced measure, giving importance to the recall metric, too. We compute these metrics using `seqeval` (Nakayama, 2018) default token-level mode, in order to grant partial scores to entities that are correctly identified and classified, even if their boundary is not correct, and mitigate the problem of annotation granularity (Ehrmann et al., 2020).

If multiple classes are present, as in our case with PERS and LOC, micro averaging is commonly used for global measures (Shang et al., 2018; Torres Aguilar, 2022) to ensure that frequent classes have an appropriate influence on the final results.

5 Results

We analyze here the test performances, in terms of micro-averaged F1-Score, Precision, and Recall, obtained by the various methods under fully supervised NER and different weak supervision settings. For each technique, we micro-average measures among classes and then, for each setting, we average among datasets and report them in percentage in Figure 1. Detailed measures are reported in Appendix A. For comparison, Stacked Embeddings (StEmb), a domain-specific architecture for supervised NER in medieval texts (Torres Aguilar, 2022), is also reported along with dictionary matching and the CuPUL DS-NER technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings as underlying classifier. We can notice that:

- Despite having a precision among the best in every setting, if we consider recall and F1-score, the Stacked Embeddings architecture is the best performer with full supervision and large dictionaries, but its performances drop quickly when the sample size is decreased.

- If trained under full supervision, the best DS-NER techniques are BOND (Liang et al., 2020) and SANTA (Si et al., 2023), with an F1-score close to the baseline supervised Stacked Embeddings architecture.

- SANTA maintains good recall with large dictionaries, but if we consider the F1-score, BOND remains the best DS-NER technique with large and medium dictionaries, overcoming the supervised baseline as well for medium dictionaries, and is the best DS-NER technique in every setting in terms of precision.

- With small dictionaries, the best technique in terms of F1-score is CuPUL (Li et al., 2025), which has the best recall even with medium dictionaries.

- Considering all three metrics, CuPUL is the least affected by the noise introduced with distant labels, with a performance that only slightly changes on the different settings. Conf-MPU (Zhou et al., 2022), which is the other PUL-based technique, and MProto (Wu et al., 2023) also have stable but lower performances.

- Except for the simpler BOND, the SSL techniques obtained generally unstable performances across settings, especially in terms of recall and F1-score.

Because CuPUL is the best performing technique in low supervision settings, and the least affected by the noise of distant labels, we use it to train (1) XLM-RoBERTa (Conneau et al., 2020) and (2) Stacked Embeddings (Torres Aguilar, 2022) on the WS-NER task. This is done by replacing the original RoBERTa classifier in CuPUL. Despite that they perform slightly worse than the baseline with full supervision and large dictionaries, we can see once again in Figure 1 that CuPUL is little affected by the noise of distant labels. Although XLM-RoBERTa (1) still has some noise-related problems, especially with small dictionaries, CuPUL allows the ad-hoc classifier (2) to have performances in the WS-NER settings that are significantly closer to that of the fully supervised task for which it has been designed.

6 Discussion

The results of our experiments show that weak supervision can achieve F1-Scores within 20% of supervised Named Entity Recognition performance on historical texts using only 10% of the annotations, making them particularly appealing for this context where large annotated datasets are often unavailable. To mitigate the linguistic challenges that characterize NER in historical texts, the experiments demonstrate that DS-NER techniques can effectively adapt domain-specific NER models to low-resource historical corpora, to achieve performance levels that are close to fully supervised methods even with additional noise introduced with distant labels. The amount of noise in the distant labels, controlled via the dictionary size, has a significant impact on the final performance, causing noticeable drops in most of the techniques. Nevertheless, CuPUL and MProto have shown good robustness, maintaining consistent performance across different dictionary sizes, and revealing particularly suitable for this scenario. Considering these results, we can now answer the research questions formulated in Section 1.

6.1 RQ1: Validity of modern text DS-NER techniques for ancient texts

The results reported in Figure 1 show that most DS-NER techniques suffer from performance drops when they are applied to NER in ancient texts, especially in low supervision settings with small dictionaries. CuPUL (Li et al., 2025), BOND (Liang et al., 2020), and SANTA (Si et al., 2023), belong-

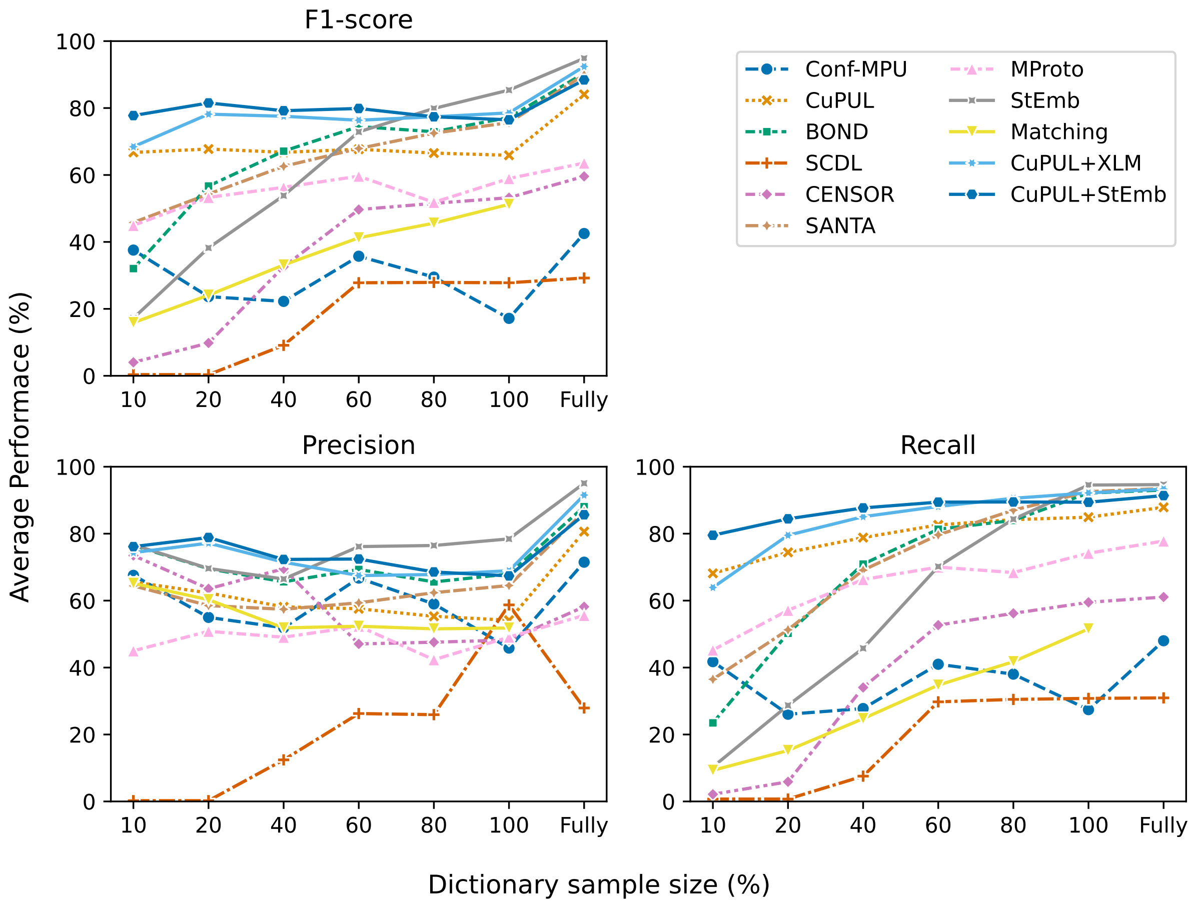

Figure 1: Test results of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. The Stacked Embeddings (StEmb) supervised architecture, dictionary matching, and the CuPUL DS-NER technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings as underlying classifier are reported using solid lines.

