**Chapter 1: Introduction**

Background: Introduce the topic, relevance, and timeliness of your research.

Problem Statement: Clearly state the problem your thesis addresses.

Objectives: Outline the specific goals of your research.

Thesis Structure: Briefly describe the structure of the thesis.

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 OCR-scanned historical newspapers. Previous studies, such as those by Do et al. (2012) and Sarimehmetoğlu and Erdem (2023), have focused on extracting book titles from visual sources like 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 newspaper texts is predominantly rule-based. This approach involves selecting segments of approximately 200 words, 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 is highly dependent on the completeness of the database, and it requires extensive manual verification. These limitations compromise the efficacy of cultural analysis over time and restrict our understanding of historical intellectual trends.

In response to these challenges, this study proposes a novel approach leveraging advancements in natural language processing (NLP). By utilizing 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 without the need for exhaustive databases.

Eisenstein (2019) describes natural language processing (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 thoroughly investigated in the context of historical newspapers. Which leads us to the central research question of this thesis: *How can Named Entity Recognition be utilized to autonomously extract book titles from OCR-scanned historical newspaper, thereby facilitating deeper cultural and literary analyses?*

**Chapter 2: Background**

Previous Work: Review past studies and theories relevant to your research.

Gaps: Highlight the gaps in the literature that your study will address.

* NER
* NER models
* Rise of Transformer Models
* Evaluation Metrics for NER
* Approaches to NER
  + Rule-based
  + Supervised learning

**Named Entity Recognition**

As introduced in the "Introduction" chapter, Named Entity Recognition (NER) is a fundamental component in the toolkit of Natural Language Processing (NLP). NER focuses on locating and categorizing textual elements into specific categories such as names of people, organizations, and geographical locations. According to Jurafsky & Martin (2023), a named entity is defined as any item that can be distinctly identified by a proper name. They further explain that NER is a subset of a broader category of tasks called sequence labelling, which involves assigning a label to each word in a sequence.

Bird et al. (2009) emphasises two major challenges in Named Entity Recognition (NER). The first challenge is the ambiguity of many named entity terms. They illustrate 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 concerns multi-word named entities, which necessitates determining the boundaries of an entity, specifically identifying its beginning and end.

**Tagging schemes**

A tagging scheme in NER is a systematic method for labelling text to identify and classify named entities. The scheme defines how tokens are marked to indicate whether they are part of a named entity and the role they play within that entity. Bird et al. (2009) describe tokens as a "technical name for a sequence of characters," which can include words, subwords, or multiple words. Thus, NER fundamentally operates as a token classification task, where each token in a text is assigned a specific label to identify and categorize it within the structured data of named entities.

According to Jurafsky & Martin (2023), the BIO tagging scheme is the standard method for sequence labelling. Developed by Ramshaw & Marcus in 1995, BIO stands for Beginning, Inside, and Outside. The 'B' tag is used for the first token of a named entity, the 'I' tag for subsequent tokens within the same entity, and the 'O' tag for tokens that do not belong to any named entity.

Jurafsky and Martin (2023) discuss two other schemes for tagging: the simpler IO and the more complex BIOES. The IO scheme is akin to BIO but omits the Beginning tag, focusing only on Inside or Outside tags. Conversely, the BIOES scheme introduces two additional tags to the BIO scheme: End and Single. The End tag marks the last token of a multi-token entity, while the Single tag is used for entities that comprise one token only.

**Evaluation Metrics**

Named Entity Recognition (NER) is a task focused on token classification, where the 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 instances where the model accurately identifies a token as belonging to a specific entity class. False Positives are instances where the model erroneously labels a token as belonging to the entity class. Conversely, True Negatives refer to instances where the model correctly identifies a token as not belonging to the entity class. Lastly, False Negatives occur when the model fails to recognize a token as part of the entity class when it actually is.

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

Accuracy = TP + TN / TP + TN+ FP + FN

When dealing with 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 cases where only 1% of all tokens represent an entity, a model could achieve an accuracy of 99% by merely predicting the majority class. Despite this seemingly high accuracy score, the model's actual performance would be quite inadequate.

Precision and recall are some another simple metrics, commonly employed as a foundational component for more sophisticated metrics. Precision measures the proportion of correctly predicted entities out of all the entities predicted by the model (Grus, 2019). It addresses the question: "Of all the tokens the model predicted as entities, what proportion was correctly identified as entities?"

Precision = TP / TP + FP

Recall, on the other hand, focuses on the proportion of actual entities that were correctly identified by the model (Grus, 2019). It seeks to answer: "Of all the actual entities, what proportion did the model correctly identify as entities?"

