Universiteit Leiden

Master Computer Science

Translation as a Bridge: Assessing the Feasibility of English BERT for Low-Resource Languages

Name: Giulia Rivetti
Student ID: s4026543

Date: [04/06/2025]

Specialisation: Artificial Intelligence

1st supervisor: Marco Spruit
2nd supervisor: Marcel Haas

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

BERT models represent a significant advancement in Natural Language Processing (NLP), establishing themselves as state-of-the-art due to their robust ability to handle unstructured text across diverse languages and domains. However, developing high-quality BERT models for non-English languages remains a major challenge, often requiring extensive computational resources and large annotated datasets, resources that are scarce for many minority or low-resource languages. A promising alternative to building separate language-specific models is to translate non-English data into English and leverage existing, pre-trained English BERT models. Because these English models are typically trained on broader and more diverse corpora, they often offer improved generalization and robustness. Although initial studies suggest that translation-based approaches can yield competitive or even superior results compared to native-language models, research in this area remains limited, with most efforts still focused on developing dedicated models for each language. This thesis investigates whether translating text into English and fine-tuning English BERT models can serve as a viable and scalable strategy for multilingual NLP. We evaluate this approach across six core NLP tasks (Sentiment Analysis, Hate Speech Detection, Question Answering, Named Entity Recognition, Part-of-Speech Tagging, and Natural Language Inference), using datasets translated from five typologically diverse languages: Bulgarian, Chinese, Dutch, Italian, and Russian. Our findings indicate that translation-based models match or surpass native-language BERT models in many cases, particularly in tasks like POS tagging and QA, where lexical semantics and structural alignment are less sensitive to translation artifacts. Performance was especially promising for Dutch, likely due to its linguistic proximity to English. In contrast, results for Chinese were consistently weaker, reflecting the greater typological distance from English and the presence of strong native models trained on extensive Chinese corpora. Moreover, tasks requiring fine-grained token-level precision or cultural nuance, such as NER and Hate Speech Detection, tended to suffer under the translation-based method, revealing clear limitations. While this approach proved effective in 57% of the evaluated cases and demonstrates real potential for scalable, resource-efficient multilingual NLP, this thesis also highlights its boundaries. The results show that translation isn't a universal solution and requires careful consideration, particularly in contexts where it could introduce ambiguity or when robust native-language models already exist. Nonetheless, the translation-based strategy remains a valuable tool for extending NLP capabilities to underrepresented languages and contributes to ongoing efforts toward linguistic inclusivity and sustainability in AI.

Contents

1 Introduction

In recent years, advances in Natural Language Processing (NLP) have led to revolutionary progress in how machines process, understand, and generate human language. Central to this growth are deep learning models like BERT (Bidirectional Encoder Representations from Transformers), which have significantly improved performance across a wide range of NLP tasks. However, much of this development has been concentrated in high-resource languages such as English, where vast amounts of annotated data and computational resources are readily available. In contrast, most of the world's languages remain largely underrepresented in NLP research and tools, reinforcing digital inequality and limiting the global applicability of state-of-the-art models.

Efforts to develop language-specific BERT models for non-English languages have shown promising results. However, training these models from scratch is resource-intensive, demanding significant computational power, specialized hardware, and extensive human annotation efforts. As a response to this challenge, recent studies have proposed a translation-based alternative: translating non-English text into English and then fine-tuning an English BERT model on the translated text. Preliminary findings suggest that this method can offer performance comparable to, or even better than, native-language models for certain tasks. The advantage of this approach lies in its potential efficiency, which comes from reusing a powerful English model instead of training individual models for each language, and its scalability across a broader range of linguistic contexts.

Despite its promise, this translation-based approach remains underexplored. Prior studies have typically focused on individual languages or isolated tasks, offering limited insight into broader generalizability. Furthermore, the effectiveness of translation varies depending on linguistic distance, cultural context, and task sensitivity, raising questions about when and why this method succeeds or fails. There is a clear need for more comprehensive research that evaluates the strengths and limitations of this strategy across diverse languages and NLP tasks.

This thesis addresses that gap by evaluating the feasibility and effectiveness of using a translation-based strategy instead of fine-tuning language-specific BERT models for multiple low-resource languages. Specifically, we investigate whether translating datasets from languages such as Bulgarian, Chinese, Dutch, Italian, and Russian into English and fine-tuning a pre-trained English BERT model can deliver performance comparable or superior to that of language-specific BERT models. We consider six key NLP tasks: Sentiment Analysis, Hate Speech Detection, Question Answering (QA), Named Entity Recognition (NER), Part-of-Speech (POS) Tagging, and Natural Language Inference (NLI). This selection allows for a broad evaluation across both classification and token-level tasks.

1.1 Research Objectives

This thesis seeks to explore the feasibility and effectiveness of leveraging a translation-based approach to enhance the performance of NLP tasks for low-resource languages. These languages often face challenges due to lack of linguistic resources and pre-trained models, which limits their NLP development. The main objective of this research is to evaluate whether translating datasets from low-resource languages into English, followed by fine-tuning a pre-trained English BERT model, can achieve performance comparable to, or even surpass, models trained on native-language data. Specifically, we aim to determine whether this method can serve as a practical and scalable solution for adapting powerful English NLP tools to underrepresented

languages.
This study evaluates the effectiveness of this approach across a variety of NLP tasks and, to ensure a comprehensive analysis, it includes languages from diverse linguistic families: Indo-European Romance (Italian), Indo-European Germanic (Dutch and English), Sino-Tibetan (Chinese), and Indo-European Slavic (Russian and Bulgarian). This diversity allows us to assess the generalizability of the translation-based method and to explore how linguistic characteristics and data availability affect performance.
Beyond scientific contributions, this project aims to make a meaningful societal impact by offering a more sustainable and inclusive approach to NLP. If translating text and fine-tuning English BERT models proves effective, it could significantly reduce the computational cost and energy consumption associated with training separate models for each language, while also accelerating NLP development for underrepresented linguistic communities.
Ultimately, this thesis seeks to lay the groundwork for adaptable, resource-efficient NLP solutions, contributing to the long-term goal of reducing language-based inequality in AI technologies.

1.2 Research Questions

In order to achieve our predefined objectives, we investigate several key research questions, designed to assess the overall effectiveness, generalizability as well as limitations of using translation as a strategy for improving NLP performance across different language families and tasks.

Main Research Question: To what extent can a translation-based approach using the English BERT model achieve robust performance across NLP tasks in low-resource languages, and potentially surpass native-language BERT models?

To address the main research question, this thesis will explore several sub-questions:

1. Does a translation-based approach using English BERT perform consistently across languages from different linguistic families?
2. Are there specific NLP tasks where this approach is more effective?
3. How does fine-tuning English BERT on translated text compare to using pre-trained native-language BERT models?

To answer these sub-questions we will apply a translation method to all datasets, and then evaluate: the consistency across languages (Q1), the effectiveness on specific NLP tasks (Q2), and finally the comparative performance with native-language models (Q3).

1.3 Thesis Organization

This thesis follows the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology [141], a structured and widely adopted framework for managing data mining projects. Introduced in 1999 to standardize data mining processes across industries, CRISP-DM has since become the most popular methodology for data mining analytics and data science projects. The CRISP-DM framework consists of six sequential phases, as illustrated in Figure 1. The

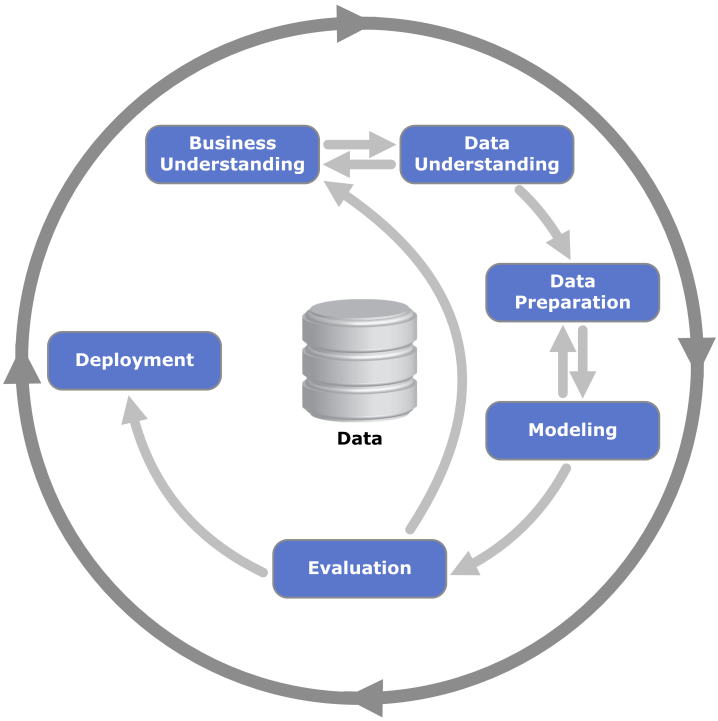

Figure 1: Phases of the CRISP-DM Process Model for Data Mining [141].

first phase, business understanding, focuses on defining the project objectives and formulating the data mining goals. This phase is addressed Section 1, where we establish the research objectives as well as the research questions that we aim to answer. Further discussion is provided in the Background (2) and Systematic Literature Review (3) sections. Section 4 also falls under the Business Understanding phase, as it outlines the experimental design decisions, including model selection, hyperparameter configuration and training strategy. The next step, data understanding, involves collecting, describing, exploring and assessing the data. This is presented in Section 5, where we report the dataset selection process and description for each task, with further exploratory analysis in Section 5.3. Next, in the data preparation phase, we apply several preprocessing techniques to make the data suitable for model training. This process is detailed in Section 5.2. The modeling phase of CRISP-DM corresponds to Section 6, where the results of our experiments are presented, while the fifth stage, evaluation, is covered in Section 7, where the results are analyzed and discussed. The final phase, deployment, involves implementing the model in a real-world application. Here, this phase is not treated as a separate step, but rather forms the core of the research itself: the work presented here can be seen as the practical application of the CRISP-DM process. The final chapters of this thesis include future work (Section 8), where we report the current limitations and propose improvements, followed by a conclusion (Section 9), that summarizes the key findings and contributions.
The code used in this research is available at `https://github.com/GiuliaRivets01/Master-Thesis`.

2 Background

This section provides essential background knowledge to understand the research presented in this thesis. In particular, we touch the fundamentals of Natural Language Processing, the set of NLP tasks considered in this work, the Transformer architecture, BERT, and machine translation.

2.1 Natural Language Processing (NLP)

Natural Language Processing is a subfield of Artificial Intelligence that enables computers to read, analyze, interpret and derive meaning from human language. This practice combines linguistics, statistics, and machine learning to allow computers to understand language in a meaningful way [35]. Over the past decade, NLP has become deeply embedded in everyday technologies: it is used to filter spam emails, get relevant results on search engines, and it is applied in the autocorrect features of messaging platforms. These are only a few examples of how NLP impacts one's daily life. By facilitating human-machine communication in natural language, NLP offers benefits across many industries and applications, such as the automation of repetitive tasks, improved data analysis and insights, and content generation [49].
Until recently, NLP tasks were typically addressed by designing separate models for each task. However, this changed with the introduction of BERT in 2018 [29], which demonstrated that a single model could achieve state-of-the-art performance across multiple NLP tasks. This marked a turning point in the field, positioning BERT as a foundational model for general-purpose language understanding.

2.2 NLP Tasks

In this thesis, we fine-tune BERT on six NLP tasks: Sentiment Analysis, Question Answering, Hate Speech Detection, Natural Language Inference, Part-Of-Speech Tagging and Named Entity Recognition. Each task is briefly introduced below to provide the necessary context.

Sentiment Analysis Also known as opinion mining, sentiment analysis refers to the task of identifying the emotional tone conveyed in a piece of text, typically categorized as positive, negative, or neutral. It is widely used in applications like product reviews, customer feedback analysis, and social media monitoring. As a classification task, it requires models to associate textual cues with sentiment labels. Key challenges include handling sarcasm, negation, and the presence of mixed sentiments (multi-polarity) [9].

Question Answering QA involves automatically providing answers to questions posed in natural language. It plays a key role in many applications such as virtual assistants, customer support systems, and search engines [20]. This thesis focuses on extractive QA, where the model identifies an exact span from a given context that answers the question. While effective for fact-based questions, extractive QA struggles with queries that require reasoning or synthesizing information across multiple sentences.

Hate Speech Detection This task aims to identify language that expresses hatred, discrimination, or violence toward individuals or groups based on characteristics like race, religion, gender, or nationality. With the rise of social media and online platforms, automatic hate speech

detection has become vital for content moderation [86]. Typically treated as a binary or multi-label classification task, it presents major challenges due to the subjective nature of hate, the use of subtle language, and cultural variability in offensive expressions.

Natural Language Inference (NLI) Also known as textual entailment, NLI determines whether a given hypothesis logically follows from a premise. Specifically, the relationships are classified as *entailment*, *contradiction*, or *neutral*. NLI plays a critical role in tasks such as fact-checking, text summarization, semantic search, and question answering [81].

Part-Of-Speech tagging (POS) POS tagging assigns grammatical categories—such as noun, verb, adjective, or adverb—to each word in a sentence. As a sequence labeling task, it provides essential information about syntactic structure and is important for many downstream tasks, including parsing and information extraction. Challenges include dealing with ambiguous word forms, polysemy, and the morphological complexity of certain languages.

Named Entity Recognition (NER) NER is a sequence labeling task, focused on identifying and classifying named entities in text, such as *persons*, *organizations*, *locations*, and *dates*. NER is widely used in tasks like information extraction, question answering, and knowledge base construction. Difficulties arise from nested entities, ambiguous boundaries, and language-specific naming conventions.

2.3 The Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. in 2017 [135], revolutionized the field of NLP by replacing the recurrent mechanisms of previous models, like RNNs and LSTMs, with a self-attention mechanism. This shift enabled models to process all input tokens simultaneously and capture long-range dependencies more efficiently.
At a high level, the Transformer consists of two main components: an encoder, which processes input text, and a decoder, which generates output text. BERT, the model used in this thesis, leverages only the encoder for language understanding tasks.
The core building blocks of the Transformer architecture, which is sketched in Figure 2, include:

1. **Input encoding**: each input token is mapped to a vector representation using an embedding matrix. Since the Transformer lacks inherent recurrence, positional encodings are added to these embeddings to capture the order of tokens.

2. **Transformer blocks**: these consists of multi-head self-attention layers, which enables the model to focus on different parts of the input sequence simultaneously, and feed-forward neural networks, applied independently to each position. Layer normalization and residual connections are used to stabilize training.

3. **Language modeling head**: in generation tasks, the final hidden states are projected back into the vocabulary space using an output matrix followed by a softmax layer to predict the next token [53].

This architecture forms the backbone of almost all modern NLP models, including BERT, GPT, and T5, enabling them to learn rich, contextual representations from large-scale corpora.

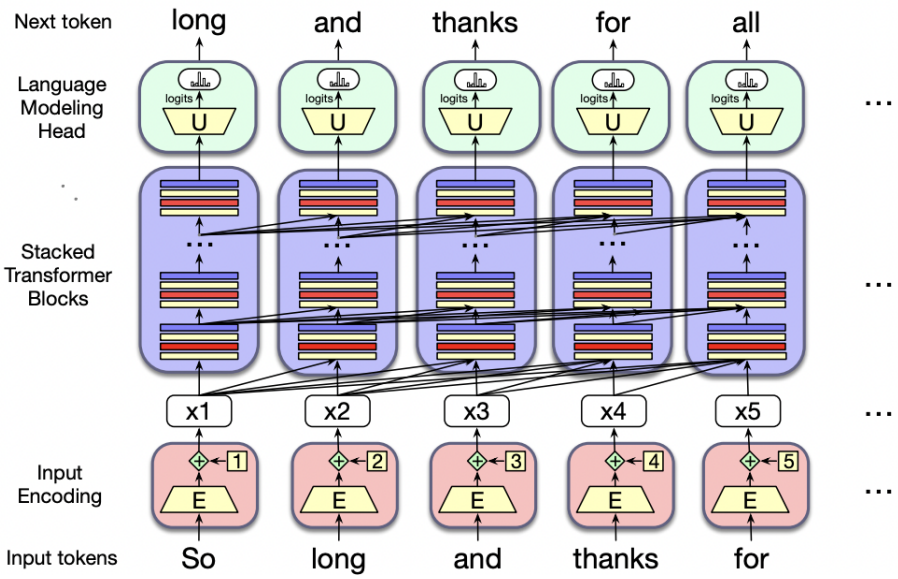

Figure 2: The architecture of a left-to-right Transformer model. Each input token is encoded and passed through a series of stacked Transformer blocks, followed by a language modeling head that predicts the next token in the sequence.

2.4 BERT

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. in 2018 [29], is one of the most influential applications of the Transformer architecture in language understanding. While the original Transformer includes both encoder and decoder stacks, BERT uses only the encoder. Its core innovation lies in its bidirectional self-attention mechanism, which allows the model to consider both the left and right context of every word in a sentence simultaneously, which is something unidirectional models like GPT cannot do. This enables BERT to capture richer semantic representations, crucial for tasks like question answering, sentiment analysis, and named entity recognition.
A key innovation behind BERT's effectiveness lies in its two-phase training strategy. The first stage is pre-training, where the model is trained on large-scale unlabeled corpora using two objectives:

- Masked Language Modeling (MLM): Random tokens are masked and the model learns to predict them based on surrounding context.
- Next Sentence Prediction (NSP): The model predicts whether a second sentence logically follows a given first sentence.

The second phase is fine-tuning: after pre-training, BERT can be adapted to specific tasks (like sentiment analysis or NLI) by adding a task-specific output layer and training on labeled data. This greatly reduces the need for training task-specific models from scratch.
Before input text (e.g., a sentence, paragraph, or document) can be processed by BERT, it must be tokenized, or in other words, it needs to be split into smaller units called tokens. These tokens can be words, subwords, or characters, depending on the tokenization method used. BERT uses WordPiece tokenization, a subword-based approach that enables it to handle rare or unseen words by breaking them into smaller known units [74]. For example, the word "apples" would be tokenized into ["apple", "s"], and "unhappiness" would be split into ["un",

"##happiness"], where "##" indicates that the subword is a continuation of the previous token. This helps the model generalize to words outside its vocabulary. Tokenization is the first step in converting text into a format the model can understand. Once tokenized, each subword is mapped to a unique integer ID from BERT's vocabulary, then converted into a vector embedding, and padded or truncated as needed.

2.5 Machine Translation

Machine Translation (MT) refers to the use of computational systems to automatically translate text from one language into another. Since its first appearance in 1947, Machine Translation has been considered one of the most complex challenges in the field of natural language processing [137]. Over the decades, MT systems have evolved from rule-based systems to statistical methods, and more recently, to deep learning approaches.

Today, Neural Machine Translation (NMT) represents the state of the art in MT. NMT systems employ deep neural networks, typically based on the Transformer architecture, to model translation as a sequence-to-sequence learning task. The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text [145]. Although initially viewed by many as an inconsistent translation tool for translating important or high-visibility content, especially at scale [117], NMT has rapidly matured over the past decade, gaining widespread adoption across industry and academia due to its ability to handle subtle linguistic nuances and deliver high-quality results.

In the context of low-resource languages, where labeled data, linguistic tools, and pretrained models are often lacking, MT offers a promising alternative. By translating data into English, researchers can exploit powerful, well-established models like BERT for downstream NLP tasks. This translation-based strategy allows for high performance without the overhead of training language-specific models from scratch. In this thesis, we adopt this approach as a practical and scalable solution to address resource limitations in multilingual NLP.

3 Systematic Literature Review (SLR)

We perform a Systematic Literature Review (SLR) to understand the current state of research on key topics, including current translation-based approaches used to extend the resources of minority languages in NLP, their limitations, and the benchmarks used in NLP for non-English languages. This assessment helps identify advancements and gaps in existing studies, while ensuring a comprehensive, standardized, and transparent analysis of relevant findings.
While systematic reviews provide valuable insights, they are also known to be time-consuming and resource-intensive processes [127], often requiring a significant level of expertise [152]. To address these challenges, this SLR will follow the SYMBALS (SYstematic review Methodology Blending Active Learning and Snowballing) procedure proposed by van Haastrecht et al.[132]. This approach can speed up the process of finding relevant papers by employing machine learning techniques combined with backward snowballing, reducing the risk of omitting important studies. The SYMBALS procedure, which is reported in Figure 3, will be followed in the next sections of this literature review.