ing to three different categories as identified in Section 2, are able to achieve good performance with large dictionaries, but still suffer substantial drops when switching to smaller dictionaries, especially the latter two techniques. CuPUL is the most robust to the noise of low supervision, thanks to a high recall as opposed to the other techniques.

6.2 RQ2: Adaptation of domain-specific supervised NER architectures using modern text DS-NER

Based on the answer to RQ1, we selected CuPUL to carry out further experiments on the adaptability to the WS-NER task in historical texts. Figure 1 shows that, especially for the Stacked Embeddings architecture (Torres Aguilar, 2022), the performance in all settings is comparable and closer to the supervised performance of the architecture, remaining above 80% of the supervised F1-Score. This indicates that CuPUL enables the supervised NER architecture to generalize well to weak supervision. The relatively low capacity of Stacked Embeddings, which is based on a BiLSTM-CRF instead of a transformer, may help in this, as it is, in principle, less prone to overfitting and easier to generalize.

7 Conclusion and Future Work

This study shows that weak supervision can effectively be used to train domain-specific models to achieve NER performance levels, on historical texts with little annotations, that are close to fully supervised methods. This is particularly important for the historical domain, where large annotated datasets are often unavailable.

Further assessment of the robustness of this framework and improvement of its performance are the two main areas of focus for future work. Additional analysis could involve examining the adaptability of the framework under various sources of noise that are characteristic of the domain, such as the use of external historical knowledge bases (Uckelman; Wrisley, 2018) to build annotation dictionaries and its application to historical documents transcribed with OCR (Ehrmann et al., 2020). These investigations will determine the framework's applicability when manual transcription of texts or compilation of entity lists is impractical or unfeasible. Additionally, similar analysis can be conducted for other information extraction tasks, including Entity linking and Relationship extraction. In order to increase the NER performance, a major development of the technique would be the integration of active learning, which offers the dual benefit of assisting the model in resolving difficult instances and exploiting domain expertise from historians and linguists where it is most needed and appropriate.

By reducing the need for large annotated datasets, these techniques support the adaptation of NER tools to a wider range of historical low-resource text collections, enabling more comprehensive analysis of historical documents (Bouillon et al., 2024), with significant implications for the fields of digital humanities and cultural heritage.

Limitations

The empirical analysis of this work is restricted to manually transcribed medieval French charter collections with two entity types (persons and locations), which limits the generalizability of the findings to other historical languages and periods, document genres, larger sets of entity categories, and OCR-transcribed documents. In order to avoid the quality of knowledge bases impacting the results, weak supervision is performed via randomly sampled dictionaries constructed from the training splits; manually curated or external bases, which could be biased towards some subset of entities, may exhibit different noise profiles and cause a different model behavior. Moreover, a fully annotated, albeit small, validation set is kept for model evaluation and hyperparameter tuning. Finally, the evaluation relies on standard NER metrics and does not include human analysis of downstream impact on digital humanities tasks.

Ethical Considerations

Historical corpora reflect the social, cultural, and political biases of the periods and institutions that produced them. The medieval charters considered in this work primarily document legal and administrative activities of municipal and ecclesiastical authorities, emphasizing the interests of elites. As a result, NER models trained on these sources could reproduce existing biases about people and places, and the usage of models and techniques designed for modern languages or pretrained on modern texts, which is often necessary due to the inherent low volume of available historical data, could amplify these biases when facing challenges specific to the domain, such as linguistic variation and orthographic inconsistency. Moreover, learning from genre-specific conventions like formulaic expressions may capture patterns that are characteristic of the genre and which could not generalize to other genres and time periods.

The systems proposed in this work are intended as assistive tools for historical research, and their outputs should be interpreted in collaboration with domain experts who can assess biases in both the sources and the models.

Acknowledgments

The Swiss National Science Foundation funds this work under grant number 215733 "Une édition sémantique et multilingue en ligne des registres du Conseil de Genève (1545-1550)". The project website is accessible at `https://www.unige.ch/registresconseilge/en`. We express our gratitude to Christophe Chazalon, who helped process the texts thanks to his historical domain expertise, Hélène de Ribaupierre and Gilles Falquet for their knowledge and insights about Information Extraction and Semantic Web, and the anonymous reviewers and editors for their comments and suggestions.

References

Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring modern named entity recognition to the historical domain: How to take the step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162, NIT Silchar, India. NLP Association of India (NLPAI).

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Pierrette Bouillon, Christophe Chazalon, Sandra Coram-Mekkey, Gilles Falquet, Johanna Gerlach, Stephane Marchand-Maillet, Laurent Moccozet, Jonathan Mutal, Raphael Rubino, and Marco Sorbi. 2024. RC-num: A semantic and multilingual online edition of the geneva council registers from 1545 to 1550. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 21–22, Sheffield, UK. European Association for Machine Translation (EAMT).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

de Hemptinne, Thérèse and Deploige, Jeroen and Kupper, Jean-Louis and Prevenier, Walter, editor. 2015. *Diplomata Belgica: les sources diplomatiques des Pays-Bas méridionaux au Moyen Âge. The Diplomatic Sources from the Medieval Southern Low Countries*. Commission royale d'Histoire/Koninklijke Commissie voor Geschiedenis.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 288–310, Berlin, Heidelberg. Springer-Verlag.

Zheng Fang, Yanan Cao, Tai Li, Ruipeng Jia, Fang Fang, Yanmin Shang, and Yuhai Lu. 2021. TEBNER: Domain specific named entity recognition with type expanded boundary-aware network. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 198–207, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach. *Preprint*, arXiv:2502.04351.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent advances in named entity recognition: A comprehensive survey and comparative study. *Preprint*, arXiv:2401.10825.

Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2017. Old content and modern tools - searching named entities in a finnish ocred historical newspaper collection 1771-1910. *Digital Humanities Quarterly*, 11.