Recall = TP / TP + FN

A more sophisticated metric that makes use of precision and recall is the F1 score. The F1 score is the harmonic mean of precision and recall, providing a single score that balances both the precision and the recall (Grus, 2019).

F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)

When adapting the F1 score for multi-class classification tasks, it can be calculated using two principal methods: micro-averaged and macro-averaged F1 scores (Grandini, Bagli, & Visani, 2020). The micro-averaged F1 score aggregates the counts of true positives, false negatives, and false positives across all classes to compute overall precision and recall. In contrast, the macro-averaged F1 score computes the F1 score for each class separately and then averages these scores. This approach gives equal importance to each class, making it particularly useful for datasets with imbalanced class distributions, as it ensures that each class contributes equally to the overall metric.

**Approaches to Named Entity Recognition**

In their survey on Named Entity Recognition, Jehangir et al. (2023) classified approaches to Named Entity Recognition into three categories: rule-based, unsupervised learning, and supervised learning.

They 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 is approach is the method currently used by the thesis supervisor to extract book titles from newspaper archives. Furthermore, Jehangir et al. (2023) identified two major drawbacks of rule-based approaches: their poor generalizability, as the rules are often domain-specific, and the requirement for advanced programming skills and significant human effort to develop these rules.

Unsupervised learning in Named Entity Recognition (NER) utilizes algorithms that can identify and categorize entities from text without any labelled training data. One such algorithm employs a dictionary of words, often known as a gazetteer-based approach. Bird et al. (2009) describe the gazetteer-based method as involving the examination of each word in the text to determine if it matches an entry in a predefined dictionary of named entities. If a match is found, the word is assigned to the corresponding entity category. This technique relies on direct lookup and comparison, making it straightforward yet dependent on the quality and completeness of the dictionary used. An example of this approach is the study by Toral et al. (2005), which explored the automatic creation and maintenance of gazetteers using Wikipedia. This research demonstrated the method's efficacy in constructing NER gazetteers for location and person categories.

Géron (2017) explains the concept of supervised machine learning, where models are trained using labelled data. This means that each piece of input data is associated with a corresponding expected output, allowing the model to learn how to map inputs to outputs during the training process. Due to the limitations inherent in rule-based and unsupervised methods and the availability of a substantial labelled dataset, this thesis will concentrate exclusively on the supervised approach. A deeper exploration of supervised models for Named Entity Recognition will be presented in the section ‘Models for Named Entity Recognition'.

**Models for Named Entity Recognition**

In their survey on Named Entity Recognition in historical documents, Ehrmann et al. (2023) distinguish between two types of supervised machine learning approaches: traditional machine learning and deep learning. They refer 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) define deep learning as the application of modern, multilayered neural networks to model complex patterns and relationships in data.

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 Named Entity Recognition, 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 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.

Collobert et al. (2008) was one of the first studies that applied deep learning for Named Entity Recognition. They successfully used a simple Convolutional Neural Network (CNN) for several Neural Language Processing tasks, including Named Entity Recognition. Géron (2017) provides a detailed explanation of CNNs, noting that in addition to their application in NLP, they have been utilized in computer vision since the 1980s. This connection is logical given that CNNs emerged from the study of the brain’s visual cortex. Convolutional Neural Networks 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 understand and interpret large and complex text datasets.

Another category of deep learning models employed for NER includes Recurrent Neural Networks (RNNs). In contrast to CNNs, RNNs are specifically designed to process sequential data (Keraghel et al. 2024). Jurafsky & Martin (2023) provide a comprehensive explanation of their 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 counter 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) highlight 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 LSTM (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 an RNN, the sequential input is fed in one by one, with the hidden state being updated at each step. By the time the final input of the sequence is processed, the hidden state is expected to encapsulate the entire context of the sequence, regardless of its length. However, as the RNN processes longer sequences, early inputs can be "forgotten" because of the limited capacity of the hidden state to retain information over time. The attention mechanism was developed as a solution to this bottleneck problem. Instead of relying solely on the final hidden state to encode the entire sequence, the attention mechanism allows the model to access information from all the hidden states throughout the sequence. This enables the RNN to focus on different parts of the input sequence as needed, thereby retaining important details and improving performance on tasks involving long sequences.

Out of the idea of the attention mechanism introduced in RNNs, the mechanism called self-attention was developed. While traditional attention mechanisms in RNNs improved performance, they still relied on sequential processing, limiting parallelization and efficiency. The self-attention mechanism, however, allows each token in the input sequence to directly interact with every other token, enabling parallel processing and better capturing long-range dependencies. This shift led to a new deep learning architecture known as the transformer model introduced by google in the paper “Attention is All you Need” (Vaswani et al., 2017) and revolutionized the field of NLP.

