**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

**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.

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 only one token.

**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>

**Appendices**