Ivano Lauriola, Fabio Aiolli, Alberto Lavelli, and Fabio Rinaldi. 2021. Learning adaptive representations for entity recognition in the biomedical domain. *Journal of Biomedical Semantics*, 12(1):10.

Yuepei Li, Kang Zhou, Qiao Qiao, Qing Wang, and Qi Li. 2025. Re-examine distantly supervised NER: A new benchmark and a simple approach. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10940–10959, Abu Dhabi, UAE. Association for Computational Linguistics.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1054–1064, New York, NY, USA. Association for Computing Machinery.

Eliana Magnani. 2020. *Des chartae au Corpus: La plateforme des CBMA - Chartae/Corpus Burgundiae Medii Aevi*, volume 27 of *Atelier de Recherche Sur Les Textes Médiévaux*, pages 57–67. Brepols Publishers.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Vit Novotny, Kristina Luger, Michal Štefánik, Tereza Vrabcova, and Ales Horak. 2023. People and places of historical Europe: Bootstrapping annotation pipeline and a new corpus of named entities in late medieval texts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14104–14113, Toronto, Canada. Association for Computational Linguistics.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Cristian Santini, Laura Melosi, and Emanuele Frontoni. 2025. Named entity recognition in historical italian: The case of giacomo leopardi's zibaldone. *Preprint*, arXiv:2505.20113.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

Shuzheng Si, Zefan Cai, Shuang Zeng, Guoqiang Feng, Jiaxing Lin, and Baobao Chang. 2023. SANTA: Separate strategies for inaccurate and incomplete annotation noise in distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3883–3896, Toronto, Canada. Association for Computational Linguistics.

Shuzheng Si, Helan Hu, Haozhe Zhao, Shuang Zeng, Kaikai An, Zefan Cai, and Baobao Chang. 2024. Improving the robustness of distantly-supervised named entity recognition via uncertainty-aware teacher learning and student-student collaborative learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5533–5546, Bangkok, Thailand. Association for Computational Linguistics.

Dominique Stutzmann, Sergio Torres Aguilar, and Paul Chaffenet. 2021. HOME-Alcar: Aligned and Annotated Cartularies. Type: dataset.

Sergio Torres Aguilar. 2022. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. 2016. Named entity recognition applied on a data base of Medieval Latin charters. The case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016)*, Krakow, Poland.

Sara L Uckelman. DRAFT: Names in the 1292 census of Paris.

David Joseph Wrisley. 2018. The literary geographies of christine de pizan. MLA.

Shuhui Wu, Yongliang Shen, Zeqi Tan, Wenqi Ren, Jietian Guo, Shiliang Pu, and Weiming Lu. 2023. MProto: Multi-prototype network with denoised optimal transport for distantly supervised named entity recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2361–2374, Singapore. Association for Computational Linguistics.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model. *Preprint*, arXiv:2508.18090.

Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Xue Mengge, and Hongbo Xu. 2021. Improving distantly-supervised named entity recognition with self-collaborative denoising learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1518–1529, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.

A Detailed Results

Tables 3 to 5 report the micro-averaged F1-Score, Precision, and Recall obtained by the different techniques, averaged on the test split of each source, in the cases of fully supervised NER and varying distant supervision settings. These measures are used to draw Figure 1.

Tables 6 to 8 report the micro-averaged F1-Score, Precision, and Recall obtained by the different techniques on the test split of each source, in the cases of fully supervised NER and varying distant supervision settings.

B Hyperparameter tuning

For RQ1, we used the default hyperparameters for every technique, computing prior class probabilities from the validation set when required. For RQ2, according to Li et al. (2025) and using grid searches, we made a first tuning step for the voters' hyperparameters, and a second step for the curriculum training hyperparameters, as we used CuPUL with Stacked embeddings (Torres Aguilar, 2022) and XLM-RoBERTa (Conneau et al., 2020) voters.

B.1 Voters' hyperparameter tuning

The tuned hyperparameters for the voters are the train epochs, drop negative, loss type, and m. The learning rate is set to 1e−3, and other hyperparameters are left at their default values. Tables 9 to 11 show the validation performances of Stacked Embeddings and XLM-RoBERTa voters with every set of hyperparameters. We select the set of hyperparameters based on the mean and standard deviation of the performances and on the performance with

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	38 ± 27	24 ± 33	22 ± 31	36 ± 27	29 ± 22	17 ± 24	43 ± 31	30 ± 8
CuPUL	**67** ± 1	**68** ± 0	**67** ± 3	68 ± 5	67 ± 5	66 ± 5	84 ± 7	**69** ± 6
BOND	32 ± 18	57 ± 13	**67** ± 5	**74** ± 5	**73** ± 5	**77** ± 11	**90** ± 4	67 ± 16
SCDL	0 ± 0	0 ± 0	9 ± 12	28 ± 39	28 ± 39	28 ± 38	29 ± 40	18 ± 12
CENSOR	4 ± 5	10 ± 7	33 ± 32	50 ± 36	52 ± 37	53 ± 38	60 ± 42	37 ± 19
SANTA	46 ± 0	54 ± 3	63 ± 3	68 ± 6	72 ± 4	76 ± 5	**90** ± 4	67 ± 13
MProto	45 ± 6	53 ± 3	56 ± 9	60 ± 10	52 ± 15	59 ± 8	64 ± 18	56 ± 5
StEmb	17 ± 11	38 ± 20	54 ± 6	73 ± 8	80 ± 3	85 ± 7	95 ± 1	63 ± 24
Matching	16 ± 9	24 ± 7	33 ± 4	41 ± 2	46 ± 3	51 ± 5		
CuPUL+XLM	68 ± 6	78 ± 3	78 ± 2	76 ± 5	**77** ± 5	**79** ± 7	**92** ± 3	78 ± 6
CuPUL+StEmb	**78** ± 3	**82** ± 3	**79** ± 1	**80** ± 4	**77** ± 6	76 ± 7	88 ± 4	**80** ± 3

Table 3: Test results in terms of F1-Score of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

low supervision settings. The selected hyperparameters are:

- Stacked Embeddings voter hyperparameters:
  - train epochs (E) = 15, drop negative (Neg) = 0.1, loss type = MPN, m = 20
- XLM-RoBERTa voter hyperparameters:
  - train epochs (E) = 5, drop negative (Neg) = 0.5, loss type = MPN-CE, m = 20

B.2 Curriculum training hyperparameter tuning

Tables 12 and 13 show the validation performances of the CuPUL curriculum train stage with Stacked Embeddings and XLM-RoBERTa, respectively, with every set of hyperparameters. We select the set of hyperparameters based on the mean and standard deviation of the performances and on the performance with low supervision settings. The selected hyperparameters are:

- CuPUL with Stacked Embeddings:
  - train epochs (E) = 20, loss type = MPN
- CuPUL with XLM-RoBERTa:
  - train epochs (E) = 10, train sub-epochs (sE) = 1, loss type = Conf-MPU-CE, learning rate (LR): 1e−5

