EXTRACTING BOOK TITLES FROM HISTORICAL NEWSPAPER ARCHIVES: A NAMED ENTITY RECOGNITION APPROACH

Niels Bijl

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Applied Data Science

Department of Information and Computing Science, Utrecht University

First examiner: Prof. dr. Joris van Eijnatten
Second examiner: Dr. Marijn Schraagen

7 July 2024

Abstract

This thesis introduces a novel method for extracting book titles from Optical Character Recognition scanned historical newspaper archives using Named Entity Recognition (NER), a task not extensively explored in existing literature. By identifying books highlighted by reviewers and journalists, we can gain insights into the evolving cultural and literary tastes of society.

Utilizing a dataset from the *Leeuwarder Courant*, the study applies various NER models, including BiLSTM-CRF and transformer-based models. The transformer-based models outperformed others, achieving an F1 score of 84.3% on the test dataset, demonstrating the effectiveness of these models in extracting text representing book titles from newspaper archives.

In addition to assessing performance on a NER level, an evaluation was conducted to measure how well the best NER model could identify the actual discussed books. This was achieved by matching the extracted book title text to the titles in the *Nederlandse Bibliografie Totaal* (*NBT*), a comprehensive compilation of all books published by Dutch publishers. Despite high NER performance, the matching process yielded a suboptimal F1 score of 59.4%. This gap was primarily due to the training data not being specifically labeled for NER purposes, making its repurposing as a NER dataset inadequate. Consequently, the model often missed subtitles, resulting in incomplete title extraction.

Further analysis showed that even with perfect NER predictions, matching titles to the *NBT* achieved an F1 score of only 65.5%. This finding highlights the need for additional information besides the main title, such as subtitles, authors, and potentially publishers, to improve the accuracy of title matching to the *NBT*.

Preface & Acknowledgments

This thesis has been a remarkable journey, filled with learning and growth. At the onset, I had never undertaken a large Natural Language Processing project and had never heard of the concept Named Entity Recognition. Over the course of this 10-week project, I have learned an incredible amount, expanding my knowledge and skills in ways I never anticipated.

I would like to express my gratitude to my supervisor, Joris van Eijnatten, for his invaluable feedback and guidance. His willingness to read and provide constructive comments on several parts of the thesis multiple times has been instrumental in shaping the final work. Additionally, I thank him for creating the *Leeuwarder Courant* dataset, as this study would not have been possible without it. I am also grateful to Philippa Linardatos for assisting with data collection and for the insightful discussions about the subject matter.

Coming from a computer science background, I was unfamiliar with the Faculty of Humanities before starting this thesis. When I first encountered the term, I had no idea what it encompassed. This project has not only introduced me to new technical concepts, but also broadened my understanding of interdisciplinary research, bridging the gap between computer science and the humanities.

The final model developed during this project is openly accessible on *Hugging Face*, adhering to the open science principles of open access and open collaboration. Additionally, the source code is available on *GitHub*, in line with the open source principle, ensuring transparency and enabling further research and development by the scientific community.

Contents

Chapter 1

Introduction

Book reviews in newspapers not only reflect the reading preferences of their time, but also offer a window into the shifting intellectual landscapes that characterize different eras. By analyzing which books were highlighted by reviewers and journalists, we can gain insights into the evolving cultural and literary tastes of society.

Notably, to date, no published research specifically targets the extraction of book titles from Optical Character Recognition (OCR)-scanned historical newspapers. Previous studies, including those by Do et al. (2012) and Sarimehmetoğlu and Erdem (2023), have focused on extracting book titles from visual sources, such as book cover images and video content. These methods, while innovative, rely on visual attributes and are fundamentally different from our text-based approach.

The method currently employed by the thesis supervisor to identify book titles in historical newspaper texts is predominantly rule-based (Van Eijnatten, 2024). This approach involves selecting segments of 600 characters, identified by a set of rules as those most likely to contain a book title, and comparing them against a database of known book titles to ascertain matches. This method suffers from several drawbacks: it yields suboptimal accuracy, it requires extensive manual verification, and specific rules are tailored for a particular newspaper, making it potentially hard to generalize to other newspapers. These limitations compromise the efficacy of cultural analysis over time and restrict our understanding of historical intellectual trends.

Despite its drawbacks, the current method has successfully produced a substantial dataset, serving as a valuable foundation for developing an improved model. In response to these challenges and leveraging the existing dataset, this study proposes a novel approach utilizing advancements in Natural Language Processing (NLP). By harnessing NLP models, we aim to develop a more robust and autonomous system capable of accurately extracting book titles from a vast corpus of newspaper text.

Eisenstein (2019) describes NLP as a set of methods for making human language accessible to computers. Key applications of NLP include language translation, sentiment analysis, speech recognition, text summarization, and named entity recognition. Named Entity Recognition (NER) is the process of identifying named entities in text. Commonly used entities are people, locations, and organizations (Jurafsky & Martin, 2023).

This thesis explores the innovative idea of categorizing book titles as a specialized form of named entity, a concept that, until now, has not been investigated in the context of historical newspapers. Which leads us to the research question of this thesis: *To what extent can Named Entity Recognition be utilized to autonomously extract book titles from OCR-scanned historical newspapers, thereby facilitating deeper cultural and literary analyses?*

The remainder of this thesis is organized as follows:

- **Chapter 2: Concepts and Literature** - This chapter introduces the essential concepts required to comprehend the thesis. It includes a review of relevant literature, comparing various NER methods based on existing research.
- **Chapter 3: Data** - This chapter details the processes involved in creating the datasets used in the study, along with a brief data analysis.
- **Chapter 4: Methodology** - This chapter describes the methodology employed to determine the most effective NER model for extracting book titles from historical newspapers.
- **Chapter 5: Results** - This chapter presents the findings obtained from the applied methodology.
- **Chapter 6: Discussion** - This chapter interprets the results, offering insights into why they turned out as they did, discussing their implications, and comparing them with claims from the existing literature.
- **Chapter 7: Conclusion** - The final chapter addresses the research question, summarizes the key findings, and proposes directions for future research.

Chapter 2

Concepts and Literature

This chapter outlines the key concepts and literature required to comprehend the thesis. We begin with an exploration of Named Entity Recognition (NER) within the field of Natural Language Processing (NLP), discussing its challenges, tagging schemes, and evaluation metrics. Next, we examine various NER approaches, including rule-based, unsupervised, and supervised methods, with a particular emphasis on supervised learning. The chapter further highlights the evolution from traditional machine learning to transformer-based Large Language Models. Finally, we delve into the methods of text representation in NER models, focusing on word embeddings.

2.1 Named Entity Recognition

As introduced in Chapter 1, titled "Introduction", NER is a fundamental task in NLP that involves identifying and classifying entities in text into predefined categories, in most cases: names of people, organizations, and locations. According to Jurafsky & Martin (2023), a named entity is defined as any item that can be distinctly identified by a proper name.

Bird et al. (2009) emphasized two major challenges in NER. The first challenge was the ambiguity of many named entity terms. The authors illustrated this issue with the examples "May" and "North", which could be part of named entities for dates and geographical locations. However, these terms often refer to a person's name. The second challenge concerned multi-word named entities, which necessitated determining the boundaries of an entity, specifically identifying its beginning and end.

Later in this chapter, we discuss a concept that tackles the ambiguity challenge in Section 2.6, titled "Text Representation". In Section 2.5, titled "Models for Named Entity Recognition", we explore models capable of identifying the beginning and end of multi-word named entities, thereby addressing the second challenge.

2.2 Tagging Schemes

To perform NER in a structured way, tagging schemes are used to label and categorize named entities in text. These schemes determine how each token is marked to indicate its role within a named entity. Bird et al. (2009) described tokens as sequences of characters, which can include words, subwords, or multiple words. NER operates as a token classification task, with each token assigned a specific label.

The BIO tagging scheme, developed by Ramshaw & Marcus (1995), is the standard method for token classification. It uses "B" for the beginning of a named entity, "I" for tokens inside the

entity, and "O" for tokens outside any entity. Other schemes include the simpler IO and the more complex BIOES, which adds tags for entity endings and single-token entities.

2.3 Evaluation Metrics

To evaluate the performance of a NER system, evaluation primarily relies on classification metrics. These metrics are best understood through the concepts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). True Positives represent, in the context of book titles, instances where the model accurately identifies a token as belonging to a book title entity class. False Positives are instances where the model erroneously labels a token as belonging to the book title entity class. Conversely, True Negatives refer to instances where the model correctly identifies a token as not belonging to the book title entity class. Lastly, False Negatives occur when the model fails to recognize a token as part of the book title entity class when it actually is.

2.3.1 Accuracy

The simplest evaluation metric is the accuracy score (Equation 2.1). Accuracy measures the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

When dealing with datasets where certain classes are significantly underrepresented compared to others (known as imbalanced datasets), relying on accuracy as an evaluation metric fails to accurately reflect the true performance of models (Grandini, Bagli, & Visani, 2020). For instance, in our dataset, only 1.4% of all tokens represent book titles (Section 3.5, titled "Data Analysis"). A model could achieve an accuracy of 98.6% by simply predicting "no book" for every token. Despite this seemingly high accuracy score, the model's actual performance in identifying book titles would be inadequate, as it would fail to detect any of the book titles present in the data.

2.3.2 Precision and Recall

In contrast, precision (Equation 2.2) and recall (Equation 2.3) are more informative metrics for evaluating model performance in this imbalanced context. Precision measures the proportion of correctly identified book titles out of all tokens predicted as book titles (Grus, 2019). This helps us understand the accuracy of the positive predictions made by the model. Recall, on the other hand, measures the proportion of actual book title tokens that were correctly identified by the model, indicating the model's ability to capture true positives (Grus, 2019).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

2.3.3 F1 Score

The F1 score, shown in Equation 2.4, is a more sophisticated metric that combines both precision and recall. It is the harmonic mean of these two measures, offering a balanced single score (Grus, 2019).

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

2.4 Approaches to Named Entity Recognition

To determine the most suitable NER approach for extracting book titles, this section explores the high-level conceptual approaches of NER. Jehangir et al. (2023) categorized NER approaches into three primary categories: rule-based, unsupervised learning, and supervised learning.

2.4.1 Rule-based

Jehangir et al. (2023) described the rule-based approach as one where a predefined set of rules is used by the system to identify and classify entities within a text. An example of this approach is the method currently used by the thesis supervisor to extract book titles from newspaper archives. Furthermore, Jehangir et al. (2023) identified three major drawbacks of rule-based approaches: poor generalizability due to their domain-specific nature, the need for advanced programming skills, and significant human effort to develop these rules. The domain-specific nature of these rules means that rules effective for one newspaper may not apply to another, necessitating custom sets of rules for different sources.

2.4.2 Unsupervised

Unsupervised machine learning is typically used for data without labels, employing methods such as association and clustering (Jehangir et al., 2023). Given that this thesis has access to a large, labeled dataset, unsupervised methods were not further explored.

2.4.3 Supervised

Supervised machine learning involves training models using labeled data, where each input has a corresponding expected output (Géron, 2017). Given the limitations of rule-based methods and the availability of a substantial labeled dataset, this thesis further focused solely on the supervised approach. A deeper exploration of supervised models for NER is presented in Section 2.5, titled "Models for Named Entity Recognition".

2.5 Models for Named Entity Recognition

To identify the most effective supervised NER model for extracting book titles, this section examines several widely-used supervised machine learning models, comparing traditional machine learning approaches with modern deep learning techniques.

In their survey on NER in historical documents, Ehrmann et al. (2023) distinguished between two types of supervised machine learning approaches: traditional machine learning and deep learning. They referred to the traditional machine learning approach as "pre-neural machine learning", which primarily utilizes algorithms that do not involve neural network architectures. Meanwhile, Jurafsky & Martin (2023) defined deep learning as the application of modern, multilayered neural networks to model complex patterns and relationships in data.

