Institute for Natural Language Processing (IMS)
University of Stuttgart
Pfaffenwaldring 5 b
70569 Stuttgart

Masterarbeit

Multilingual Prompt Engineering Via Large Language Models: An Approach To Sentiment Analysis

Pascal Huszár

Course of Study:	M. Sc. Informatik
Examiner:	Prof. Dr. Roman Klinger
Supervisor:	Prof. Dr. Roman Klinger Prof. Dr. Jeremy Barnes
Commenced:	31.07.2023
Completed:	31.01.2024

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die einge- reichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.

Statement of Authorship

This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

Pascal Huszár

Abstract

Exploring the efficacy of multilingual prompt engineering for sentiment analysis reveals a promising avenue for extending the adaptability of large language models (LLMs) beyond the confines of the primary predominant English. The core ambition revolves around devising strategies for transferring adept English instructions into the target language. These strategies exploit the remarkable capability of large language models to extract information and learn new task by the context of a few demonstrations – known as in-context learning. In this research, the strategies leverage both monolingual and cross-lingual prompt templates, augmented with demonstrations. Furthermore, the process of instruction generation is supported by an iterative rephrasing approach that refines instructions into their optimal counterparts.

The investigation unfolds through a careful analysis of how multilingual instruction generation benefits from incorporating demonstrations, either in English or the target language, within the prompt template. Results substantiate that iteratively rephrasing instructions further improves the effectiveness of the instruction generation process, underscoring the proficiency of large language models to follow the request.

Through this exploration, it emerges that the automatic prompt engineering methods exhibit potential in multilingual contexts. The findings advocate for a broader utilization of demonstration learning and iterative refinement techniques in multilingual prompt engineering, aiming to universalize the application of large language model across diverse communities and languages. This study not only fills the gap identified in previous research regarding the effectiveness of automatic prompt engineering methods for non-English languages but also facilitates broader access for linguistic communities to generative AI.

Kurzfassung

Die Erforschung der Wirksamkeit von mehrsprachigem Prompt-Engineering für die Sentiment-Analyse zeigt einen vielversprechenden Weg auf, um die Anpassungsfähigkeit von großen Sprachmodellen (LLMs) über die Grenzen des primär vorherrschenden Englisch hinaus zu erweitern. Das Hauptanliegen besteht darin, Strategien für die Übertragung geschickter englischer Anweisungen in die Zielsprache zu entwickeln. Diese Strategien nutzen die bemerkenswerte Fähigkeit großer Sprachmodelle, Informationen zu extrahieren und neue Aufgaben aus dem Kontext einiger weniger Demonstrationen zu lernen - bekannt als kontextbezogenes Lernen. In dieser Forschung nutzen die Strategien sowohl einsprachige als auch sprachübergreifende Aufforderungsvorlagen, die durch Demonstrationen ergänzt werden. Darüber hinaus wird der Prozess der Anweisungsgenerierung durch einen iterativen Umformulierungsansatz unterstützt, der die Anweisungen zu ihren optimalen Gegenstücken verfeinert.

Im Rahmen der Untersuchung wird sorgfältig analysiert, inwieweit die Generierung mehrsprachiger Instruktionen von der Einbeziehung von Demonstrationen, entweder in Englisch oder in der Zielsprache, in die Instruktionsvorlage profitiert. Die Ergebnisse belegen, dass die iterative Umformulierung von Anweisungen die Effektivität des Anweisungserstellungsprozesses weiter verbessert und die Fähigkeit großer Sprachmodelle unterstreicht, der Aufforderung zu folgen.

Durch diese Untersuchung wird deutlich, dass die Methoden zur automatischen Erstellung von Eingabeaufforderungen in mehrsprachigen Kontexten Potenzial aufweisen. Die Ergebnisse sprechen für einen breiteren Einsatz von Demonstrationslernen und iterativen Verfeinerungstechniken bei der Entwicklung mehrsprachiger Prompts mit dem Ziel, die Anwendung großer Sprachmodelle in verschiedenen Gemeinschaften und Sprachen zu universalisieren. Diese Studie schließt nicht nur die Lücke, die in der bisherigen Forschung in Bezug auf die Effektivität automatischer Prompt-Engineering-Methoden für nicht-englische Sprachen identifiziert wurde, sondern erleichtert auch einen breiteren Zugang zu generativer KI für Sprachgemeinschaften.