C Online Resources

The source code is available online at https://github.com/msorbi/hwsner

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	68 ± 23	55 ± 41	52 ± 41	67 ± 24	59 ± 29	46 ± 41	71 ± 21	60 ± 8
CuPUL	66 ± 2	62 ± 2	58 ± 4	58 ± 6	55 ± 6	54 ± 6	81 ± 7	62 ± 8
BOND	**76** ± 18	**70** ± 8	66 ± 5	**69** ± 6	**66** ± 10	**68** ± 16	**88** ± 6	**72** ± 7
SCDL	0 ± 0	0 ± 0	12 ± 17	26 ± 37	26 ± 36	59 ± 42	28 ± 39	22 ± 18
CENSOR	73 ± 33	64 ± 28	**70** ± 24	47 ± 34	48 ± 34	48 ± 34	58 ± 41	58 ± 10
SANTA	64 ± 10	59 ± 8	57 ± 5	59 ± 7	62 ± 5	65 ± 8	87 ± 5	65 ± 9
MProto	45 ± 7	51 ± 7	49 ± 7	52 ± 10	42 ± 14	49 ± 8	56 ± 21	49 ± 4
StEmb	77 ± 4	70 ± 9	66 ± 1	76 ± 8	76 ± 7	78 ± 11	95 ± 1	77 ± 8
Matching	65 ± 17	60 ± 9	52 ± 7	52 ± 10	52 ± 9	52 ± 9		
CuPUL+XLM	74 ± 3	77 ± 4	71 ± 4	67 ± 6	68 ± 7	**69** ± 10	**92** ± 3	**74** ± 7
CuPUL+StEmb	**76** ± 2	**79** ± 4	**72** ± 3	**72** ± 6	**69** ± 8	67 ± 10	86 ± 5	**74** ± 5

Table 4: Test results in terms of Precision of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	42 ± 30	26 ± 37	28 ± 39	41 ± 34	38 ± 32	27 ± 39	48 ± 35	36 ± 7
CuPUL	**68** ± 3	**74** ± 2	**79** ± 6	**83** ± 6	84 ± 5	85 ± 5	88 ± 6	**80** ± 6
BOND	23 ± 16	50 ± 19	71 ± 14	81 ± 8	84 ± 5	92 ± 1	93 ± 2	71 ± 22
SCDL	1 ± 1	1 ± 1	8 ± 9	30 ± 41	30 ± 41	31 ± 42	31 ± 42	19 ± 13
CENSOR	2 ± 3	6 ± 4	34 ± 38	53 ± 39	56 ± 40	59 ± 42	61 ± 43	39 ± 22
SANTA	37 ± 4	51 ± 4	69 ± 3	80 ± 6	**87** ± 4	**93** ± 2	**94** ± 3	73 ± 19
MProto	45 ± 5	57 ± 5	66 ± 14	70 ± 12	68 ± 15	74 ± 10	78 ± 9	66 ± 10
StEmb	10 ± 7	29 ± 19	46 ± 8	70 ± 8	84 ± 3	95 ± 1	95 ± 1	61 ± 29
Matching	9 ± 6	15 ± 5	25 ± 4	35 ± 2	42 ± 3	52 ± 2		
CuPUL+XLM	64 ± 9	80 ± 5	85 ± 3	88 ± 3	**91** ± 2	**92** ± 2	**93** ± 2	85 ± 9
CuPUL+StEmb	**80** ± 5	**84** ± 2	**88** ± 1	**89** ± 1	89 ± 1	89 ± 2	91 ± 3	**87** ± 4

Table 5: Test results in terms of Recall of Fully supervised NER and Weakly supervised NER varying the dictionary sample size. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	55	0	0	43	37	0	71	30 ± 26
CuPUL	**66**	**67**	**62**	63	63	64	75	**66** ± 4
BOND	29	43	60	**70**	**75**	**86**	**90**	65 ± 19
SCDL	1	1	1	1	1	1	1	1 ± 0
CENSOR	0	0	0	0	0	0	0	0 ± 0
SANTA	46	53	59	59	68	79	85	64 ± 12
MProto	37	49	47	46	31	47	39	42 ± 6
StEmb	32	66	59	83	83	92	93	73 ± 19
Matching	27	33	38	41	44	51		
CuPUL+XLM	77	**82**	**80**	75	77	**82**	**89**	80 ± 4
CuPUL+StEmb	**80**	81	**80**	**84**	**80**	80	83	**81** ± 1

Conf-MPU	0	71	67	64	51	51	56	52 ± 21
CuPUL	**68**	68	69	65	63	61	89	**69** ± 8
BOND	55	**74**	72	73	66	61	86	**69** ± 9
SCDL	0	0	26	82	82	**82**	86	51 ± 35
CENSOR	1	12	**75**	**85**	**85**	76	91	61 ± 33
SANTA	45	59	64	71	72	68	**94**	68 ± 13
MProto	51	57	69	64	62	64	83	64 ± 9
StEmb	6	22	57	66	76	76	96	57 ± 28
Matching	5	17	33	39	43	45		
CuPUL+XLM	65	74	75	72	**71**	**69**	**93**	74 ± 8
CuPUL+StEmb	**80**	**78**	**78**	**74**	69	67	91	**77** ± 7

Conf-MPU	57	0	0	0	0	0	0	8 ± 19
CuPUL	**66**	**68**	69	75	73	73	88	**73** ± 6
BOND	12	53	**70**	**81**	**78**	**85**	**95**	68 ± 24
SCDL	0	0	0	0	0	0	0	0 ± 0
CENSOR	11	17	22	64	69	83	87	51 ± 28
SANTA	46	51	65	74	**78**	80	91	69 ± 14
MProto	47	53	53	70	63	66	69	60 ± 8
StEmb	14	27	46	70	81	88	95	60 ± 27
Matching	15	23	29	44	50	57		
CuPUL+XLM	64	78	78	**83**	**84**	**85**	**95**	81 ± 8
CuPUL+StEmb	**73**	**86**	**80**	82	83	83	91	**82** ± 5