2.5.1 Traditional Machine Learning

Ehrmann et al. (2023) identified the Conditional Random Field (CRF) classifier as the most prevalent traditional machine learning model for NER. Meanwhile, in a separate survey focused

on NER, Jehangir et al. (2023) explored additional traditional machine learning models, including Decision Trees, Naive Bayes, Hidden Markov Models, Maximum Entropy models, and Support Vector Machines. Ehrmann et al. (2023) concluded that the performance of traditional machine learning approaches in NER is generally significantly lower than that of deep learning approaches. Supporting this, Wang et al. (2016) demonstrated superior performance of deep learning in their study, where a Recurrent Neural Network significantly outperformed a traditional CRF model. They also highlighted how traditional models rely heavily on manual feature engineering, which can be very time-consuming when creating a custom NER system.

2.5.2 Deep Learning

Collobert et al. (2008) pioneered the use of deep learning for NER with a Convolutional Neural Network (CNN). Géron (2017) explained that CNNs are utilized in NLP to process text by applying convolutional layers. These layers employ multiple filters that slide over the input text data sequentially. Each filter is designed to capture specific local patterns or features within the text, such as suffixes, prefixes, and combinations of words that are indicative of named entities (Keraghel et al. 2024). This enables the CNN to extract meaningful features from the text, and by aggregating these local features, CNNs can effectively extract named entities from large and complex texts.

Recurrent Neural Networks (RNNs) represent another deep learning model for NER. In contrast to CNNs, RNNs are specifically designed to handle sequential data (Keraghel et al., 2024). Sequential data refers to data where the order of elements matters, such as a sequence of words from a text. Jurafsky & Martin (2023) provided a comprehensive explanation of the RNN's functionality. RNNs process a sequence of words by iterating through the words one at a time while maintaining an "internal memory" (hidden state) that captures information about the sequence seen so far. They also note a key limitation of RNNs: the vanishing gradient problem. As the length of the data sequence increases in RNN, the issue of vanishing gradients often arises. This occurs when the gradients, which represent how much and in what direction the weights of the neural network should be adjusted during training, become excessively small. This extremely small gradient effectively prevents the weights from changing, thereby stalling the network's learning process. To address this issue, the Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997) was developed. The LSTM uses gating mechanisms to tackle the vanishing gradient problem. These gates decide which information is important to keep or discard in its "internal memory", avoiding the exponential shrinking that is characteristic of the vanishing gradient problem.

Yang and Xu (2020) highlighted that the BiLSTM-CRF model, an adapted LSTM network, was considered the state-of-the-art for NER as of 2020. The BiLSTM-CRF network modifies the traditional LSTM architecture by processing data both forwards and backwards, a method known as bidirectional (Jurafsky & Martin, 2023). The outputs from the BiLSTM are then fed into a Conditional Random Field (CRF), which serves as the final output layer.

Apart from the vanishing gradient problem, traditional RNNs also suffer from what is known as the bottleneck problem (Jurafsky & Martin, 2023). In RNNs, each item from a sequential input updates the hidden state, which is expected to encapsulate the entire sequence by the final item. However, for longer sequences, early inputs can be "forgotten" due to the limited capacity of the hidden state. The attention mechanism addresses this by allowing the model to access information from all hidden states, enabling it to focus on different parts of the sequence and retain important details.

Building on the attention mechanism, self-attention was developed to further enhance performance. Unlike the traditional attention mechanisms that still relied on sequential processing, self-attention allows each token to interact directly with every other token, enabling parallel

processing and better capturing long-range dependencies. This innovation led to the transformer architecture, introduced by Google in the paper "Attention is All You Need" (Vaswani et al., 2017).

While Yang and Xu (2020) identified the BiLSTM-CRF as state-of-the-art, Labusch et al. (2019) had already shown how BERT (Devlin et al., 2018), an early transformer model, outperformed the BiLSTM-CRF in NER tasks on historical OCR-scanned German text. Ehrmann et al. (2023) and Sun et al. (2021) also noted that transformer-based networks are now surpassing BiLSTM models, including in the biomedical domain, which, like this study, deals with specialized custom named entities.

2.5.3 Large Language Models

After analyzing the most popular models on Hugging Face, a collaborative platform for machine learning models and datasets, we found that Large Language Models (LLMs) are the preferred method for NER. Chockalingam et al. (n.d.) defined LLMs as "deep learning algorithms that can recognize, extract, summarize, predict, and generate text based on knowledge gained during training on very large datasets". They further explained how LLMs differ from all other language models that can perform NLP tasks. LLMs are considered large because of two reasons: they are trained on large amounts of data, and they comprise a huge number of trainable parameters.

Initially, LLMs were predominantly based on RNNs. However, the transformer architecture has become the preferred choice (Chockalingam et al., n.d.) due to its built-in attention mechanisms, which effectively capture long-range dependencies and allow for parallel computations. This results in faster training times, the ability to handle larger datasets, and the capability to train models with more parameters, leading to superior performance in NLP tasks (Amaratunga, 2023).

2.5.4 Transformer-based Large Language Models

There are numerous transformer-based LLMs, each differing from one another primarily in their architecture and training methods (Amaratunga, 2023). To understand these differences, we first need to explore the transformer's training process, as outlined by Jurafsky & Martin (2023). The training of transformer models occurs in two main phases: pretraining and fine-tuning.

During the pre-training phase, transformer models use self-supervised learning rather than the supervised training methods discussed in Section 2.5, titled "Models for Named Entity Recognition". In self-supervised learning, the model learns to predict parts of its input data without relying on explicitly human-labeled data. Two common pre-training tasks are Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, a percentage of the input tokens are masked, and the model is trained to predict these masked tokens. In NSP, the model is given pairs of sentences and must determine if the second sentence follows the first in the original document.

After pre-training, the transformer model is fine-tuned on human-labeled data for specific tasks, such as NER. The purpose of pre-training is to teach the model the meanings of words and their relationships within the language, enabling easier learning of the specific task during fine-tuning.

According to Ravichandiran (2021), one of the most popular transformer models that revolutionized the field of NLP is BERT, short for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018). BERTs training process relies on both MLM and NSP during pre-training. Its architecture is based on the original transformer model introduced by Vaswani et al. (2017).

Liu et al. (2019) identified that BERT was significantly undertrained and introduced an improved version called RoBERTa (Robustly optimized BERT approach). Major enhancements in RoBERTa include training on a much larger dataset and eliminating the NSP task from the pre-training process.

2.6 Text Representation

To understand text processing in NER models, it is essential to explore how text is transformed into a machine-readable format. Text representation bridges human language and machine learning by converting text into numerical formats that models can interpret and utilize.

Patil et al. (2023) illustrated how early techniques, such as One Hot Encoding, Bag of Words, Term Frequency, and Inverse Document Frequency, focused on word frequency. These methods struggled with high-dimensional vector representations in large vocabularies, leading to inefficiencies in training machine learning models. This limitation was addressed by advanced representations, called word embeddings.

Word embeddings are derived using neural networks trained in a self-supervised manner, such as predicting the next word. This method uses a continuous vector space to represent words as low-dimensional arrays of real numbers. These word embeddings capture both the semantic and syntactic aspects of words by considering their context. This allows them to identify relationships, such as synonyms, antonyms, and analogies (Patil et al., 2023). Notable examples of word embeddings include Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017).

The previously mentioned word embeddings are categorized as static word embeddings. Static word embeddings map each word to a single fixed vector representation, meaning they cannot account for the different meanings a word might have in various contexts. This limitation is addressed by dynamic word embeddings, which produce context-dependent representations, allowing the same word to have different vector representations depending on the context.

BERT (Devlin et al., 2019) uses WordPiece tokenization, developed by Google, which breaks words into subwords to handle rare or complex words and manage out-of-vocabulary words. These subwords are converted into numerical representations via a lookup table. BERT's pre-training involves creating its own embedding layer, transforming numerical input into dynamic word embeddings. The exact implementation of WordPiece tokenization remains undisclosed by Google (WordPiece Tokenization - Hugging Face NLP Course, n.d.).

Chapter 3

Data

This chapter outlines the data used in this study, detailing its collection, preparation, and transformation into a usable NER format. Additionally, we analyze the dataset's composition and evaluate the quality of the OCR.

3.1 Data Collection

The dataset provided was derived from the *Leeuwarder Courant* (henceforth: *LC*), a Dutch newspaper with a digital archive dating back to 1752. For this study, only the issues of the *LC* from 1962 to 1995 were provided. Due to its national prominence, the *LC* offers a valuable resource for studying book reviews. The newspapers were digitized using OCR, which tends to be less accurate for older issues (Section 3.5, titled "Data Analysis"). An example of erroneous OCR output from the *LC* is visualized in Figure 3.2, while an accurate OCR output from the *LC* is shown in Figure 3.3.

In the provided dataset (Van Eijnatten, 2024), the book titles were extracted by the thesis supervisor using a rule-based process involving three main steps:

1. A classical algorithm identified consistent tokens within each book review, extracting a 600-character segment (title pericope) likely containing the book title. This segment was compared to the *Nederlandse Bibliografie Totaal* (henceforth: *NBT*) database, which compiles all books, periodicals, and newspapers published by a Dutch publisher. Matches were identified based on the intersection of words between the title pericope and the *NBT* titles.
2. To address false negatives from faulty OCR, the title pericopes were processed using a Large Language Model, specifically ChatGPT 4. The results were again matched with the *NBT* titles, improving the identification rate.
3. Manual correction ensured the accuracy of the dataset, involving verifying the identified titles and checking for false positives and negatives. This thorough and time-consuming process resulted in a precise dataset of book titles extracted from the *LC*.

3.2 Locating Book Titles in Newspaper Text

A significant challenge in this study was converting the provided dataset into a format suitable for NER. The original dataset contained book reviews along with a list of book titles derived from the *NBT* database. However, these book titles rarely appeared directly in the review texts, presenting a major obstacle. Ideally, manual annotation of the data directly in an NER format

would have been the best solution, but due to limited time, this was not feasible. Consequently, we developed an implementation to locate the given *NBT* titles within the text of the book reviews. However, due to the complexity of the task, this implementation is imperfect and often only partially locates full book titles.

The first step involved preprocessing the text. Both the newspaper content and the book titles were cleaned to remove extra spaces and converted to lowercase for consistency. Following this, the algorithm attempted direct matching. Initially, it searched for the exact given title within the newspaper text. If the exact title was not found, the algorithm checked for partial matches. It split the title at various delimiters, such as colons, semicolons, equals signs, and commas to extract the main title, in the hope that this main title was present in the text.

For titles that did not match through direct methods, a fuzzy matching technique was employed. The algorithm split the title and the newspaper text into tokens and then formed segments from the text with the same number of tokens as the title. Each segment was compared to the title using a similarity score. The segment with the highest similarity score was selected as the best match.

To ensure the accuracy of this fuzzy matching approach, these matched titles were manually verified. The matched text sometimes missed one or two words from the title in the newspaper text, but these instances were retained to maintain data quantity. However, matches that missed too many words of the main title or were almost unrecognizable due to faulty OCR were marked and later excluded from the dataset. Some examples obtained from the fuzzy matching approach are shown in Table 3.1.

Table 3.1: *Examples of fuzzy matching results for book titles in newspaper content.*

Main book title	**Fuzzy matched text in newspaper content**
het weerlicht op de kimmen	het weerlicht op de kimmen;
kosmos vogelveldgids van europa	ders: “vogelveldgids van europa”.
beter blote jan dan dode jan, en andere uitspraken van louis paul boon	beter blote jan dode jan en andere “ltsp,eng_ran louis paul boon.”
knotsgekke uitvindingen van de 19e eeuw	knotsgekke uitvindingen van de 19de eeuw.