Contents

1 Introduction **11**

2 Background and Related Work **19**

2.1 Multilingual Natural Language Processing 19

2.1.1 The Significance 19

2.1.2 Challenges and Techniques 20

2.1.3 Language-Specific and Multilingual Language Models 23

2.2 Text Classification 25

2.2.1 Challenges and Wins 25

2.2.2 Overview Techniques 25

2.2.3 Multilingual Text Classification 27

2.2.4 Few-Shot Classification 28

2.2.5 Zero-Shot Classification 30

2.2.6 Sentiment Analysis 30

2.3 Prompting Methods 33

2.3.1 Prompt Engineering 33

2.3.2 In-context Learning 36

2.3.3 Large Language Models 37

3 Methods **39**

3.1 Automatic Multilingual Prompt Engineering 39

3.2 Instruction Generation 39

3.2.1 Prompt Templates 42

3.2.2 Instruction Selection 44

3.2.3 Rephrasing the Instructions 44

3.2.4 Data Collection 46

3.2.5 Selection of Language Models 47

4 Results **49**

4.1 Effectiveness of Demonstrations on Target Language Instruction Generation . . . 49

4.1.1 Inclusion versus Exclusion 50

4.1.2 English versus Target Language Demonstrations 50

4.2 Effectiveness of Demonstrations on Instruction Rephrasing 54

5 Discussion **59**

5.1 Limitations 61

6 Conclusion and Future Work **63**

A Prompt Templates **65**

B Language Codes **69**

Chapter 1

Introduction

Sentiment Analysis signifies a computational approach intrinsic to Natural Language Processing (NLP) that seeks to decode, understand and classify text-based sentiments. The varied emotions - positive, negative, neutral, and beyond - weave fundamental insights into individual or collective perspectives related to various subjects or events. This revolutionary computational tool has rendered valuable contributions across diverse spheres, including market analysis (Rao and Srivastava, 2012), social media (Stieglitz and Dang-Xuan, 2013), politic campaigns (Bhaumik et al., 2023), customer opinions on goods (Keung, Lu, Szarvas, et al., 2020). By filtering and examining data from tweets, articles, reviews or user comments, subjective details emerge that help companies gain critical insights into public sentiment trends and customer opinions.

Although Sentiment Analysis has repeatedly demonstrated significant applicability, it is inevitably accompanied by specific challenges and limitations that researchers need to acknowledge. A principal challenge is the intrinsic richness and diversity of human language, including elements such as sarcasm, irony, context-dependent meanings and cultural nuances, which often constitute obstacles. (Blanco and Lourenço, 2022). Misunderstandings may arise from lexical ambiguities - words with different possible interpretations depending on their usage or phrases whose sentiment prone to vary based on additions or absence of certain words (Pang and Lee, 2008). Moreover, translating sentiments across languages remains another considerable challenge. Translation complexity due to different grammar rules, vocabulary variances across languages and lack of standard translation for slangs or colloquial terms complicates sentiment analysis. Another limitation involves handling neutral instances in multi-class Sentiment Analysis tasks. These instances often get mistakenly classified into either positive or negative categories due to insufficient representation in training datasets (Koppel and Schler, 2006).

Sentiment Analysis historically commenced with Rule-based approaches, leveraging sets of predefined rules or heuristics tailored to identify sentiment within the text. These methods typically rely on sentiment lexicons—comprehensive lists of words and phrases annotated with their corresponding sentiment polarity values. For instance, the presence of words such as “happy“,“thrilled“ or“excellent“ within a text would be indicative of a positive sentiment, whereas words like “disappointed“, “terrible“ or “sad“ suggest a negative sentiment. A concrete example of a rule-based sentiment analysis system is the use of VADER, which not only accounts for sentiment lexicons but also incorporates grammatical and syntactical rules to gauge sentiment intensity of texts, especially short social media postings (Hutto and Gilbert, 2014). Although Rule-based models are straightforward and interpretable, their major limitation lies in their rigidity and the extensive manual effort required to create and update the rule sets and lexicons for different domains or languages, potentially limiting their adaptability and scalability.

Machine learning approaches in sentiment analysis leverage algorithms and statistical methods to process data and generate a model. These techniques employ various types of data, including bag of words, n-grams, and part-of-speech tags to train on a corpus of labeled sentiment data and to develop models that predict sentiment on unseen data. One notable implementation of a machine learning approach is Support Vector Machines (SVM) (Sun et al., 2009), which has been demonstrated to produce strong performance in categorizing text sentiment (Pang and Lee, 2008). Machine learning techniques can capture complex patterns that might be missed by rule-based systems. Yet, they suffer from bias and inaccuracies if the training data suffers from imbalance, inconsistency, or is not representative of the target text to analyze (Golbeck et al., 2011).

Deep learning techniques for sentiment analysis offer more sophisticated algorithms and models that can pick up deeper syntactic and semantic characteristics from text data, enhancing the performance particularly on complex tasks. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) represent two notable deep learning algorithms employed in sentiment analysis (Kim, 2014; Qian et al., 2017). Deep learning models are capable of processing large amounts of data and can learn both short and long-term dependencies in sequences of words. Furthermore, they are proficient at understanding negation and deciphering sentiment in complex phrases (Tang et al., 2016). However, an essential limitation resides in their need for extensive labeled training data and computational resources. The ”black box”nature of these models also hampers their interpretability, thereby presenting challenges for practical implementation (Guidotti et al., 2018).