Table 6: Test results in terms of F1-Score of Fully supervised NER and Weakly supervised NER varying the dictionary sample size on each of the three datasets (CBMA, CDBE, HOME), respectively. Measures are micro-averaged among classes. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	52	0	0	48	39	0	63	29 ± 24
CuPUL	64	62	54	55	53	54	70	59 ± 6
BOND	57	59	65	**69**	**73**	**80**	**89**	**70** ± 10
SCDL	1	1	1	1	1	1	1	1 ± 0
CENSOR	**100**	**100**	**100**	0	0	0	0	43 ± 46
SANTA	51	50	52	50	58	70	80	59 ± 10
MProto	35	42	41	40	23	39	28	35 ± 6
StEmb	74	80	68	85	83	89	93	82 ± 8
Matching	74	66	58	52	52	53		
CuPUL+XLM	**77**	**78**	73	66	68	**74**	**87**	75 ± 6
CuPUL+StEmb	76	76	**75**	**78**	**72**	73	79	**76** ± 2
Conf-MPU	**100**	65	56	52	38	37	51	57 ± 18
CuPUL	64	60	57	52	49	46	86	59 ± 12
BOND	73	**72**	60	62	52	46	80	63 ± 11
SCDL	0	0	37	**78**	77	**75**	83	50 ± 32
CENSOR	28	32	**67**	**78**	**79**	67	90	63 ± 21
SANTA	74	68	57	60	59	53	**92**	**66** ± 12
MProto	49	54	58	52	49	51	77	56 ± 9
StEmb	74	57	66	66	66	64	96	70 ± 11
Matching	41	48	42	41	41	40		
CuPUL+XLM	70	72	66	61	**59**	**55**	**93**	68 ± 11
CuPUL+StEmb	**79**	**76**	**69**	**63**	57	54	90	**70** ± 11
Conf-MPU	50	**100**	**100**	**100**	**100**	**100**	**100**	**93** ± 16
CuPUL	69	65	63	66	63	62	85	68 ± 7
BOND	**100**	78	72	77	71	78	95	82 ± 10
SCDL	0	0	0	0	0	**100**	0	14 ± 33
CENSOR	93	59	42	63	64	78	85	69 ± 15
SANTA	68	57	63	67	70	70	88	69 ± 8
MProto	51	57	48	65	56	58	62	57 ± 5
StEmb	83	72	66	78	80	82	96	79 ± 8
Matching	80	67	55	65	62	62		
CuPUL+XLM	**76**	82	**76**	75	**76**	**78**	**94**	**80** ± 6
CuPUL+StEmb	73	**85**	74	**76**	**76**	75	88	78 ± 5

Table 7: Test results in terms of Precision of Fully supervised NER and Weakly supervised NER varying the dictionary sample size on each of the three datasets (CBMA, CDBE, HOME), respectively. Measures are micro-averaged among classes. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

Sample (%)	10	20	40	60	80	100	Fully	MEAN
Conf-MPU	59	0	0	39	35	0	81	31 ± 28
CuPUL	**68**	**74**	**73**	**74**	77	78	80	**75** ± 3
BOND	20	33	55	70	77	**92**	**92**	63 ± 24
SCDL	2	2	2	2	3	3	3	2 ± 0
CENSOR	0	0	0	0	0	0	0	0 ± 0
SANTA	42	55	67	71	**82**	90	90	71 ± 15
MProto	40	60	53	54	49	61	67	55 ± 8
StEmb	20	56	52	82	83	95	93	69 ± 24
Matching	17	22	28	34	39	49		
CuPUL+XLM	76	**86**	**87**	85	**88**	**92**	**91**	86 ± 4
CuPUL+StEmb	**84**	**86**	86	**90**	**88**	88	87	**87** ± 2
Conf-MPU	0	**78**	83	84	79	82	63	67 ± 26
CuPUL	**72**	**78**	87	86	88	88	92	**84** ± 6
BOND	45	77	**90**	88	89	91	92	82 ± 15
SCDL	0	0	21	87	89	90	90	54 ± 38
CENSOR	0	8	87	**92**	**93**	89	93	66 ± 37
SANTA	33	52	73	86	92	**95**	**96**	75 ± 21
MProto	53	61	85	81	84	85	90	77 ± 12
StEmb	3	13	50	66	88	94	96	59 ± 33
Matching	3	10	27	37	45	53		
CuPUL+XLM	61	77	87	87	**90**	**90**	**94**	84 ± 10
CuPUL+StEmb	**82**	**81**	**90**	**89**	89	88	93	**87** ± 4
Conf-MPU	**66**	0	0	0	0	0	0	9 ± 22
CuPUL	64	**72**	**77**	**87**	**87**	89	91	**81** ± 9
BOND	6	40	68	85	86	**94**	**95**	68 ± 29
SCDL	0	0	0	0	0	0	0	0 ± 0
CENSOR	6	10	15	66	76	90	90	50 ± 33
SANTA	34	46	67	82	**87**	93	94	72 ± 20
MProto	43	50	60	75	72	77	77	65 ± 12
StEmb	8	17	35	63	81	94	95	56 ± 32
Matching	8	14	19	33	42	53		
CuPUL+XLM	55	75	81	**92**	**94**	**95**	**96**	84 ± 13
CuPUL+StEmb	**73**	**86**	**87**	89	92	92	94	**87** ± 6

Table 8: Test results in terms of Recall of Fully supervised NER and Weakly supervised NER varying the dictionary sample size on each of the three datasets (CBMA, CDBE, HOME), respectively. Measures are micro-averaged among classes. The MEAN column reports the performance averaged across settings with its standard deviation. The best results among DS-NER techniques for each setting are reported in bold. For reference, the Stacked Embeddings (StEmb) supervised architecture and the pure dictionary matching are reported, too. The bottom rows report performances of the CuPUL technique paired with XLM-RoBERTa (XLM) and Stacked Embeddings, respectively, as underlying classification model.