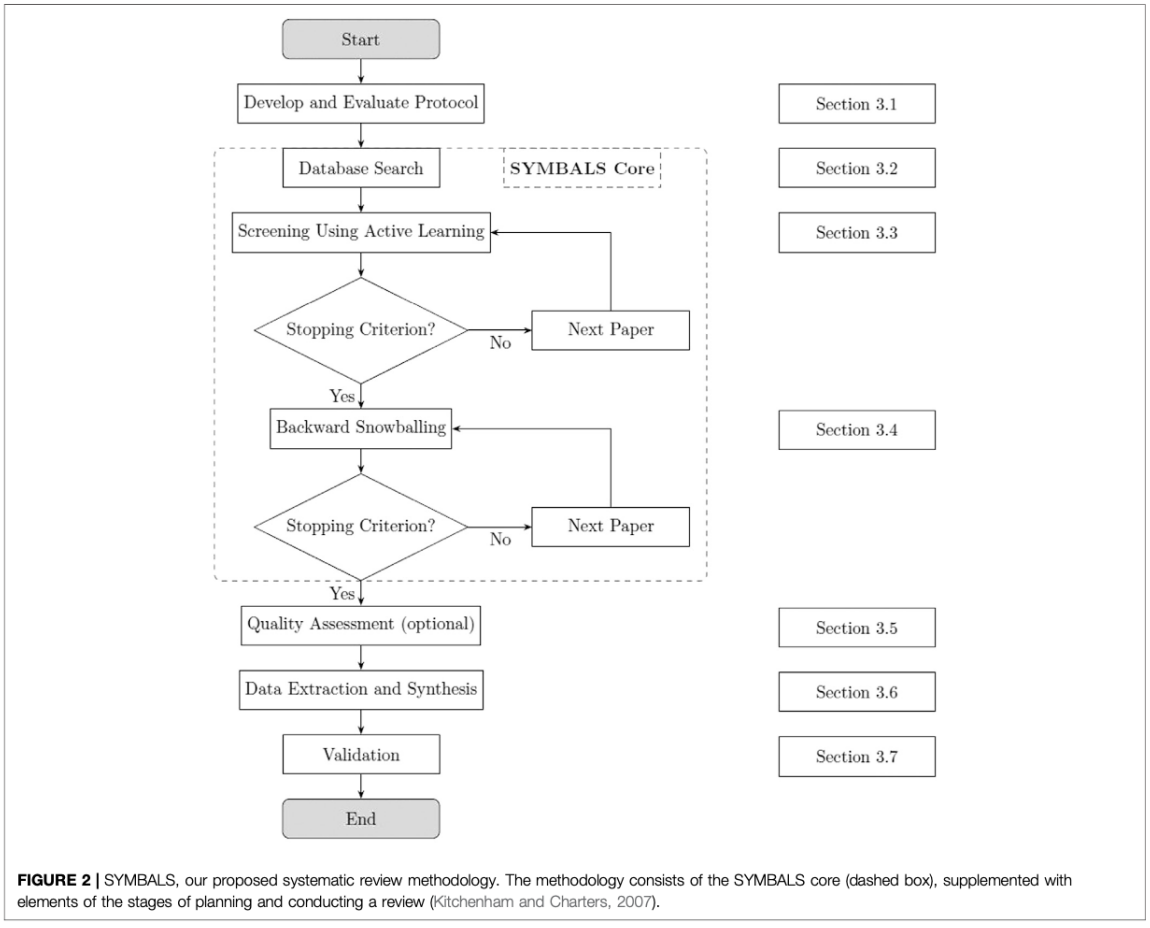

FIGURE 2 | SYMBALS, our proposed systematic review methodology. The methodology consists of the SYMBALS core (dashed box), supplemented with elements of the stages of planning and conducting a review (Kitchenham and Charters, 2007).

Figure 3: SYMBALS pipeline. The methodology consists of the SYMBALS core (dashed box), supplemented with elements of the stages of planning and conducting a review [132].

3.1 Develop and Evaluate Protocol

The first step of SYMBALS involves defining a protocol and formulating research questions that justify the need for the systematic literature review. This SLR aims to investigate the current state of research on BERT models for non-English languages and translation-based approaches in NLP. The review seeks to identify advancements and research gaps, while evaluating the potential of translation-based methods as an alternative to native-language BERT models. The insights gained will play a crucial role in defining the scope and direction of this thesis. To guide the review process, the following research questions have been formulated:

1. What is the current state of research on translation-based approaches for extending English BERT models to NLP tasks in low-resource languages?
2. What are the primary challenges in performing NLP tasks in low-resource languages with BERT models?
3. Which benchmarks and datasets have been used in prior research for the NLP tasks and the languages considered in this thesis?

Search strategy The search process begins by identifying relevant keywords related to BERT models, NLP tasks, translation methods and the target languages. Using these keywords, structured queries will be formulated and applied to three major academic databases: *IEEE Xplore*[1], *ACM Digital Library*[2] and *Web of Science*[3]. This will provide an initial collection of relevant studies. Next, the ASReview framework [130] will be employed to perform Active Learning with a machine learning model, systematically identifying the most relevant studies from the initial pool. To further ensure proper coverage, backward snowballing will be applied, incorporating additional relevant studies cited in the selected papers. Since the final number of selected papers is expected to remain manageable, a final validation step is deemed unnecessary. Therefore, we will follow the SYMBALS methodology until the backward snowballing step.

Selection criteria Throughout the screening process, both during Active Learning and backward snowballing, specific selection criteria will be applied to determine the relevance of each study:

- **SLR Question 1**: Papers were included if they used translation from a specific language into English to apply English BERT models for NLP tasks.
- **SLR Question 2**: Papers were selected if they focused on challenges in NLP tasks for low-resource languages, particularly those involving BERT models. Papers addressing NLP tasks in low-resource languages without discussing associated challenges were excluded.
- **SLR Question 3**: Studies were excluded if they did not cover the selected languages or if they involved tasks different from the target tasks. For instance, studies on Aspect-Based Sentiment Analysis (ABSA) and implicit sentiment analysis were not considered.

These criteria will be applied to the titles and abstracts of each paper during the review process. A complete overview of the adopted SLR protocol is provided in Figure 4.

[1] `https://ieeexplore.ieee.org/Xplore/home.jsp`
[2] `https://dl.acm.org/`
[3] `https://www.webofscience.com/wos/woscc/basic-search`

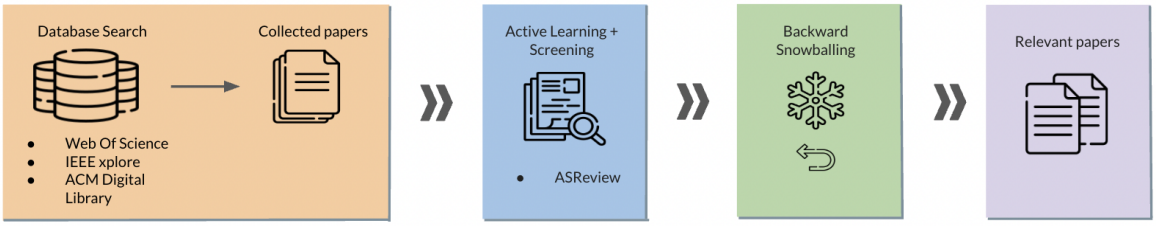

Figure 4: The SLR protocol adopted, which follows the SYMBALS methodology. The process is composed of a database search, active learning and screening, and backward snowballing.

SLR Question 4 Unlike the first three SLR questions, SLR Question 4 does not require a separate database search. Instead, we will extract relevant information from the papers retrieved during the database search and screening process for SLR Questions 1, 2, and 3. This approach ensures that preprocessing techniques are examined within the context of studies already deemed relevant to our research, reducing redundancy and maintaining consistency across the review process.

3.2 Database Search

The second step in the SYMBALS pipeline involves conducting a database search to identify relevant studies. As previously mentioned, we employed three academic databases: IEEE Xplore, Web Of Science and ACM Digital Library. These databases provide advanced search functionalities, allowing us to apply well-structured queries to retrieve papers relevant to the SLR research questions. The queries used for each SLR question are as follows:

SLR Q1 query:

```
(
    ("Abstract": BERT OR "Abstract": mBERT OR "Abstract": transformer*)
    AND ("Abstract": minority language* OR "Abstract": low−resource language*
        OR "Abstract": low−resource NLP OR "Abstract": underrepresented language*)
    AND ("Abstract": translated)
    AND ("Abstract": NLP OR "Abstract": natural language processing)
)
```

SLR Q2 query:

```
(
    ("Abstract":BERT OR "Abstract":mBERT)
    AND ("Abstract":minority language OR "Abstract": low−resource language
        OR "Abstract": Italian OR "Abstract": Dutch OR "Abstract": Mandarin
        OR "Abstract": Russian OR "Abstract": Bulgarian)
    AND ("Abstract":NLP OR "Abstract": natural language processing OR "Abstract": NER
        OR "Abstract": POS OR "Abstract": Part−Of−Speech OR "Abstract": Tatoeba
        OR "Abstract": sentiment analysis OR "Abstract": textual entailment
        OR "Abstract": Hate Speech Detection OR "Abstract": Question Answering)
)
```

SLR Q3 query:

```
(
    ("Abstract":BERT)
    AND ("Abstract":Question Answering OR "Abstract":Textual Entailment
        OR "Abstract":Natural Language Inference OR "Abstract":NLI
        OR "Abstract":Sentiment Analysis OR "Abstract":Hate Speech Detection)
    AND ("Abstract":Italian OR "Abstract":Bulgarian OR "Abstract":Dutch
        OR "Abstract":Russian OR "Abstract":Chinese
)
```

The database queries were specifically designed to retrieve studies related to BERT and its multilingual version (mBERT), as these are the models used in this thesis. As a result, BERT variants such as RoBERTa, ALBERT, or DistilBERT were not explicitly included in the search criteria, to ensure alignment between the literature reviewed and the model architecture used in the experiments. Future work could expand the queries to include additional transformer-based models for a broader comparative analysis.

For SLR Question 3, not all target tasks were included in the query. For Part-Of-Speech tagging, Named Entity Recognition and Sentence alignment we will rely on datasets from the XTREME benchmark [47], which already covers the five languages considered in this study, namely Bulgarian, Chinese, Dutch, Italian and Russian. The choice of these five languages was guided by two main criteria: the linguistic diversity, aiming to include languages from different language families to better assess cross-lingual generalization; and the dataset availability, as these languages are among those consistently supported across multiple tasks in the XTREME benchmark.

Table 1 presents the number of papers retrieved from each database for the first three research questions. In total, we obtained 29 papers for the first research question, 230 for the second and 171 for the third after removing duplicates.

Database	SLR Q1	SLR Q2	SLR Q3
ACM Library	18	32	25
Web Of Science	4	142	96
IEEE xplore	7	83	62
Total (without duplicates)	29	230	171

Table 1: Number of papers found after the database search on the three considered databases for the first three SLR research questions.

3.3 Screening Using Active Learning

The third step in the SYMBALS pipeline involves active learning, a machine learning method where the algorithm selects the most informative data points to learn from [132]. This approach is particularly effective for systematic literature reviews, as it allows the model to achieve high accuracy with fewer training samples by prioritizing the most relevant papers [114]. By using active learning, researchers can significantly reduce the number of papers they need to manually review, which is especially valuable when dealing with a large number of related studies.

To implement this, we used ASReview [130], an open-source machine learning tool designed for screening and labeling large amounts of data. The screening process for the first three research questions began by exporting the papers retrieved during the database search, and organizing them using Mendeley[4], a reference manager software. Since ASReview requires initial labeled examples for classification, we manually labeled five to ten papers as relevant or not relevant for each query to establish prior knowledge. A full list of these prior-labeled papers is provided in Appendix B. For model training, we employed the default ASReview configuration, as it has been shown to outperform alternative configurations on multiple datasets [37]. The configuration includes: **TF-IDF** as feature extraction technique, **Naïve Bayes** classifier as the machine learning model, **Maximum** as the query strategy and **Dynamic Resampling (double)** as the balance strategy.
Once the model was trained, we screened the papers based on their titles and abstracts to confirm their relevance to our research. A stopping criterion was applied to determine when to stop the screening process: for the first two SLR questions, screening stopped after finding ten consecutive non-relevant papers. For the third SLR question, this criterion proved insufficient, as the query was more specific, resulting in a higher proportion of relevant papers. Had we followed the same stopping rule, we would have needed to review all retrieved papers, as it was unlikely to encounter ten consecutive irrelevant ones. Instead, we set a fixed limit of 100 reviewed papers for this question. Table 2 summarizes the number of papers obtained after the active learning screening phase.

Number of Papers	SLR Q1	SLR Q2	SLR Q3
Labeled papers during screening	31	59	100
Relevant papers after screening	10	11	72

Table 2: Number of papers found after active learning for the three SLR research questions.

Challenges for SLR Question 3 During this phase, we encountered a major challenge related to the third SLR question: the results obtained after active learning lacked diversity. The majority of relevant papers identified by the ASReview model focused on sentiment analysis, particularly for the Chinese language. This created an imbalance, with a lack of studies on other target languages, particularly for Bulgarian, which has significantly fewer NLP resources compared to Chinese. It is important to note that this bias was already present in the initial literature retrieved from the databases, where the majority of available studies concerned Chinese NLP tasks, especially sentiment analysis. ASReview subsequently amplified this bias, as the prior knowledge provided to the model predominantly consisted of studies focused on Chinese sentiment analysis.
Figure 5 shows the distribution of relevant papers across task-language combinations at this stage, with more than 80% of papers focusing on Chinese NLP. To address this issue, we have modified the search query (see Appendix A) and retrained the ASReview model, carefully selecting a more diverse set of papers as prior knowledge. Indeed the studies selected as prior knowledge for the original SLR Q3 query mostly focused on sentiment analysis, which have contributed to bias the model towards this category. However, even with these modifications, the issue persisted, as the model continued prioritizing studies on Chinese sentiment analysis. At this point, it became evident that while the SYMBALS approach effectively structures the

[4]`https://www.mendeley.com/reference-management/reference-manager`

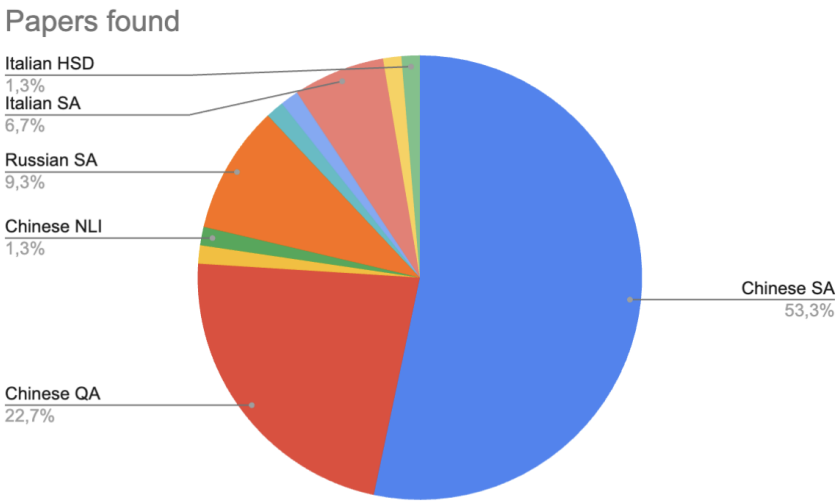

Figure 5: Pie chart depicting the proportion of papers found during screening for different task-language combinations in SLR Q3. The covered datasets include: Sentiment Analysis (SA), Question Answering (QA), Hate Speech Detection (HSD) and Natural Language Inference (NLI).

review process, it alone cannot fully counteract the underlying biases present in the available literature—particularly when the initial seed papers used to train the ASReview model are themselves unbalanced. In retrospect, selecting a more diverse and representative set of seed papers might have mitigated the model's tendency to prioritize studies on Chinese sentiment analysis. However, due to the dominance of such studies in the initial search results, identifying a sufficiently varied set of seed papers would have required additional manual screening—precisely the kind of effort that SYMBALS and active learning aim to reduce. This underscores a key trade-off in using automated review tools: while they enhance efficiency, their performance is still sensitive to the quality and diversity of the initial prior knowledge.

Alternative strategy for SLR Question 3 To attenuate this limitation, we took the following steps. First, we analyzed the relevant papers identified during active learning and selected only those that used publicly available datasets. Since the dataset list was still incomplete, we manually reviewed the first 100 papers outputted by ASReview after applying the modified query. This helped identify additional relevant studies. To mitigate this, we reviewed the first 100 papers outputted by the ASReview model using the modified query for the third SLR question, identifying additional relevant papers. Although this step expanded the list of relevant studies, the dataset selection remained biased toward sentiment analysis and question answering in Chinese and Russian, leaving other tasks and languages underrepresented. To address this issue, we conducted a targeted search for each language-task pair that lacked sufficient datasets. The final dataset selection results are presented in Appendix D, which comprises tables listing the datasets obtained for sentiment analysis, question answering, hate speech detection, and natural language inference across the target languages.

3.4 Backward Snowballing

As previously mentioned, the SYMBALS methodology combines active learning with backward snowballing, to identify additional relevant studies by examining the references of relevant papers found after the active learning phase. This step helps expand the pool of relevant literature, but since it increases the number of papers to review, it requires an appropriate stopping criterion. In SYMBALS, the proposed stopping criterion is based on three parameters: N_r, which is the number of most recent references checked; r_r, representing the number of newly identified relevant papers within N_r references; and S, which is the minimum number of snowballed papers required before stopping. The screening process stops when, in the last N_r references, the number of new relevant additions r_r is less than some constant C, provided that at least S papers have already been snowballed [132]. For this literature review, we set the parameters as follows: we have snowballed at least $S = 3$ papers and then we stopped when, in the last $N_r = 50$ references, fewer than 5 relevant papers were found.
All relevant studies identified through this process are reported in the Appendix C. As previously mentioned, given the limited number of relevant studies retrieved, we decided not to proceed with the quality assessment step of the SYMBALS methodology, as the dataset remained manageable without further filtering.

3.5 Related Work

Developing effective NLP systems for low-resource languages remains a difficult challenge due to the lack of annotated data, linguistic tools, and language-specific pre-trained models. This section synthesizes the findings from our Systematic Literature Review, which aimed to answer two core questions: To what extent have translation-based approaches been used to apply English BERT models to non-English or low-resource languages? What are the main challenges in applying BERT-based models to such languages?

3.5.1 Challenges in NLP for Low-Resource Languages (SLR Q2)

Numerous studies have highlighted the limitations that low-resource languages face in NLP development. As observed by Joshi et al. [52], there is an evident disparity between high-resource and low-resource languages, with many languages having little to no digital presence. This issue extends beyond rarely spoken tongues: even widely spoken languages such as Spanish face data scarcity in specific domains [17]. Similarly, morphologically rich or structurally complex languages, such as Turkish [109], Thai [67, 56], and Amharic [149], present significant challenges for cross-lingual transfer and pretraining.
A major obstacle is the lack of annotated datasets for supervised learning, especially in specialized domains. For instance, in biomedical NLP, languages like Italian and Spanish remain under-resourced despite their widespread use [14, 16]. In other domains such as cultural heritage, fake review detection [19], and hate speech detection [4, 92], similar challenges are evident. Sentiment analysis and NLI tasks are also hindered by the high cost of manually compiling lexicons [26], which leaves many low-resource languages without essential linguistic resources.
While multilingual models such as mBERT offer some relief, their performance is often suboptimal for truly underrepresented languages. Wu and Dredze [144] showed that mBERT's performance declines sharply for languages with minimal training data. Even monolingual models built for these languages often fail to reach competitive accuracy without extensive

pretraining or domain-specific corpora [72]. The situation is further complicated in tasks like intent detection and slot filling, where data labeling remains both expensive and inconsistent [112]. Notably, even high-resource languages like Russian remain disadvantaged when models are optimized for English-centric domains, such as social media sentiment analysis [65].
In summary, the main challenges in applying BERT-based models to low-resource languages include insufficient labeled data, underdeveloped domain-specific resources, limitations in cross-lingual transfer, and high costs of linguistic resource development. These barriers collectively motivate the exploration of translation-based approaches as a more scalable alternative.

3.5.2 Translation-Based Approaches to Extend English BERT (SLR Q1)

To address these challenges, an increasing number of studies have explored the potential of using machine translation to bridge the resource gap. A key strategy involves translating datasets from low-resource languages into English, enabling the use of powerful English-language models such as BERT. This method has shown encouraging results across several NLP tasks, suggesting it may serve as a practical alternative to developing language-specific models.
In sentiment analysis, for instance, Balahur et al. [10] found that machine-translated corpora yielded accuracy within 8% of native-language models, demonstrating the viability of this approach. Similarly, Demirtas et al. [26] reported that translation noise had minimal effect on polarity classification. Refaee and Rieser [106] further showed that translated Arabic tweets, when analyzed using English sentiment tools, could outperform native baselines, illustrating how English-centric resources can be extended to support underrepresented languages.
In domain-specific contexts, translation has proven particularly effective. Borchert et al. [14] and Buonocore et al. [16] successfully translated English biomedical corpora into French, Spanish, Dutch, and Italian, leading to significant performance gains over native-language training. Gallego et al. [40] also leveraged translation to expand concept recognition across multiple languages without requiring costly manual annotation.
Moreover, MT has been used not only to generate training data but also to enhance semantic diversity through auxiliary inputs. Amplayo et al. [6] demonstrated that using multiple translations of the same input can enrich model training, while Sohn et al. [119] showed that transformer models can compensate for moderate translation errors through contextual embeddings. Pamungkas et al. [92] similarly observed gains when combining MT with fine-tuned transformers for hate speech detection. Hybrid and ensemble methods further underscore this trend: Miah et al. [79] achieved high accuracy (86%) by combining translation-based preprocessing with a fusion of transformers and large language models.
However, not all studies report consistent success. The effectiveness of translation-based approaches is highly dependent on task complexity and translation quality. For example, while Balahur et al. [10] found translation useful for basic sentiment tasks, they also noted performance drops in traditional models like SVMs when translation noise affected feature quality. In code-mixed and morphologically rich languages, the semantic integrity of texts often suffers during translation. Pravalika et al. [100] highlighted how syntactic mismatches degrade sentiment analysis outcomes in such settings.
In more structure-sensitive tasks like question answering, translation introduces alignment issues. Canete et al. [17] reported that nearly half of the examples in the MLQA dataset had mismatched answer spans due to poor translation, undermining model evaluation. Similarly, Schuster et al. [112] found that translated training data was less effective than multilingual embeddings or contextual encoders when limited target-language data was available. Yamaguchi

et al. [148] confirmed this finding in prompt-based models, noting that translation-induced noise reduced performance for languages like Swahili and Japanese.
Generative and lexical alignment tasks also face translation issues. Soni et al. [120] found that translating QA blueprints introduced errors that hindered model training. Meng et al. [77] observed that bilingual sentiment lexicons created via MT had limited coverage and ambiguity, reducing their effectiveness. In biomedical NLP, Dorendahl et al. [30] reported suboptimal results when using English tools like MetaMap on German texts translated with MT. Even dataset augmentation via translated corpora, as shown by Demirtas et al. [26], does not always lead to performance gains due to corpus divergence and translation inaccuracies.
Taken together, these findings suggest that while translation-based approaches offer a promising and resource-efficient alternative for extending English BERT to low-resource languages, their effectiveness is task- and language-dependent. This thesis builds on this line of work by systematically evaluating translation-driven BERT fine-tuning across multiple NLP tasks and linguistic families, seeking to determine when and how such methods can match or exceed the performance of native-language models.