Pre-training language models on a large scale marked a revolutionary shift in sentiment analysis and related areas, moving away from the traditional approaches. Transformer-based models, as introduced by Vaswani et al. (2017), when trained on vast amounts of text data, demonstrated remarkable performance across a wide range of NLP tasks (X. Zhao et al., 2023) (Brown et al., 2020) (Bang et al., 2023). Initially, researchers and developers focused on pre-training language models following fine-tuning on downstream tasks such as named entity recognition (Devlin et al., 2019), sentiment analysis (Hoang et al., 2019) or machine translation (Zhu et al., 2023), among others.

Nonetheless, the capabilities of language models significantly increase with the scaling up of model size, extensive pre-training on larger text collections, and the utilization of sophisticated algorithms. These improvements have led to the addition of the prefix *large* to the term *language model* (W. X. Zhao et al., 2023). Typical characteristics of large language models are emergent abilities that occur above a certain level of scale. In-context learning is one of these abilities and refers to the adaption and learning from demonstrations provided within the input without requiring explicit retraining or fine-tuning on specialized datasets (Brown et al., 2020). By processing and using the information presented in the immediate text, these models can generate responses or continue text in a coherent manner. Different to the fine-tuning process, prompting a large language model with a natural language prompts involves no parameter updates, thus needs far less computational power and resources. This approach offers a more natural interaction with the underlying large language model. Users devise a prompt that encapsulates the task using natural language and optionally augment it with demonstrations illustrating the task’s resolution.

The act of directing a large language model's output through contextualization using prompts highlights the criticality of devising meaningful prompts, a process termed *prompt engineering* (P. Liu et al., 2023). This process is central to the efficacy of models in handling various downstream tasks, as the quality and nature of the prompts can significantly influence the performance of the language model. The input of descriptive and well-designed prompts enables the model to generate more accurate and relevant responses. To produce accurate and coherent responses, the language model must effectively adhere to the instructions provided within the given prompt. Utilizing a combination of multi-task datasets complemented with natural language instructions describing each task — a process known as instruction tuning — has demonstrated to improve the instruction-following capability of large language model on novel tasks (Ouyang et al., 2022; Sanh et al., 2021; Wei et al., 2022). Through this instruction tuning process, large language models acquire the capability to comprehend and follow instructions of unfamiliar tasks, thereby exhibiting an improved generalization. Numerous studies have demonstrated that large language models, when appropriately contextualized with prompts, exhibit a remarkable ability to solve a wide range of challenging NLP tasks, such as reading comprehension, question answering, mathematical reasoning, among others (Brown et al., 2020; Shin et al., 2020; Zhou et al., 2023). This ability is particularly remarkable because these models achieve such performance without undergoing any training for the tasks mentioned above.

The capability of large language models to solve NLP tasks is substantially influenced by their natural language comprehension, instruction following capability but also by the quality of prompts. Prompt engineering entails the systematic exploration and refinement of prompts, ensuring that the language model's text output is guided by the given context to accomplish a defined task. This refinement can be carried out through an intuitive, manual approach or through automated means. The manual method entails a series of trials, leveraging human intuition in formulating meaningful instructions. Brown et al. (2020) showcased GPT-3's proficiency in handling diverse tasks, such as translation and question answering, through the manual formulation of a task description and its integration into a prompt alongside the test instance. However, this approach comes with its downsides, including the possibility of errors, cognitive biases, the element of guesswork, and a significant investment of time (P. Liu et al., 2023). To overcome these limitations and fully exploit the potential of large language models in generating coherent, context-sensitive texts, research has shifted to automating the prompt engineering process. The underlying drive is to minimize human-induced errors and tap into the capacity of language models to craft natural language prompts autonomously.

The process of engineering prompts can be categorized into two distinct types: The first involves the creation of discrete, or 'hard', prompts, which take the form of text strings. Alternatively, there is the generation of continuous, or 'soft' prompts, which are represented directly within the embedding space of the language model itself. Figure 1.1 illustrates the difference of the two types.

The objective of designing prompts is to devise a strategy that enables a large language model to effectively execute a task, with the primary goal being performance rather than human readability. Hence, there is no need to confine the prompt to forms comprehensible in natural language. Consequently, researchers explore the concept of continuous prompts which involve initiating the prompting process within the model's embedding space itself. Prefix-tuning is a parameter-efficient approach for adapting large language models to specific tasks without the need to fine-tune all the