3.3 Formatting Data for Named Entity Recognition

After obtaining the locations of book titles in the newspaper text, the next step was to format the data for NER using a tagging scheme (Section 2.2, titled “Tagging Schemes”). Archana et al. (2023) examined methods for handling imbalanced data in NER and demonstrated the effectiveness of the IO format in addressing this issue. Given that this study also faces substantial data imbalance (Section 3.5, ”Data Analysis”), the IO tagging format was chosen to be utilized.

To implement this, the subsequent step involved tokenizing the newspaper text by splitting it into individual words, resulting in an array of words. Additionally, an array of labels was generated based on the location of book titles within the text. Each word in the array was assigned a label: ”I” (Inside) if the word was part of a book reference, or ”O” (Outside) if it was not. For example, in the sentence ”Recent heb ik het boek De paarden van oranje gelezen”, the words would be tokenized and labeled as follows: [”Recent”, ”heb”, ”ik”, ”het”, ”boek”, ”De”, ”paarden”, ”van”, ”oranje”, ”gelezen”] would have the corresponding labels [”O”, ”O”, ”O”, ”O”, ”O”, ”I”, ”I”, ”I”, ”I”, ”O”].

3.4 Labels from Other Newspapers

In addition to the book titles provided from the *LC* newspaper, I manually labeled book titles from other newspapers to evaluate the developed NER system's generalizability to other newspapers. This involved detecting articles containing book reviews from the newspapers *Het Parool* and *Trouw*. Due to time constraints, only the articles that were most clearly book reviews were selected, resulting in 115 book reviews from *Trouw* and 193 from *Het Parool*. An article was clearly a book review if it contained the characters "blz", "ISBN", and "f" (the symbol for the Dutch guilder). The labeling process was streamlined using a Label Studio environment, an open-source data labeling tool, to facilitate NER labeling.

A significant difference between the *Trouw* and *Het Parool* datasets compared to the *LC* dataset is that the *Trouw* and *Het Parool* datasets are specifically labeled in a NER format. Instead of matching the given titles to their most likely positions in the book reviews, this approach labels each occurrence of the title directly in the text. Consequently, in these *Trouw* and *Het Parool* datasets, the title always includes the full book title along with the subtitle. In contrast, the *LC* data often required discarding the subtitle to accurately identify the title's position. Additionally, if a title appears multiple times in *Het Parool* or *Trouw*, it is labeled each time it is mentioned, unlike in the *LC* dataset, where difficulties in determining the book title positions prevented this.

3.5 Data Analysis

In this section, we conduct a quantitative analysis of our datasets and attempt to quantify the OCR quality.

3.5.1 Leeuwarder Courant

As explained in Section 3.2, titled "Locating Book Titles in Newspaper Text", we manually selected 729 fuzzy matches of book titles as insufficient. If a book title within a review was marked as insufficient, the entire review was excluded. This step is crucial to avoid a dataset where some tokens that represent book titles in the review text were incorrectly annotated as non-book titles, which would have misled the model and potentially decreased its performance. Consequently, 594 book reviews were removed from the initial dataset.

From the *LC* data, we compiled a total of 12,535 book reviews, encompassing 23,529 book titles. This dataset contains a total of 7,643,958 tokens, with 110,018 of these tokens being book titles. As a result, book titles represent only 1.4% of the total tokens, indicating a significant class imbalance.

3.5.2 Het Parool & Trouw

In Section 3.4, titled "Labels from Other Newspapers", we discussed the creation of two secondary datasets, which includes 115 book reviews from *Trouw* and 193 from *Het Parool*. The *Het Parool* dataset consists of 119,597 tokens in total, with 2,931 of these tokens representing book titles, while the *Trouw* dataset consists of 101,640 tokens in total, with 1,333 of these tokens representing book titles. Similar to the *LC* dataset, these secondary datasets exhibit a significant class imbalance, with book titles accounting for 2.5% of the total tokens in the *Het Parool* dataset and 1.3% in the *Trouw* dataset.

3.5.3 Optical Character Recognition Quality

To further assess the quality of our dataset, we attempted to quantify the OCR accuracy. This was done by splitting each review into individual words and determining the percentage of words that are present in dictionaries. For Dutch words, we used the dictionary from opentaal.org, which contains a total of 164,105 words. Additionally, since some German books are reviewed, we included a German dictionary from openthesaurus.de, containing 164,913 words. Figure 3.1 shows the results of the percentage of words found in the dictionaries over time. This analysis provides an indication of the OCR accuracy, but it is important to note that the dictionaries do not include every Dutch and German word or their derivatives. These were the most complete open-source dictionaries available for download.

Figure 3.1 indicates that the OCR performance is generally better for the *Trouw* data compared to *LC*. This may be attributed to the method used to distinguish book review articles from other articles in the newspapers *Het Parool* and *Trouw*, which relies on the presence of the key terms "blz", "ISBN", and "f" in the text. Articles with OCR errors in these key terms were ignored, resulting in the exclusion of some poorly OCR'd articles.

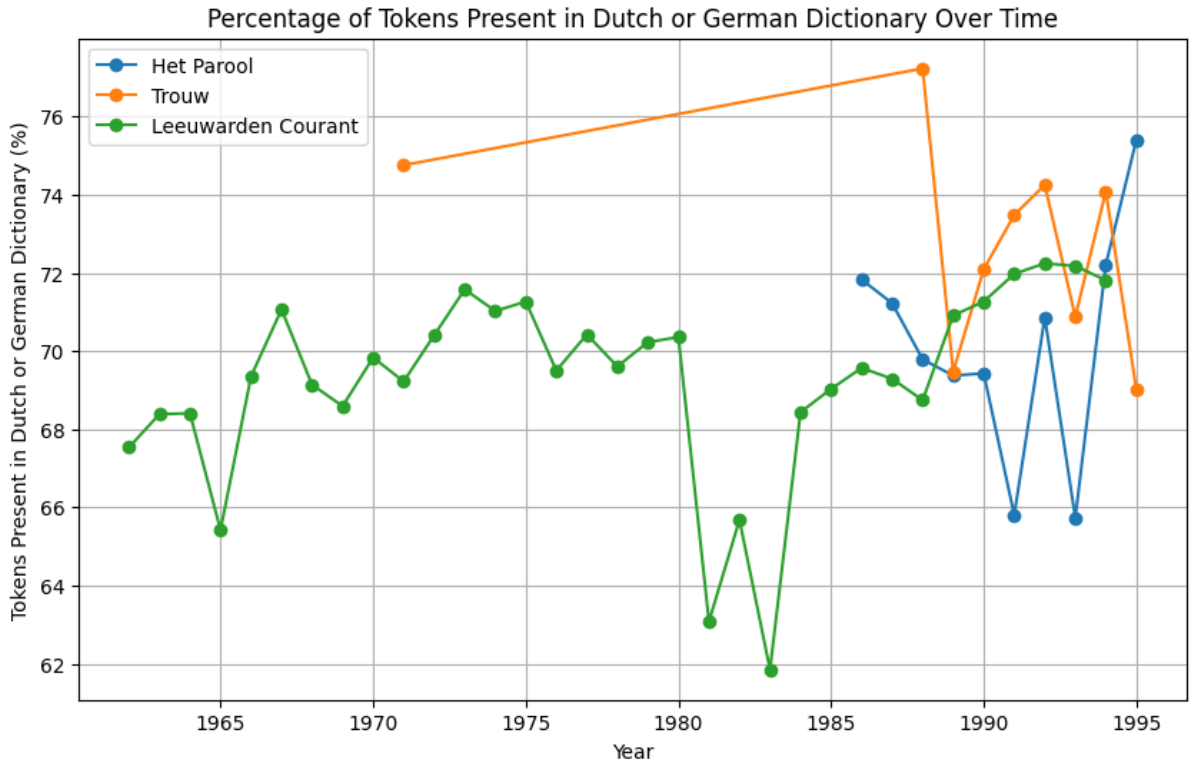


Figure 3.1: *Percentage of tokens from the book review articles present in Dutch or German dictionaries over time for the newspapers Het Parool, Trouw, and Leeuwarder Courant.*

INGEKOMEN BOEKEN

UITG. AD. DONKER, ROTTERDAM:

Joseph Roth. Biecht van een moordenaar. Uit het Duits vertaald. Tweede druk. Een kleine roman, waarin een man in één nacht van zijn leven vertelt en van zijn misdadige hartstochten, van de verzoekingen van de duivel, die hem tot spion en moordenaar maakten. Boeiend geschreven, indringend van psychologie. Ing., 144 blz., f 1,50. (Donker-pockets, nr. 20.)

INGEKOMEN BOEKEN
UITG. AD. DONKER, ROTTERDAM-
JOS6?_It h^ thV_^ eeit Tan 'en moordenaar.
elÏ k^ln^ ltS T«*aal°. Tweede d"k.
-ïn ,!__?_2 e roman' "aarin ««" man in
Sn SL™. ? Jn teve" vertelt eTvan
foëkiS^^ hartstochten, van de ver-
sokrn^n m^ .f* duivel- di« hem tot
g^^en^ri^^3^"- Boelend
Ing M^ 7 m? r 1 m«<m<k v? n Psychologie,
ing., I*4 wz„ / i oo (Donker-pockets. nr.

Figure 3.2: *Example of inaccurate OCR from the Leeuwarder Courant (30-06-1958) showing the original text (left) and the erroneous OCR output (right).*

Ingeboekt

★ HANS VERVOORT - *„Een zomer apart"*. De hoofdpersoon in deze paperback (166 blz., prijs *f* 26,50) doet een poging te ontsnappen aan de kantoorsleur en de knellende banden van het huwelijk: hij wil een roman schrijven, die alle romans overbodig maakt. De Arbeiderspers, Amsterdam.

Ingeboekt * HANS VERVOORT - „Een zomer apart". De hoofdpersoon in deze paperback (166 blz., prijs ƒ 26,50) doet een poging te ontsnappen aan de kantoorsleur en de knellende banden van het huwelijk: hij wil een roman schrijven, die alle romans overbodig maakt. De Arbeiderspers, Amsterdam.

Figure 3.3: *Example of accurate OCR from the Leeuwarder Courant (04-08-1982) showing the original text (left) and the accurate OCR output (right).*

Chapter 4

Methodology

This chapter outlines the experiments conducted to identify the best NER model for extracting book titles from book review articles, as well as the methodologies used to evaluate model performance. Initially, the models employed in these experiments are described in detail.

4.1 Models

In this section, we introduce the NER models utilized in this study and outline the rationale behind their selection, starting with the simplest model and progressing to the most complex. Additionally, we discuss the loss function used in the training process.

4.1.1 Baseline and SpaCy

To develop a NER system for extracting book titles from historical newspapers, we established a baseline model as a performance benchmark. This ensures that improvements from more complex models are meaningful.

For our baseline, we utilized SpaCy, an open-source NLP library, specifically its "nl_core_news_lg" model. This pre-trained model is trained on Dutch text, including news articles, and is capable of identifying a range of entities within Dutch texts. For our purposes, we focused on the "work_of_art" entity, which includes titles of books, songs, and other artistic works. While the baseline model may not achieve the highest performance compared to more specialized models, its ease of use and accessibility make it a valuable initial tool.

Additionally, we leveraged SpaCy's capability to fine-tune the pre-trained model on our custom training data. This adaptability could enhance the recognition of our specific entity. However, SpaCy's streamlined training process has limitations, restricting the ability to customize various training aspects, which may limit performance improvements. SpaCy's NER model architecture is based on a Convolutional Neural Network (CNN), as detailed in its official documentation.