3.6 Conclusion

The Systematic Literature Review conducted for this thesis provides a foundational understanding of the research landscape surrounding translation-based NLP strategies and the application of BERT to low-resource languages. Employing the SYMBALS methodology allows for a thorough and efficient retrieval of relevant studies, combining the precision of active learning with the extensive coverage of backward snowballing. A notable challenge during the review process was the limited diversity in studies retrieved for the third SLR question, where an overwhelming focus on Chinese sentiment analysis created a skew in the literature pool. This was partly a reflection of the field's current research bias, and partly a side-effect of the active learning model reinforcing dominant patterns. To counter this, we adapted our approach by modifying search queries and manually diversifying the dataset list. Despite these challenges, the insights gained through the SLR played a crucial role in shaping the direction of this thesis. They informed the selection of languages and tasks used in the experimental design and highlighted key considerations, such as typological proximity and task sensitivity, that became central to our evaluation framework. Ultimately, this phase not only deepened our understanding of current research trends but also clarified the methodological and practical gaps that this thesis aims to address.

4 Methodology

This section presents the experimental framework adopted in this thesis, corresponding to Business Understanding phase of the CRISP-DM methodology. In particular, we first introduce the models used in our experiments, then we describe the hyperparameter tuning process used to optimize model performance, followed by details on fine-tuning and the evaluation metrics applied.

4.1 Models

This section outlines the models used in our experiments. We first describe the translation system used to convert non-English datasets into English. Then, we detail the English BERT model employed as a core component of our translation-based approach. Finally, we present the language-specific BERT models used to fine-tune directly on the original non-English datasets, which serve as baselines for comparison.

4.1.1 Translation Model

To translate non-English datasets into English, we employ the **Helsinki-NLP/opus-mt** models from the OPUS-MT project [126], an initiative focused on developing accessible and high-quality machine translation tools, especially for low-resource and minority languages. The models are built using the Marian-NMT framework, a production-ready neural machine translation system optimized for efficiency and scalability. Architecturally, OPUS-MT models adopt a standard Transformer setup, consisting of 6 encoder and 6 decoder layers with 8 attention heads each. They are trained on large-scale parallel corpora sourced from the OPUS bitext repository, which provides diverse and multilingual text data. OPUS-MT models offer competitive translation quality in the open-source landscape, often achieving performance comparable to state-of-the-art commercial systems for many language pairs [88]. Although commercial systems such as DeepL[5] are frequently reported to outperform open-source alternatives in terms of translation fluency and contextual accuracy, we chose OPUS-MT for this study due to its open-source availability, ease of integration, and ability to run locally without API limitations or cost. These qualities make it a practical and reproducible choice for this research.

4.1.2 English BERT

To perform hyperparameter tuning and then fine-tuning on the translated datasets, we employ the pre-trained BERT model, specifically **BERT-Base**. This variant consists of 12 transformer layers and 110 million parameters and was pre-trained on a large English corpus. Although the larger *BERT-Large* model, with 24 transformer layers and 340 million parameters, is able to capture more complex contextual information, we opt for BERT-Base due to its lower computational requirements and strong performance on a wide range of tasks [75]. BERT comes in two variations: *cased* and *uncased*. The cased model is sensitive to letter casing (e.g., distinguishing between "Dutch" and "dutch"), while the uncased model is not. Depending on the requirements of each task, we selected the most appropriate BERT variant.

[5] `https://www.deepl.com/nl/translator`

4.1.3 Non-English BERT Models

To validate the effectiveness of the translation-based approach, we run experiments also with the original, non-translated datasets using language-specific BERT models. These serve as baselines to assess whether translation and English fine-tuning offer performance advantages over native processing. We selected publicly available models from Hugging Face for each language, ensuring comparability in size and training architecture. Below is a breakdown of the models chosen for each language:

- **Bulgarian**: we use *bert-web-bg*, a cased model with 109M parameters developed by Marinova et al.[70], that is shown to outperform *bert-base-bg*, another cased model that we found on Hugging face. For the uncased setting, we employ *AIaLT-IICT/bert_bg_lit_web_base_uncased*, which closely follows the original BERT architecture and objective.

- **Dutch**: we use *GroNLP/BERTje*, a standard BERT base architecture trained from scratch on Dutch corpora. It represents the most widely used and validated BERT variant for Dutch, comparable in size and training setup to bert-base-cased. De Vries et al. showed that BERTje was able to outperform the multilingual BERT model on several downstream NLP tasks, including part-of-speech tagging, named-entity recognition, semantic role labeling, and sentiment analysis [24]. Due to the absence of a modern uncased BERT model for Dutch, and because alternatives like GysBERT are based on historical texts, we used BERTje for both cased and uncased scenarios to maintain consistency.

- **Italian**: we use models developed by the *dbmdz* team, which offers standard BERT base architecture pretrained from scratch on large Italian corpora including Wikipedia and OPUS. Both the cased (*bert-base-italian-cased*) and uncased (*bert-base-italian-uncased*) models were proved to outperform both Multilingual BERT (M-BERT) and XLM-RoBERTa [113].

- **Chinese**: we employ *bert-base-chinese*, the official Google BERT model trained using WordPiece tokenization. Since Chinese does not use case distinctions, the same model was used in both cased and uncased configurations.

- **Russian**: for the cased model, we use *DeepPavlov/rubert-base-cased*, a widely adopted Russian BERT variant pretrained on Russian Wikipedia and news texts. This model was shown to outperform M-BERT on several NLP tasks [58]. For the uncased setting, we selected *deepvk/bert-base-uncased*, one of the few available uncased models for Russian.

A summary of the selected models is provided in the following table:

4.2 Hyperparameter Tuning

Hyperparameter tuning plays a crucial role in improving model performance by identifying the optimal configuration of parameters set before training begins[22].
To automate this process, we employ Optuna[3], an efficient hyperparameter optimization framework based on Bayesian optimization. For each dataset, we defined a search space over key hyperparameters, including learning rate, batch size, number of training epochs and weight decay. During each optimization trial, Optuna samples a set of hyperparameters, trains the

Language	Cased	Uncased
Bulgarian	bert-web-bg [6]	bert_bg_lit_web_base_uncased [7]
Chinese	bert-base-chinese [8]	bert-base-chinese [9]
Dutch	bert-base-dutch-cased [10]	bert-base-dutch-cased [11]
Italian	bert-base-italian-cased [12]	bert-base-italian-uncased [13]
Russian	rubert-base-cased [14]	bert-base-uncased [15]

Table 3: Hugging Face BERT models employed on the original, non-translated datasets.

model, and evaluates its F1 score on a validation set. This score is used as the objective metric. Optuna then refines its sampling strategy based on past trials, balancing exploration with exploitation.
Hyperparameter tuning is performed for question answering, hate speech detection, POS tagging and NER, while the NLI and sentiment analysis datasets are only fine-tuned. Even though this may not guarantee a fair comparison across tasks, we deemed it necessary due to time limits, especially since the datasets found for sentiment analysis and NLI are among the biggest among the considered data.
We perform 10 trials per task, selecting the configuration that yields the highest validation F1 score. This approach enables a consistent and reproducible tuning process across languages and tasks.

4.3 Fine-tuning

We fine-tune all BERT-based models using **Hugging Face's Trainer API**, which simplifies the process of training and evaluating transformer-based models. The *Trainer* class automates key steps, including forward passes, backpropagation, optimization, evaluation, and checkpoint management. This training configuration allow us to maintain a clean and modular training pipeline while ensuring reproducibility and consistency across experiments.

4.4 Evaluation metrics

We evaluate model performance using standard classification metrics widely adopted in NLP research:

- **Accuracy**: it represents the proportion of correct predictions over all predictions.
- **Recall**: also known as True Positive Rate, it indicates the proportion of actual positive instances correctly identified.
- **Precision**: it measures the proportion of predicted positive instances that are actually positive.
- **F1-score**: this is the harmonic mean of precision and recall, balancing both metrics in a single value.
- **Exact Match (EM)**: this metric, used for extractive question answering, measures the percentage of predictions that exactly match the ground truth answer string[59].

4.5 Experimental Settings

This section outlines the computational resources and general experimental configuration used throughout this study. The setup was chosen to ensure reproducibility, scalability, and efficient handling of multilingual NLP tasks involving large pre-trained transformer models.
All experiments in this study are conducted using a combination of local and high-performance computing resources. For translation tasks, we make use of the resources of the Data Science Lab of LIACS, equipped with 256 GB of RAM, 24 Intel Xeon Silver 4214 cores, and two NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory each. For more compute-intensive tasks, including hyperparameter tuning and fine-tuning of transformer models, we rely on the Dutch national supercomputer Snellius, which offers access to AMD CPUs and GPGPU accelerators. Model training is implemented using the Hugging Face Trainer API, which allows for streamlined fine-tuning and evaluation across tasks. Hyperparameter tuning is performed using the Optuna framework with Bayesian optimization. For most tasks, this includes up to 10 optimization trials; however, for sentiment analysis and natural language inference, we employ a fixed set of hyperparameters to reduce computational cost, with a learning rate of $5e-5$, a batch size of 16, weight decay of 0.01, and 3 training epochs. All experiments use the AdamW optimizer, as implemented in the Trainer API. Because of its popularity, we employ *NLTK* to handle preprocessing. Only later did we realize that NLTK does not support Bulgarian, so for that language we use *Stanza*. We did not switch all preprocessing to Stanza because NLTK was already included in the experimental pipeline of the translated datasets. All models employ Hugging Face tokenizers and custom PyTorch Dataset classes for input preparation. Training, validation and test sets are created using an 80/10/10 split. To ensure reproducibility across all experiments, the random seed is fixed to 42.

5 Data Understanding and Preparation

Following the CRISP-DM methodology, we proceed with the data understanding and preparation phase. This stage is crucial for ensuring that the data used in our experiments is well-suited for the research objectives. First, we select and describe the datasets for each task-language pair and then we apply preprocessing techniques to prepare the data for model training and evaluation. Finally, we conduct an Exploratory Data Analysis (EDA) to gain insights into the structure, distribution and additional characteristics of the datasets. This step helps to identify potential challenges such as class imbalances, missing data, or inconsistencies that may impact model performance.

5.1 Dataset Selection and Description

In this section, we provide an overview of the datasets identified during the literature review and justify the selection of the most appropriate ones for each task-language pair considered in this study. The dataset selection process is based on multiple factors, including dataset size, structure, and availability of information about the dataset creation and intended use.
During systematic literature review, we found the XTREME benchmark [47], which is characterized by POS tagging and NER datasets for all the five languages that we are considering in this thesis. Therefore, for these two tasks, we have chosen datasets belonging to such benchmark.

5.1.1 Sentiment Analysis

We identify and evaluate publicly available sentiment analysis datasets for each language. The selection is based on several criteria, among which the availability of clear sentiment polarity labels (positive, negative, and optionally neutral); general-domain applicability (i.e., avoiding domain-specific or emotion-focused datasets); sufficient documentation about dataset creation. Additionally, when multiple datasets satisfy the criteria, the largest is selected unless deemed computationally impractical. Table 24 in Appendix C provides a full list of the considered datasets. Table 4 includes representative examples of selected data samples. Below, we briefly describe the selected datasets for each language.

Bulgarian We select the *Cinexio Movie Reviews* dataset [16], which belongs to the bgGLUE benchmark [55], and which contains 9,827 movie reviews labeled as positive, neutral, or negative. It is preferred over alternatives due to its clearer documentation and well-defined structure. This dataset will be referred to as **Cinexio** in the remainder of the thesis.

Dutch We choose the *Dutch Book Reviews Dataset (**DBRD**) v3.0* [17] [131], which provides 21,895 labeled book reviews for binary sentiment classification. It is selected over the Dutch Sentiment Analysis dataset due to its larger size and inclusion in the *DUMB* benchmark.

[16] https://bgglue.github.io/tasks/task_info/cinexio/
[17] https://github.com/benjaminvdb/DBRD/tree/master?tab=readme-ov-file

Dataset	Review	English Translation	Label
Cinexio	Това беше един от най-яките филми!!!	It was one of the coolest movies !!!	Positive
DBRD	Dit is vast een impopulaire mening maar jeetje wat een saai boek! Een hoop irritaties... De schrijfstijl, de personages. Ik dacht de hele tijd mens doe iets, kom voor jezelf op! Het eind was gelukkig wel oké en maakte het eea goed maar nee wat mij betreft geen aanrader.	This is probably an unpopular opinion but gosh what a boring book! A lot of irritations... The writing style, the characters. I kept thinking woman do something, stand up for yourself! The ending was fortunately okay and made up for it but no, as far as I'm concerned, not recommended.	Negative
Weibo Senti 100k	好样儿的！严重的支持！[顶]@李小孩儿_生如夏花: [鼓掌][赞]	Attaboy! Serious support! Lee Child: [applause] [applause]	Positive
Italian Tweets Dataset	Volevo anche segnalare che OfficialRadja in tutto ciò ha anche sospeso il suo profilo Instagram. #Nainggolan	I also wanted to point out that OfficialRadja has also suspended his Instagram profile in all of this. #Nainggolan	Neutral
RuReviews	Заказ не пришёл, жду возврат средств	The order didn't come, I'm waiting for the money back.	Negative

Table 4: Sentiment Analysis Datasets Examples.

Italian The **Italian Tweets Dataset** [18] [69] is selected, comprising 165,815 tweets labeled using AWS Comprehend API. It is the only dataset meeting the labeling and general-domain criteria, as others are either emotion-based or focused on aspect-level sentiment. Tweets are classified into four categories: positive, negative, neutral or mixed.

Chinese We select **Weibo Senti 100k** [123] [19], containing 119,988 labeled posts from Sina Weibo with balanced binary sentiment classes. It is the most suitable in terms of size and label clarity, while others are excluded for domain specificity or emotion-based labels.

Russian The **RuReviews** dataset [20] [118] is selected and it contains 180,000 product reviews derived from a major Russian e-commerce site. It strikes a balance between size and general applicability. Larger datasets (e.g., *RuTweetCorp*) are avoided due to computational constraints.

[18] https://github.com/charlesmalafosse/open-dataset-for-sentiment-analysis/tree/master

[19] https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb

[20] https://github.com/sismetanin/rureviews/blob/master/women-clothing-accessories.3-class.balanced.csv

5.1.2 Question Answering

We select one question answering dataset per target language, prioritizing datasets with an open-ended question format and a clear structure that includes the question, context, and answer. Datasets based on translations or without full context information are excluded to maintain consistency and quality. When multiple datasets satisfy the criteria, the largest suitable one is selected unless impractical to use.
Representative samples for each dataset are reported in Table 5, and a full list of the datasets found appears in Table 25.

Bulgarian : We select the *Multi-Subject High School Examinations*, or **EXAMS** dataset, from the bgGLUE benchmark [21] [45], which contains 3,349 QA pairs from high school exams. Although originally formatted as multiple-choice, we adapt it to an open-ended format by extracting the correct answer and its context. Other candidates are excluded due to pre-translation or incompatible formats.

Dutch : The Vraag-en-antwoord dataset Rijksportaal Personeel (**P-Direkt** dataset for short) [22] [134] is selected, comprising 323 real-world QA examples from a government contact center. It includes question, answer, and context fields and follows a structure similar to SQuAD. Other datasets lack context or are created via translation.

Italian : We select **QA-ITA-200k**, which includes 202,471 QA pairs primarily sourced from Wikipedia. It is the largest suitable Italian QA dataset found, with well-structured triplets and comprehensive coverage.

Chinese : The **CMRC2018** dataset [23] [41] is chosen, consisting of 14,363 QA pairs annotated on Wikipedia paragraphs. It is selected over others due to its span-extraction format and inclusion of both context and manually curated answers, whereas alternatives lack context or are inaccessible.

Russian : We select **SberQuAD** [24] [33], a reading comprehension dataset with 74,300 QA entries. It includes both answerable and unanswerable questions based on Wikipedia texts. It is preferred over smaller alternatives like *RuBQ* and *XQuAD*.

Dataset	QA Example
EXAMS	**Context:** Хемоглобин. Хемоглобинът или хемоглобулинът е белтък с четвъртична структура и молекулна маса около 66000–68000 Da. ... **Question:** Кое от изброените химични съединения е белтък? **Answer:** хемоглобин

Continued on next page...

[21] https://bgglue.github.io/tasks/task_info/exams/
[22] https://data.overheid.nl/dataset/vraag-en-antwoord-dataset-rijksportaal-personeel
[23] https://github.com/ymcui/cmrc2018
[24] https://huggingface.co/datasets/kuznetsoffandrey/sberquad

Dataset	QA Example
EXAMS translated	**Context:** Hemoglobin or hemoglobin is a protein with a fourth structure and a molecular mass of about 66000–68000 Da. ... **Question:** Which of the listed chemical compounds is protein? **Answer:** Haemoglobin
CMRC2018	**Context:** 广茂铁路是中国广东省一条起自广州市广州西站，向西跨越北江、西江，经佛山、三水、肇庆、云浮、阳江至茂名市茂名站的铁路，全长364.6公里， **Question:** 广茂铁路全长多少公里? **Answer:** 364.6公里
CMRC2018 translated	**Context:** The Guangmao Railway, which runs 364.6 kilometres from Guangdong Province, China's Guangzhou West Station, ... **Question:** How many kilometers is the Hiroshima Railway? **Answer:** 364.6 kilometres
P-Direkt	**Context:** Deelname aan de PAS-regeling heeft geen gevolgen voor uw wettelijke vakantie-aanspraak. **Question:** Wat zijn de gevolgen voor mijn vakantie-uren als ik gebruik maak van de PAS-regeling? **Answer:** Deelname aan de PAS-regeling heeft geen gevolgen voor uw wettelijke vakantie-aanspraak
P-Direkt translated	**Context:** Participation in the PAS scheme has no consequences for your statutory holiday entitlement. **Question:** What are the consequences for my holiday hours if I use the PAS scheme? **Answer:** Participation in the PAS scheme has no consequences for your statutory holiday entitlement
QA-ITA-200k	**Context:** Alien - La clonazione Alien - La clonazione (Alien Resurrection) è un film del 1997 diretto da Jean-Pierre Jeunet. Dopo un'anteprima a Parigi il 6 novembre 1997 ... **Question:** Chi è il regista del film 'Alien Resurrection'? **Answer:** Il regista del film 'Alien Resurrection' è Jean-Pierre Jeunet.
QA-ITA-200k translated	**Context:** Alien - Cloning Alien - Cloning (Alien Resurrection) is a 1997 film directed by Jean-Pierre Jeunet. After a preview in Paris on November 6, 1997 ... **Question:** Who is the director of the film 'Alien Resurrection'?

Continued on next page...

Dataset	QA Example
	Answer: The director of the film 'Alien Resurrection' is Jean-Pierre Jeunet.
SberQuAD	**Context:** Троллейбусы используются преимущественно в городах, но также существуют междугородные и пригородные троллейбусы ... **Question:** Где преимущественно используют троллейбусы? **Answer:** Троллейбусы используются преимущественно в городах
SberQuAD translated	**Context:** Trolleybuses are mainly used in urban areas, but there are also long-distance and suburban trolleybuses ... **Question:** Where are the trolley buses mostly used? **Answer:** Trolleybuses are mostly used in cities

Table 5: Question Answering Datasets Examples.

5.1.3 Hate Speech Detection

We select one hate speech detection dataset per target language, prioritizing datasets that include the full text of labeled content. Datasets that only provide content IDs (e.g., Tweet IDs) are excluded due to API limitations and the inability to retrieve text at scale. Among datasets with accessible content, we select the largest suitable dataset, unless major quality concerns existed.
All datasets that we found during the SLR are listed in Table 26, and representative examples are shown in Table 6.

Bulgarian: We select the *Hate Speech Classification in Bulgarian* dataset [25] [104], which includes over 100,000 manually annotated comments from Bulgarian online forums. It is chosen over smaller alternatives due to its size and accessible text. For simplicity, this dataset will be referred to as **HS-bg** in the remainder of this thesis.

Dutch: The **Dutch HateCheck** dataset [26] [48], which is part of the *Multilingual HateCheck (MHC)* suite [107], is selected. It contains 3,765 manually designed test cases created by native speakers. Other Dutch datasets are excluded as they provide only Tweet IDs without full text access.

Chinese: We select the *Chinese Offensive Language Dataset (**COLD**)* [27] [27], which contains 37,480 annotated comments labeled for offensive language relating to race, gender, and region. It is preferred over larger alternatives like *TOCAB* due to concerns over low-content entries (e.g., single-word comments).

[25] http://www.pfeffer.at/data/bulgarian/
[26] https://huggingface.co/datasets/Paul/hatecheck-dutch
[27] https://github.com/thu-coai/COLDataset/tree/main/COLDataset

Italian: We select the *Multilingual Hate Speech Dataset* [28] [83], which includes 6,839 Italian entries translated from Arabic. While the translation-based origin may introduce bias, it is the only viable dataset with fully accessible content. This dataset will be referred to as **HS-it** in the remainder of this thesis.

Russian : The **Russian South Park** [29] [108] dataset is chosen, with 15,875 annotated examples derived from television transcripts. It is selected based on its size and availability of text, whereas other datasets are excluded for small scale or misaligned objectives.

Dataset	Review	English Translation	Label
HS-bg	за мен вера си остава боклук	For me, vera is still trash	Hateful
COLD	任天堂新主机Nintendo Switch 是否有潜力成为成功的主机?	Does Nintendo Switch have the potential to be a successful host?	Non-hateful
Dutch HateCheck	Zwermen transgenders infesteren onze regering.	Transgenders are infesting our government.	Hateful
HS-it	@user I soliti immigrati che o rubano o diversamente delinquono,ma per i nostri governanti va bene così'.	@user The usual immigrants who either steal or otherwise delinquote, but for our rulers it's okay.	Hateful
South Park	Молодец, непросто было доказывать свою правоту в такой ситуации.	Well done, it wasn't easy to prove yourself right in a situation like this.	Non-hateful

Table 6: Hate Speech Detection Datasets Examples.

5.1.4 Natural Language Inference (NLI)

For NLI, we prioritize datasets with a standard premise-hypothesis-label structure and select one dataset per language. When multiple options exist, we choose the largest suitable dataset, excluding translated versions if a higher-quality or benchmark-supported version is available.

Bulgarian, Chinese and Russian For these languages, we select the *XNLI* dataset [30], which is part of the *XTREME* benchmark [47]. *XNLI* is a professionally translated subset of *MultiNLI*, covering 14 languages. Each language subset contains 400,202 examples (see Table 10), ensuring consistency and comparability across languages. For the remainder of this thesis, we will refer to Bulgarian XNLI, Chinese XNLI and Russian XNLI as **XNLI-bg**, **XNLI-zh**, **XNLI-ru**, respectively.