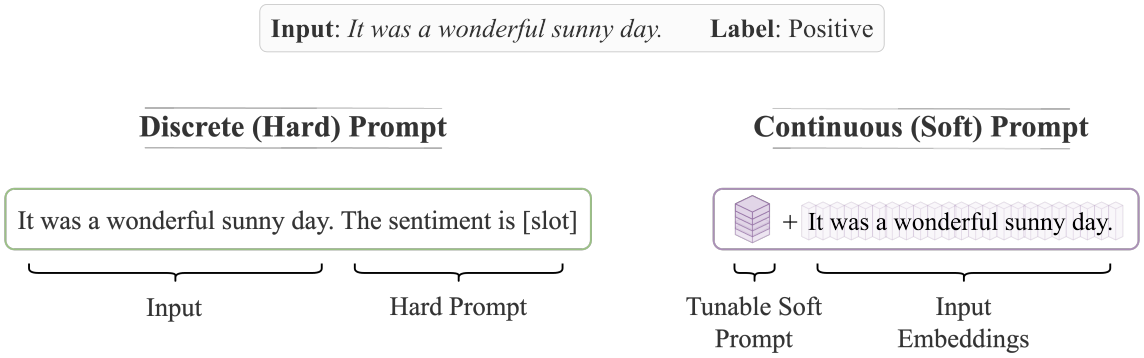


Figure 1.1: Comparison between discrete and continuous prompts. Discrete prompts process the input by inserting it into a template (illustrated by the green box), accompanied by a task description or instruction. The "[slot]" marks where the model begins its completion and is not part of the input string. Meanwhile, the continuous prompt, using the same input example but embedded, entails the insertion of a tunable embedding prior to the input. Both are combined an act as one input (illustrated by the purple box).

model parameters. Originally introduced by X. L. Li and Liang (2021) this technique involves a small set of trainable parameters, known as continuous prompts or prefixes, that are prepended to the input. Creating prompts in the continuous space ease some constraints such as the embedding of the prompt in natural language. Nevertheless, finding and optimizing prompts in the continuous space becomes less optimal at scale, as computing gradients requires more computational power (Zhou et al., 2023). Furthermore, continuous prompt are unintelligible for humans.

Research focusing on the identification of discrete prompts primarily seeks to discover representations within a discrete space, associated with phrases in natural language. Several approaches have been proposed for the creation of discrete prompts, including the paraphrasing of existing prompts (Haviv et al., 2021; Jiang et al., 2020) and the generation by large language models (T. Gao et al., 2021). The automated method of AutoPrompt (Shin et al., 2020) generates discrete prompts by combining the original input and *trigger* tokens. This uniform set of trigger tokens is found and optimized through a modified gradient-based search technique. The approach outperforms fine-tuning in low-data settings and demonstrates the ability of large language models to solve a diverse set of natural language processing tasks when properly prompted. In addition, compared to the fine-tuning process, prompting language models is more storage-efficient since it eliminates the need for extensive disk space to save model checkpoints. The research conducted by Zhou et al. (2023) illustrates the capacity for using large language models not only to generate natural language instructions but also to assess and select the most efficient ones through a method known as Automatic Prompt Engineering (APE). The authors guide the search for the best instructions for various NLP tasks employing in-context learning techniques. They construct a prompt template by hand, which directs the model to produce an instruction for a particular task. This template is enhanced with input-output demonstrations. The underlying assumption is that providing context to the language models with these demonstrations will steer the model towards producing more

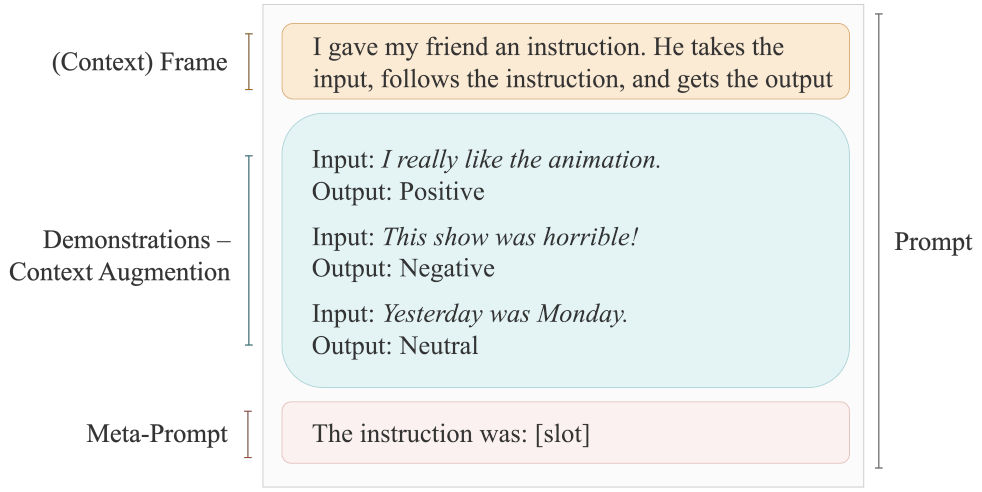


Figure 1.2: Depicts a prompt designed to create an instruction for a sentiment analysis task, illustrating a few-shot learning scenario. The language model receives the implicit task of deducing the intended task from provided completions demonstrations and context. In this case, a meta-prompt defines the implicit task or instruction that the model will follow. The symbol "[slot]" serves as an output indicator, marking the starting point for the model's output. Importantly, this output indicator is not part of the actual prompt. The example presented draws inspiration from the approach known as APE (Zhou et al., 2023).

effective instructions, potentially leading to improved performance. Figure 1.2 illustrates a prompt, contextualizing the language model on input-output pairs. The model is implicitly tasked to complete with an instruction appropriate to the input-output pairs.