4.1.2 BiLSTM-CRF

Despite the recognition of transformer-based models as the state-of-the-art in NER, their performance in contexts with substantial class imbalance remains underexplored. Therefore, this study also incorporated the previous state-of-the-art, the BiLSTM-CRF model, for a comparative analysis.

Training the BiLSTM-CRF model requires converting text into numeric representations, as detailed in Section 2.6, titled "Text Representation". This study exclusively used pre-trained

Dutch word embeddings for this transformation. Various embeddings were evaluated based on the proportion of unknown tokens in the training data. GloVe embeddings (Pennington et al., 2014) classified 24% of tokens as unknown, while FastText embeddings (Bojanowski et al., 2017) marked only 2% as unknown. Consequently, FastText was chosen for this study.

4.1.3 Transformer-based Large Language Models

In addition to the BiLSTM-CRF, this research employed the current state-of-the-art: transformer-based models. Liu et al. (2021) demonstrated that domain-specific fine-tuning of transformer models enhances performance if the model has been previously fine-tuned on another NER task. Hugging Face hosts numerous fine-tuned NER transformer models. For this study, several popular Dutch models from Hugging Face, already fine-tuned on NER tasks, were further fine-tuned on our dataset using the default hyperparameters as specified in the Hugging Face token classification tutorial (Token Classification, n.d.).

- ***WikiNEuRal* (Tedeschi et al., 2021)**: A multilingual BERT model fine-tuned on NER tasks across nine languages, including Dutch and German.
- ***xlm-roberta-large-finetuned-conll03-english***: An XLM-RoBERTa model (Conneau et al., 2019) pre-trained on 2.5 terabytes of data in 100 languages, including Dutch, and fine-tuned on the CoNLL-2003 NER dataset (Sang et al., 2003).
- ***robbert-v2-dutch-ner* (Delobelle et al., 2020)**: A RoBERTa-based model fine-tuned specifically for Dutch NER tasks.
- ***BERTje* (De Vries et al., 2019)**: A BERT-based model pre-trained on Dutch data, including contemporary and historical fiction, and the Multifaceted Dutch News Corpus (Ordelman et al., 2007). After pre-training, it was fine-tuned on the CoNLL-2002 dataset (Sang, 2002).

4.1.4 Loss Function

The loss function is essential for training neural networks, measuring how well the model's predictions align with actual target values. During training, the neural network adjusts its weights to minimize the loss function. Initially, the transformer-based models and the BiLSTM-CRF model struggled to learn effectively due to significant class imbalance in the training data (Section 3.5, titled "Data Analysis"). The default Cross-Entropy loss function led to models predicting that all tokens were "no book."

Nemoto et al. (2024) addressed this issue by developing a loss function capable of handling class imbalance. They introduced the *Majority or Minority (MoM)* loss function, which outperformed other commonly used loss functions for NER tasks with class imbalance in their comparative analysis. Therefore, the *MoM* loss function was used for training all transformer-based models and the BiLSTM-CRF model. However, the "user-friendly" SpaCy model did not easily accommodate a custom loss function.

4.2 Experimental Setup

This section outlines the experiments conducted to tune the BiLSTM-CRF model, select the best model from various transformer models, and finally compare these two models with the baseline and the trained Spacy model with the goal to determine the overall superior NER model. Detailed descriptions of the models used in each experiment can be found in Section 4.1, titled "Models".

The datasets used for these experiments are further detailed in Chapter 3, titled "Data". The *LC* dataset was divided into training (70%), validation (15%), and test (15%) sets. This division was performed at the review level, ensuring that a review cannot be present in more than one split, thereby preventing any overlap between the training, validation, and test sets. This approach allows for a realistic evaluation of the model's performance. The training set was used to train the NER models, the validation set was utilized to determine the hyperparameters and for model selection, and the test set provided an unbiased evaluation of the final model's performance. The *Het Parool* and *Trouw* datasets were retained as single units to function as secondary test sets, intended to evaluate the generalizability of the NER models across different newspapers.

For each experiment involving a training procedure, the maximum batch size was employed to optimize computational efficiency. Training was conducted on a GPU to ensure that the models were trained within a feasible timeframe. Details of the computational resources used in these experiments are provided in Appendix A.

4.2.1 Limitations

In this study, we did not compare the developed models to the existing rule-based approach detailed in Section 3.1, "Data Collection". The primary reason was that the rule-based approach involved numerous manual actions within its algorithm, making it unfair to compare our autonomous models to this manually intensive method. Moreover, the performance measurements of the rule-based approach were recorded in a limited manner, preventing an accurate determination of an F1 score. Therefore, this study did not focus on determining whether the autonomous algorithm outperformed the manual intensive approach, but rather on evaluating the effectiveness of an autonomous approach in performing this task.

A second limitation of this study was that it relied solely on the provided OCR-scanned text and did not involve performing OCR on images independently.

4.2.2 Tuning the BiLSTM-CRF

The first experiment involved a hyperparameter search to determine the optimal number of memory units in the BiLSTM-CRF model. Three variations of the model, each with 50, 100, or 200 memory units, were trained for 20 epochs with a batch size of 64. The F1 score on the validation set was recorded at each epoch, enabling the determination of the optimal number of training epochs through a callback mechanism. The final BiLSTM-CRF model selected was the one with the best overall performance on the validation set.

4.2.3 Comparing the Transformer-based Large Language Models

The second experiment evaluated multiple transformer-based LLMs to identify the best performer. Each LLM was trained for 20 epochs with batch sizes as shown in Table 4.1. As with the BiLSTM-CRF model, the F1 score on the validation set was recorded at each epoch to determine the optimal number of epochs through a callback mechanism. The LLM with the highest overall F1 score on the validation set was chosen as the final transformer-based LLM for this study.

4.2.4 SpaCy

No experimentation was needed to optimize the SpaCy approaches. The pre-trained SpaCy model (baseline) required no hyperparameter tuning. For the fine-tuned SpaCy model on our training data, default settings were used.

Table 4.1: *Batch sizes used during training for each transformer-based large language model.*

Model	Batch size
WikiNEuRal	16
xlm-roberta-large-finetuned-conll03-english	2
robbert-v2-dutch-ner	16
BERTje	16

4.2.5 Final Comparison

The final experiment was conducted after selecting the best transformer-based LLM and the best BiLSTM-CRF model, both determined using the validation set from the *LC* data. In this experiment, all final NER models were compared: the SpaCy model (baseline), the fine-tuned SpaCy model, the best transformer-based LLM, and the optimized BiLSTM-CRF model. These comparisons were made using the *LC* validation set, *LC* test set, *Het Parool* dataset, and *Trouw* dataset. The evaluation methods for model performance are detailed in Section 4.3, titled "Performance Evaluation".

4.3 Performance Evaluation

This section outlines the methods used to evaluate the model's performance, assessed through two primary approaches: token classification performance and the accuracy of book identification. For the experiments aimed at identifying the best performing NER model, only token classification performance was considered. After selecting the best NER model, the performance of the identified books was evaluated.

4.3.1 Token Classification Performance

The token classification evaluation measured how accurately tokens were identified as either "I" (indicating a book title) or "O" (indicating no book title). This assessment utilized the metrics recall, precision, and F1 score, which are detailed in Section 2.3, titled "Evaluation Metrics".

4.3.2 Book Identification Performance

Beyond token-level performance, this study primarily focuses on the accurate identification of books mentioned within book review articles. This evaluation involved matching the extracted tokens to the most similar book title in the *NBT* database. Using the *LC* dataset, which contains the exact titles from the *NBT* for each book review, recall, precision, and F1 scores were calculated by comparing the set of identified book titles with the set of actual book titles.

Matching to the *NBT* has several advantages, as it reveals a wealth of additional information about the actual book, such as the ISBN number and genre, thereby enriching the dataset and providing more context for further analysis

In addition to the *LC* data, Philippa Linardatos, a fellow student working on a similar study, utilized the NER datasets from the newspapers *Het Parool* and *Trouw* to manually identify the reviewed book titles from the *NBT*. This effort extends the datasets beyond the NER format to match the original format of the *LC* dataset, allowing for performance evaluation of book identification in other newspapers. A limitation of this approach is that the newspapers *Het Parool* and *Trouw* review several non-Dutch books not included in the *NBT*, which were ignored as this study is restricted to the *NBT*.

The matching of extracted book title tokens to titles in the *NBT* was performed by calculating the Levenshtein distance between the NER output and the titles in the *NBT*. The *NBT* title with the highest similarity score was selected. Given the computational expense of this process, considering the *NBT* from 1930 until 1995 contains 1,954,801 books, we restricted our analysis to a subset of books published within three years before the book review's publication. This approach is justified as 98% of the books in the training data were published within this timeframe.

Chapter 5

Results

In this chapter, the results of the experiments are presented. The training history of all models that required a training procedure is depicted in Appendix B, showing the F1 score on the validation data and the loss on the training data over epochs.

5.1 BiLSTM-CRF

The results from the hyperparameter tuning to determine the number of memory units for the BiLSTM-CRF are shown in Table 5.1. The table highlights three different configurations of memory units: 50, 100, and 200. As the number of memory units increases, the total number of parameters (model size) also increases from 141,010 to 804,010.

The F1 score on the *LC* validation data shows a consistent improvement with the increase in memory units, going from 66.9% for 50 units to 68.8% for 200 units. However, the training time also varies with the number of memory units, with 100 units yielding the shortest training time of 278 minutes, whereas 50 and 200 units require 310 and 318 minutes, respectively. This varying training time is influenced by other computations and computer programs running simultaneously during the training of the models, which could explain why the smallest model is not the fastest.

Table 5.1: *Results from hyperparameter tuning to determine the number of memory units for the BiLSTM-CRF model, including total number of parameters, F1 score on Leeuwarder Courant validation dataset, and training time.*

Number of Memory Units	**Total Number of Parameters (model size)**	**Leeuwarder Courant Validation F1 score (%)**	**Training Time (minutes)**
50	141,010	66.9	310
100	322,010	68.0	278
200	804,010	68.8	318

5.2 Transformer-based Large Language Models

The results from training several transformer-based LLMs are presented in Table 5.2. Among the models tested, the *xlm-roberta-large-finetuned-conll03-english* model achieved the highest F1 score on the *LC* validation dataset, with a score of 83.9%. However, this model also required the longest training time at 624 minutes, which is substantially longer compared to the other models.

This extended training duration is attributable to the model's larger number of parameters that needed to be adjusted during training. Additionally, hardware limitations, particularly the GPU's limited VRAM (Appendix A), restricted the maximum batch size for training this larger model to 2 (Table 4.1), whereas the other models were trained with a batch size of 16. This smaller batch size further contributed to the increased training time.

Table 5.2: *Results from training several transformer-based large language models, including total number of parameters, F1 score on Leeuwarder Courant validation dataset, and training time.*

Hugging Face Transformer model	**Total Number of Parameters (model size)**	**Leeuwarder Courant Validation F1 score (%)**	**Training Time (minutes)**
WikiNEuRal	177,264,386	80.3	108
xlm-roberta-largefinetuned-conll03-english	558,842,882	83.9	624
robbert-v2-dutch-ner	116,173,058	78.7	110
BERTje	108,548,354	78.5	111

5.3 Final Comparison

This section compares the final models from previous experiments based on their token classification performance. Additionally, for the best model, an analysis was conducted to determine how many books were correctly identified by matching predictions to the *NBT*.

5.3.1 Token Classification

Table 5.3 presents the final token classification results on the *LC* validation set and *LC* test set from four models: the baseline model, the trained SpaCy model, the best BiLSTM-CRF model, and the best transformer-based LLM. Table 5.4 shows the results from these same models on the *Trouw* and *Het Parool* datasets.

