**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 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 describe 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 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.

**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>

**Appendices**