The recent advancements in automatic prompt engineering have sparked interest and opened up new possibilities for automating the process of generating high-quality instructions. This development shows promise in enhancing the engineering process by leveraging the expertise of large language models, ultimately leading to more efficient and successful interactions with these models.

However, these developments have predominantly focused on English corpora, resulting in task solutions and established applications primarily in English. Although large language models demonstrate multilingual capabilities, only few work addressed prompt engineering in other languages than English (Fu et al., 2022; Lin et al., 2022; Tu et al., 2023; Winata et al., 2021). As a result, significant number of languages, especially those with scarce resources, remain under-explored in the realm of prompt engineering research. To date, there has been no exploration of strategies and methods to automatically generate discrete instructions in languages other than English. Ahuja et al. (2023) have conducted a comprehensive benchmark study assessing the multilingual capabilities of language models and proposed prompt strategies for in-context learning or the language of the prompt template. Research has yet to investigate whether these strategies are effective in automatically generating instructions for multiple non-English languages. P. Agrawal et al. (2023) emphasize the use of cross-lingual prompting, utilizing demonstrations in the target language combined with a template in English. Their analysis reveals improved outcomes with cross-lingual templates, show-

casing English as a good *bridge* language. However, their method aligns with continuous prompt tuning. Therefore fails to produce text that is understandable by humans, making it impossible to analyze.

This research centers around an in-context learning setting and investigates how in-lingual and cross-lingual prompt templates, augmented with demonstrations, influence the generation of target language instructions for multilingual sentiment analysis. The objective is to identify the optimal setup for a prompt template, augmented by example demonstrations. Such demonstrations, akin to the method known as APE (Zhou et al., 2023), showcase pairs of input and output for sentiment analysis. They can be obtained from a language with abundant resources or directly from the target language itself. It is believed that this method of learning within context can efficiently steer the creation of accurate sentiment analysis instructions for a target language. Furthermore, in-context learning with target language demonstrations is expected to improve the generation. An extensive evaluation and analysis are conducted to assess the effectiveness of prompts, both cross-lingual and in the target language, in producing instructions for this target language. Expanding research in the field of multilingual prompt engineering is crucial to ensure that the advancements in generative artificial intelligence are accessible and beneficial to all language communities equally (Ahuja et al., 2023).

The research is guided by the following central questions:

RQ1: Is there an improvement in the generation of target language instructions when demonstrations are integrated into prompt templates, relative to templates that exclude such demonstrations? Previous studies demonstrated that in-context learning benefits in NLP tasks (Brown et al., 2020; S. Gao et al., 2023). Integrating input-output demonstrations into the prompt template optimizes the automatic search and generation of optimal natural language instructions (Zhou et al., 2023). The subsequent evaluation and analysis expose well-crafted and refined instructions. This demonstrates that large language models possess the capability to deduce the underlying task from demonstrations, generate instructions out of the context, often leading to comparable or superior task performance This research also tackles the question of whether demonstrations advantage the instruction generation, but for some target language. To this end, comparing the generated target language instructions with their English counterparts will provide valuable insights.

RQ2: Does the use of target language demonstrations within prompt templates enhance the effectiveness of instruction generation in the same language as opposed to utilizing demonstrations in English? P. Agrawal et al. (2023) highlights the use of cross-lingual prompting. This experiment will address the questions whether the instruction generation for some target language can be refined by augmenting the prompt template with input-output demonstrations from the same language. This is contrasted by the generation with template augmented with English-only demonstrations.

RQ3: Is rephrasing of target language instructions with prompt templates that include demonstrations more effective in comparison to templates without such demonstrations? This question challenges the premise that generation of instructions benefits from integrating demonstrations of suboptimal instructions produced through previously mentioned methods. Zhou et al. (2023) suggested a slight enhancement when prompting language models with instruction rephrasing. Although, their approach concentrated on scenario without any in-context learning, so without any demonstrations. This segment of the study examines how few-shot learning with demonstrations might affects the iterative rephrasing of target language instructions.

The thesis is organized into multiple sections that lay the groundwork of the subject before delving into the core aspects of the thesis. Starting with an overview in section 2 over the theoretical background and previous work that this thesis builds upon. Section 3 proposes the approach and methods for automatic multilingual prompt engineering. Section 4 presents the results, which will be discussed in section 5. Finally, the thesis concludes and suggests a direction for future work in Chapter 6.