In the *LC* validation set, the baseline model performed poorly, whereas the other models demonstrated much better performance, indicating that training on our own dataset enhances results. The BiLSTM-CRF model achieved an F1 score of 68.8%, precision of 73.8%, and recall of 64.4%, outperforming the trained SpaCy model. The transformer-based model, *xlm-roberta-large-finetuned-conll03-english*, exhibited the highest performance with an F1 score of 83.9%, precision of 82.9%, and recall of 85.0%, demonstrating its superior effectiveness on this dataset.

In the *LC* test set, performance trends were similar to the validation set, with all models outperforming the baseline. The *xlm-roberta-large-finetuned-conll03-english* model again showed the highest performance.

In both the *Het Parool* and the *Trouw* datasets, the baseline model performed better compared to the other *LC* sets, but still lagged behind the more complex models. The transformer-based model continued to lead, achieving an F1 score of 54.2%, precision of 80.6%, and recall of 40.8% on the *Het Parool* dataset, and an F1 score of 60.4%, precision of 74.8% and recall of 50.6% on the *Trouw* dataset. All models showed notably high precision relative to recall. This suggests that when a model identifies a token as a book title, it is usually correct (high precision), but several book title tokens are missed (lower recall). This pattern indicates that the models are conservative in their predictions, opting to classify a token as a book title only when they

are confident. This conservative approach reduces false positives, but results in several true positives being missed.

Table 5.3: *Final token classification results from the best model in each category, including F1 score, precision, and recall on the Leeuwarder Courant validation set and Leeuwarder Courant test set.*

Named Entity Recognition model	Leeuwarder Courant validation			Leeuwarder Courant test		
	F1 (%)	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**
Baseline (SpaCy)	9.6	12.8	7.7	9.1	12.5	7.2
Trained Spacy	64.9	74.7	57.3	63.9	74.3	56.0
BiLSTM-CRF (200 memory units)	68.8	73.8	64.4	69.0	74.8	64.0
xlm-roberta-large-finetuned-conll03-english	83.9	82.9	85.0	84.3	83.4	85.2

Table 5.4: *Final token classification results from the best model in each category, including F1 score, precision, and recall on the Trouw dataset and the Het Parool dataset.*

Named Entity Recognition model	Het Parool			Trouw		
	F1 (%)	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**
Baseline (SpaCy)	25.7	36.6	19.9	10.0	17.1	7.1
Trained Spacy	27.6	70.0	17.2	40.7	74.4	28.1
BiLSTM-CRF (200 memory units)	34.7	79.6	22.2	34.3	71.8	22.5
xlm-roberta-large-finetuned-conll03-english	54.2	80.6	40.8	60.4	74.8	50.6

5.3.2 Book Identification

The transformer-based LLM, *xlm-roberta-large-finetuned-conll03-english*, demonstrated superior NER performance on the *LC* validation, *LC* test, *Het Parool*, and *Trouw* datasets. Consequently, this model was selected for further assessment. In addition to its token classification performance, the model's accuracy in identifying books mentioned in book review articles was evaluated. Table 5.5 presents the metric scores, and Table 5.6 displays the true positives, false positives, and false negatives in absolute numbers.

The model demonstrated similar performance in book identification across all three newspaper datasets. Notably, the *Trouw* dataset achieved a higher precision score of 68.8%, compared to 65.9% for the *LC* test set. These findings indicate that the model generalizes reasonably well to newspapers it was not trained on.

Table 5.5: *Performance of the xlm-roberta-large-finetuned-conll03-english model in identifying books mentioned in book review articles. The model's F1 score, recall, and precision on the Leeuwarder Courant test, Het Parool, and Trouw datasets.*

Dataset	**F1 (%)**	**Recall (%)**	**Precision (%)**
Leeuwarder Courant test	59.4	54.1	65.9
Trouw	58.8	51.4	68.8
Het Parool	54.2	48.2	61.9

Table 5.6: *Performance of the xlm-roberta-large-finetuned-conll03-english model in identifying books mentioned in book review articles. The model's true positives, false positives, and false negatives on the Leeuwarder Courant test, Het Parool, and Trouw datasets.*

Dataset	**True Positives**	**False Positives**	**False Negatives**
Leeuwarder Courant test	1880	974	1598
Trouw	55	25	52
Het Parool	26	16	28

Chapter 6

Discussion

This chapter examines the potential impact of faulty Optical Character Recognition (OCR) on extracting book titles from historical texts. We analyze the predictions of a transformer-based model, noting its strength in NER, but its difficulty in accurately matching titles to the *NBT*. Finally, we compare our results with existing literature to contextualize our findings.

6.1 Impact of Faulty Optical Character Recognition

A qualitative analysis of the provided book reviews reveals the inherent imperfections of OCR. Although a quantitative method to assess OCR accuracy is outlined in Section 3.5, titled "Data Analysis", pinpointing an exact accuracy rate remains challenging. The quantitative analysis offers only an approximation, indicating that while OCR accuracy is improving over time, it is still not flawless.

A relevant study conducted by Hamdi et al. (2019) investigated the impact of faulty OCR on NER tasks using an LSTM-CRF model. They introduced OCR noise to determine its effect on performance. Their findings demonstrated that increasing the Word Error Rate (WER) from 1% to 7% and the Character Error Rate (CER) from 8% to 20% in OCR outputs resulted in a substantial drop in the F1 score from 90% to 60%. WER quantifies the percentage of incorrectly recognized words, while CER measures the percentage of incorrectly recognized characters. Their study conclusively showed how faulty OCR can drastically affect NER performance.

Unfortunately, since we cannot determine the exact OCR accuracy, we cannot precisely measure the impact of faulty OCR on extracting book titles from historical newspapers. However, it is likely that our results were impacted by the OCR.

6.2 Manual Analysis of Named Entity Recognition Predictions

This section further analyzes the best performing transformer-based LLM, emphasizing its strengths and weaknesses based on a manual review of its predictions on the test datasets to explain the observed F1, recall, and precision metrics. Several predictions used for this analysis are visualized in Appendix C.

To begin with, it stands out how the model accurately classifies tokens as book titles, resulting in very few false positives, which aligns with its high precision score.

Additionally, the model consistently identifies book titles without duplicating them when they appear multiple times in the text. It marks only one instance of each title, reflecting the structure of the training data in which each book title was annotated just once (Section 3.2,

"Locating Book Titles in Newspaper Text"). Consequently, the annotated positions do not always match where the model marks that same title in the text. This discrepancy contributes to the model's lower recall score. For the purpose of this research, it ultimately has no impact if a book is mentioned multiple times in the text, as we only need to identify it once. However, this characteristic does impact the measured NER performance.

Moreover, the tokens predicted as book titles often exclude the full title, frequently omitting subtitles. This is again logical, as subtitles had to be frequently omitted to accurately locate the book title in the text (Section 3.2, titled "Locating Book Titles in Newspaper Text"). This observation contributes to the lower recall score for the test examples that did include the subtitle.

Finally, there is a substantial difference in the model's performance when applied to the *LC* test dataset compared to the *Het Parool* and *Trouw* datasets. This discrepancy arises because the *LC* data was initially not annotated directly in a NER format, whereas the Het Parool and Trouw datasets were. In the *Het Parool* and *Trouw* datasets, each occurrence of a book title was annotated, leading to multiple annotations for the same title. In contrast, the *LC* data only provides a single annotation per title. As a result, the model was trained to predict only a single instance of each title, which causes a much lower recall score for the *Het Parool* and *Trouw* datasets compared to the *LC* dataset. Despite this, the precision score remains high for both datasets, likely because the single instance predicted by the model tends to be accurate.

To address these shortcomings, a potential solution is to annotate the training data directly in a NER format, rather than inferring the most likely positions of book titles from the *NBT*. This approach ensures the highest quality of labels by correctly annotating the entire title, including subtitles, and consistently marking each occurrence of a title within the text.

6.3 Challenges in Book Identification through NBT Matching

Despite achieving a high performance at the NER level, the results of matching to the *NBT* to evaluate the model's ability to identify books mentioned in texts were not as satisfactory as anticipated. An analysis was conducted to understand the cause of the suboptimal performance, using the actual labels to assess whether perfect NER predictions would improve matching to the *NBT*. However, this analysis revealed that even with perfect NER predictions, the match to the *NBT* only yielded an F1 score of 65.5%. This is because the main title of a book alone is insufficient for accurate matching to the *NBT*. Including additional information, such as the subtitle, author, and possibly the publisher would likely enhance the matching quality. Here again, arises the issue that the training data was not initially labeled in an NER format.

6.4 Transformer-based NER Models Confirmed as State-of-the-art

The results of this study support the findings of Labusch et al. (2019), Ehrmann et al. (2023), and Sun et al. (2021), which assert that transformer-based NER models are state-of-the-art. Despite a notable gap in the literature regarding the performance of these models in contexts with substantial class imbalances, this study demonstrates that transformer-based NER models maintain their state-of-the-art status even under such challenging conditions. This study further confirms why transformers are preferred over RNNs in the context of LLMs, beyond their superior performance. While the BiLSTM-CRF model, with 141,010 parameters, required 310 minutes to train, the transformer model, with 177,264,386 parameters, completed training in just 108 minutes. Consequently, transformers enable the development of even larger models and facilitate training on more extensive datasets in less time, making the LLMs even larger.

Chapter 7

Conclusion & Future Work

This thesis aimed to answer the question: *"To what extent can Named Entity Recognition be utilized to autonomously extract book titles from OCR-scanned historical newspapers, thereby facilitating deeper cultural and literary analyses?"*. The findings reveal that a transformer-based large language model can accurately and autonomously extract text representing book titles from book reviews within historical newspapers. The results demonstrate that transformer-based models outperform both the BiLSTM-CRF model and SpaCy models, showcasing their superior performance in Named Entity Recognition (NER) tasks for this specific application. Specifically, the transformer model achieved an F1 score of 84.3% on the test dataset.

In addition to assessing performance on a NER level, an evaluation was conducted to measure how well the best NER model could identify the actual discussed books. This was achieved by matching the extracted book title text to the titles in the *Nederlandse Bibliografie Totaal* (*NBT*), a comprehensive compilation of all books published by Dutch publishers. However, although accurate NER performance, these results showed suboptimal performance with an F1 score of 59.4%. This discrepancy was primarily due to the training data not being explicitly labeled for NER applications, and efforts to repurpose it as an NER dataset proved insufficient. Consequently, the model frequently missed subtitles, leading to incomplete title extraction. Further analysis indicated that even with perfect NER predictions, the match to the *NBT* yielded an F1 score of only 65.5%. This low score underscores the necessity for additional information beyond the main title, such as subtitles, authors, and possibly publishers, to achieve accurate title matching to the *NBT*.

For future work, it is strongly recommended to annotate a dataset directly for NER to enhance both NER performance and its matching with the *NBT*. Ideally, this dataset should be formatted for Nested Named Entity Recognition (nested NER), a method that identifies hierarchical entities within the text. This approach enables the extraction of sub-entities within larger entities. In the context of book titles, nested NER would distinguish the main title, subtitle, author, and publisher as separate entities within a single bibliographic entry. Moreover, the labeling process for NER can be simplified by using the tool Label Studio, which offers an intuitive interface for annotating data efficiently.

Additionally, it would be valuable to explore various methods for post-processing faulty OCR to potentially improve overall performance. By refining the OCR text before applying NER, the accuracy of extracted book titles could be significantly increased, thereby enhancing the quality of cultural and literary analyses.

References

Amaratunga, T. (2023). Understanding large language models: Learning Their Underlying Concepts and Technologies. Apress.