[28] https://huggingface.co/datasets/ysenarath/moosa2022multilingual
[29] https://github.com/Sariellee/Russan-Hate-speech-Recognition
[30] https://huggingface.co/datasets/facebook/xnli

Italian : The SLR identified two Italian datasets: a manually translated version of *RTE-3* [116] and a machine-translated version of *LingNLI* [116]. Due to the limited size of the RTE-3 dataset, we select **LingNLI** [31], acknowledging the fact that it was machine translated to Italian.

Dutch : We select **SICK-NL** [32], used in the *DUMB* benchmark [25]. This dataset is a manually corrected translation of the original *SICK* dataset, ensuring high quality while preserving the premise-hypothesis-label format required for NLI.

Dataset	Premise	Hypothesis	Label
XNLI-bg	дейв хенсън , ти никога не си умрял !	дейв хенсън , ти умря в края на краищата .	Contradiction
	Dave Hanson, you're never dead!	Dave Hanson, you died after all.	
XNLI-zh	'变化即将到来.	什么都不会改变.	Contradiction
	'Change is coming.	Nothing's gonna change.	
SICK-NL	De man rent op de weg	Een hond rent op de weg	Neutral
	The man is running on the road	A dog is running on the road	
LingNLI	Potrebbe essere una strategia a lungo termine, ma Dole è molto indietro.	Dole è molto lontano dal 1 ° posto.	Entailment
	It could be a long-term strategy, but Dole is way behind.	Dole is very far from the 1st place.	
XNLI-ru	Ну конечно . Дэниэл посмотрел вокруг .	Дэниелс взгляд не был устойчивым .	Entailment
	Of course, Daniel looked around.	Daniels' eyes weren't steady.	

Table 7: NLI Datasets Examples.

5.1.5 POS tagging

We select the Universal Dependencies v2.5 treebanks [87] [33], used in the *XTREME* benchmark, where each word is assigned one of 17 universal POS tags. Among the datasets present for each language, we prioritize datasets of larger but also similar size.

[31] https://huggingface.co/datasets/maximoss/lingnli-multi-mt
[32] https://huggingface.co/datasets/maximedb/sick_nl
[33] https://universaldependencies.org/

Bulgarian : We select **UD_Bulgarian-BTB** [91], based on the HPSG-based BulTreeBank and composed by sentences mainly from Bulgarian newspapers, but also from fiction and administrative documents.

Chinese : As our traditional Chinese Universal Dependencies Treebank, we select **UD_Chinese-GSD** [1], annotated and converted by Google.

Dutch : We select **Alpino** [15], containing samples from various treebanks annotated at the University of Groningen using the Alpino annotation tools and guidelines.

Italian : We choose **PoSTWITA-UD** [110], a collection of Italian tweets annotated in Universal Dependencies.

Russian : We select **Taiga** [68], a Universal Dependencies treebank based on data samples extracted from *Taiga Corpus* and *MorphoRuEval-2017* and *GramEval-2020* shared tasks collections. It is characterized by sentences from several domains: blogs and social media, poetry, news and Wikipedia.

Dataset	POS Example
UD_Bulgarian-BTB	**Sentence:** Да не би да съм закъснял ? **Annotated sentence:** ('Да': AUX), ('не': PART), ('би': AUX), ('да': AUX), ('съм': AUX), ('закъснял': VERB), ('?': PUNCT) **Translated sentence:** Am I late ? **Annotated translated sentence:** ('Am': AUX), ('I': PRON), ('late': ADJ), ('?': PUNCT)
PoSTWITA-UD	**Sentence:** RT @user : Non esiste una strada verso la felicità . La felicità è la strada . **Annotated sentence:** ('RT': SYM), ('@user': SYM), (':': PUNCT), ('Non': ADV), ('esiste': VERB), ('una': DET), (strada: 'NOUN'), ('verso': ADP), ('la': DET), ('felicità': NOUN), ('.': PUNCT), ('La': DET), ('felicità': NOUN), ('è': AUX), ('la': DET), ('strada': NOUN), ('.': PUNCT) **Translated sentence:** RT @user : There is no way to happiness . Happiness is the way . **Annotated translated sentence:** ('RT': PROPN), ('@user': PROPN), (':': PUNCT), ('There': PRON), ('is': VERB), ('no': DET), ('way': NOUN), ('to': ADP), ('happiness': NOUN), ('.': PUNCT), ('Happiness': NOUN), ('is': AUX), ('the': DET), ('way': NOUN), ('.': PUNCT)
UD_Chinese-GSD	**Sentence:** 島嶼長度約為11 公里，寬度6 公里。 *Continued on next page...*

Dataset	POS Example
	Annotated sentence: ('島嶼': NOUN), ('長度': NOUN), ('約': ADV), ('為': AUX), ('11': NUM), ('公里': NOUN), (',': PUNCT) ('寬度': NOUN), ('6': NUM), ('公里': NOUN), ('。': PUNCT)
	Translated sentence: The island is about 11 km long and 6 km wide .
	Annotated translated sentence: ('The': DET), ('island': NOUN), ('is': AUX), ('about': ADV), ('11': NUM), ('km': NOUN), ('long': ADJ), ('and': CCONJ), ('6': NUM), ('km': NOUN), ('wide': ADJ), ('.': PUNCT)
Alpino	**Sentence:** Avondvluchten gingen wel redelijk op tijd weg .
	Annotated sentence: ('Avondvluchten': NOUN), ('gingen': VERB), ('wel': ADV), ('redelijk': ADJ), ('op': ADP), ('tijd': NOUN), ('weg': ADV), ('.': PUNCT)
	Translated sentence: Evening flights did leave fairly in time .
	Annotated translated sentence: ('Evening': NOUN), ('flights': NOUN), ('did': AUX), ('leave': VERB), ('fairly': ADV), ('in': ADP), ('time': NOUN), ('.': PUNCT)
Taiga	**Sentence:** ума ни в какие помышления "
	Annotated sentence: ('ума': NOUN), ('ни': PART), ('в': ADP), ('какие': DET), ('помышления': NOUN), ('"': PUNCT)
	Translated sentence: I'm not thinking about anything . "
	Annotated translated sentence: ('I': PRON), ("'m": AUX), ('not': PART), ('thinking': VERB), ('about': ADP), ('anything': PRON), ('.': PUNCT), ('"': PUNCT)

Table 8: POS Dataset Examples.

5.1.6 NER

Following the *XTREME* benchmark, we select the *Wikiann* dataset [34] [93], which contains named entities from Wikipedia automatically annotated in IOB2 format. This dataset is used for all five considered languages: Bulgarian (**wikiann-bg**), Chinese (**wikiann-zh**), Dutch (**wikiann-nl**), Italian (**wikiann-it**), Russian (**wikiann-ru**).

5.1.7 Datasets Dimensions

Table 10 compares the sizes of the selected datasets for each task-language pair. We can observe that the sizes vary significantly both across tasks and across languages. The Question Answering task shows large datasets for Italian and Russian, while Dutch has a notably small dataset in this task, with only 323 examples, which may affect model performance and generalization. Similarly, NLI datasets are large for Bulgarian, Chinese, and Russian (all over

[34] https://huggingface.co/datasets/unimelb-nlp/wikiann

Language	NER Example
wikiann-bg	**Tokens:** 'пренасочване', 'Мащеха', '(', 'теленовела', ')' **Translated Tokens:** 're-routing', 'Stepmother', '(', 'telenovela', ')' **Tags:** "O", "B-ORG", "I-ORG", "I-ORG", "I-ORG"
wikiann-zh	**Tokens:** 前 **Translated Tokens:** 'Front', 'Town.', 'Zone' **Tags:** "B-LOC", "I-LOC", "I-LOC"
wikiann-nl	**Tokens:** '2', 'etappes', 'in', 'Ronde', 'van', 'Frankrijk' **Translated Tokens:** '2', 'stages', 'in', 'Round', 'of', 'France' **Tags:** "O", "O", "O", "B-ORG", "I-ORG", "I-ORG"
wikiann-it	**Tokens:** "'", "'", 'Sandra', 'Cecchini', '(', 'campionessa', ')' **Translated Tokens:** "'", "'", 'Sandra', 'Snipers', '(', 'sample', ')' **Tags:** "O", "O", "B-PER", "I-PER", "O", "O", "O"
wikiann-ru	**Tokens:** 'Паро', '(', 'река', ')' **Translated Tokens:** 'Pair', '(', 'River', ')' **Tags:** "B-LOC", "I-LOC", "I-LOC", "I-LOC"

Table 9: NER Datasets Examples.

400k examples), but much smaller for Dutch and Italian. The Hate Speech task also shows wide variation: Bulgarian has over 100k samples, while Dutch and Italian have fewer than 7k. In contrast, the NER datasets are balanced across languages, with exactly 40k samples each, since part of the same benchmark. POS tagging datasets are relatively large and more comparable across languages. Finally, Sentiment Analysis datasets range broadly, from under 10k in Bulgarian to over 165k in Italian. These discrepancies in dataset sizes are important to consider, as they may impact the difficulty of the tasks across languages and the relative performance of models trained under different data availability conditions.

Task	Bulgarian	Chinese	Dutch	Italian	Russian
Sentiment Analysis	9,827	119,988	51,069	165,815	90,000
Question Answering	3,349	14,363	323	94,105	74,300
Hate Speech	102,750	37,480	3,765	6,839	15,875
NLI	400,202	400,202	9,840	34,878	400,202
POS	156,149	123,291	208,747	129,668	197,001
NER	40,000	40,000	40,000	40,000	40,000

Table 10: Number of rows in each dataset for different NLP tasks and languages.

5.2 Data Preprocessing

Text preprocessing is a crucial step in NLP that involves cleaning and transforming raw text data into a format suitable for analysis and machine learning models [2]. Effective preprocessing not only helps cleaning and standardizing the data, but also enhances the model's performance by removing noise, ensuring uniformity, and facilitating better generalization. The preprocessing pipeline adopted in this work is illustrated in Figure 6 and consists of the following key steps:

1. **Dataset splitting**: each dataset is divided into training (80%), validation (10%) and test (10%) subsets.
2. **Handling missing values**: any missing data, whether originally present or introduced during translation, is addressed by removing examples containing *NaN* values.
3. **Dropping unnecessary columns**: non-essential columns that are not required for BERT fine-tuning are removed.
4. **Text cleaning**: dataset-specific cleaning procedures are applied to the text fields to remove unwanted characters, artifacts, or formatting issues.
5. **Label adjustment**: labels are reviewed and converted into integer format if necessary, ensuring compatibility with model training.
6. **Class imbalance handling**: techniques are applied to mitigate skewed class distributions where appropriate.
7. **Tokenization**: text is tokenized in preparation for input into the BERT model.

After outlining the general preprocessing pipeline, we now describe in more detail some of its key steps. The following subsections provide task-specific insights into how data cleaning, label alignment, and other adjustments are handled to ensure appropriate inputs for fine-tuning.

5.2.1 Text Cleaning

For NER, we observe that translation often hallucinates by appending punctuation marks to translated words, that, however, are not present in the original text. This issue arises because the translation model is trained on full sentences, but for NER, we translate the text word-by-word, leading to potential translation artifacts. To address this, we remove such spurious

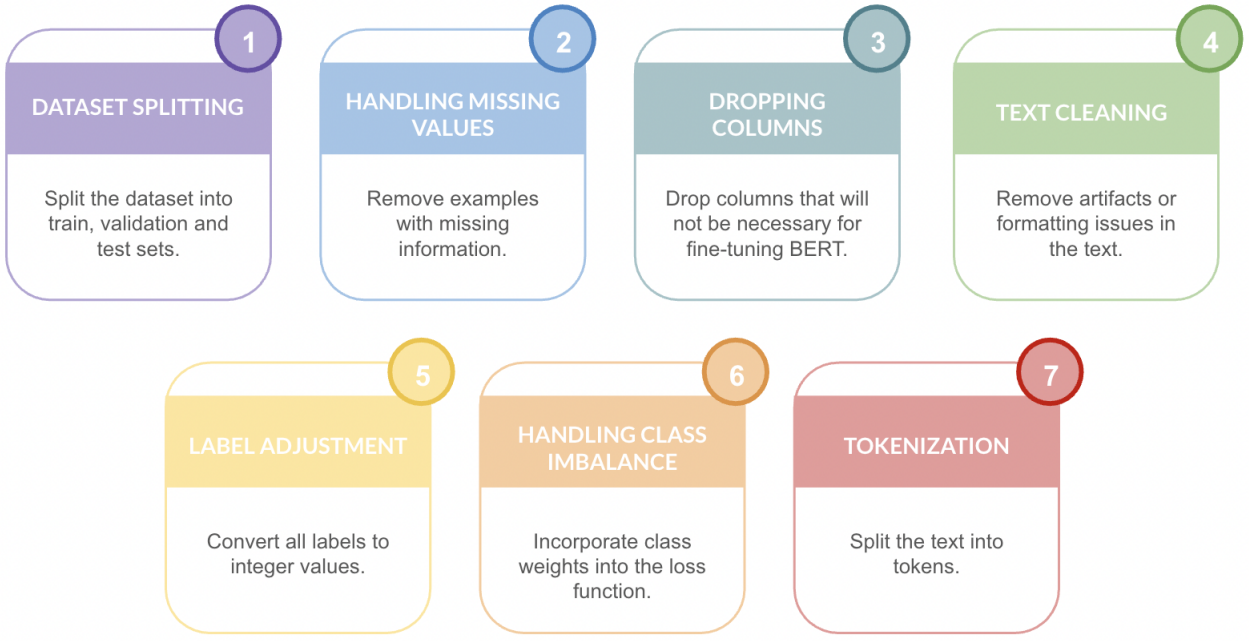

Figure 6: Data processing pipeline.

characters, while also ensuring that the number of translated tokens matches the number of original NER tags.
For Question Answering, we reformat all datasets to match the structure of the SQuAD dataset. This includes introducing a consistent format for questions, contexts, and answers, as well as adding a unique identifier to each example to facilitate further processing.
For POS tagging, we observe that the translation model occasionally ignores whitespace delimiters. For example, punctuation marks that are treated as separate tokens in the original text, are sometimes merged with adjacent words in the translation. Therefore, we post-process the translated text to separate such punctuation, ensuring alignment with the original token structure.
In the cases of hate speech detection and sentiment analysis, we apply a more intensive data cleaning process. These datasets contain noise such as hashtags and user mentions, which may hinder model performance. Therefore, we apply the following preprocessing steps specifically to the datasets of these two tasks:

- Lowercasing, which converts all text to lowercase, thus reducing the vocabulary size. This step ensures that variations such as "Word" and "word" are treated as the same entity [2].
- Removing URLs, special characters, and HTML tags, to eliminate non-linguistic elements that can introduce noise into the dataset, thus improving sentiment or hate classification accuracy.
- Removing non-ASCII characters, to clean potential artifacts resulting from translations or encoding mismatches.
- Removing extra whitespace, to standardize text formatting and ensure consistent spacing.

- Removing excessive punctuation, while ensuring that critical elements like negations (e.g., "can't") are preserved.
- Handling missing values, by removing examples with missing information, to addresses inconsistencies in data.
- Removing usernames, (i.e., words preceded by "@"). This steps is applied to the Italian Tweets Dataset, where usernames are frequent but irrelevant for sentiment classification.

5.2.2 Label Adjustment

The original *HS-bg* dataset includes six distinct classes: one neutral, and five classes representing different forms of hate speech—namely, *sexism*, *racism*, *profanity*, *rudeness*, and *others*. To simplify the classification task and to align with a binary hate speech detection framework, we group all hateful categories into a single class labeled as "1", indicating hate speech. The neutral texts are assigned the label "0", representing non-hateful content. This binary labeling allows for a more straightforward and consistent training process.
Additionally, for datasets such as *Dutch HateCheck*, *South Park*, *Italian Tweets Dataset* and *RuReviews*, categorical labels are first mapped to integer values to ensure compatibility with model training requirements.

5.2.3 Handling Class Imbalance

Class imbalance is a common issue in machine learning, where some classes are significantly underrepresented compared to others. Training models on imbalanced datasets can lead to biased predictions, as the model may favor the majority class and perform poorly on the minority class. To address this, we compute class weights and incorporate them into the loss function during training for those datasets that are found to have an unbalanced distribution. Class weights are calculated based on the frequency of each class in the dataset. Specifically, less frequent classes are assigned higher weights, while more common classes receive lower weights, to ensure that the model pays more attention to underrepresented classes. Once computed, these class weights are incorporated into the loss function used during training. By doing so, the model is encouraged to treat all classes more equally, preventing it from simply optimizing for the most frequent class. This strategy is applied only to datasets where the class distribution ratio is approximately 60:40 or more imbalanced. By incorporating class weights into the loss computation, we aim to improve the model's generalization and robustness.

5.2.4 Tokenization

Tokenization is a fundamental step in preparing textual data for transformer-based models like BERT. Depending on the task, different strategies and levels of granularity are required to ensure correct alignment between input tokens and labels or target spans.
For the Question Answering task, each training example consists of a question-context pair. The model must predict the start and end positions of the answer span within the context. During tokenization, if the context is too long to fit within the model's maximum input length, it is divided into overlapping chunks using a sliding window mechanism known as stride. This approach ensures that potential answer spans are not truncated at chunk boundaries. After tokenization, two additional steps are necessary: tracking which original example each tokenized chunk corresponds to; and aligning token indices with the character-level answer

span to correctly map the start and end positions of the answer within the tokenized input. This alignment is crucial, as the model relies on it to accurately predict the answer span in the context.
For hate speech detection, each input consists of a single piece of text - such as a tweet, or a comment - which is tokenized using BERT's tokenizer and converted to tensors. Then, the corresponding label is attached to each tokenized input. Since this is a sentence-level classification task, there is no need to align tokens with individual word-level labels.
In the case of NER, word-level tokenization is performed. Because BERT's tokenizer may split words into multiple subwords, it becomes necessary to align each original entity label with the corresponding subwords. This is done by assigning the entity label to the first subword and replicating it for the remaining subwords. This alignment ensures that the model learns to make predictions at the appropriate positions in the sequence.
The POS Tagging task follows a similar tokenization strategy to NER. Each sentence is tokenized at the word level, and when a word is split into subwords, all resulting subwords are assigned the same POS tag. This approach maintains the consistency of syntactic labeling across the tokenized inputs.
In the sentiment analysis task, the input consists of a single sentence or paragraph of text (e.g., a review or social media post). Each tokenized sequence is then paired with a single sentiment label, making label alignment straightforward.
For the NLI task, each example consists of a pair of sentences: a premise and a hypothesis. The tokenizer processes both sequences simultaneously, using BERT's special token format that separates the two texts with a [SEP] token. This format is essential for enabling BERT to capture inter-sentence dependencies effectively.

5.2.5 Additional Preprocessing

In addition to general preprocessing steps applied across all tasks, certain task-specific challenges require specialized handling. Below, we detail the additional preprocessing steps that we perform for the Question Answering and POS Tagging tasks.

Question Answering An essential step in the preprocessing pipeline of the question answering datasets that were translated to English, is to recompute the answer start index within the translated context. This step is crucial because models like BERT depend on accurate character-level start and end indices to learn span-based predictions. However, translation often alters sentence structure, word order, or phrasing, invalidating the original indices defined in the source language. To address this, we use Hugging Face's *transformers* library and apply a Question Answering pipeline built on the pretrained model *distilbert-base-uncased-distilled-squad*. For each translated question-context pair, the pipeline predicts the most likely answer span, from which both the answer text and its new starting index were extracted. These recomputed values are then used to fine-tune the BERT-Base model. In cases where the model fails to return a valid result, the samples are discarded.
Some datasets, such as *EXAMS*, *QA-ITA-200k* and *CMRC2018*, are missing answer start indices even in their original form. For these, we first attempt a direct substring match of the answer within the context. When that fails, we apply fuzzy string matching to locate a similar phrase in the context. If a match is found above a similarity threshold, we compute the corresponding start and end indices. This approach achieves full alignment for *CMRC2018* and good coverage in *EXAMS*, where the answer start index is found for 92% of the entries.

However, it proves ineffective for *QA-ITA-200k*, where approximately 78% of examples lacks a detectable match. As a result, we adopt the same strategy used for translated datasets: applying a QA pipeline based on the *xlm-roberta-large-squad2* model to infer the answer span and starting index. For datasets such as *SberQuAD* and *P-Direkt*, the original annotations already includes answer start indices. However, since the test set of *SberQuAD* lacks answer annotations, we use only the training and validation sets, which contain complete information, and then re-split them to create new training, validation, and test sets.
When using the QA pipeline, so for the cases of the translated datasets and the Italian original dataset, the method always predicts a valid answer span, allowing us to recompute and replace the original annotations in 100% of cases. No examples are discarded in these scenarios. However, it is important to note that while the pipeline consistently returns results, the correctness of these predictions is not guaranteed. As such, some degree of noise or misalignment may persist despite the technical completeness of the preprocessing step.

POS Tagging For the POS tagging task, a critical challenge emerges from the translation of datasets that were originally annotated with Universal POS (UPOS) tags at the word level. The translation model used (*Helsinki-NLP/Opus-MT*), is optimized for sentence-level translation. While this improves fluency and reduces hallucinations, it also means that the translated sentences no longer have a one-to-one correspondence with the original tokenized forms. As a result, the original UPOS annotations become misaligned or unusable. To overcome this, we opt to recompute POS tags directly on the translated English text using the *spaCy* library. Each translated sentence is processed with spaCy's pretrained English pipeline, which outputs POS tags for each token based on linguistic analysis. This ensures that the number of tokens matches the number of predicted POS tags, restoring alignment and making the dataset suitable for supervised training.
For POS tagging, since spaCy's English pipeline assigns a POS tag to every token, the re-annotation process succeeded on 100% of the translated examples, without the need to discard or manually adjust any samples. However, also in this case we acknowledge the fact that, even though all tokens are annotated with this method, it is still subject to errors.