Chapter 2

Background and Related Work

2.1 Multilingual Natural Language Processing

Reflecting the dynamics of our globalized world, Natural Language Processing (NLP) has evolved beyond mono-lingual horizons, embracing multilingualism. This transition is crucial given the rich diversity of languages worldwide, necessitating broader reach and enhanced inclusion of NLP applications. As such, multilinguality emerged to a key and indispensable aspect of current NLP studies. It tracks the evolution from initial monolingual models to the present advanced multilingual paradigms, reflecting significant advances in technological development. However, the realm of multilingual NLP is not devoid of challenges. From the confrontation with cultural nuances, interpreting context-based idioms to code-switching among multilingual communities (El Bolock et al., 2020), several challenges follow this evolution. These are met with innovative techniques, including the use of parallel corpora for training models (Ramesh et al., 2022; Zeroual and Lakhouaja, 2020), alignment methods (Schuster et al., 2019; Cao et al., 2020) or exploiting cross-lingual language understanding by pre-trained language models (Conneau, Khandelwal, et al., 2020). Nonetheless, certain limitations persist, such as data scarcity for low-resource languages and a consistent bias towards high-resource ones. Figure 2.1 illustrates the distribution of the top seven content languages for websites on the internet compared to the top seven languages by number of worldwide speakers. The figure highlights a discrepancy: a substantial proportion (>50%) of online content is in English (W3Techs, 2024), despite only a fifth of the global population having comprehensive understanding of the language (Gary F. et al., 2023). But apart from these hurdles, the countless advantages of multilingual NLP underscore its importance. It eases language barriers as applications become more accessible through machine translation. It initiates novel pathways for global human-computer interaction and intercultural exchanges, and ultimately, stands as a catalyst for expanding the horizons of NLP.

2.1.1 The Significance

Globalization: Multilingualism, a natural consequence of our globalized world, is increasingly pivotal in the realm of NLP. As globalization intensifies, it fosters a world where multilingual individuals not only gain proficiency in additional languages but also immerse themselves in diverse cultures. This phenomenon enhances their openness and adaptability, contributing to a more interconnected global community (Nandi, 2022). In the academic realm, the multilingual approach to education, particularly in technical fields, is seen as essential. It equips future engineers with multilingual communicative competence, making them more competitive and effective in the global labor market. This reflects a broader educational trend where multilingualism is integrated into

professional training, acknowledging the strategic importance of language skills in a globalized workforce (Prokhorova, 2020). Therefore, the abundance of multilingual data on the internet, social media, and various domains necessitates the evolution of NLP systems capable of understanding and processing multiple languages efficiently.

Language Diversity: The disparity in language resources, as highlighted by Joshi et al. (2020) poses a significant challenge in multilingual NLP, particularly in representing the 7000+ global languages in new applications and evolving fields. They emphasize the need for a fair distribution of resources and question the ability of current models to deal with different languages. Building on this, Razumovskaia et al. (2022) emphasize the shortage of datasets for many non-English languages, which restricts the creation of truly multilingual NLP systems.

Code-Switching and Cross-lingual Text: Code-switching, the practice of alternating between languages within a single conversation or text, is a significant phenomenon in multilingual communities. It's not just about language choice, it reflects complex social and cultural dynamics (Poplack, 2001). El Bolock et al. (2020) demonstrated that factors like background and social context influence when and why individuals engage in code-switching, particularly in Egyptian Arabic-English contexts. This behavior imposes unique challenges for NLP systems, as highlighted by Ahmad et al. (2022), who researched the nuances in sentiment analysis of code-mixed social media text in India. The complexity of code-switching necessitates advanced NLP techniques to effectively process and understand the subtleties present in multilingual communication.

Ethical Considerations: In multilingual NLP, ethical considerations have evolved over time. Initially, the focus was on the social responsibility of NLP, highlighting the need to address biases like demographic disparities in data collection and developing de-biasing methods (Tsvetkov et al., 2018). More recently, research has identified the impact of high-resource languages influencing grammatical structures multilingual BERT, affecting fluency in less resourced languages (Papadimitriou et al., 2023). Evaluating bias and fairness in multilingual NLP is complex. It necessitates increased transparency in model documentation and expanding the scope of bias beyond gender to include various cultural contexts. This approach is crucial to address the nuanced power dynamics and consequences in developing large language models (Talat et al., 2022).

2.1.2 Challenges and Techniques

Multilingual Data Dealing with linguistic variations and cultural context represents a significant challenge in the domain of multilingual NLP. Cultural nuances and language-specific characteristics add layers of complexity to the interpretation of text. As all semantic characteristics reside within textual data and its representation, one of the primary focus in multilingual NLP lies in the realm of data: its collection, annotation, and processing. In the comprehensive study conducted by X. Yu et al. (2022), an exploration into the origins and methods for gathering and generating multilingual datasets was carried out. The authors identified that the primary reservoir of multilingual text data in various languages stems from news media, with web corpora and Wikipedia following closely behind. Regarding the creation of data for specific multilingual NLP tasks, they noted two prevalent approaches for dataset development: one involves automatically induced datasets, and the other widely adopted method is with help of native crowdworkers.