Archana, S. M., Prakash, J., Singh, P. K., & Ahmed, W. (2023). An Effective Biomedical Named Entity Recognition by Handling Imbalanced Data Sets Using Deep Learning and Rule-Based Methods. SN Computer Science/SN Computer Science, 4(5). https://doi.org/10.1007/s42979-023-02068-6

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Chockalingam, A., Patel, A., Verma, S., & Yeung, T. (n.d.). A Beginner's Guide to Large Language Models. NVIDIA Corporation.

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582.

Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.06286.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Do, Y., Kim, S. H., & Na, I. S. (2012). Title Extraction from Book Cover Images Using Histogram of Oriented Gradients and Color Information. International Journal Of Contents/Journal Of Contents, 8(4), 95–102. https://doi.org/10.5392/ijoc.2012.8.4.095

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents: A Survey. ACM Computing Surveys, 56(2), 1–47. https://doi.org/10.1145/3604931

Eisenstein, J. (2019). Introduction to Natural Language Processing. MIT Press.

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. http://cds.cern.ch/record/2699693

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Grus, J. (2019). Data Science from Scratch: First Principles with Python. O'Reilly Media.

Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., & Doucet, A. (2019, June). An analysis of the performance of named entity recognition over OCRed documents. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 333-334). IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on Named Entity Recognition — datasets, tools, and methodologies. Natural Language Processing Journal, 3, 100017. https://doi.org/10.1016/j.nlp.2023.100017

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. arXiv preprint arXiv:2401.10825.

Labusch, K., Kulturbesitz, P., Neudecker, C., & Zellhöfer, D. (2019, October). BERT for named entity recognition in contemporary and historical German. In Proceedings of the 15th conference on natural language processing, Erlangen, Germany (pp. 8-11).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Z., Jiang, F., Hu, Y., Shi, C., & Fung, P. (2021). NER-BERT: a pre-trained model for low-resource entity tagging. arXiv preprint arXiv:2112.00405.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nemoto, S., Kitada, S., & Iyatomi, H. (2024). Majority or Minority: Data Imbalance Learning Method for Named Entity Recognition. arXiv preprint arXiv:2401.11431.

Ordelman, R. J., de Jong, F. M., van Hessen, A. J., & Hondorp, G. H. W. (2007). TwNC: a multifaceted Dutch news corpus. ELRA Newsletter, 12(3-4).

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A Survey of Text Representation and Embedding Techniques in NLP. IEEE Access, 11, 36120–36146. https://doi.org/10.1109/access.2023.3266377

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Ramshaw, L. A., & Marcus, M. P. (1995). Text Chunking using Transformation-Based Learning. arXiv.org. https://arxiv.org/abs/cmp-lg/9505040v1

Ravichandiran, S. (2021). Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd.

Sang, E. F. T. K. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. arXiv.org. https://arxiv.org/abs/cs/0209010

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

Sarimehmetoğlu, B., & Erdem, H. (2023). Extracting Book Titles From Book Recommendation Videos Using a Deep Learning Approach. MANAS Journal Of Engineering, 11(2), 229–234. https://doi.org/10.51354/mjen.1369636

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica. D, Nonlinear Phenomena, 404, 132306. https://doi.org/10.1016/j.ph

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2021). Biomedical named entity recognition using BERT in the machine reading comprehension framework. Journal Of Biomedical Informatics, 118, 103799. https://doi.org/10.1016/j.jbi.2021.103799

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., & Navigli, R. (2021, November). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2521-2533).

Token classification. (n.d.). https://huggingface.co/docs/transformers/tasks/token_classification

Toral, A., & Munoz, R. (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources.

Van Eijnatten, J. (2024). Dutch intellectual culture between 1962 and 1995, or, using classical algorithms and LLMs to efficiently extract data with imperfect OCR [Journal-article]. https://2024.dhbenelux.org/wp-content/uploads/2024/05/DHB24_paper_van_Eijnatten_Dutch-intellectual-culture-between-1962-and-1995.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. arXiv (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Wang, W., Bao, F., & Gao, G. (2016, November). Mongolian named entity recognition with bidirectional recurrent neural networks. In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 495-500). IEEE.

WordPiece tokenization - Hugging Face NLP Course. (n.d.). https://huggingface.co/learn/nlp-course/chapter6/6

Yang, G., & Xu, H. (2020). A Residual BiLSTM Model for Named Entity Recognition. IEEE Access, 8, 227710–227718. https://doi.org/10.1109/access.2020.3046253

Appendix A

Hardware specifications

Details of the hardware used for the experiments.

Processor: AMD Ryzen 7 3700X 8-Core
RAM: 32 GB
GPU: NVIDIA GeForce RTX 3060 12GB

Appendix B

Training history

Plots of the training history of the models.