5.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in any data-driven project. It involves systematically examining datasets to uncover underlying patterns, identify anomalies, test hypothesis and verify assumptions. This process is typically carried out using a combination of summary statistics and graphical representations [76]. The primary goal of EDA is to gain a comprehensive understanding of the data's structure, distribution, and quality, insights that are essential for guiding preprocessing choices, feature engineering, and model selection. In the context of this thesis, EDA is performed separately for each task to highlight dataset-specific characteristics and challenges.

5.3.1 Sentiment Analysis

Figure 7, depicting the label distributions in the sentiment analysis datasets, reveals varying degrees of class balance and imbalance across the corpora. The Cinexio dataset is notably skewed, with a dominant proportion of positive samples (68.4%), while negative and neutral sentiments are underrepresented at approximately 16.41% and 15.2%, respectively. In contrast, the Weibo Senti 100k and DBRD datasets are almost perfectly balanced between positive and negative classes. The Italian Tweets Dataset shows a strong skew toward the neutral class (81.5%), with relatively few positive (14.2%) and even fewer negative (3.9%) or mixed (0.3%) samples, which could present challenges for detecting minority sentiments and may bias models toward predicting the majority class. Finally, RuReviews is constructed with perfect class balance across positive, negative, and neutral sentiments (each 33.3%). This plot suggests that imbalanced datasets like Cinexio or the Italian Tweets require special rebalancing techniques.

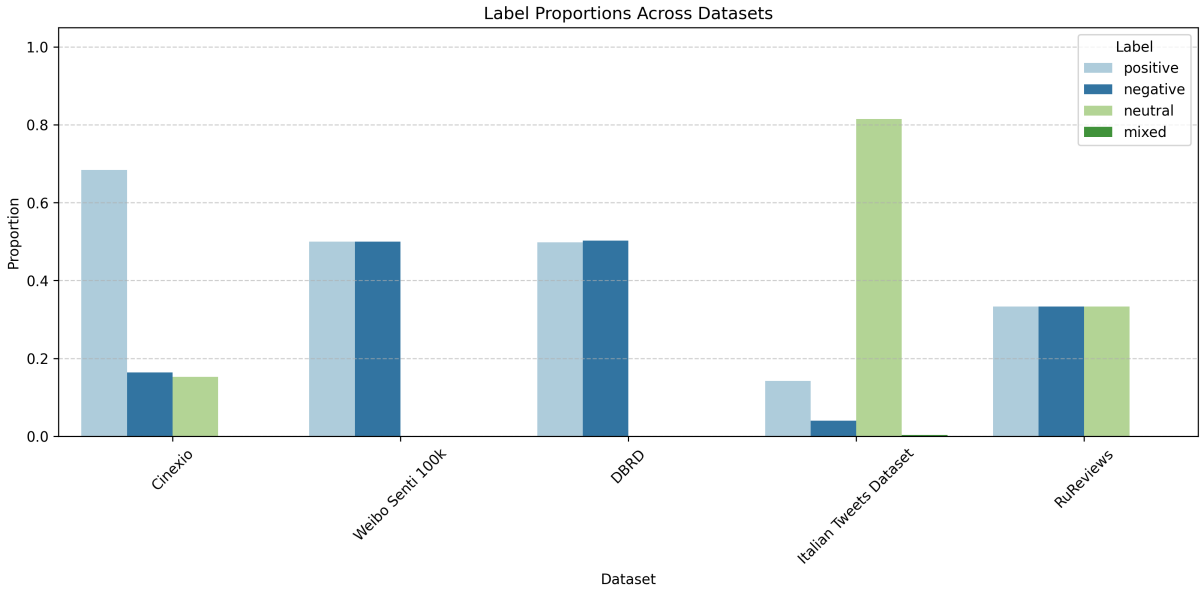

Figure 7: Label distribution across the considered sentiment analysis datasets.

5.3.2 Question Answering

Table 11 reports the average lengths of answer spans and contexts across the various Question Answering datasets used in this thesis. As shown, the average context length varies signifi-

cantly, ranging from 80 words in the Dutch HateCheck dataset to 191 words in the HS-bg dataset. Similarly, answer lengths also vary, with HS-bg and South Park containing relatively short answers (averaging 3–4 tokens), while HS-it has notably longer answer spans, averaging 44 tokens. Recognizing these dataset-level differences is important for interpreting model performance, as context and answer lengths can affect how well the model captures and predicts the correct spans. Additionally, after inspecting the distribution of context lengths across the QA datasets, we found it necessary to address cases where the context exceeded the model's maximum input length. This was done by applying a sliding window mechanism, as previously described in Section 5.2.4.

Dataset	**Avg answer length**	**Avg context length**
EXAMS	3	191
CMRC2018	10	143
P-Direkt	21	80
QA-ita-200k	44	141
SberQuAD	4	105

Table 11: Average length of answers and contexts for each Question Answering dataset.

5.3.3 Hate Speech Detection

As part of our exploratory data analysis for the hate speech detection task, we examine the class distribution of labels across the five selected datasets: HS-bg, Dutch HateCheck, COLD, HS-it, South Park. Figure 8 presents the proportion of hateful and non-hateful comments for each dataset. The y-axis represents the proportion (ranging from 0 to 1), while the x-axis displays two bars per dataset, indicating the relative frequency of the two classes.
HS-bg is highly imbalanced, with 98.2% of comments labeled as non-hateful and only 1.8% labeled as hateful. This skewness may pose challenges for model training, as classifiers tend to favor the majority class in such settings. In contrast to Bulgarian, the Dutch dataset, Dutch HateCheck, exhibits a strong imbalance in the opposite direction. 70.1% of the comments are labeled as hateful, while only 29.9% are non-hateful. The class distribution of COLD is relatively balanced, with 48.1% hateful and 51.9% non-hateful comments, The HS-it dataset also presents a moderate imbalance, with 58.5% of comments labeled as non-hateful and 41.5% as hateful. South Park leans toward non-hateful content, with 67.2% of comments in the non-hateful class and 32.8% labeled as hateful.
Overall, these results underscore the importance of accounting for class distribution in model development. For highly imbalanced datasets, such as Bulgarian or Dutch, standard training may lead to suboptimal performance on the minority class. To address this, we perform class weighting as explained earlier in Section 5.2.3.

5.3.4 POS Tagging

To understand the label distribution for the POS tagging task, we conduct an exploratory analysis across the five datasets. Figure 9 presents a grouped bar chart showing the proportion

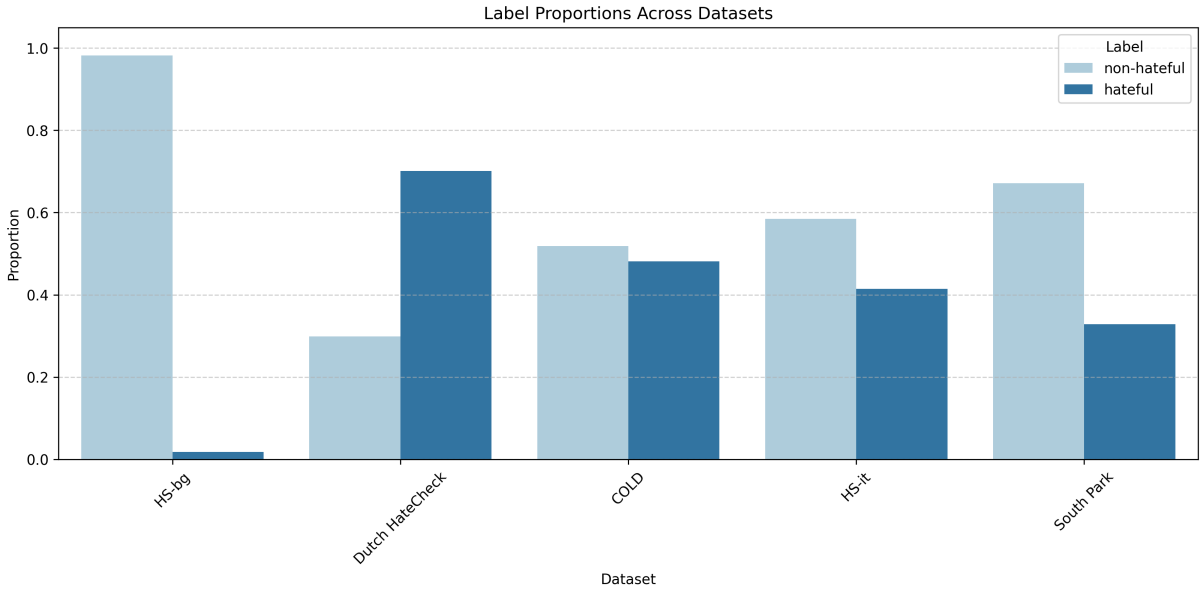

Figure 8: Label distribution across the considered hate speech detection datasets.

of each POS tag within each dataset. Each group corresponds to a dataset (UD_Bulgarian-BTB, UD_Chinese-GSD, Alpino, PoSTWITA-UD, Taiga), and within each group, individual bars represent the relative frequency of POS tags such as NOUN, VERB, ADJ, etc. The full set of POS labels used follows the Universal POS tagset standard [35], which includes 17 unique tags. Overall, the distribution of POS tags is reasonably consistent across languages. Common categories like NOUN, VERB, and PUNCT are among the most frequent in all datasets, with NOUN consistently occupying a large proportion (e.g., 18.3% in UD_Bulgarian-BTB, 19.1% in UD_Chinese-GSD). Some variation is observed for less frequent categories like INTJ, SYM, and X, which appear sparsely or are nearly absent in several datasets. Due to this skewness, class balancing techniques are applied, as reported in Section 5.2.3.

5.3.5 NER

The plot in Table 10 presents the distribution of NER labels across the WikiAnn datasets for Bulgarian (bg), Dutch (nl), Italian (it), and Russian (ru). Each bar represents the proportion of tokens assigned to a specific label, such as B-PER, I-ORG, or O. All datasets show a strong class imbalance, with the O label (non-entity tokens) dominating, and ranging from approximately 56.5% in wikiann-ru to 74.0% in wikiann-nl. Despite this, differences in the distribution of entity labels are evident. For example, wikiann-ru shows a notably high proportion of I-PER tokens (10.7%), indicating that person entities in Russian tend to span multiple tokens. Similarly, Italian and Russian both exhibit relatively high proportions of I-ORG tokens (over 10%), suggesting a prevalence of multi-token organization names. In contrast, Dutch has fewer B-PER and B-ORG tokens, which may suggest either fewer named entities or shorter entity spans in the dataset. Bulgarian displays a higher proportion of B-LOC tokens compared to other languages, indicating more frequent mentions of locations. These differences are important, as they can affect the learning dynamics of NER models, particularly in terms of entity recall and boundary detection. Models trained on such imbalanced and varied datasets may become

[35] https://universaldependencies.org/u/pos/

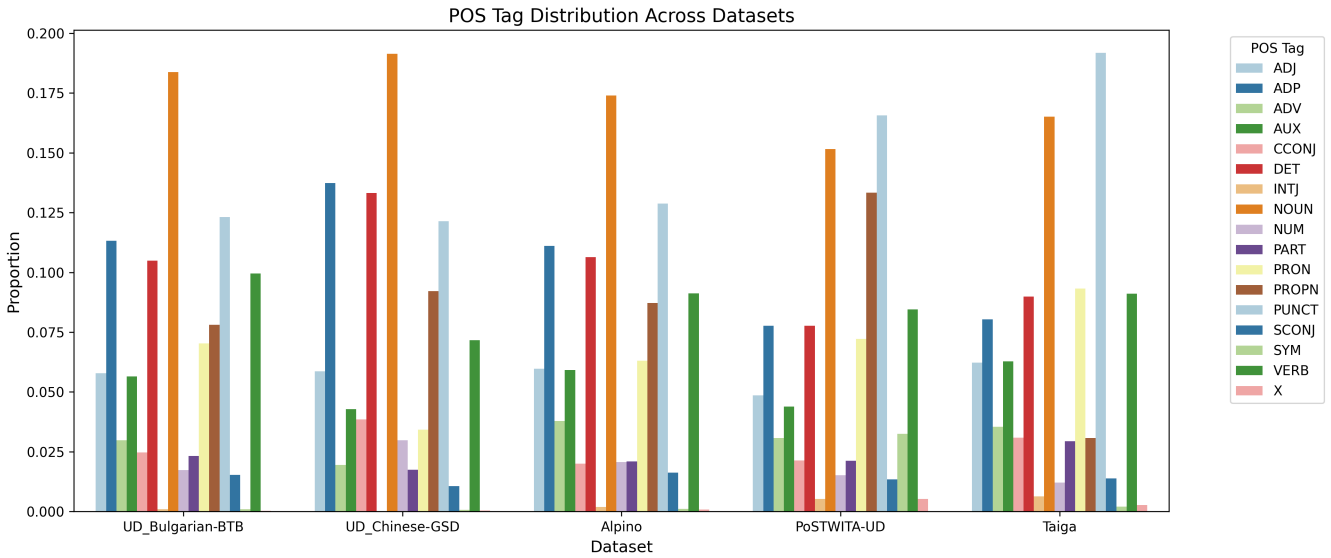

Figure 9: Proportion of each POS Tag in the five considered datasets.

biased toward predicting the dominant O label or struggle with correctly identifying long, multi-token entities. For this reason, dataset balancing techniques are applied to these datasets (see Section 5.2.3).

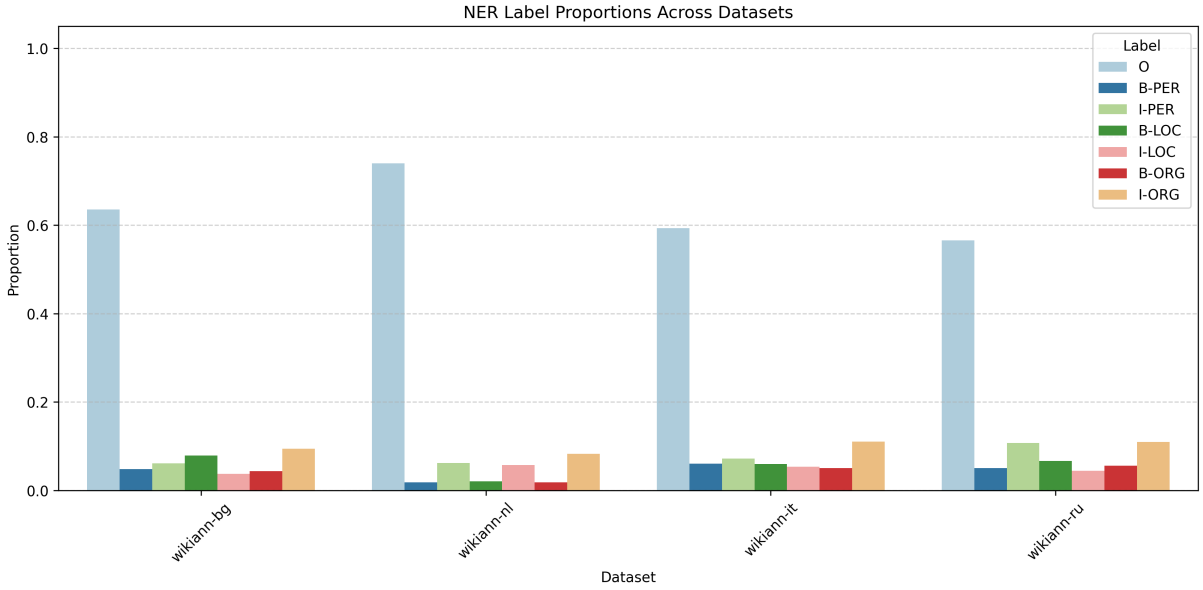

Figure 10: Length distribution of each text across the considered NER datasets.

5.3.6 NLI

Figure 11 shows the class distribution for the NLI datasets used in this study. As seen in the plot, the XNLI datasets (Bulgarian, Chinese, and Russian) and LingNLI exhibit balanced label distributions, with each class accounting for approximately one-third of the examples. This balanced setup is ideal for training models without introducing bias toward a specific label. In contrast, the SICK-NL dataset displays a notable imbalance: the neutral class dominates,

followed by entailment and contradiction. Since this skewed distribution could potentially affect model performance by biasing predictions toward the majority class, we handle this issue as proposed in Section 5.2.3.

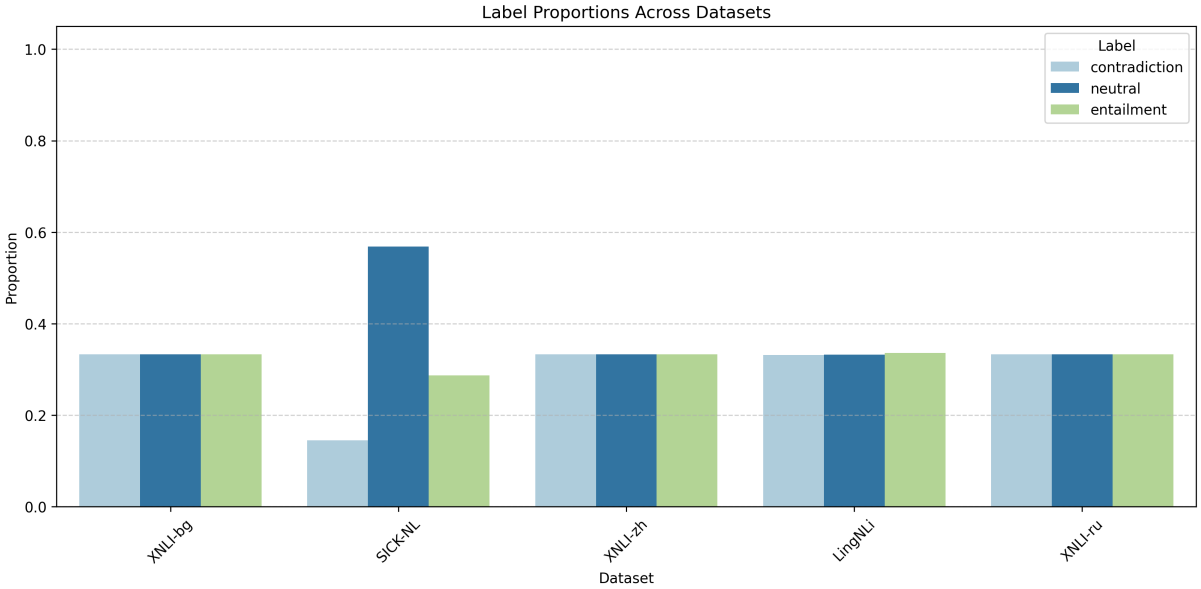

Figure 11: Length distribution across the NLI datasets.

6 Results

This section, which corresponds to the Modeling phase of CRISP-DM, presents the results obtained from fine-tuning BERT on the translated datasets and their original-language counterparts. By comparing the performance of English BERT models trained on translated text with that of native-language BERT models, we aim to evaluate the extent to which translation can serve as a viable alternative to language-specific modeling. For each table we have also reported the signed difference in F1 score (expressed in percentage points), between the original and translated methods, where a negative value indicates lower performance from the translation approach.

Hate Speech Detection : Table 12 presents the results for the hate speech detection task across the five considered languages, reported in terms of accuracy, precision, recall, and F1-score. Overall, these results reveal varying levels of success in using BERT to detect hate speech, with performance largely dependent on the language and dataset characteristics.
For Bulgarian (HS-bg), both the original and translated models achieve very high accuracy (above 97.8%), suggesting strong overall classification capability. However, the considerably lower precision, recall, and F1-scores (all below 0.74) indicate difficulty in correctly identifying hate speech instances, likely due to class imbalance. This means that while the model is good at predicting the majority class, it struggles with minority (i.e. hate) cases.
The largest performance gap is observed for Chinese (COLD), where the translated model significantly underperforms the original-language BERT by more than 7% across all metrics. The translated model reaches an F1-score of 0.8303, compared to 0.9063 in the original.
For Dutch (HateCheck), the original-language model demonstrates the best performance across all metrics, achieving an F1-score of 0.9829, the highest among all evaluated hate speech datasets. The translated model also performs well, suggesting that both configurations are quite effective. These strong results, especially in recall, suggest that the model can reliably identify hate speech in this dataset.
In the case of Italian (HS-it), both models show more modest results, with F1-scores below 0.79. While the original-language model performs slightly better (by about 3–4% in recall and F1), the overall performance suggests moderate effectiveness. The models appear to capture hate speech patterns to some extent but with noticeable limitations in distinguishing subtle cases.
For Russian (South Park), the differences between the translated and original versions are relatively minor, with the translated model reporting a drop of only about 1.5% across most metrics. Both versions yield F1-scores in the 0.86–0.88 range, indicating solid performance.

Question Answering Table 13 presents the results for the Question Answering task, evaluated using Exact Match (EM) and F1-score. The EM score measures the percentage of predictions that match the ground truth answer exactly, while the F1-score accounts for partial matches by evaluating the overlap between the predicted and true answers. Overall, the results reveal significant variability in performance across languages and datasets, driven by factors such as dataset size, context complexity, and preprocessing challenges related to answer alignment.
For Bulgarian (EXAMS), the performance is low across both metrics, regardless of whether the original or translated version is used. The highest EM recorded is just 21.1%, and the F1 score reaches 33.3%. These modest scores suggest that the model struggles with this dataset,

Lang	Dataset	Accuracy	Precision	Recall	F1-Score	ΔF1 (p.p.)
bg	HS-bg Translated	0.9817	0.7163	0.6889	0.7016	-2.36
	HS-bg Original	0.9782	0.7104	0.7424	0.7252	
zh	COLD Translated	0.8303	0.8303	0.8307	0.8303	-7.6
	COLD Original	0.9064	0.9065	0.9066	0.9063	
nl	Dutch HateCheck Translated	0.9469	0.9574	0.9081	0.9291	-5.38
	Dutch HateCheck Original	**0.9867**	**0.9911**	**0.9755**	**0.9829**	
it	HS-it Translated	0.7596	0.7522	0.7511	0.7516	-3.71
	HS-it Original	0.8002	0.8008	0.7833	0.7887	
ru	South Park Translated	0.8797	0.8560	0.8685	0.8617	-1.34
	South Park Original	0.8911	0.8683	0.8835	0.8751	

Table 12: Results on the test sets of the **Hate Speech Detection** datasets obtained after fine-tuning.

possibly due to the relatively small size of the dataset (3,349 examples) and the fact that the dataset was originally created for multiple choice answers. These factors likely make it difficult for the model to accurately locate and extract answer spans. The low EM and F1 scores reflect this challenge, though we do not measure the average positional error directly; further span alignment analysis could clarify whether the errors are due to mislocation or semantic mismatch.