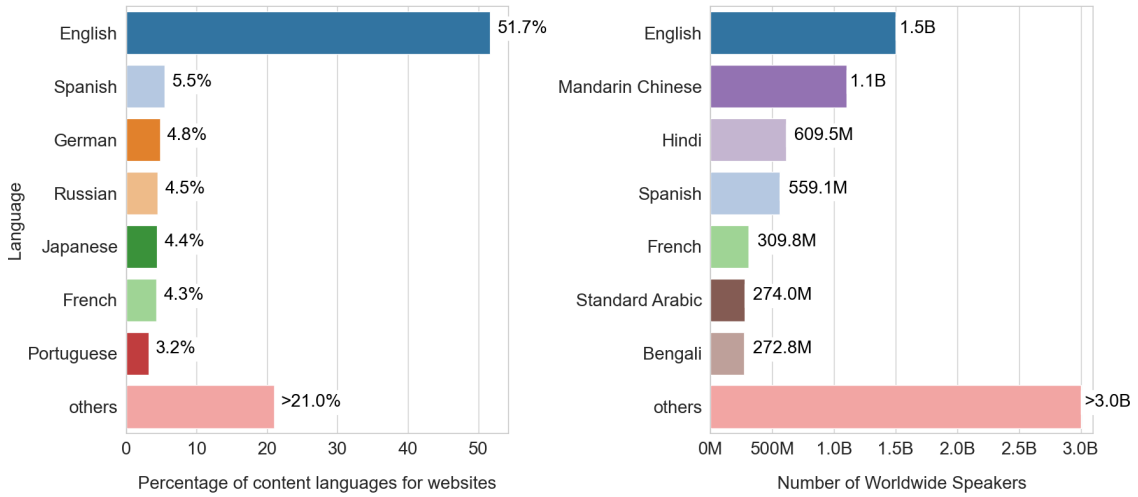


Figure 2.1: Comparison of the top 7 languages based on either their prevalence in web content (W3Techs, 2024) and the number of worldwide speakers (Gary F. et al., 2023). The bar chart on the left represents the proportionate distribution of content languages used on websites as of January 2024. English is the most prevalent language, making up over 50% of websites content. In comparison, the chart on right shows that English is spoken, as a first or second language, by 1.5 billion people. That is less than 20% of the worlds population. Furthermore, languages like Hindi or Arabic are under-represented in the web content languages.

The use of machine translation provides an alternative approach for generating multilingual datasets. It involves translating pre-existing datasets from high-resource languages to targeted low-resource languages. This builds an automatic approach for the creation of parallel datasets across diverse languages and decreases dependency on the small group of linguistically proficient annotators. Although this approach enables the automated translation of high-resource datasets for the creation of parallel data, it isn't free from challenges. The efficacy of the translation heavily relies on the techniques used, often presenting concerns such as a decline in quality for long sequences of text and translation inaccuracies as detailed by Clark et al. (2020).

The combination of automated methods for dataset creation and human-based annotations can yield datasets of immense significance for NLP applications. One major contribution to the research in Neural Machine Translation (NMT) and its applications was the release of the broadest parallel corpus encompassing 11 Indic languages. Through rigorous data collation from web-crawling and the employment of multilingual models to align sentences with the pivot language English, Ramesh et al. (2022) created a robust model for translating between the Indic languages and English. Other notable parallel corpora useful for multilingual NLP tasks include Europarl (Koehn, 2005), Tatoeba (Tiedemann, 2020), among others.

The Unified Multilingual Sentiment Analysis Benchmark (UMSAB) dataset (Barnes et al., 2018) encompasses a collection of tweets in multiple languages, notably including a blend of English and Hindi in Latin script. UMSAB's development aligns with the growing need for diverse, multilingual datasets that cater to the nuanced requirements of sentiment analysis in different linguistic contexts.

By combining data from social media platforms like Twitter, UMSAB offers a realistic and practical dataset for assessing and fine-tuning language models in sentiment analysis tasks. This dataset not only serves as a valuable resource for multilingual NLP but also aids in the advancement of research and applications in areas such as multilingual prompt engineering and in-context learning.

Multilingual Learning Robust multilingual or cross-lingual datasets for training and evaluating machine learning models represent a critical component of multilingual NLP. Nonetheless, in instances where access to annotated data is limited or human resources are scarce, sophisticated algorithms and techniques optimizing the utilization of existing data help to tackle these challenges. Multilingual applications have historically depended on labor-intensive processes, such as the creation of dictionaries or phrase tables, without any substantial use of contextual or syntactical information in their early iterations (Fung and Schultz, 2008). The idea of representing words as embeddings within vector spaces has steadily enhanced the ability of models to better generalize and estimate within the context. Aligning word embeddings in the embedding space for enabling cross-lingual transfer (for e.g. machine translation) was body of early research in multilingual NLP (Mikolov et al., 2013) (Smith et al., 2016). Complex models like multilingual BERT (Devlin et al., 2019), XLM-R (Conneau, Khandelwal, et al., 2020), LaBSE (Feng et al., 2022) or MUSE (Yang et al., 2020) enable to represent words context-dependent. This unified embedding space allows to richly capture the nuances and meaning of words, facilitating the transfer of context across various languages.