B.1 SpaCy

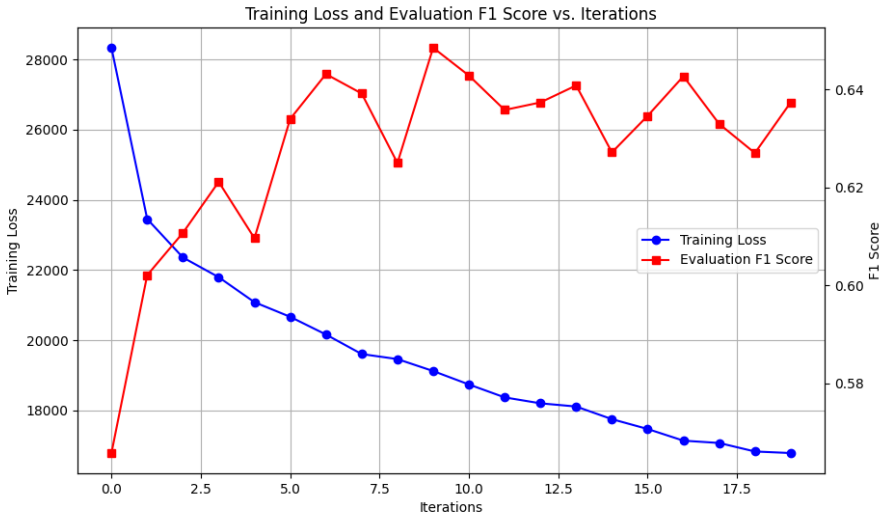


Figure B.1: *Training history of SpaCy model.*

B.2 BiLSTM-CRF

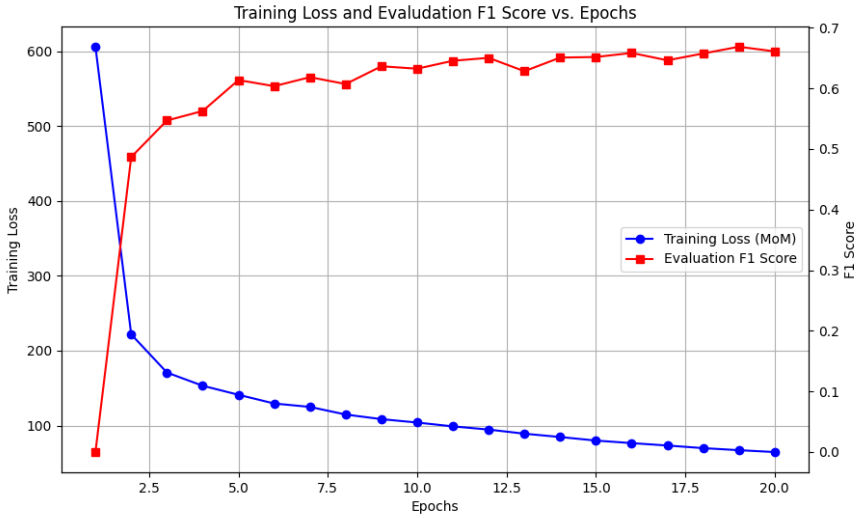


Figure B.2: *Training history of BiLSTM-CRF model (50 memory units).*

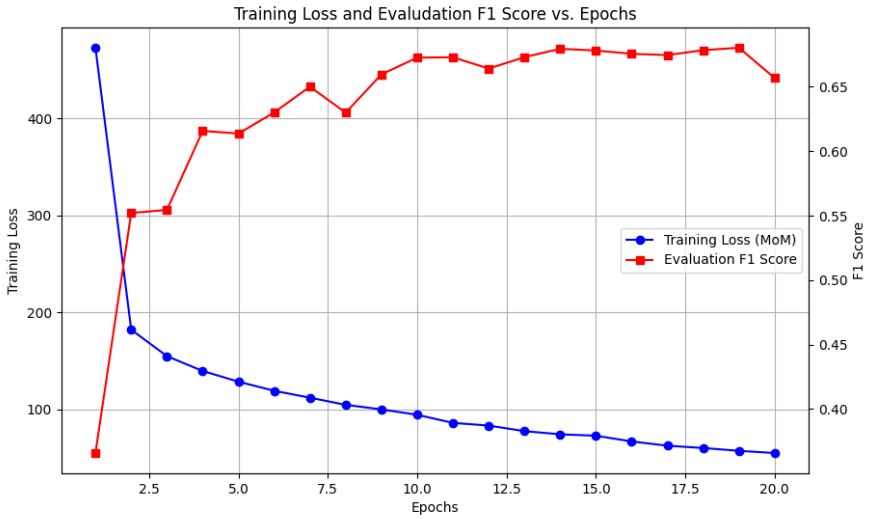


Figure B.3: *Training history of BiLSTM-CRF model (100 memory units).*

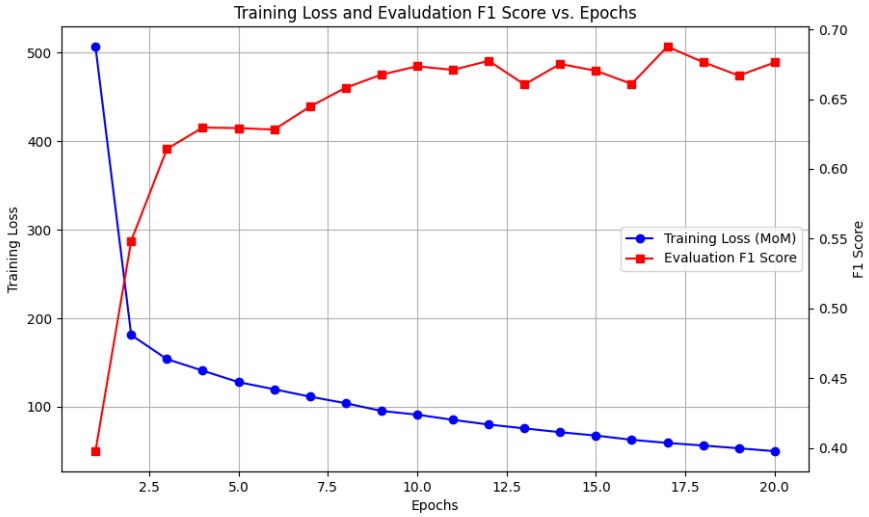


Figure B.4: *Training history of BiLSTM-CRF model (200 memory units).*

B.3 Transformer-based Large Language Models

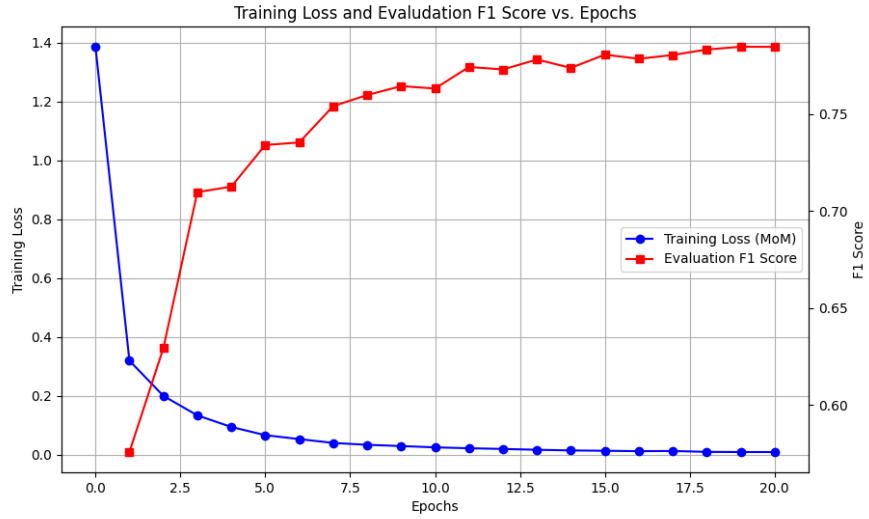


Figure B.5: *Training history of BERTje.*

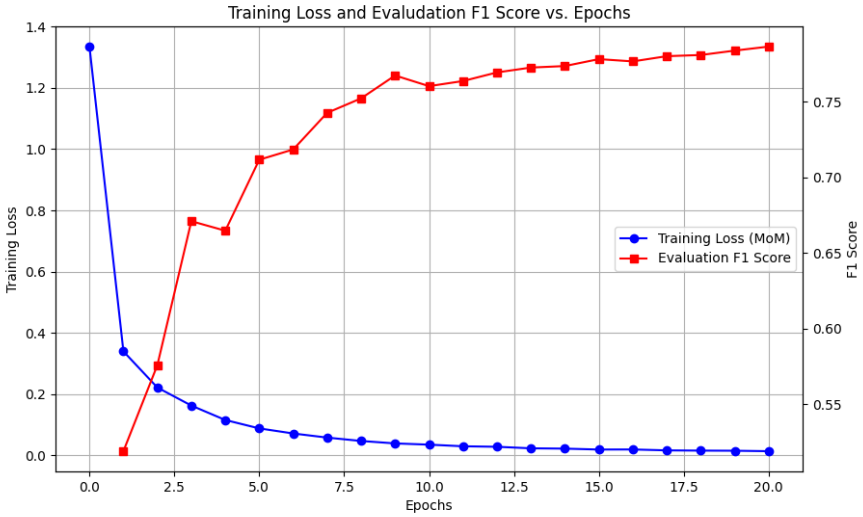


Figure B.6: *Training history of robbert-v2-dutch-ner.*

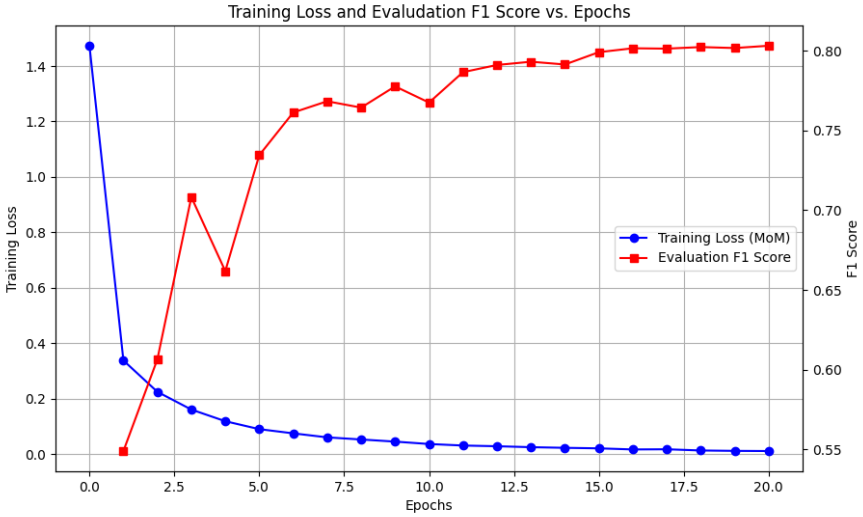


Figure B.7: *Training history of wikineural-multilingual-ner.*

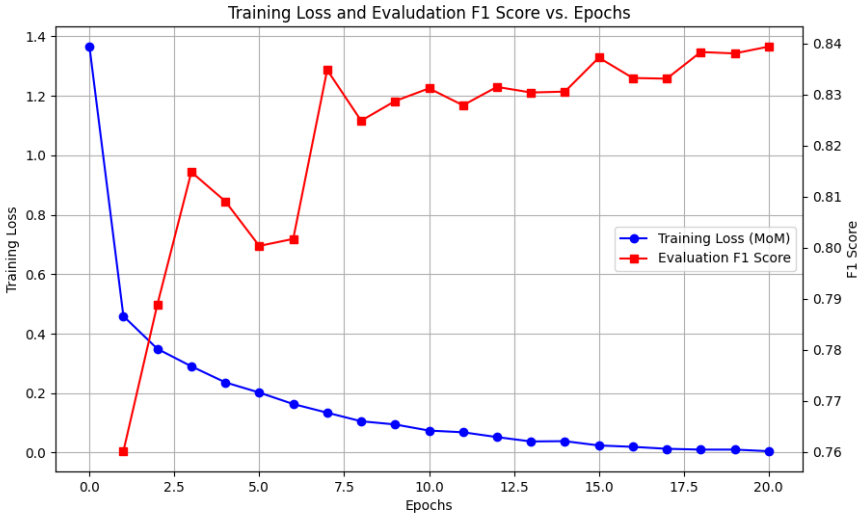


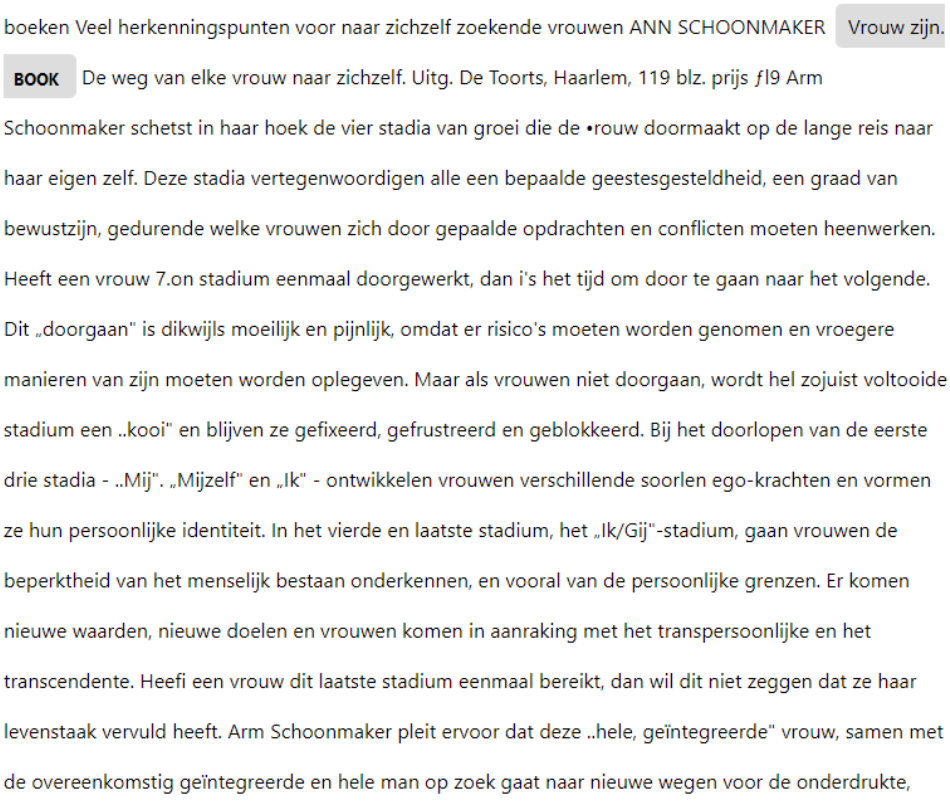
Figure B.8: *Training history of xlm-roberta-large-finetuned-conll03-english.*

Appendix C

Prediction examples on unseen data

Examples of predictions made by the final model (xlm-roberta-large-finetuned-conll03-english) on unseen data.

C.1 Leeuwarder Courant



Een reus ontwaakt Leo Klatser. „China nu BOOK — een reus ontwaakt". Paperback, 175 blz., geïll., ƒ7.50. Uitgave „De Bezige By", Amsterdam. De journalist Leo Klatser heeft de gelegenheid gehad een reis door Conimunistisch-China te maken. Tot verrassende onthullingen heeft dat bezoek niet geleid; meer dan enkele grote steden kon hy niet bezichtigen. En by wat hij zag, en met wie hy sprak, er was steeds een dienstwillige, maar zeer gezagsgetrouwe tolk-gids aanwezig. Als hy alleen zijn persoonlijke waarnemingen op napier had gesteld, zou hy niet veel meer dan een geretoucheerd beeld van het huidige China hebben kunnen geven. Klatser heeft echter ernstig geprobeerd door de uiterlykheden heen te zien en zyn eigen indrukken te mengen met een aanzienlijke dosis informatie, die uit andere bronnen over de ontwikkeling van China bekend is. Het resultaat is een intelligent geschreven beschouwing over een ontzaglijk groot land, „dat ons van uur tot uur nieuwe verrassingen zal bezorgen". Uiteraard besteedt de schrijver ruime aandacht aan de mislukking van de „grote sprong", waarmee China op korte termijn de status van industriële natie wilde veroveren. Ten dele is die mislukt door Chinese misrekeningen, ten dele ook door het wegvallen van de beloofde Russische hulp en door natuurrampen, die de voedselvoorziening desorganiseerden. „Ondanks alles", zegt Klatser, „beschikt China thans over een industrieel potentieel, dat honderden malen groter is dan het ooit in de geschiedenis bezat. Sinds de mislukking van de grote sprong heeft men de vijfjarenplannen afgeschaft en is overgegaan tot oen soort prognoseplanning.

INGEBOEKT ■ir LEONHARD REINISCH (Herausfeber) — Vom Sinn der Tradition. BOOK aperback uit de uitstekende Schwarze Reihe met tien essays die in eerste instantie deel uitmaakten van een serie-programma van de Nachtstudio van de Beierse radio. De bijdragen zijn afkomstig van Leszek Kolakowski, Ernst Bloch, Arnold Toynbee, Raimund Panikkar, Ernst Simon, Hans Urs von Balthasar, Saburo Ichn, Ignazio Silone, Bertrand d'Astorg en Walter von Cube. Aan het boek werden biografische aantekeningen toegevoegd. Uitgave: Verlag C. H. Beek, München; 181 blz., DM 10.80. * ANDRÉ MALRAUX — Het menselijk tekort. BOOK Deze roman over de Chinese burgeroorlog van de latere Franse politicus, gaullist en minister dateert uit 1933. Het boek wordt sindsdien gerekend tot de grote romans van deze eeuw. De Nederlandse vertaling ervan maakte E. du Perron, een goede vriend van Malraux die de roman ook aan hem opdroeg. Dex eerste zeven drukken verschenen tot en met 1956 bil de Wereldbibliotheek, Het Spectrum bracht vervolgens de achtste druk ln 1958. Deze negende druk verscheen bij weer een andere uitgever: Contact, Amsterdam; 260 blz., gebonden ƒ19.75. + JAMES BARLOW — De bewijslast. BOOK Uit het Engels (door M. Kok) vertaalde politie- en misdaadroman met een kritische tendens. Het is de auteur er mede om te doen te wijzen op de als bedrijf georganiseerde en beoefende misdaad waarbij de grote bazen achter de schermen vaak ongemoeid worden gelaten dan wel niet te vangen zijnen de kleine handlangers voor de bul saan Het speelt in Londen in de hukten met de speel- en nachtclubs.