For Chinese (CMRC2018), results are noticeably better. The original-language model achieves an EM of 46.1 and an F1-score of 46.9, demonstrating moderate success. In contrast, the translation-based approach suffers a noticeable drop in EM (to 29.5), although the F1-score remains relatively close (42.5). These results suggest that the model can handle Chinese QA with reasonable competence in its native form, but translating to English introduces degradation, possibly due to changes in sentence structure or alignment errors in long answers.

For Dutch (P-Direkt), the translated model achieved an EM of 21.4 and an F1-score of 50.9. The original-language model, by contrast, produces no exact matches, highlighting alignment or formatting issues in the original version. Given the translated model's performance, particularly its solid F1-score, the model demonstrates reasonable capability for Dutch QA despite the dataset's limited size.

In Italian (QA-ITA-200k), the results are especially strong for the translated version, which achieves the second-highest F1-score overall (70.1) and a high EM of 58.3. The original-language model performs substantially worse, likely due to alignment issues. The translated setup, supported by reliable English QA span inference using distilBERT, enables effective model training and prediction.

Russian (SberQuAD) delivers the best results among all datasets, with the translated version reaching an EM of 65.2 and an F1-score of 76.3. Even the original version performs well, suggesting the dataset's high quality and suitability for span-based QA. The short average

answers (4 tokens) and moderate context length (105 tokens) likely help the model focus its predictions.

Lang	Dataset	Exact Match	F1-Score	ΔF1 (p.p.)
bg	EXAMS Translated	21.1155	29.9204	+3.38
	EXAMS Original	16.1812	33.2970	
zh	CMRC2018 Translated	29.4774	42.4524	+4.45
	CMRC2018 Original	46.0682	46.9045	
nl	P-Direkt Translated	21.4286	50.9458	-4.50
	P-Direkt Original	0.0	46.4448	
it	QA-ITA-200k Translated	58.2868	70.1191	-13.3
	QA-ITA-200k Original	8.6920	56.7849	
ru	SberQuAD Translated	**65.2105**	**76.3440**	-2.14
	SberQuAD Original	53.5835	74.2025	

Table 13: Results on the test sets of the **Question answering** datasests obtained after fine-tuning.

Sentiment Analysis Table 14 presents the results for the Sentiment Analysis task. Performance varies across languages and datasets, shaped by classification setups, label distributions, and dataset characteristics.

For Bulgarian (Cinexio), the model demonstrates moderate performance, with an original-model accuracy of 79.5% and an F1-score of 66.6%, which is lower likely due to the imbalanced nature of the dataset. The translated version shows a decline across all metrics, particularly in precision (from 69.6% to 63.3%), suggesting more false positives.

In Chinese (Weibo Senti 100k), the model achieves near-perfect accuracy and F1-score in the original-language setting. The dataset is balanced between positive and negative sentiments and is large in size, making it ideal for training. The translated version sees a notable drop, but performance remains high overall. This demonstrates that the model performs exceptionally well on this dataset in the original language and reasonably well post-translation.

Dutch (DBRD) also yields very strong performance, with the original and translated models achieving nearly identical results, both around 92.9% accuracy and F1. The dataset is binary and balanced, further supporting the model's success.

For the Italian Tweets Dataset, results are more nuanced. The translated model achieves slightly higher accuracy, but both models converge at an F1-score around 56%. The low F1 is likely influenced by the extreme class imbalance: over 81% of samples are labeled neutral, while mixed and negative classes together make up less than 5%. This imbalance reduces recall and precision for minority classes. Despite this, the relatively high accuracy and consistent F1 suggest the model performs adequately, though class-level performance (especially for rare sentiments) may be weak.

Lastly, for Russian (RuReviews), the model shows solid performance in both configurations, with the original achieving accuracy and F1-score above 77%. The dataset is evenly split

across three sentiment labels, and the model maintains balanced performance across classes. The translated version shows only a modest drop in all metrics, maintaining an F1 of 74.4%. These scores reflect that the model is effective in performing sentiment analysis on this dataset, even after translation.

Lang	Dataset	Accuracy	Precision	Recall	F1-Score	ΔF1 (p.p.)
bg	Cinexio Translated	0.7620	0.6332	0.6432	0.6380	+2.76
	Cinexio Original	0.7945	0.6963	0.6632	0.6656	
zh	Weibo Senti 100k Translated	0.8294	0.8297	0.8293	0.8293	+15.49
	Weibo Senti 100k Original	**0.9842**	**0.9846**	**0.9843**	**0.9842**	
nl	DBRD Translated	0.9283	0.9284	0.9282	0.9283	+0.07
	DBRD Original	0.9293	0.9295	0.9290	0.9290	
it	Italian Tweets Dataset Translated	0.8341	0.5745	0.5861	0.5603	+0.05
	Italian Tweets Dataset Original	0.7993	0.5063	0.6815	0.5608	
ru	RuReviews Translated	0.7417	0.7621	0.7414	0.7439	+3.1
	RuReviews Original	0.7719	0.7833	0.7715	0.7749	

Table 14: Results on the test sets of the **Sentiment Analysis** datasests obtained after fine-tuning.

POS Tagging Table 15 presents the results for the POS tagging task, showing varying effects of machine translation across different languages and datasets. Across all datasets, the models generally achieved high performance.

In Chinese, translation significantly improved performance. The F1-score rose from 80.4% with the original data to 93.7% with the translated version, indicating a substantial gain of over 13 points. Accuracy and all other metrics also increased notably, making this the dataset with the most pronounced improvement from translation.

The Dutch dataset Alpino also benefited from the translation approach. The F1-score increased from 78.3% in the original to 87% in the translated version, along with a boost in overall accuracy.

In Bulgarian (UD_Bulgarian-BTB), performance remained relatively stable across original and translated versions. The F1-scores were similar, and accuracy showed only a slight drop.

By contrast, PoSTWITA-UD saw a significant performance decline when using the translated dataset. The F1-score dropped from 93.3% to 78.8%, and accuracy fell from 96.2% to 74.9%. This represents the largest drop in performance among all languages in this task.

Russian Taiga also experienced a drop in all metrics with the translated data. The F1-score decreased from 89.4% to 81.2%, and accuracy fell from 94.1% to 86.8%, though the drop was

less drastic than in the Italian case.

Lang	Dataset	Accuracy	Precision	Recall	F1-Score	ΔF1 (p.p.)
bg	UD_Bulgarian-BTB Translated	0.9745	0.9114	0.9528	0.9285	+0.28
	UD_Bulgarian-BTB Original	0.9764	0.9144	**0.9806**	0.9257	
zh	UD_ChineseGSD Translated	**0.9819**	0.9230	0.9573	**0.9369**	+13.31
	UD_ChineseGSD Original	0.8569	0.7637	0.8791	0.8038	
nl	Alpino Translated	0.9643	0.8505	0.9138	0.8699	+8.68
	Alpino Original	0.8442	0.7741	0.8116	0.7831	
it	PoSTWITA-UD Translated	0.7492	0.8057	0.8258	0.7881	-14.47
	PoSTWITA-UD Original	0.9621	**0.9256**	0.9411	0.9328	
ru	Taiga Translated	0.8681	0.8052	0.8445	0.8122	-8.13
	Taiga Original	0.9407	0.8678	0.9338	0.8935	

Table 15: Results on the test sets of the **POS** datasests obtained after fine-tuning.

NER Table 16 presents the results for the Named Entity Recognition task.
Looking at the overall results for each language, the model demonstrates strong performance on the original datasets for Bulgarian, Italian, and Russian, with F1-scores above 87%. This indicates that the model can effectively perform NER when trained on high-quality, original language data. Bulgarian and Italian both show solid results, with F1-scores of 90.25% and 87.42% respectively, suggesting reliable entity recognition. The translated datasets for Bulgarian and Italian, despite showing some performance degradation compared to the originals, still maintain reasonably strong F1-scores . This implies that the model retains some capacity to perform NER on these translated datasets, though with reduced reliability. Russian translated version suffers a drastic drop, indicating that translation severely impacts model effectiveness for this language.
For Dutch, the model achieves moderate performance on both original and translated datasets, with a slight improvement on the translated version (77.01% vs. 74.26%). This suggests that the model is relatively robust to translation noise in Dutch, and can even benefit from translated data in some cases.
Chinese presents lower performance, with F1-scores of 59.79% on the original dataset and 52.25% on the translated one. These results indicate challenges in modeling NER for Chinese, potentially due to the complexity of the language or dataset characteristics, and suggest that the model struggles to generalize well in this case regardless of translation.

Lang	Dataset	Accuracy	Precision	Recall	F1-Score	ΔF1 (p.p.)
bg	wikiann-bg Translated	0.9291	0.6906	0.8388	0.7575	-14.5
	wikiann-bg Original	**0.9513**	**0.8829**	**0.9230**	**0.9025**	
zh	wikiann-zh Translated	0.8296	0.4063	0.7317	0.5225	-7.54
	wikiann-zh Original	0.8578	0.4803	0.7917	0.5979	
nl	wikiann-nl Translated	0.9363	0.7439	0.7982	0.7701	+2.75
	wikiann-nl Original	0.8961	0.7011	0.7892	0.7426	
it	wikiann-it Translated	0.9141	0.7421	0.8030	0.7713	-10.29
	wikiann-it Original	0.9486	0.8526	0.8969	0.8742	
ru	wikiann-ru Translated	0.8172	0.2685	0.6446	0.3791	-50.05
	wikiann-ru Original	0.9450	0.8593	0.9008	0.8796	

Table 16: Results on the test sets of the **NER** datasests obtained after fine-tuning.

NLI Table 17 shows the results for the Natural Language Inference task, which demonstrate relatively stable model performance across both original and translated datasets, with only modest variations in most cases.

For Bulgarian, the model performs well on both the original and translated versions of XNLI, with F1-scores of 78.48% and 77.73%, respectively. The difference is minimal, indicating that the model can maintain robust inference capabilities regardless of whether the data is in the original language or translated.

In Chinese, performance is slightly lower overall, but the model still reaches an F1-score of 77.29% on the original dataset and 73.82% on the translated version. While the drop is more pronounced here than for Bulgarian, the results remain reasonably strong, suggesting that the model can still effectively perform NLI, though with some sensitivity to translation quality or linguistic complexity.

Dutch also shows strong and consistent results, with F1-scores of 82.90% on the original and 82.48% on the translated dataset. Interestingly, while accuracy is slightly higher on the original data, the translated version has a slight edge in precision.

For Italian, the translated dataset (LingNLI) actually outperforms the original, with an F1-score of 64.34 % compared to 62.07%. Though both scores are lower than in other languages, this small improvement implies that translation may help mitigate some limitations in the original data quality or structure. Nevertheless, the relatively modest scores suggest that the model

struggles somewhat more with this dataset overall.
Russian is another case where the translated version marginally outperforms the original: F1-scores are 77.04% (translated) versus 75.91% (original). This again reflects the model's resilience to translation in the NLI setting and suggests effective generalization.

Lang	Dataset	Accuracy	Precision	Recall	F1-Score	Δ F1 (p.p.)
bg	XNLI-bg Translated	0.7771	0.7776	0.7771	0.7773	-0.75
	XNLI-bg Original	0.7847	0.7868	0.7849	0.7848	
zh	XNLI-zh Translated	0.7390	0.7384	0.7389	0.7382	-3.47
	XNLI-zh Original	0.7723	0.7749	0.7722	0.7729	
nl	SICK-NL Translated	0.8333	**0.8166**	0.8377	0.8248	-0.42
	SICK-NL Original	**0.8394**	0.8100	**0.8558**	**0.8290**	
it	LingNLI Translated	0.6451	0.6434	0.6440	0.6434	+2.27
	LingNLI Original	0.6210	0.6256	0.6207	0.6207	
ru	XNLI-ru Translated	0.7702	0.7707	0.7702	0.7704	+1.13
	XNLI-ru Original	0.7599	0.7600	0.7600	0.7591	

Table 17: Results on the test sets of the **NLI** datasests obtained after fine-tuning.

7 Discussion

This section synthesizes the results presented in Section 6 by addressing the research questions outlined earlier. The goal is to evaluate the consistency and reliability of a translation-based approach using English BERT across different languages and NLP tasks. We analyze how linguistic factors and task characteristics influence performance, and assess whether this method can serve as a viable alternative to native-language models.

7.1 RQ1: Does a translation-based approach using English BERT perform consistently across languages from different linguistic families?

To address the first research question, we classify the languages in our study according to their linguistic families and genera, using the WALS (World Atlas of Language Structures) phylogenetic tree [31]. This typologically informed framework groups languages based on shared historical and structural features. According to this classification, Bulgarian and Russian belong to the Slavic genus within the Indo-European family; Dutch and English are part of the Germanic genus, also within Indo-European; Italian belongs to the Romance genus, again within the same family. Chinese, by contrast, falls under the Sino-Tibetan family, entirely distinct from Indo-European languages.

The results reported in Section 6 indicate that performance is not uniform across languages, though certain patterns emerge when considering linguistic relatedness. Most notably, Chinese consistently underperforms in the translation-based setting. This likely stems from two key factors. First, the typological distance between Chinese and English, spanning differences in syntax, morphology, and word order, poses challenges for machine translation systems, which may fail to preserve linguistic features critical to downstream tasks. Second, Chinese benefits from a high-quality monolingual BERT model, trained on abundant native data and optimized for its unique linguistic properties, which can outperform English BERT applied to translated text, especially when translation introduces noise.

In contrast, Dutch consistently shows strong results with the translation-based approach, often matching or outperforming native-language models, particularly in QA, POS tagging, and NER. This can likely be attributed to the shared linguistic lineage between Dutch and English. As members of the same genus (Germanic) and family (Indo-European), they exhibit similar syntactic structures and morphosyntactic features that are more likely to be preserved in translation, making English BERT's learned representations more transferable.

However, linguistic proximity alone does not guarantee consistent outcomes. Bulgarian and Russian, though both Slavic languages, display divergent performance across tasks. This inconsistency may arise from differences in dataset size, morphological complexity, or the quality of the translation systems employed for each language. These findings suggest that while shared phylogenetic roots can support effective transfer, other factors, such as translation quality and language-specific characteristics, also play a critical role.

7.2 RQ2: Are there specific NLP tasks where this approach is more effective?

The effectiveness of the translation-based approach varies substantially across NLP tasks. Some tasks show strong performance transfer when using English BERT on translated text,

while others suffer significant degradation. Below, we evaluate each task with respect to its suitability for cross-lingual transfer via translation.

Question Answering QA yields some of the most promising results for the translation-based approach. In particular, datasets such as QA-ITA-200k, SberQuAD, and P-Direkt showed improved or comparable performance when using translated data with English BERT, often outperforming native-language models. A likely explanation is the strength of English BERT models on QA, largely due to extensive pretraining on large-scale datasets like SQuAD.
Moreover, in cases like Dutch and Russian, where original datasets included labeled answer spans, the translated versions, despite requiring automatic re-annotation, still outperformed the native versions. This suggests that even with possible span alignment errors, the robust generalization capabilities of English BERT compensate effectively.
Chinese presents a notable exception: the translated version underperformed significantly. This may stem from several issues: limitations in our method of computing post-translation answer spans, and the overall strength of the native Chinese BERT model, which benefits from language-specific pretraining. Additionally, the considerable typological distance between Chinese and English likely increases translation difficulty, introducing semantic and syntactic inconsistencies that degrade QA performance.
In summary, QA appears particularly well-suited to translation-based transfer. Its format aligns well with English BERT's strengths, and the use of automatically annotated datasets minimizes disparities introduced through translation.

Part-of-Speech Tagging The translation-based models also performed well in POS tagging, particularly for Dutch and Chinese. English BERT's pretraining includes extensive syntactic exposure, which appears to translate effectively even when the data is non-native. These results suggest that lower-level linguistic tasks like POS tagging are relatively robust to translation-induced distortions.

Sentiment Analysis For sentiment analysis, translation-based models achieved performance roughly comparable to native-language models in most cases. While native models generally had a slight edge, the differences were often minor, especially in datasets with binary classification and balanced class distributions (e.g., DBRD, RuReviews).
However, in more complex datasets like Italian Tweets or Cinexio, which feature multiple sentiment categories and imbalanced classes, the translation-based approach showed a noticeable decline. This highlights its sensitivity to semantic nuance and label granularity, where even slight mistranslations can shift sentiment cues.

Named Entity Recognition and Hate Speech Detection NER and hate speech detection were the least compatible with the translation-based strategy. NER requires precise token-level alignment and boundary preservation. Even minor changes introduced by translation, such as altered phrasing or reordering, can break entity spans, degrading model performance. This issue was particularly evident in the WikiANN datasets, where our word-by-word translation approach introduced artifacts that distorted named entity boundaries, leading to lower F1-scores. Hate speech detection similarly relies on nuanced semantics, idiomatic expressions, and cultural context, all elements highly susceptible to distortion in translation. Subtle shifts in tone or meaning can obscure offensive language, making this task ill-suited

to a translation-based approach. These results suggest that tasks dependent on fine-grained linguistic precision or cultural interpretation are poor candidates for cross-lingual transfer via translation.

Natural Language Inference Natural language inference (NLI) generally produced favorable outcomes in the translation setting. Translated premise-hypothesis pairs preserved the relational structure of inference tasks well enough that English BERT could still perform effectively. In many cases, performance matched or exceeded that of native-language models. This may reflect both the clarity of the NLI format and the strength of English BERT on inference tasks, which involve pattern recognition more than domain-specific semantics or token-level precision.

Summary In summary, translation-based approaches are most effective for tasks that are less sensitive to token-level structure (e.g., POS tagging) and that are aligned with English BERT's pretraining strengths (e.g., QA, NLI), They are less effective for tasks that require exact word boundaries (e.g., NER), depend on subtle semantics or cultural specificity (e.g., hate speech detection), and involve complex or imbalanced label spaces (e.g., multi-class sentiment analysis).

7.3 RQ3: How does fine-tuning English BERT on translated text compare to using pre-trained native-language BERT models?

To summarize the relative performance of the translation-based approach, Table 18 categorizes results by language and task. Each language is classified into three categories per task: translation-based results were worse than native-language BERT (✗), comparable (∼), or better (✓). The performance of the translation-based approach is considered comparable to that of native-language BERT models when it degrades by no more than 2–3 percentage points across all evaluation metrics.
As shown, the translation-based method using English BERT performed comparably or better than native models in 56.7% of all cases. This suggests that translation is a viable alternative in over half of the settings studied, especially when native resources are limited or English BERT has strong task-specific pretraining. Notably, languages such as Dutch and Bulgarian achieved comparable or better performance in the majority of tasks.
However, in 43.3% of cases, native-language models outperformed the translation approach. This was particularly evident for Chinese, which performed worse in 5 out of 6 tasks. This underperformance reinforces the importance of considering typological distance, translation quality, and the availability of strong native-language models when choosing between approaches.

7.4 Main Research Question

To what extent can a translation-based approach using English BERT obtain comparable or better performance than native-language BERT models?
Our findings suggest that the translation-based approach can serve as a viable alternative to native-language BERT models, but its effectiveness is highly dependent on both the language and the specific NLP task. Across all evaluated settings, the translation-based method achieved comparable or superior performance in approximately 56.7% of cases, indicating that it can match or even exceed the performance of language-specific models under certain conditions.

	Bulgarian	Chinese	Dutch	Italian	Russian
Hate Speech Detection	∼	✗	✗	✗	∼
Question Answering	∼	✗	✓	✓	✓
Sentiment Analysis	✗	✗	∼	∼	∼
POS Tagging	∼	✓	✓	✗	✗
Named Entity Recognition	✗	✗	✓	✗	✗
Natural Language Inference	∼	✗	∼	✓	✓

Table 18: Performance of translation-based models compared to native-language models across tasks and languages. Green ticks indicate better performance, red Xs indicate worse performance, and tildes indicate comparable results.

This performance advantage is most evident in languages closely related to English, such as Dutch, where shared linguistic structures like syntax and word order are more likely to be preserved during translation. For Slavic languages like Russian and Bulgarian, the approach was still competitive, with comparable or better results in four out of six tasks, despite greater typological distance. On the other hand, Chinese consistently showed degraded performance with the translation-based method. This can be attributed to the significant structural differences between English and Chinese, the potential loss of semantic nuances during translation, and the high quality of the native Chinese BERT model, which likely captures language-specific features more effectively.

The effectiveness of the translation-based approach also varied by task. It was particularly well-suited to: question answering, where English BERT's strong pretraining and robust span-prediction capabilities transferred well; POS tagging, which relies more on syntactic patterns than semantic nuance; and Natural Language Inference, where the premise-hypothesis format is relatively translation-stable. In contrast, the method proved less effective for NER and hate speech detection: NER is sensitive to token-level disruptions caused by translation, especially when word boundaries and entity spans are not preserved, while hate speech detection, is characterized by subtle, often culturally grounded expressions of offensive language that were frequently lost or neutralized in translation.

In summary, while translation-based fine-tuning with English BERT cannot universally replace native-language models, it is a promising approach in resource-constrained scenarios, especially for structurally similar languages and tasks less reliant on precise lexical or cultural features. Its success depends on the linguistic proximity to English, the task's sensitivity to translation artifacts, and the availability and quality of native-language resources.

8 Future Work

While this study has demonstrated the potential and limitations of translation-based cross-lingual transfer using English BERT, several aspects remain open for further exploration and improvement:

- **Statistical Significance Testing**: One limitation of the current analysis is the absence of statistical significance testing. While performance differences were often clear, formal significance tests would provide stronger evidence for the reliability of observed trends. Incorporating statistical testing would help validate whether improvements or degradations in performance are meaningful and consistent across multiple runs.

- **Expanding Language Coverage**: The experiments were limited to five target languages spanning a subset of language families. Future work could broaden the analysis to include a more diverse set of languages, particularly those from underrepresented or low-resource families such as Afro-Asiatic, or Dravidian. This would enable a more comprehensive understanding of how typological and genealogical factors influence translation-based cross-lingual performance.