E	Neg	Loss	m	10	20	40	60	80	100	Fully	MEAN
10	0.1	MPN	10	53 ± 37	54 ± 38	55 ± 39	54 ± 38	83 ± 5	82 ± 6	92 ± 2	68 ± 15
10	0.1	MPN	20	76 ± 5	**82** ± 5	81 ± 3	81 ± 3	80 ± 6	80 ± 7	90 ± 2	81 ± 4
10	0.1	MPN-CE	10	75 ± 1	79 ± 2	83 ± 3	83 ± 3	84 ± 6	83 ± 7	93 ± 2	**83** ± 5
10	0.1	MPN-CE	20	73 ± 5	79 ± 3	81 ± 2	82 ± 3	81 ± 6	80 ± 8	92 ± 1	81 ± 5
10	0.3	MPN	10	54 ± 38	53 ± 37	55 ± 39	54 ± 38	83 ± 5	82 ± 7	92 ± 3	67 ± 15
10	0.3	MPN	20	74 ± 9	81 ± 6	81 ± 4	80 ± 3	80 ± 5	78 ± 7	90 ± 3	80 ± 4
10	0.3	MPN-CE	10	75 ± 2	79 ± 1	83 ± 3	83 ± 3	83 ± 5	83 ± 7	93 ± 1	**83** ± 5
10	0.3	MPN-CE	20	73 ± 4	80 ± 2	82 ± 2	82 ± 3	81 ± 6	81 ± 8	92 ± 1	81 ± 5
15	0.1	MPN	10	53 ± 37	55 ± 39	**87** ± 2	**86** ± 3	85 ± 5	83 ± 6	92 ± 2	77 ± 14
15	**0.1**	**MPN**	**20**	**80** ± 2	81 ± 7	84 ± 2	81 ± 2	79 ± 6	80 ± 8	91 ± 1	82 ± 3
15	0.1	MPN-CE	10	69 ± 5	75 ± 4	81 ± 3	85 ± 5	84 ± 7	84 ± 8	**94** ± 0	82 ± 7
15	0.1	MPN-CE	20	71 ± 5	75 ± 6	79 ± 0	83 ± 4	82 ± 6	82 ± 8	92 ± 1	81 ± 6
15	0.3	MPN	10	51 ± 36	53 ± 37	**87** ± 2	85 ± 3	85 ± 5	**85** ± 7	92 ± 2	77 ± 15
15	0.3	MPN	20	78 ± 2	81 ± 6	83 ± 1	82 ± 3	79 ± 6	81 ± 7	90 ± 1	82 ± 3
15	0.3	MPN-CE	10	69 ± 4	75 ± 2	81 ± 3	85 ± 5	84 ± 7	84 ± 8	**94** ± 1	82 ± 7
15	0.3	MPN-CE	20	70 ± 5	75 ± 6	81 ± 2	83 ± 4	82 ± 6	81 ± 9	92 ± 1	81 ± 6
20	0.1	MPN	10	52 ± 37	53 ± 38	84 ± 3	**86** ± 3	**86** ± 5	84 ± 6	93 ± 1	77 ± 15
20	0.1	MPN	20	77 ± 2	81 ± 7	84 ± 2	83 ± 3	82 ± 6	82 ± 8	91 ± 1	**83** ± 3
20	0.1	MPN-CE	10	67 ± 6	71 ± 6	80 ± 3	84 ± 5	84 ± 6	**85** ± 9	**94** ± 1	81 ± 8
20	0.1	MPN-CE	20	68 ± 7	73 ± 6	79 ± 2	83 ± 5	82 ± 6	83 ± 9	**94** ± 1	80 ± 7
20	0.3	MPN	10	51 ± 36	53 ± 38	**87** ± 3	**86** ± 3	**86** ± 5	**85** ± 7	92 ± 1	77 ± 15
20	0.3	MPN	20	79 ± 0	81 ± 7	82 ± 1	83 ± 4	81 ± 6	81 ± 8	91 ± 2	**83** ± 3
20	0.3	MPN-CE	10	67 ± 5	72 ± 5	81 ± 5	84 ± 6	84 ± 6	**85** ± 9	**94** ± 1	81 ± 8
20	0.3	MPN-CE	20	68 ± 7	73 ± 6	80 ± 2	83 ± 5	82 ± 7	83 ± 9	**94** ± 1	81 ± 7

Table 9: Validation results in terms of F1-Score of Stacked Embeddings voters. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold.

E	Neg	Loss	m	10	20	40	60	80	100	Fully	MEAN
10	0.1	MPN	10	40 ± 8	68 ± 1	81 ± 3	$\mathbf{87} \pm 4$	90 ± 3	$\mathbf{92} \pm 1$	92 ± 3	78 ± 16
10	0.1	MPN	20	39 ± 4	69 ± 3	79 ± 5	84 ± 7	90 ± 3	89 ± 4	92 ± 4	77 ± 16
10	0.1	MPN-CE	10	25 ± 5	40 ± 7	56 ± 8	77 ± 5	87 ± 2	88 ± 3	94 ± 3	67 ± 23
10	0.1	MPN-CE	20	32 ± 11	44 ± 5	60 ± 9	77 ± 2	87 ± 0	91 ± 0	93 ± 4	69 ± 21
10	0.1	MPU	10	69 ± 11	55 ± 19	47 ± 1	45 ± 2	43 ± 3	36 ± 7	42 ± 4	48 ± 9
10	0.1	MPU	20	65 ± 17	50 ± 10	56 ± 15	52 ± 14	39 ± 1	39 ± 1	48 ± 8	50 ± 8
10	0.1	MPU-CE	10	64 ± 9	60 ± 16	61 ± 7	54 ± 13	51 ± 12	46 ± 12	40 ± 16	54 ± 7
10	0.1	MPU-CE	20	58 ± 13	55 ± 13	54 ± 10	45 ± 15	41 ± 10	36 ± 12	43 ± 9	47 ± 7
10	0.3	MPN	10	31 ± 7	69 ± 3	80 ± 3	86 ± 6	90 ± 3	90 ± 2	93 ± 2	77 ± 19
10	0.3	MPN	20	35 ± 9	67 ± 0	81 ± 4	86 ± 5	88 ± 5	89 ± 4	91 ± 3	77 ± 17
10	0.3	MPN-CE	10	27 ± 7	38 ± 6	59 ± 6	77 ± 4	86 ± 2	88 ± 2	94 ± 2	67 ± 23
10	0.3	MPN-CE	20	32 ± 11	44 ± 5	62 ± 6	78 ± 3	87 ± 0	91 ± 0	94 ± 3	70 ± 21
10	0.3	MPU	10	$\mathbf{70} \pm 11$	52 ± 14	46 ± 3	47 ± 0	47 ± 11	48 ± 9	44 ± 3	50 ± 8
10	0.3	MPU	20	63 ± 20	57 ± 17	53 ± 15	55 ± 13	38 ± 0	37 ± 2	41 ± 1	49 ± 9
10	0.3	MPU-CE	10	65 ± 9	59 ± 13	64 ± 7	55 ± 14	52 ± 11	38 ± 19	51 ± 9	55 ± 8
10	0.3	MPU-CE	20	58 ± 11	59 ± 12	57 ± 8	45 ± 15	40 ± 12	34 ± 15	56 ± 14	50 ± 9
10	0.5	MPN	10	36 ± 11	67 ± 3	80 ± 3	85 ± 2	90 ± 2	$\mathbf{92} \pm 1$	93 ± 3	78 ± 18
10	0.5	MPN	20	40 ± 8	71 ± 5	$\mathbf{82} \pm 7$	85 ± 6	90 ± 3	88 ± 3	93 ± 3	78 ± 16
10	0.5	MPN-CE	10	26 ± 6	39 ± 5	61 ± 5	78 ± 3	86 ± 2	87 ± 3	94 ± 2	67 ± 23
10	0.5	MPN-CE	20	33 ± 11	44 ± 3	63 ± 5	77 ± 2	85 ± 2	91 ± 1	$\mathbf{95} \pm 2$	70 ± 21
10	0.5	MPU	10	69 ± 12	61 ± 21	43 ± 2	44 ± 1	44 ± 1	48 ± 8	45 ± 2	51 ± 9
10	0.5	MPU	20	63 ± 17	56 ± 19	57 ± 13	44 ± 8	41 ± 3	35 ± 6	45 ± 7	49 ± 9
10	0.5	MPU-CE	10	63 ± 9	61 ± 12	62 ± 11	52 ± 13	48 ± 11	39 ± 18	55 ± 21	54 ± 8
10	0.5	MPU-CE	20	56 ± 15	55 ± 16	54 ± 11	44 ± 12	42 ± 13	31 ± 18	51 ± 15	48 ± 8

Table 10: Validation results in terms of F1-Score of XML-RoBERTa voters – Part I. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold.