While Yang and Xu (2020) highlighted the BiLSTM-CRF as state-of-the-art as of 2020, Labusch et al. (2019) demonstrated the superior performance of an early transformer model, BERT (Devlin et al., 2018), over the BiLSTM-CRF in NER tasks involving historical OCR-scanned German text. Furthermore, Ehrmann et al. (2023) noted in their survey on NER in historical documents that transformer-based networks are surpassing BiLSTM models in the deep learning landscape. Similarly, Sun et al. (2021) studied NER in the biomedical domain, noting that although BiLSTM-CRF models were once considered state-of-the-art, transformer-based models have since surpassed them in performance within the biomedical domain. The importance of Sun et al. (2021) lies in the fact that the biomedical domain, like this study, focuses on very specialized custom named entities.

**Large Language Models**

Chockalingam et al. (n.d.) define Large Language Models (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 explain how Large Language Models (LMM) 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.

In the past, LLMs were predominantly based on architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. However, the recent development of the transformer architecture has made it the preferred choice for LLMs (Chockalingam et al., n.d.).

As described in the section “Models for Named Entity Recognition,” RNNs face limitations when processing longer sequences, as early inputs can be "forgotten" due to the limited capacity of the hidden state. While attention mechanisms can be added to RNNs and LSTMs to mitigate this issue, doing so often introduces significant complexity. Transformers, on the other hand, are inherently more suitable for LLMs due to their built-in attention mechanisms, which facilitate the capture of long-range dependencies. Additionally, transformers process all tokens in a sequence simultaneously, allowing for massive parallelization and significantly faster training times. In contrast, RNNs and LSTMs process data sequentially, making parallel computation challenging.

Transformers’ ability to parallel compute results in faster training times, the capability to train models with more parameters, and the handling of larger datasets, all leading to better performance on various natural language processing tasks (Amaratunga, 2023).

**Transformer-based Large Language Models**

Transformer-based Large Language Models mostly differ from each other based on architecture, and the way they are trained (Amaratunga, 2023).

Jurafsky & Martin (2023) explain how transformer models are trained and categorises the training process in two phases: the pre-training and fine-tuning.

The pre-training of a transformer model is a bit different than most supervised machine learning model earlier explain in section “Models for Named Entity Recognition”. Instead of supervised training, the pre-training of transformer models relies on the concept of self-supervised learning. This is a type of machine learning approach where the model learns to predict part of its input data from other parts, without needing explicit human-labeled data. One such pre-training task is Masked Language Modeling (MLM). Here a percentage of the input tokens are masked, and the model is trained to predict these masked tokens. Another pre-training task is Next Sentence Prediction (NSP), in this task the model is given pairs of sentences and must determine if the second sentence is the actual next sentence in the original document. Another pre-training task is Causal Language Modeling (CLM), here to model is trained to predict the next word in a sentence.

After pre-training, the transformer model can be fine-tuned for specific tasks such as Named Entity Recognition. The purpose of the pre-training phase is to teach the model the meanings of words and their relationships within the language. This foundational knowledge enables the model to learn the specific final task more easily during fine-tuning. This concept is an instance of what is called transfer learning, where knowledge gained from one task is transferred to improve performance on a different but related task.

BERT (Devlin et al., 2019), short for Bidirectional Encoder Representations from Transformers, is an example of transformer model that relies on both MLM and NSP during pre-training. BERT is the one of the most popular used transformer model and revolutionized the world of NLP (Ravichandiran, 2021). Its architecture is based on the original implementation from Vaswani et al. (2017), which introduced the concept of the transformer model.

**Text representation**

TODO describe word embeddings etc.

**Chapter 3: Methodology**

Data Collection: Describe how and where you collected your data (e.g., sources of book reviews).

Tools and Technologies: Detail the LLMs and other tools used, focusing on their relevance to NER tasks.

Implementation: Discuss the development and implementation of your NER model.

* Limitations and restrictions
* Dataset Construction & Analysis
* Detailed explanation of chosen NER models
* Research procedure
  + Evaluation
  + Experiments

**Chapter 4: Results**

Model Performance: Present the effectiveness of your model in identifying book titles.

Comparisons: Compare your results with existing methods, if applicable.

Discussion: Interpret the results and discuss their implications.

**Chapter 5: Conclusion and Future Work**

Summary of Findings: Recap the key findings and their significance.

Contributions: State the contributions of your research to the field of applied data science.

Limitations: Acknowledge any limitations encountered during the study.

Future Research: Suggest areas for future investigation.

**References**

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

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

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>

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>

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

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

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.

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>

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.

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>

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>

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.

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).

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

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.physd.2019.132306>

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

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>

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).

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>

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>

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.

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

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

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

**Appendices**