- **Task Coverage in Hyperparameter Tuning**: Hyperparameter tuning was performed for four out of six tasks. A natural extension would be to fine-tune English BERT on translated training data for all tasks, ensuring a uniform comparison across settings. This could help clarify whether observed limitations stem from the translation process itself or from the absence of task-specific adaptation.

- **Model and Translation Quality**: Further improvements may also come from exploring different pretrained models or alternative translation techniques. This study relied on OPUS-MT due to its open-source availability and ease of integration, but these models may underperform compared to commercial systems like DeepL, particularly for complex sentence structures or low-resource languages. While we did not perform a systematic evaluation of translation accuracy, we observed notable artifacts in specific tasks. For example, in the Named Entity Recognition task, where the input consists of isolated words rather than full sentences, OPUS-MT frequently produced inconsistent or inaccurate translations. This is likely due to the model being optimized for sentence-level translation, making it ill-suited for word-level inputs. Future work could investigate the impact of using higher-quality translation tools to improve the overall performance of the methodology used in this thesis.

- **Exploring Large Language Models:** Although this study did not incorporate proprietary large language models (LLMs) such as ChatGPT or GPT-4, future work could investigate their capabilities in zero-shot or few-shot settings. These models have shown strong performance in various NLP tasks without task-specific training. However, several reasons justified their exclusion from this study. To begin with, the experimental setup was particularly targeted towards evaluating translation-based transfer learning, i.e., training and comparing fine-tuned models over original and translated data. This level of experimental control is not achievable with proprietary LLMs, which operate through prompt-based inference and not through supervised fine-tuning. In addition, most of the tasks within this thesis, including POS tagging, Named Entity Recognition, and Question Answering, require token-level or span-aligned predictions. These are not

entirely backed by ChatGPT, which provides no structured outputs such as token indices or IOB tags, and would require complex prompt engineering. GPT-based systems also are not open-source, which limits reproducibility and transparency, essentially the foundation of academic research. Their behavior may change over time due to backend updates, and their outputs are stochastic in nature, so repeated testing is challenging. Pragmatic constraints such as cost, API quotas, and privacy concerns regarding the data also decrease the feasibility of relying on GPT-based systems for this study. For these reasons, open-source models like multilingual BERT and OPUS-MT were adopted, offering greater flexibility, full offline operation, and controlled fine-tuning. It is important to notice that BERT and similar encoder-based architectures were originally developed and optimized for downstream NLP tasks, while models like GPT are primarily designed for generative tasks.

In summary, this study lays a foundation for understanding the efficacy of translation-based transfer across languages and tasks, but also highlights the need for more rigorous and comprehensive evaluation in future research.

9 Conclusion

This thesis investigated the central research question: "Does a translation-based approach using English BERT perform consistently across languages from different linguistic families and across different NLP tasks?"
To address this, we evaluated the effectiveness of fine-tuning the English BERT model on machine-translated data for six NLP tasks: Sentiment Analysis, Hate Speech Detection, Question Answering, Named Entity Recognition, Part-of-Speech Tagging, and Natural Language Inference. For each task, we considered datasets from five different languages: Bulgarian, Chinese, Dutch, Italian, and Russian. The goal was to assess whether translation can serve as a viable cross-lingual strategy in scenarios where native-language models or resources are limited.
The results showed that the translation-based approach does not perform consistently across all languages or tasks. In practice, its effectiveness is influenced by both the linguistic proximity of the target language to English and the nature of the NLP task. This aligns partially with theoretical expectations from prior work, which suggest that typological similarity can enhance transfer learning. However, our findings also highlight exceptions and limitations.

9.1 Subquestion 1

Does a translation-based approach using English BERT perform consistently across languages from different linguistic families?
The results indicate inconsistent performance across languages. For instance, Dutch, a Germanic language closely related to English, benefited the most from translation, achieving comparable or superior results to native-language models in most tasks. In contrast, Chinese, from the Sino-Tibetan family, consistently underperformed, especially in tasks requiring precise syntactic or token-level alignment. This suggests that linguistic distance and structural divergence negatively impact the effectiveness of translation-based transfer. Nevertheless, languages within the same family (e.g., Bulgarian and Russian) still showed varied outcomes, indicating that other factors, such as translation quality and model robustness, also play a role.

9.2 Subquestion 2

Are there specific NLP tasks where this approach is more effective?
Yes. Translation-based models were most effective in tasks with lower reliance on precise semantic or token-level information, such as Question Answering, POS tagging, and Natural Language Inference. These tasks often preserved enough structural information through translation for English BERT to perform well. In contrast, tasks such as Named Entity Recognition and Hate Speech Detection suffered significant degradation, largely due to alignment errors and the loss of culturally or contextually grounded information. Therefore, task characteristics, such as reliance on exact boundaries or cultural nuance, strongly influence transfer effectiveness.

9.3 Subquestion 3

How does fine-tuning English BERT on translated text compare to using pre-trained native-language BERT models?

When compared to native-language BERT models, the translation-based approach achieved comparable or better results in approximately 56.7% of evaluated cases, while native models outperformed it in 43.3% of cases. This suggests that while translation is a viable alternative when native resources are unavailable, it is not a universal substitute. Particularly for Chinese, native models consistently outperformed the translated approach, highlighting the need for language-specific models in certain cases.

9.4 Main Research Question

In conclusion, a translation-based approach using English BERT can be effective, but its performance is not consistent across languages or tasks. Its success depends on multiple factors including typological proximity to English, task-specific requirements, and the quality of translation. While it offers a practical alternative in low-resource settings, especially for syntactically compatible languages and less semantically demanding tasks, it should not be assumed to be universally applicable. Careful consideration must be given to both linguistic and task-specific characteristics when adopting this strategy in cross-lingual NLP applications.

References

[1] GSD - Traditional Chinese Universal Dependencies Treebank. https://github.com/UniversalDependencies/UD_Chinese-GSD/blob/master/README.md. Accessed: 02-05-2025.

[2] Sheikh Muhammad Abdullah. Text Preprocessing | NLP | Steps to Process Text. https://www.kaggle.com/code/abdmental01/text-preprocessing-nlp-steps-to-process-text/notebook, 2024. Accessed: 12-05-2025.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[4] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*, 2020.

[5] Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert. In *CLEF (Working Notes)*, pages 1–15, 2019.

[6] Reinald Kim Amplayo, Kyungjae Lee, Jinyeong Yeo, and Seung-won Hwang. Translations as additional contexts for sentence classification. *arXiv preprint arXiv:1806.05516*, 2018.

[7] Tejaswini Ananthanarayana, Priyanshu Srivastava, Akash Chintha, Akhil Santha, Brian Landy, Joseph Panaro, Andre Webster, Nikunj Kotecha, Shagan Sah, Thomastine Sarchet, et al. Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing (TACCESS)*, 14(4):1–30, 2021.

[8] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.

[9] AWS. What is Sentiment Analysis? https://aws.amazon.com/what-is/sentiment-analysis/. Accessed: 16/05/2025.

[10] Alexandra Balahur and Marco Turchi. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60, 2012.

[11] Emirhan Balcı and Esra Saraç. Automated depression detection from tweets: a comparison of nlp techniques. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE, 2024.

[12] Alessandro Lo Bello. Italian tripadvisor reviews comment dataset. https://www.kaggle.com/datasets/alessandrolobello/italian-tripadvisor, 2023. Accessed: January 27, 2025.

[13] Nikolay Bogoychev and Rico Sennrich. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*, 2019.

[14] Florian Borchert, Ignacio Llorca, Roland Roller, Bert Arnrich, and Matthieu-P Schapranow. xmen: a modular toolkit for cross-lingual medical entity normalization. *JAMIA open*, 8(1):ooae147, 2025.

[15] Gosse Bouma and Gertjan van Noord. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden, May 2017. Association for Computational Linguistics.

[16] Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431, 2023.

[17] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*, 2023.

[18] Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. Dalc: the dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*, online, August 2021. Association for Computational Linguistics.

[19] Rosario Catelli, Luca Bevilacqua, Nicola Mariniello, Vladimiro Scotto Di Carlo, Massimo Magaldi, Hamido Fujita, Giuseppe De Pietro, and Massimo Esposito. A new italian cultural heritage data set: Detecting fake reviews with bert and electra leveraging the sentiment. *IEEE Access*, 11:52214–52225, 2023.

[20] Yllias Chali, Sadid A Hasan, and Shafiq R Joty. Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. *Information Processing & Management*, 47(6):843–855, 2011.

[21] Nicola Cirillo. Italian sentiment analysis. `https://github.com/nicolaCirillo/italian-sentiment-analysis`, 2018. Accessed: January 27, 2025.

[22] Google Cloud. Hyperparameter tuning overview. `https://cloud.google.com/bigquery/docs/hp-tuning-overview#:~:text=In%20machine%20learning%2C%20hyperparameter%20tuning,a%20linear%20model%20are%20learned.` Accessed: 16/05/2025.

[23] Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[24] Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*, 2019.

[25] Wietse de Vries, Martijn Wieling, and Malvina Nissim. Dumb: A benchmark for smart evaluation of dutch models. *arXiv preprint arXiv:2305.13026*, 2023.

[26] Erkin Demirtas and Mykola Pechenizkiy. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8, 2013.

[27] Jiawen Deng, Jingyan Zhou, Hao Sun, Fei Mi, and Minlie Huang. Cold: A benchmark for chinese offensive language detection. pages 11580–11599, December 2022.

[28] Leon Derczynski. Hate speech data. `https://github.com/leondz/hatespeechdata`, 2020. Accessed: January 27, 2025.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[30] Antje Dörendahl, Nora Leich, Benedikt Hummel, Gilbert Schönfelder, and Barbara Grune. Overview of the clef ehealth 2019 multilingual information extraction. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, 2019.

[31] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online (v2020.4)*. Zenodo, 2013.

[32] Kevin Duh, Akinori Fujino, and Masaaki Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 429–433, 2011.

[33] Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. Sberquad – russian reading comprehension dataset: Description and analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 3–15. Springer International Publishing, 2020.

[34] DA Evseev. Query generation for answering complex questions in russian using a syntax parser. *Scientific and Technical Information Processing*, 49(5):310–316, 2022.

[35] Hugging Face. BERT 101 State Of The Art NLP Model Explained. `https://huggingface.co/blog/bert-101`, 2022. Accessed: 28/03/2025.

[36] Hong Fang, Guangjie Jiang, and Desheng Li. Sentiment analysis based on chinese bert and fused deep neural networks for sentence-level chinese e-commerce product reviews. *Systems Science & Control Engineering*, 10(1):802–810, 2022.

[37] Schram R. De Bruin J. Bagheri A. Oberski D. L. Tummers L. Van de Schoot R. Ferdinands, G. Active learning for screening prioritization in systematic reviews – A simulation study. `https://asreview.nl/project/simulation_study_1/`, 2020. Accessed: 28/03/2025.

[38] Lifang Fu and Shuai Liu. A syntax-based bsgcn model for chinese implicit sentiment analysis with multi-classification. In *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–7. IEEE, 2022.

[39] Muhammad Jauharul Fuadvy and Roliana Ibrahim. Multilingual sentiment analysis on social media disaster data. In *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, volume 6, pages 269–272. IEEE, 2019.

[40] Fernando Gallego and Francisco J Veredas. Recognition and normalization of multilingual symptom entities using in-domain-adapted bert models and classification layers. *Database*, 2024:baae087, 2024.

[41] Ming Gao, Mengshi Li, Tianyao Ji, Nanfang Wang, Guowu Lin, and Qinghua Wu. Key technologies of intelligent question-answering system for power system rules and regulations based on improved bertserini algorithm. *Processes*, 12(1):58, 2023.

[42] AA Golubev and NV Loukachevitch. Use of bert neural network models for sentiment analysis in russian. *Automatic Documentation and Mathematical Linguistics*, 55:17–25, 2021.

[43] Saroj Gopali, Faranak Abri, Akbar Siami Namin, and Keith S Jones. The applicability of llms in generating textual samples for analysis of imbalanced datasets. *IEEE Access*, 2024.

[44] Mika Hämäläinen, Khalid Alnajjar, and Thierry Poibeau. Video games as a corpus: Sentiment analysis using fallout new vegas dialog. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, pages 1–4, 2022.

[45] Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Ves Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. bgglue: A bulgarian general language understanding evaluation benchmark. *arXiv preprint arXiv:2306.02349*, 2023.

[46] Glenn Hiemstra. Dutch sentiment analysis. `https://github.com/Glender/DutchSentimentAnalysis`, 2021. Accessed: January 27, 2025.

[47] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

[48] HuggingFace. Datasets: Paul/hatecheck-dutch. `https://huggingface.co/datasets/Paul/hatecheck-dutch`. Accessed: February 23, 2025.

[49] IBM. What is NLP (natural language processing)? . `https://www.ibm.com/think/topics/natural-language-processing`, 2024. Accessed: 28/03/2025).

[50] Amsterdam Internships. Automatic-answering-of-city-council-questions. `https://github.com/Amsterdam-Internships/Automatic-Answering-of-City-Council-Questions`, 2023. Accessed: January 27, 2025.

[51] Lee Hao Jie, Ramesh Kumar Ayyasamy, Anbuselvan Sangodiah, Norazira Binti A Jalil, Kesavan Krishnan, and P Chinnasamy. The role of ernie model in analyzing hotel reviews using chinese sentiment analysis. In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE, 2023.

[52] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.

[53] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[54] Anna Jurek, Maurice D Mulvenna, and Yaxin Bi. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4:1–13, 2015.

[55] Borislav Kapukaranov and Preslav Nakov. Fine-grained sentiment analysis for movie reviews in Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274, Hissar, Bulgaria, September 2015. INCOMA Ltd. Shoumen, BULGARIA.

[56] Nattawat Khamphakdee and Pusadee Seresangtakul. An efficient deep learning for thai sentiment analysis. *Data*, 8(5):90, 2023.

[57] Maksim A Kosterin and Ilya V Paramonov. Neural network sentiment classification of russian sentences into four classes. *Automatic Control and Computer Sciences*, 57(7):727–739, 2023.

[58] Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*, 2019.

[59] Labelf. What is Accuracy, Precision, Recall and F1 Score? `https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score`. Accessed: 23/05/2025.

[60] Jian Lan, Wei Liu, YangYang Hu, and JunJie Zhang. Semantic parsing and text generation of complex questions answering based on deep learning and knowledge graph. In *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, pages 201–207. IEEE, 2021.

[61] Hongjing Li and Lin Li. A joint model of entity recognition and predicate mapping for chinese knowledge base question answering. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6. IEEE, 2020.

[62] Mingzheng Li, Lei Chen, Jing Zhao, and Qiang Li. Sentiment analysis of chinese stock reviews based on bert model. *Applied Intelligence*, 51:5016–5024, 2021.

[63] Xinlu Li, Yuanyuan Lei, and Shengwei Ji. Bert-and bilstm-based sentiment analysis of online chinese buzzwords. *Future Internet*, 14(11):332, 2022.

[64] Daniele Licari and Giovanni Comandè. Italian-legal-bert models for improving natural language processing tasks in the italian legal domain. *Computer Law & Security Review*, 52:105908, 2024.

[65] Sun Lina, Konstantin A Aksyonov, and Wu Shiying. Based on runewscorp: Improving accuracy in long text classification. In *2024 International Russian Automation Conference (RusAutoCon)*, pages 589–594. IEEE, 2024.

[66] Shuang Liu, Nannan Tan, Yaqian Ge, and Niko Lukač. Research on automatic question answering of generative knowledge graph based on pointer network. *Information*, 12(3):136, 2021.

[67] Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*, 2021.

[68] Olga Lyashevkaya, Kira Droganova, Daniel Zeman, Maria Alexeeva, Tatiana Gavrilova, Nina Mustafina, and Elena Shakurova. *Universal dependencies for russian: A new syntactic dependencies tagset*. SSRN, 2017.

[69] Charles Malafosse. Open dataset for sentiment analysis. `https://github.com/charlesmalafosse/open-dataset-for-sentiment-analysis`, 2018. Accessed: January 27, 2025.

[70] Iva Marinova, Kiril Simov, and Petya Osenova. Transformer-based language models for bulgarian. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 712–720, 2023.

[71] Ilia Markov, Lisa Hilte, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. Facebook metadata dataset LiLaH-HAG, 2022. Slovenian language resource repository CLARIN.SI.

[72] Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34, 2022.

[73] Antonio Martínez-García, Toni Badia, and Jeremy Barnes. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, 2021.

[74] Anmol Kohli (Medium). Notes on bert tokenizer and model. `https://medium.com/@anmolkohli/my-notes-on-bert-tokenizer-and-model-98dc22d0b64`. Accessed: February 18, 2025.

[75] Hitech BPO Medium. Which bert model is best for your entity annotation project? `https://hitechbpo.medium.com/which-bert-model-is-best-for-your-entity-annotation-project-f69bd576a3fd`, 2023. Accessed: February 27, 2025.

[76] Prasad Patil Medium. What is exploratory data analysis? `https://medium.com/towards-data-science/exploratory-data-analysis-8fc1cb20fd15`, 2018. Accessed: February 4, 2025.

[77] Xinfan Meng, Furu Wei, Ge Xu, Longkai Zhang, Xiaohua Liu, Ming Zhou, and Houfeng Wang. Lost in translations? building sentiment lexicons using context based machine translation. In *Proceedings of COLING 2012: Posters*, pages 829–838, 2012.

[78] Stefano Menini, Rachele Sprugnoli, and Antonio Uva. "who was pietro badoglio?" towards a qa system for italian history. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 430–435, 2016.

[79] Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603, 2024.

[80] Anamaria-Monica MIGEA, Vlad-Andrei NEGRU, TOMA Sebastian-Antonio, Camelia LEMNARU, and Rodica POTOLEA. Cook smarter not harder: Enhancing learning capacity in smart ovens with supplementary data. In *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 1–8. IEEE, 2024.

[81] Miquido. What is Natural Language Inference (NLI)? `https://www.miquido.com/ai-glossary/natural-language-inference/#:~:text=Natural%20Language%20Inference%20(NLI)%20is%20a%20foundational%20task%20in%20machine,%2Dchecking%2C%20and%20text%20summarization`. Accessed: 16/05/2025.

[82] Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *CLEF (Working Notes)*, pages 179–203, 2022.

[83] Wajid Hassan Moosa and Najiba. Multi-lingual hatespeech dataset, 2022.

[84] Hielke Muizelaar, Marcel Haas, Koert van Dortmont, Peter van der Putten, and Marco Spruit. Extracting patient lifestyle characteristics from dutch clinical text with bert models. *BMC medical informatics and decision making*, 24(1):151, 2024.

[85] Vlad-Andrei Negru, Vasile Suciu, Alex-Mihai Lăpușan, Camelia Lemnaru, Mihaela Dînșoreanu, and Rodica Potolea. Assessing language models' task and language transfer capabilities for sentiment analysis in dialog data. *Computer Speech & Language*, 89:101704, 2025.

[86] Utkarsh Mittal Neha Keshari, Durga Malladi. Hate Speech Detection Using Natural Language Processing. *Stanford CS224N Custom Project*.

[87] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*, 2020.

[88] Opus NLP. OPUS-MT Dashboard: Compare Models. `https://opus.nlpl.eu/dashboard/compare.php?model1=opusmt%2FOPUS-MT-models%2Fit-en%2Fopus-2019-12-18&model2=external%2Fhuggingface%2Ffacebook%2Fm2m100_418M`. Accessed: 19/05/2025.

[89] Debora Nozza, Federico Bianchi, and Dirk Hovy. "HONEST: Measuring hurtful sentence completion in language models". In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics.

[90] Anaïs Ollagnier, Elena Cabrio, and Serena Villata. Unsupervised fine-grained hate speech target community detection and characterisation on social media. *Social Network Analysis and Mining*, 13(1):58, 2023.

[91] Petya Osenova and Kiril Simov. Btb-tr05: Bultreebank stylebook ą 05. Technical report, 2004.

[92] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544, 2021.

[93] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1946–1958, 2017.

[94] Petko Petkov. Question answering with books translated to bulgarian. `https://huggingface.co/datasets/petkopetkov/QABGB`. Accessed: January 27, 2025.

[95] Truong HV Phan and Phuc Do. Ner2ques: combining named entity recognition and sequence to sequence to automatically generating vietnamese questions. *Neural Computing and Applications*, 34(2):1593–1612, 2022.

[96] Antonio Piizzi, Donatello Vavallo, Gaetano Lazzo, Saverio Dimola, and Elvira Zazzera. A natural language processing model for the development of an italian-language chatbot for public administration. *International Journal of Advanced Computer Science & Applications*, 15(9), 2024.

[97] Matúš Pikuliak, Marián Šimko, and Mária Bieliková. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765, 2021.

[98] Ruben Piperno, Luca Bacco, Felice Dell'Orletta, Mario Merone, and Leandro Pecchia. Cross-lingual distillation for domain knowledge transfer with sentence transformers. *Knowledge-Based Systems*, page 113079, 2025.

[99] Marco Pota, Mirko Ventura, Rosario Catelli, and Massimo Esposito. An effective bert-based pipeline for twitter sentiment analysis: A case study in italian. *Sensors*, 21(1):133, 2020.

[100] A Pravalika, Vishvesh Oza, NP Meghana, and S Sowmya Kamath. Domain-specific sentiment analysis approaches for code-mixed social network data. In *2017 8th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–6. IEEE, 2017.

[101] Pavel Přibáň, Jakub Šmíd, Josef Steinberger, and Adam Mištera. A comparative study of cross-lingual sentiment analysis. *Expert Systems with Applications*, 247:123247, 2024.

[102] Ekaterina Pronoza, Polina Panicheva, Olessia Koltsova, and Paolo Rosso. Detecting ethnicity-targeted hate speech in russian social media texts. *Information Processing & Management*, 58(6):102674, 2021.

[103] Ruilin Qi, Hang Li, and Qingchen Zhang. Attention-based brcnn for chinese medical question answering. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3933–3937. IEEE, 2023.

[104] Radoslav Ralev and Jürgen Pfeffer. Hate speech classification in bulgarian. In *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 49–58, 2022.