E	Neg	Loss	m	10	20	40	60	80	100	Fully	MEAN
5	0.1	MPN	10	55 ± 2	74 ± 14	**82** ± 7	86 ± 6	91 ± 2	90 ± 2	93 ± 3	81 ± 12
5	0.1	MPN	20	55 ± 2	81 ± 6	77 ± 11	78 ± 8	80 ± 11	82 ± 8	89 ± 5	78 ± 9
5	0.1	MPN-CE	10	59 ± 9	75 ± 1	77 ± 2	82 ± 2	87 ± 2	87 ± 3	93 ± 3	80 ± 10
5	0.1	MPN-CE	20	64 ± 9	75 ± 5	75 ± 7	81 ± 6	83 ± 4	85 ± 5	92 ± 4	79 ± 8
5	0.1	MPU	10	67 ± 12	62 ± 8	64 ± 10	49 ± 3	44 ± 1	43 ± 6	36 ± 7	52 ± 11
5	0.1	MPU	20	67 ± 15	65 ± 14	59 ± 11	57 ± 8	45 ± 8	43 ± 9	37 ± 7	53 ± 10
5	0.1	MPU-CE	10	60 ± 11	62 ± 14	56 ± 17	51 ± 9	40 ± 14	41 ± 2	34 ± 0	49 ± 9
5	0.1	MPU-CE	20	56 ± 13	55 ± 16	53 ± 14	47 ± 8	41 ± 6	28 ± 3	32 ± 1	45 ± 10
5	0.3	MPN	10	59 ± 2	77 ± 10	**82** ± 8	86 ± 5	90 ± 3	90 ± 2	92 ± 2	**82** ± 10
5	0.3	MPN	20	59 ± 5	**82** ± 8	76 ± 11	79 ± 9	80 ± 12	82 ± 8	89 ± 4	78 ± 8
5	0.3	MPN-CE	10	59 ± 9	76 ± 2	78 ± 1	84 ± 2	86 ± 1	87 ± 3	92 ± 3	80 ± 9
5	0.3	MPN-CE	20	63 ± 9	76 ± 6	75 ± 7	81 ± 5	83 ± 5	85 ± 5	91 ± 5	79 ± 8
5	0.3	MPU	10	67 ± 13	54 ± 2	49 ± 8	45 ± 5	42 ± 4	41 ± 9	43 ± 10	49 ± 8
5	0.3	MPU	20	67 ± 15	64 ± 16	60 ± 12	48 ± 3	47 ± 3	43 ± 10	41 ± 5	53 ± 9
5	0.3	MPU-CE	10	61 ± 12	61 ± 16	57 ± 16	54 ± 10	45 ± 16	40 ± 8	34 ± 4	50 ± 9
5	0.3	MPU-CE	20	58 ± 12	56 ± 16	52 ± 14	46 ± 9	42 ± 7	26 ± 1	32 ± 0	44 ± 10
5	0.5	MPN	10	48 ± 10	76 ± 11	81 ± 8	86 ± 5	**92** ± 3	91 ± 1	92 ± 2	81 ± 14
5	0.5	MPN	20	57 ± 1	**82** ± 7	77 ± 11	79 ± 9	80 ± 12	82 ± 8	89 ± 5	78 ± 9
5	0.5	MPN-CE	10	60 ± 8	76 ± 2	76 ± 2	82 ± 3	86 ± 2	87 ± 3	93 ± 4	80 ± 9
5	**0.5**	**MPN-CE**	**20**	65 ± 8	76 ± 6	75 ± 7	81 ± 5	82 ± 4	85 ± 4	92 ± 4	80 ± 7
5	0.5	MPU	10	66 ± 15	54 ± 1	52 ± 3	48 ± 4	42 ± 4	39 ± 7	35 ± 2	48 ± 9
5	0.5	MPU	20	**70** ± 18	67 ± 11	61 ± 12	46 ± 2	46 ± 4	43 ± 9	41 ± 6	53 ± 10
5	0.5	MPU-CE	10	59 ± 12	60 ± 13	57 ± 17	53 ± 11	51 ± 11	38 ± 11	42 ± 10	51 ± 7
5	0.5	MPU-CE	20	59 ± 13	56 ± 14	53 ± 13	49 ± 9	41 ± 9	32 ± 6	46 ± 12	48 ± 8

Table 11: Validation results in terms of F1-Score of XML-RoBERTa voters – Part II. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold.

E	Loss	10	20	40	60	80	100	Fully	MEAN
15	Conf-MPU-CE	77 ± 1	80 ± 6	82 ± 3	81 ± 4	79 ± 7	79 ± 9	90 ± 1	81 ± 4
15	Conf-MPU	**81** ± 1	83 ± 5	81 ± 3	83 ± 4	80 ± 8	78 ± 8	89 ± 3	82 ± 3
15	MPN-CE	72 ± 5	77 ± 3	80 ± 2	82 ± 4	82 ± 7	80 ± 8	92 ± 1	81 ± 5
15	MPN	78 ± 2	81 ± 5	82 ± 2	82 ± 3	82 ± 7	80 ± 7	90 ± 3	82 ± 3
15	MPU-CE	64 ± 12	66 ± 12	59 ± 14	51 ± 14	44 ± 14	29 ± 18	43 ± 21	51 ± 12
15	MPU	59 ± 26	66 ± 23	62 ± 23	60 ± 20	58 ± 17	47 ± 20	58 ± 33	59 ± 5
20	Conf-MPU-CE	77 ± 2	78 ± 8	81 ± 2	81 ± 4	81 ± 7	80 ± 9	91 ± 1	81 ± 4
20	Conf-MPU	77 ± 2	82 ± 6	83 ± 2	82 ± 5	80 ± 6	79 ± 8	90 ± 3	82 ± 4
20	MPN-CE	69 ± 5	74 ± 6	80 ± 1	83 ± 5	82 ± 6	**83** ± 9	**94** ± 1	80 ± 7
20	**MPN**	80 ± 0	83 ± 6	**85** ± 3	**84** ± 4	82 ± 6	82 ± 7	90 ± 3	**84** ± 3
20	MPU-CE	63 ± 11	63 ± 17	58 ± 15	51 ± 12	43 ± 15	27 ± 18	38 ± 19	49 ± 12
20	MPU	65 ± 18	64 ± 26	74 ± 9	68 ± 10	63 ± 11	60 ± 12	61 ± 29	65 ± 4
25	Conf-MPU-CE	76 ± 3	81 ± 5	82 ± 3	82 ± 4	82 ± 7	81 ± 9	91 ± 1	82 ± 4
25	Conf-MPU	79 ± 2	82 ± 6	84 ± 3	82 ± 5	79 ± 7	80 ± 8	90 ± 3	82 ± 3
25	MPN-CE	70 ± 5	75 ± 4	78 ± 1	82 ± 4	**83** ± 6	**83** ± 9	93 ± 1	81 ± 6
25	MPN	77 ± 3	**84** ± 6	84 ± 2	**84** ± 3	**83** ± 7	82 ± 8	91 ± 2	**84** ± 3
25	MPU-CE	65 ± 12	61 ± 17	58 ± 16	50 ± 15	44 ± 16	27 ± 18	36 ± 20	49 ± 12
25	MPU	70 ± 13	67 ± 23	76 ± 6	68 ± 8	65 ± 9	61 ± 9	63 ± 26	67 ± 4

Table 12: Validation results in terms of F1-Score of CuPUL curriculum train with Stacked Embeddings. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold.