C.2 Het Parool

Chinezen lachen om andermans leed door ANGIE BERG Adriaan van Dis: Een barbaar in China. BOOK Eer reis door Centraal-Azië, ui tg.: Meulenhoff. Amsterdam, 1987, 98 blz., ƒ 18,50 Peter Yung: Xinjiang. BOOK The Silk Road. Islam's Overland Route to China, uitg.: Oxford University Press, Oxford, 1986, 16C blz., ƒ 75,- (ISBN 019 584121 2), importeur: Nilsson & Lamm, Amsterdam Al voor het begin van onze jaartelling baanden de Chinezen zich een weg door het onherbergzame westen om handelsbetrekkingen aan te knopen met andere beschavingen. Zo ontstonden de zogenaamde zijdewegen, waarlangs karavanen door de woestijnen trokken. Het gebied is nog steeds grimmig. Duyvendak, onze eminentste sinoloóg, constateerde nog maar een halve eeuw geleden dat de zijderoutes het enige gebied ter wereld zijn dat 'thans' minder toegankelijk is dan in de oudheid (China tegen de Westerkim, 1927, herdrukt in 1933 en 1948). Sindsdien is er in dit opzicht niet zo erg veel veranderd, al hebben de communisten natuurlijk wel een aantal verbindingen tot stand gebracht om hun greep op het uitgestrekte land te vestigen en te versterken. Vorig jaar heeft Peking de voornaamste zijdeweg, namelijk die van Xian via Lanzhou, Liuyan en Turpan naar Kashgar, voor buitenlanders opengesteld. Dat het reizen in dit gebied nog steeds geen sinecure is, blijkt uit de impressies van Adriaan van Dis van zijn tocht via deze legendarische route, die door de islam werd gebruikt om in Centraal-Azië door te dringen. De zijdeweg mag dan wel voor toeristen van over de grens toegankelijk zijn gemaakt, maar dat wil nog niet zeggen dat de Chinese

Het zoveelste nutteloze boek over Afrika Preston King: An African Winter. BOOK With a note on Ecology and Famine by Richard Leakey, Penguin Books, Harmondsworth, 1986,249 blz. ƒ 22,95, importeur: Nilsson & Lamm, Amsterdam (ISBN 0-14-052365-0) door LOUISE FRESCO Op mijn bureau stapelen de boeken, rapporten en kranteknipsels over Afrika zich op. Woorden zyn helaas niet eetbaar. Ik ben bang dat de meeste mensen in Afrika er, in ieder geval op korte termijn, weinig aan hebben, dat iedere zich zelf respecterende 'deskundige' zijn/haar visie op de problemen wil verwoorden. Zelf heb ik in dit opzicht ook wel boter op mijn hoofd, maar ik krijg eigenlijk steeds vaker het onbehaaglijke gevoel dat er iets niet klopt. Het gaat me daarbij niet zo zeer om de wetenschappelijke merites, of de zorgvuldig verantwoorde in- en uitleidingen van dit soort publicaties. Het is meer dat al die abstracte cijfers en generalisaties maar een deel van de werkelijkheid dekken. De werkelijkheid van de gemiddelde Afrikaan, ten zuiden van de Sahara, zoals er dan zogenaamd nauwkeurig bij wordt gemeld. Maar de gemiddelde Afrikaan bestaat natuurlijk net zo min als de gemiddelde Nederlander. En net zo als het standaard-beeld van Nederlanders, gezien door een outsider als Derek Philips op grote weerstanden stuit, zou iets dergelijks ook wel eens voor Afrika kunnen opgaan. Wat denken die Afrikanen, die rapporten onder ogen krijgen met titels als DE Afrikaanse crisis, of DE problemen van DE Afrikaanse voedselvoorziening ? Of erger nog, wat zouden ze denken bij de nieuwste westerse acties als luxe-koffiebroodjesvoor-Kenia, of 'kilo's voor Afrika' (afvallen als

boeken Twee grote journalisten laten de Apartheid voor zichzelf spreken door ROB WOUTERS Anthony Sampson: Black and Gould. BOOK Tycoons, Revolutionaries and Apartheid, uitg.: Hodder & Stoughton, Londen enz. 1987, 280 blz., ƒ 52,20, importeur: Nilsson & Lamm, Weesp (ISBN 0-340-39542-9) Joseph Lelyveld: Move your shadow. BOOK South Africa Black and White, uitg.: Michael Joseph, Londen, 19851 X + 390 blz., ƒ65,-, importeur: Nilsson & Lamm (ISBN 0 7181 2661 0) Van politieke zulthoofden die in Amsterdam een 'vijandige' bibliotheek over Zuid-Afrika in de gracht kieperen moet ik net zo min iets hebben als van het eenzijdige en vaak ook nog gesubsidieerde pamflettisme van sommige actiegroepen tegen de Apartheid. Maar een goed boek over Zuid-Afrika, geschreven door een onverdachte auteur in de beste tradities van het hoor- en wederhoor, beschouw ik als een geschenk uit de hemel. De laatste tijd zijn er twee van zulke behartigenswaardige boeken verschenen. Het ene, Black and Gold, komt van de schrijftafel van de Britse journalist Anthony Sampson en behandelt de relatie tussen de vertegenwoordigers van het grootkapitaal - de kurk waarop de Zuidafrikaanse economie drijft - en de leiders van het zwarte verzet. Het andere boek is van de Amerikaanse New Vork Times-journalist Joseph Lelyveld, voormalig correspondent van zijn krant in Zuid- Afrika. Lelyvelds boek heet Move Your Shadow ('weg met je schaduw' - schiet op), de kortaangebonden woorden waarmee een blanke golfer zijn zwarte bediende 'aanspoort. Het aardige en het prettige van deze boeken is dat ze elkaar ongewild aanvullen en

C.3 Trouw

Trouw ONDER REDACTIE VAN MONIC SLINGERLAND BOEKEN Iemoires las hield lang vast in partijtop Djilas, Jahre der Macht. BOOK f9oslawischen Kraftespiel. J°iren 1945-1966. DTV, Münf'472 blz. -ƒ29,25. ISBN 9 *23 303040. (imp. Van Dit- lil et Joegoslavië van Tito in de ten onder is gegaan, . ''id aangebroken voor een her'«tic op het communistische jiclen. Milovan Djilas, de be°ste Joegoslavische dissident, 'n 'Jahre der Macht' terug op ■ Nslukte experiment van het 'fisch socialisme, toen hij lid f/l h > ne* Politburo wa, minister en Sweling van Tito.' 'ö°ek is een pijnlijke getuigenis J orden van zijn eigen jarenlange voor de zwakke kanten de communistische heilstaat als vanzelfsprekend na de val lazi-Duitsland ontstond. „Er niet overlegd wie er aan de ' * zou komen, dat waren de schrijft Djilas. Hiern'kent hij ook zijn eigen rol Toen een afgevaardigde van e 9roep jonge intellectuelen om .Zeggenschap vroeg hoonde I sm Weg. „Heeft u soms ook '9 divisies", vroeg Djilas? De 9errian verliet angstig de verga.."""te; de top van de commuc e partij, diep verankerd in ,s Partizanenleger, bleef la* achter. pij I de partijtheoreticus geeft nu ■i at toen al enige zaken hem f etT>rnend voorkwamen. Het L e besluit van het centrale co-1 van de communistische partij de bouw van een enorm, vii modern gevangeniscomplex, i uWd door bijzonder ijverige 31 e krijgsgevangenen. Daarop i de zuivering van andersden- eY eh, monarchisten, kapitalisten m cisten geopend worden. .I'9duizend mensen die hun , Ucht hadden gezocht in de jS6n en bergen van het land, j[en in een massale klopjacht , n9edreven en afgevoerd. ,|r °ok andere politieke maatre,n strookten niet met

Geschiedenis Columbus, een bezeten dromer Gianni Granzotto, Christoffel Columbus, **BOOK** een biografie. Vert. M. G de Boer. Kwadraat, Utrecht. 376 blz., ISBN 90 6481 159 8. Ottho Kikkert, Taino, Columbus' verstoorde paradijs, Kok-Agora, Kampen. 154 blz. -ƒ 25,-. ISBN 90 391 0035 7, Michaei Dorris en Louise Erdrich, De kroon van Columbus, **BOOK** roman, vert. Cora Polet. De Prom, Baarn, 400 blz., ISBN 90 6801 188 X De geschiedenis is een gesprek zonder einde, beweerde Geyl, en in dit jaar van Columbus wordt de lezer die tracht de stroom van nieu- we boeken bij te houden, steeds wanhopiger, want wat de waarheid is over die geheimzinnige man ontgaat hem steeds meer terwijl hij leest en leest. Voor de lezer die zich niet wil verdiepen in de eindeloze haarkloverij over de details, maar die een duidelijk beeld van Columbus wil hebben, is er geen beter boek dan dat van de Italiaanse schrijver Granzotto. Het is vreemd dat het nog niet eerder in het Nederlands is vertaald, het is toch al van 1984. Maar nu is het er dan, voortreffelijk vertaald, geschreven met vaart en verbeelding en persoonlijke betrokkenheid en toch met een wijze distantie tegenover al de dwaze legenden die er rondom Columbus zijn gesponnen. Het is niet moralistisch vanuit de vragen van onze tijd geschreven, maar probeert de ontdekker te begrijpen vanuit zijn geestelijke wereld. Het heeft daarbij de these dat Columbus een ziener was, een bezeten zonderling, die met een raar mengsel van profetieën en begeerten in staat was te doen wat een logisch en rationeel renaissance-mens nooit had kunnen doen. Kortom een prachtig boek, het beste om de lezer van

Republiek eigenlijk alleen miraculeus nog bestond en elk ogenblik kon verzinken, vindt men dikwijls in buitenlandse bronnen, van William Temple tot Herder. Zo hebben we in deze Engels-Amerikaanse historicus weer eens een goede pleiter, we kunnen zijn boek met patriottisch gevoel lezen. Wat zien buitenlanders soms veel in ons! Maar het is wel wijs om kritisch te lezen, uit sporadische gegevens worden soms te makkelijk grote conclusies getrokken. Het gevaar dat de historicus met fantasie gauw bedreigt is dat hij te lichtvaardig interpreteert en te snel verschijnselen uniek acht. Maar al is de schrijver misschien niet altijd aan dat gevaar ontkomen, zijn boek is toch een verfrissende en belangrijke bijdrage tot onze Hollandse geschiedenis. En het is haast ongelooflijk dat een vreemdeling zo diep en grondig thuis weet te raken in ons verleden. Zelden laat hij een steekje vallen, zo waar hij het grafmonument van Van Wassenaar-Obdarn in Amsterdam plaatst i.p.v. in Den Haag, of waar hij Sinterklaas op 6 december laat vieren. En een enkele keer vermoed ik dat zijn dapper tentoongespreide kennis van Nederlands en Latijn niet zo heel degelijk is. Maar een kniesoor die daarop let. Ik wil deze bespreking toch eindigen met een woord van grote bewondering. Wat er misschien aan nauwkeurigheid ontbreekt wordt zo ruimschoots goed gemaakt door een moedige fantasie en een wezenlijk inzicht. J. L Price, Nederlandse Cultuur in de Gouden Eeuw, **BOOK** Aula-boeken, uitg. Het Spectrum, Utrecht, 216 blz., / 39,90. ISBN 90 274 5839 1. Simon Schama, The Embarrassment of riches, **BOOK** an interpretation of Dutch culture in the Ooiden Age, uitg. Collins. Londen 1987. 698 blz., ƒ 76,40 zeer rijk geïllustreerd, ISBN 0 0021 7801.