[105] A Ramponi, B Testa, S Tonelli, and E Jezek. Addressing religious hate online: from taxonomy creation to automated detection, peerj computer science 8 (2022) e1128. *URL: https://doi. org/10.7717/peerj-cs*, 1128.

[106] Eshrag Refaee and Verena Rieser. Benchmarking machine translated sentiment analysis for arabic tweets. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: student research workshop*, pages 71–78, 2015.

[107] Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.

[108] Kamil Saitov and Leon Derczynski. Abusive language recognition in russian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 20–25, 2021.

[109] Mustafa Salıcı and Üyesi Ercan Ölçer. Impact of transformer-based models in nlp: An in-depth study on bert and gpt. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–6. IEEE, 2024.

[110] Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. Postwita-ud: an italian twitter treebank in universal dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[111] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), May 2018, Miyazaki, Japan*, pages 2798–2895, 2018.

[112] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*, 2018.

[113] Stefan Schweter. Italian bert and electra models, November 2020.

[114] Burr Settles. *Active Learning*. Springer Cham, 2012.

[115] Deming Sheng and Jingling Yuan. An efficient long chinese text sentiment analysis method using bert-based models with bigru. In *2021 ieee 24th international conference on computer supported cooperative work in design (cscwd)*, pages 192–197. IEEE, 2021.

[116] Maximos Skandalis, Richard Moot, Christian Retoré, and Simon Robillard. New datasets for automatic detection of textual entailment and of contradictions between sentences in French. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12173–12186, Torino, Italy, May 2024. ELRA and ICCL.

[117] Smartling. Why use neural machine translation and where can you start? `https://www.smartling.com/blog/neural-machine-translation`, 2025. Accessed: June 4, 2025.

[118] Sergey Smetanin and Mikhail Komarov. Deep transfer learning baselines for sentiment analysis in russian. *Information Processing & Management*, 58(3):102484, 2021.

[119] Hajung Sohn and Hyunju Lee. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE, 2019.

[120] Arpita Soni. Enhancing multilingual table-to-text generation with qa blueprints: Overcoming challenges in low-resource languages. In *2024 International Conference on Signal Processing and Advance Research in Computing (SPARC)*, volume 1, pages 1–7. IEEE, 2024.

[121] SophonPlus. Chinese nlp corpus. `https://github.com/SophonPlus/ChineseNlpCorPus`, 2018. Accessed: February 2, 2025.

[122] Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 51–59, 2018.

[123] Fuhong Tang and Kwankamol Nongpong. Chinese sentiment analysis based on lightweight character-level bert. In *2021 13th International Conference on Knowledge and Smart Technology (KST)*, pages 27–32. IEEE, 2021.

[124] N Donald Jefferson Thabah, Aiom Minnette Mitri, Goutam Saha, Arnab Kumar Maji, and Bipul Shyam Purkayastha. A deep connection to khasi language through pre-trained embedding. *Innovations in Systems and Software Engineering*, pages 1–15, 2022.

[125] Cuk Tho, Yaya Heryadi, Iman Herwidiana Kartowisastro, and Widodo Budiharto. A comparison of lexicon-based and transformer-based sentiment analysis on code-mixed of low-resource languages. In *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, volume 1, pages 81–85. IEEE, 2021.

[126] Jörg Tiedemann and Santhosh Thottingal. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation, 2020.

[127] Alexander Tsertsvadze, Yen-Fu Chen, David Moher, Paul Sutcliffe, and Noel McCarthy. How to conduct systematic reviews more expeditiously? *Systematic reviews*, 4:1–6, 2015.

[128] TsvetoslavVasev. Ontology for toxic language and filters for automatic toxic language detection in bulgarian text. `https://github.com/TsvetoslavVasev/toxic-language-classification/blob/main/README.md`, 2024. Accessed: February 23, 2025.

[129] Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6034–6038. IEEE, 2018.

[130] Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133, 2021.

[131] Benjamin van der Burgh and Suzan Verberne. The merits of universal language model fine-tuning for small datasets - a case with dutch book reviews. *CoRR*, abs/1910.00896, 2019.

[132] Max van Haastrecht, Injy Sarhan, Bilge Yigit Ozkan, Matthieu Brinkhuis, and Marco Spruit. Symbals: A systematic review methodology blending active learning and snowballing. *Frontiers in research metrics and analytics*, 6:685591, 2021.

[133] Kiet Van Nguyen, Nhat Duy Nguyen, Phong Nguyen-Thuan Do, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. Vireader: A wikipedia-based vietnamese reading comprehension system using transfer learning. *Journal of Intelligent & Fuzzy Systems*, 41(1):1993–2011, 2021.

[134] Chaïm van Toledo, Marijn Schraagen, Friso van Dijk, Matthieu Brinkhuis, and Marco Spruit. Exploring the utility of dutch question answering datasets for human resource contact centres. *Information*, 13(11):513, 2022.

[135] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[136] David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In *WASSA 2015, the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–8. Association for Computational Linguistics, 2015.

[137] Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. Progress in machine translation. *Engineering*, 18:143–153, 2022.

[138] Juan Wang, Li Hou, Yunhan Li, Yueping Sun, Jiaming Li, and Li Yang. Classifying public health questions using large language models. In *Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science*, pages 735–740, 2024.

[139] Ziniu Wang, Zhilin Huang, and Jianling Gao. Chinese text classification method based on bert word embedding. In *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence*, pages 66–71, 2020.

[140] Haryo Akbarianto Wibowo, Tatag Aziz Prawiro, Muhammad Ihsan, Alham Fikri Aji, Radityo Eko Prasojo, Rahmad Mahendra, and Suci Fitriany. Semi-supervised low-resource style transfer of indonesian informal to formal language with iterative forward-translation. In *2020 International Conference on Asian Language Processing (IALP)*, pages 310–315. IEEE, 2020.

[141] Rüdiger Wirth and Jochen Hipp. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000.

[142] Hanqian Wu, Zhike Wang, Feng Qing, and Shoushan Li. Reinforced transformer with cross-lingual distillation for cross-lingual aspect sentiment classification. *Electronics*, 10(3):270, 2021.

[143] Jiaye Wu, Jie Liu, and Xudong Luo. Few-shot legal knowledge question answering system for covid-19 epidemic. In *Proceedings of the 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–6, 2020.

[144] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*, 2020.

[145] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[146] Yunze Xiao, Houda Bouamor, and Wajdi Zaghouani. Chinese offensive language detection: Current status and future directions. *arXiv preprint arXiv:2403.18314*, 2024.

[147] Weijia Xu, Batool Haider, and Saab Mansour. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*, 2020.

[148] Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. *arXiv preprint arXiv:2402.10712*, 2024.

[149] Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11):275, 2021.

[150] Zeynep Yirmibeşoğlu and Tunga Güngör. Morphologically motivated input variations and data augmentation in turkish-english neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–31, 2023.

[151] Reza Zadkamali, Saeedeh Momtazi, and Hossein Zeinali. Intent detection and slot filling for persian: Cross-lingual training for low-resource languages. *Natural Language Processing*, pages 1–16.

[152] He Zhang and Muhammad Ali Babar. Systematic reviews in software engineering: An empirical investigation. *Information and software technology*, 55(7):1341–1354, 2013.

[153] Lei Zhang, Pengfei Xia, Xiaoxuan Ma, Chengwei Yang, and Xin Ding. Enhanced chinese named entity recognition with multi-granularity bert adapter and efficient global pointer. *Complex & Intelligent Systems*, 10(3):4473–4491, 2024.

[154] Lingli Zhang, Yadong Wu, Qikai Chu, Pan Li, Guijuan Wang, Weihan Zhang, Yu Qiu, and Yi Li. Sa-model: Multi-feature fusion poetic sentiment analysis based on a hybrid word vector model. *CMES-Computer Modeling in Engineering & Sciences*, 137(1), 2023.

[155] Shunxiang Zhang, Hongbin Yu, and Guangli Zhu. An emotional classification method of chinese short comment text based on electra. *Connection Science*, 34(1):254–273, 2022.

[156] Xiuhao Zhao, Zhao Li, Shiwei Wu, Yiming Zhan, and Chao Zhang. Deep text matching in medical question answering system. In *Proceedings of the 2021 ACM International Conference on Intelligent Computing and its Emerging Applications*, pages 134–138, 2021.

[157] Meifeng Zhou, Jindian Tan, Song Yang, Haixia Wang, Lin Wang, and Zhifeng Xiao. Ensemble transfer learning on augmented domain resources for oncological named entity recognition in chinese clinical records. *IEEE Access*, 11:80416–80428, 2023.

[158] Tianyi Zhou, Qingchun Hu, Junzhe Li, and Linhao Wu. Design and development of a sentiment analysis system for chinese online comment texts. In *2023 3rd International Conference on Electronic Information Engineering and Computer (EIECT)*, pages 312–317. IEEE, 2023.

[159] Yuan Zhuang, Boyan Liu, Xiaotao Lin, and Canhao Xu. V-sbert: A mixture model for closed-domain question-answering systems based on natural language processing and deep learning. In *2023 6th International Conference on Data Science and Information Technology (DSIT)*, pages 328–333. IEEE, 2023.

A Modified Query

As explained in Section 3, we have modified the query for SLR Question 3 to obtain more diverse results, since most of the retrieved studies focused Chinese sentiment analysis, while almost no relevant papers were found for Bulgarian tasks. The modified query we employed is the following:

SLR Q3 modified query:

```
(
    ("Abstract":dataset)
    AND ("Abstract":Question Answering OR "Abstract":Textual Entailment
        OR "Abstract":Natural Language Inference OR "Abstract":NLI
        OR "Abstract":Sentiment Analysis OR "Abstract":Hate Speech Detection)
    AND ("Abstract":Italian  OR "Abstract":Bulgarian OR "Abstract":Dutch
        OR "Abstract":Russian OR "Abstract":Chinese
)
```

B Prior Knowledge Papers

In this section, we provide the list of papers manually labeled as relevant or not relevant in order to supply prior knowledge to the ASReview model during the active learning phase of the systematic literature review. These labels are used to guide the machine learning algorithm in identifying relevant literature. The first column indicates whether each paper was selected to support the first or the second SLR research question.

Table 19: List of papers manually labeled as prior knowledge to initialize the ASReview model. These include both relevant and not relevant examples, used to guide the initial training phase of the active learning process.

Used for	Paper Title	Label
SLR Q1	Extracting patient lifestyle characteristics from Dutch clinical text with BERT models [84]	Relevant
	MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT [5]	Relevant
	Deep Learning Methods for Sign Language Translation [7]	Not Relevant
	Recognition and normalization of multilingual symptom entities using in-domain-adapted BERT models and classification layers [40]	Relevant
	NER2QUES: combining named entity recognition and sequence to sequence to automatically generating Vietnamese questions [95]	Not Relevant
	New Italian Cultural Heritage Data Set: Detecting Fake Reviews With BERT and ELECTRA Leveraging the Sentiment [19]	Relevant
	The Applicability of LLMs in Generating Textual Samples for Analysis of Imbalanced Datasets [43]	Not Relevant

Used for	Paper Title	Label
SLR Q2	Unsupervised fine-grained hate speech target community detection and characterisation on social media [90]	Not Relevant
	ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain [64]	Not Relevant
	A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection [92]	Relevant
	Automated Depression Detection from Tweets: a Comparison of NLP Techniques [11]	Not Relevant
	Assessing language models' task and language transfer capabilities for sentiment analysis in dialog data [85]	Not Relevant
	Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP Tasks in Telugu Language [72]	Relevant
	An Efficient Deep Learning for Thai Sentiment Analysis [56]	Relevant
	Introducing Various Semantic Models for Amharic: Experimentation and Evaluation with Multiple Tasks and Datasets [149]	Relevant
SLR Q3	Design and Development of a Sentiment Analysis System for Chinese Online Comment Texts [158]	Relevant
	Semantic Parsing and Text Generation of Complex Questions Answering Based on Deep Learning and Knowledge Graph [60]	Relevant
	A Syntax-based BSGCN Model for Chinese Implicit Sentiment Analysis with Multi-classification [38]	Relevant
	ViReader: A Wikipedia-based Vietnamese reading comprehension system using transfer learning [133]	Not Relevant
	V-SBERT: A Mixture Model for Closed-Domain Question-Answering Systems Based on Natural Language Processing and Deep Learning [159]	Relevant
	Neural Network Sentiment Classification of Russian Sentences into Four Classes [57]	Relevant
	Exploring the Utility of Dutch Question Answering Datasets for Human Resource Contact Centres [134]	Relevant
	An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian [99]	Relevant
	Enhanced Chinese named entity recognition with multi-granularity BERT adapter and efficient global pointer [153]	Non Relevant
	Few-Shot Legal Knowledge Question Answering System for COVID-19 Epidemic [143]	Non Relevant

C Relevant Papers

In this section we report the tables showing the list of relevant papers obtained after applying the SYMBALS methodology for the first two SLR questions.

ID	Paper Title	Number of relevant references
Q1.1	*Extracting patient lifestyle characteristics from Dutch clinical text with BERT models [84]	1
Q1.2	*MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT [5]	2
Q1.3	Recognition and normalization of multilingual symptom entities using in-domain-adapted BERT models and classification layers [40]	3
Q1.4	Cross-lingual distillation for domain knowledge transfer with sentence transformers [98]	6
Q1.5	Intent detection and slot filling for Persian: Cross-lingual training for low-resource languages [151]	3
Q1.6	Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation [140]	1
Q1.7	A Comparison of Lexicon-based and Transformer- based Sentiment Analysis on Code-mixed of Low-Resource Languages [125]	5
Q1.8	Morphologically Motivated Input Variations and Data Augmentation in Turkish-English Neural Machine Translation [150]	0
Q1.9	Enhancing Multilingual Table-to-Text Generation with QA Blueprints: Overcoming Challenges in Low-Resource Languages [120]	N/A
Q1.10	Cook Smarter Not Harder: Enhancing Learning Capacity in Smart Ovens with Supplementary Data [80]	N/A

Table 20: List of relevant papers obtained for SLR Q1 after active learning. The number of relevant references found during backward snowballing for each paper is also reported. *N/A* indicates that the stopping criterion for backward snowballing was met.

From paper	Papers Found
Q1.2	• Translations as additional contexts for sentence classification [6] • Overview of the CLEF eHealth 2019 Multilingual Information Extraction [30]

Continued on next page

From paper	Papers Found
Q1.3	• Overview of DisTEMIST at BioASQ: automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources [82] • xMEN: a modular toolkit for cross-lingual medical entity normalization [14] • Spanish pre-trained BERT model and evaluation data [17]
Q1.4	• No language left behind: Scaling human-centered machine translation [23] • An empirical study on cross-lingual vocabulary adaptation for efficient language model inference [148] • A comparative study of cross-lingual sentiment analysis [101] • Localizing in-domain adaptation of transformer-based biomedical language models [16] • Cross-lingual learning for text processing: A survey [97] • Reinforced transformer with cross-lingual distillation for cross-lingual aspect sentiment classification [142]
Q1.5	• Cross-lingual transfer learning for multilingual task-oriented dialog [112] • (Almost) zero-shot cross-lingual spoken language understanding [129] • End-to-end slot alignment and recognition for cross-lingual NLU [147]
Q1.6	• Domain, translationese, and noise in synthetic data for neural machine translation [13]

Continued on next page

From paper	Papers Found
Q1.7	• Multilingual sentiment analysis on social media disaster data [39] • Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora [136] • Domain-specific sentiment analysis approaches for code-mixed social network data [100] • Improved lexicon-based sentiment analysis for social media analytics [54] • Benchmarking Machine Translated Sentiment Analysis for Arabic Tweets [106]

Table 21: List of relevant papers obtained for SLR Q1 after backward snowballing. Column *From paper* indicates the paper ID (see Table 20) that, during backward snowballing, yielded to the list of papers on the right.

From paper	Papers Found
Q1.2	• Deep learning models for multilingual hate speech detection [4] • Cross-lingual polarity detection with machine translation [26] • Is machine translation ripe for cross-lingual sentiment classification? [32] • The state and fate of linguistic diversity and inclusion in the NLP world [52] • Lost in translations? building sentiment lexicons using context based machine translation [77]
Q1.3	• Are all languages created equal in multilingual BERT? [144]
Q1.4	• WangchanBERTa: Pretraining Transformer-Based Thai Language Models [67]

Table 23: List of relevant papers obtained for SLR Q2 after backward snowballing. Column *From paper* indicates the paper ID (see Table 20) that, during backward snowballing, yielded to the list of papers on the right.

ID	Paper Title	Number of relevant references
Q2.1	A New Italian Cultural Heritage Data Set: Detecting Fake Reviews With BERT and ELECTRA Leveraging the Sentiment [19]	0
Q2.2	A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection [92]	5
Q2.3	Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP Tasks in Telugu Language [72]	1
Q2.4	An Efficient Deep Learning for Thai Sentiment Analysis [56]	1
Q2.5	Introducing Various Semantic Models for Amharic: Experimentation and Evaluation with Multiple Tasks and Datasets [149]	N/A
Q2.6	A deep connection to Khasi language through pre-trained embedding [124]	N/A
Q2.7	Ensemble transfer learning on augmented domain resources for oncological named entity recognition in Chinese clinical records [157]	N/A
Q2.8	MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations [119]	N/A
Q2.9	A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM [79]	N/A
Q2.10	Impact of Transformer-Based Models in NLP: An In-Depth Study on BERT and GPT [109]	N/A
Q2.11	Based on RuNewsCorp: Improving Accuracy in Long Text Classification [65]	N/A

Table 22: List of relevant papers obtained for SLR Q2 after active learning. The number of relevant references found during backward snowballing for each paper is also reported. *N/A* indicates that the stopping criterion for backward snowballing was met.

D Datasets

This section reports the list of datasets obtained after conducting the systematic literature review for the third SLR question. Each table shows the datasets, along with the papers in which they were found, for a specific language, reporting the results of sentiment analysis, hate speech detection, natural language inference and question answering.

Table 24: Sentiment Analysis Datasets found during SLR.

Language	Dataset	Source	Found During
Bulgarian	Cinexio Moview Reviews	[46]	Targeted Search
	Sentiment Analysis Data for the Bulgarian Language	[73]	Targeted Search
Chinese	Weibo Dataset	[158]	Active Learning
	ChnSentiCorp	[51]	Active Learning
	Chinese Stock Reviews Dataset	[62]	Active Learning
	waimai_10k, online_shopping, simplify-weibo_4_moods	[121]	Active Learning
	weibo_senti_100k3	[123]	Active Learning
	Internet News Sentiment Analysi	[115]	Active Learning
	Chinese Buzzwords	[63]	Active Learning
	Book Reviews Dataset	[36]	Active Learning
	bigboNed	[155]	Active Learning
	THUNLPAIPoet	[154]	Active Learning
Dutch	Dutch Book Reviews Dataset v3.0	[25]	Targeted Search
	Dutch Sentiment Analysis	[46]	Targeted Search
Italian	Fallout Dataset	[44]	Active Learning
	Italian tweets dataset	[69]	Targeted Search
	Italian TripAdvisor Reviews Comment Dataset	[12]	Targeted Search
	Italian Sentiment Analysis	[21]	Targeted Search
Russian	ROMIP2012, SentiRuEval-2015-banks, SentiRuEval-2015-telecoms, SentiRuEval-2016-banks, SentiRuEval-2016-telecoms	[42]	Active Learning
	RuTweetCorp, Twitter Sentiment for 15 European Languages, Kaggle Russian_twitter_sentiment, RuSentiment, Kaggle Russian News Dataset, Kaggle Sentiment Analysis Dataset, RuReviews	[118]	Active Learning

Table 25: Question Answering Datasets found during SLR.

Language	Dataset	Source	Found During
Bulgarian	QABGB	[94]	Targeted Search
	Sentiment Analysis Data for the Bulgarian Language	[73]	Targeted Search
Chinese	FAQ Dataset	[159]	Active Learning
	DuReader	[139]	Active Learning
	CMedQA, CMedQA2	[103]	Active Learning
	WebQA	[66]	Active Learning
	NLPCC 2016 KBQA	[61]	Active Learning
	CHIP-STS	[156]	Active Learning
	CMRC2018 (Chinese Machine Reading Comprehen- sion 2018), DRCD (Delta Reading Comprehension Dataset)	[41]	Active Learning
	CMID (Chinese Medical Intent Dataset)	[138]	Active Learning
Dutch	Question and answer dataset National Personnel Portal	[134]	Targeted Search
	Automatic Quesion Answering of City Council Question	[50]	Targeted Search
	SQuAD-NL	[25]	Targeted Search
Italian	SQuAD-IT	[96]	Active Learning
	QUANDHO	[78]	Targeted Search
Russian	RuBQ Dataset	[34]	Active Learning
	XQuAD dataset	[8]	Targeted Search
	SberQuAD	[33]	Targeted Search

Table 26: Hate Speech Detection Datasets found during SLR.

Language	Dataset	Source	Found During
Bulgarian	Hate Speech Classification in Bulgarian	[104]	Targeted Search
	Toxic Language Classification	[128]	Targeted Search
Chinese	COLD, TOCP, TOCAB, SWSR, CoLA, TOXICN	[146]	Targeted Search

Continued on next page...

Language	QA Example	Source	Found During
Dutch	LiLaH-HAG	[71]	Active Learning
	DALC v1.0	[18]	Targeted Search
	DALC v2.0	[25]	Targeted Search
Italian	Religious Hate Speech	[105]	Active Learning
	HONEST	[89]	Targeted Search
	IHSC (Italian Hate Speech Corpus)	[111]	Targeted Search
	Whatsapp Dataset	[122]	Targeted Search
Russian	RuEthnoHate	[102]	Active Learning
	Automatic Toxic Comment Detection in Social Media for Russian, Detection of Abusive Speech for Mixed So- ciolects of Russian and Ukrainian Languages, Russian South Park	[28]	Targeted Search

Table 27: NLI Datasets found during SLR.

Language	Dataset	Source	Found During
Bulgarian	Cross-lingual Natural Language Inference	[45]	Targeted Search
	XNLI	[47]	Targeted Search
Chinese	XNLI	[47]	Targeted Search
Dutch	SICK-NL	[25]	Targeted Search
Italian	RTE-3, LingNLI	[116]	Targeted Search
Russian	XNLI	[47]	Targeted Search