E	sE	Loss	LR	10	20	40	60	80	100	Fully	MEAN
1	1	Conf-MPU	1	28 ± 24	36 ± 31	40 ± 27	42 ± 33	43 ± 37	42 ± 38	44 ± 40	39 ± 5
1	1	Conf-MPU	3	51 ± 29	50 ± 33	57 ± 28	57 ± 32	58 ± 30	53 ± 35	56 ± 36	55 ± 3
1	1	Conf-MPU-CE	1	29 ± 25	43 ± 33	39 ± 28	36 ± 34	38 ± 35	39 ± 34	44 ± 41	38 ± 4
1	1	Conf-MPU-CE	3	67 ± 15	68 ± 13	61 ± 17	60 ± 18	60 ± 17	66 ± 14	74 ± 17	65 ± 5
1	1	MPU-CE	1	27 ± 23	37 ± 31	38 ± 26	31 ± 28	30 ± 27	26 ± 23	33 ± 30	32 ± 4
1	1	MPU-CE	3	59 ± 15	57 ± 6	46 ± 7	53 ± 10	37 ± 5	34 ± 6	27 ± 12	45 ± 11
1	2	MPU-CE	1	47 ± 17	53 ± 11	48 ± 16	46 ± 13	45 ± 10	43 ± 5	35 ± 5	45 ± 5
1	2	MPU-CE	3	53 ± 12	61 ± 6	54 ± 9	52 ± 11	35 ± 7	42 ± 7	30 ± 10	47 ± 10
5	1	Conf-MPU	1	70 ± 8	76 ± 9	69 ± 14	73 ± 14	74 ± 17	79 ± 12	79 ± 14	74 ± 3
5	1	Conf-MPU	3	69 ± 0	80 ± 4	79 ± 7	86 ± 3	83 ± 7	89 ± 1	92 ± 3	82 ± 7
5	1	Conf-MPU-CE	1	69 ± 5	78 ± 10	74 ± 11	76 ± 7	76 ± 8	80 ± 5	91 ± 3	78 ± 6
5	1	Conf-MPU-CE	3	70 ± 7	80 ± 2	81 ± 1	84 ± 3	88 ± 1	87 ± 1	93 ± 2	83 ± 6
5	1	MPU-CE	1	56 ± 9	55 ± 11	52 ± 14	49 ± 13	46 ± 12	39 ± 7	39 ± 13	48 ± 6
5	1	MPU-CE	3	52 ± 12	67 ± 3	57 ± 8	57 ± 1	47 ± 4	50 ± 5	42 ± 11	53 ± 7
5	2	Conf-MPU	1	66 ± 9	79 ± 4	81 ± 3	88 ± 1	85 ± 7	90 ± 2	92 ± 3	83 ± 8
5	2	Conf-MPU	3	61 ± 3	**84** ± 0	80 ± 6	87 ± 3	89 ± 1	**91** ± 1	94 ± 2	**84** ± 9
5	2	Conf-MPU-CE	1	**73** ± 14	81 ± 3	81 ± 4	78 ± 8	83 ± 5	88 ± 0	93 ± 3	83 ± 6
5	2	Conf-MPU-CE	3	64 ± 14	77 ± 3	81 ± 4	81 ± 7	84 ± 5	90 ± 0	94 ± 3	82 ± 8
5	2	MPU-CE	1	65 ± 7	63 ± 3	56 ± 4	53 ± 5	49 ± 10	44 ± 6	57 ± 12	55 ± 7
5	2	MPU-CE	3	63 ± 3	60 ± 13	50 ± 17	66 ± 5	59 ± 10	44 ± 21	65 ± 21	58 ± 7
10	1	Conf-MPU	1	70 ± 5	79 ± 5	80 ± 5	85 ± 4	84 ± 6	88 ± 3	93 ± 2	83 ± 6
10	1	Conf-MPU	3	65 ± 10	78 ± 2	82 ± 4	89 ± 0	87 ± 4	89 ± 2	92 ± 3	83 ± 8
10	**1**	**Conf-MPU-CE**	**1**	**73** ± 11	82 ± 2	81 ± 4	80 ± 8	85 ± 4	89 ± 0	92 ± 5	83 ± 5
10	1	Conf-MPU-CE	3	62 ± 16	75 ± 6	**86** ± 0	83 ± 5	86 ± 2	90 ± 0	94 ± 3	82 ± 9
10	1	MPU-CE	1	62 ± 8	63 ± 11	54 ± 8	51 ± 10	51 ± 6	47 ± 3	60 ± 13	55 ± 5
10	1	MPU-CE	3	57 ± 7	63 ± 17	52 ± 20	51 ± 20	60 ± 1	51 ± 11	56 ± 31	56 ± 4
10	2	Conf-MPU	1	61 ± 8	83 ± 2	85 ± 1	86 ± 3	91 ± 1	90 ± 1	92 ± 4	**84** ± 9
10	2	Conf-MPU	3	54 ± 9	83 ± 2	**86** ± 0	**90** ± 0	**92** ± 1	**91** ± 1	91 ± 4	**84** ± 12
10	2	Conf-MPU-CE	1	67 ± 17	77 ± 3	78 ± 7	79 ± 8	85 ± 4	**91** ± 1	**95** ± 2	82 ± 8
10	2	Conf-MPU-CE	3	57 ± 17	80 ± 3	82 ± 1	80 ± 8	87 ± 1	90 ± 0	**95** ± 2	82 ± 11
10	2	MPU-CE	1	60 ± 13	55 ± 9	62 ± 3	54 ± 11	55 ± 9	46 ± 6	67 ± 19	57 ± 6
10	2	MPU-CE	3	67 ± 8	62 ± 14	67 ± 7	63 ± 15	72 ± 5	59 ± 17	63 ± 14	65 ± 4

Table 13: Validation results in terms of F1-Score of CuPUL curriculum train with XLM-RoBERTa. Measures are micro-averaged among classes and then, for each setting, averaged among datasets and reported in percentage with their standard deviation. The MEAN column reports the performance averaged across settings with its standard deviation. The best results for each setting are reported in bold. Selected hyperparameters are in bold. Learning rate is in the scale of $1e{-5}$