Delving into the techniques that facilitate multilingual NLP, the concept of transfer learning takes center stage. Within the various difficulties encountered, transfer learning has emerged as a prominent technique that enables machine learning models to deal with multiple languages. The ability to extract information from one domain and apply it to another is the essence of transfer learning (Weiss et al., 2016). In the domain of natural language, this implies that model are able to learn and abstract from one language (e.g. English) to another language (e.g. Thai). This paradigm yields notable improvements by recognizing and applying common language patterns across varied linguistic contexts. It provides these models with a starting point, leveraging pre-established knowledge from high-resource languages rather than expecting the model to start from zero. The adaptability of language models to perform well in problem setups with limited or no amount of data is referred to few-shot or zero-shot learning.

Zero-shot learning refers to the model's capacity to make predictions for multilingual tasks, in which the model is trained on high-resource languages and applied to unknown languages. Few-shot learning describes the model's proficiency in rapidly adapting to new tasks and languages with only a handful of examples. This often depicts the reality where a small of amount of labeled target language data is available. In their research, X. Chen et al. (2019) developed a versatile multilingual model that uniquely combines language-invariant and language-specific features using adversarial networks and a mixture-of-experts model. This approach enables effective cross-lingual transfer learning, particularly for low-resource languages. Cross-lingual transfer learning enables to train models for a target language using labeled data from various, often high-resource, languages. Significantly, the model operates efficiently in target languages with no data (zero-shot), demonstrating broad applicability to text classification and sequence tagging tasks across languages.

2.1.3 Language-Specific and Multilingual Language Models

With the advent of pre-trained language models, a wide range of NLP tasks experienced an outstanding improvement boost. BERT (Devlin et al., 2019) and later RoBERTa (Y. Liu et al., 2019) demonstrated the strong performance of transformer-based model to coherently process and learn the context from input. Initially these models were based on English-only language representation but fast researchers applied the architectures and models to the multilingual setting.

Multilingual Models Not long after the release of BERT, the same authors released a multilingual version pre-trained on 104 languages. Plenty studies demonstrated that fine-tuned models on downstream task such as NER (Pires et al., 2019) or POS-tagging (S. Wu and Dredze, 2019) can extract relevant features from the multilingual representations. Proposed by Conneau and Lample (2019) in their work on cross-lingual language model, their model XLM extends the capabilities of language representation beyond monolingual settings. A key extension to BERT's masked language model pre-training approach is its Translation Language Modeling (TLM) objective. This objective is designed to enhance the model's understanding of language by learning representations that are not just language-specific but also cross-lingual. The TLM objective achieves this by using parallel sentences in different languages, allowing the model to learn contextual relationships both within and across languages. This approach has been shown to significantly improve performance in machine translation and cross-lingual classification, as it enables the model to transfer knowledge and context from one language to another effectively.

The XLM-R model, introduced by Conneau, Khandelwal, et al. (2020) represents a significant enhancement over its predecessor, XLM. It was specifically designed to improve the performance of cross-lingual language understanding tasks by pre-training on a vast and diverse dataset consisting of one hundred languages. This expansive training approach led to substantial performance gains in a variety of cross-lingual benchmarks. Barbieri et al. (2022) pre-trained an XLM-R model on millions of tweets in over thirty languages, making it particularly adept at handling the unique linguistic characteristics of Twitter's diverse and informal text. One of its key features is the inclusion of unified sentiment analysis Twitter datasets in eight different languages, demonstrating its proficiency in capturing and analyzing sentiment across various languages and cultural contexts on social media platforms.

Given the recent success of prompting methods on multilingual large language models and the effect of scaling these models, an array of innovative models keeps extending the scope of multilingual NLP. An exemplary model that merits special mention is BLOOM, appreciated especially for its open-access multilingual approach (Workshop et al., 2023). This 176 billion parameter model was trained on the ROOTS corpus, which includes text in over 40 languages and various programming languages. BLOOM stands out for its competitive performance across various benchmarks, particularly after multitask prompted fine-tuning.

Numerous large language models, pre-trained on datasets of varying multilingual breadth, are currently available. The spectrum of these models continues to expand, reflecting advancements in the field. The HuggingFace model hub[1] exemplifies this trend by providing a comprehensive

[1] https://huggingface.co/models

repository of (multilingual) machine learning models (Wolf et al., 2020). Other prominent examples but not open-source is the GPT-n series (Radford, J. Wu, et al., 2019; Brown et al., 2020; OpenAI et al., 2023) or Google's Bard[2].

Language-Specific Conneau, Khandelwal, et al. (2020) was able to illustrate that a model pre-trained with a fixed capacity on an expanding set of languages only amplifies its cross-lingual performance to a certain extent, beyond which, a decline in performance was recorded. This occurrence has been termed as the *curse of multilinguality*. Moreover, other researchers have indicated that a more focused approach may yield more favorable results. Precisely, concentrating on a singular language and fine-tuning the model for specific operations within that language can generate more refined and promising results for subtasks. The prospect of overcoming the curse of multilinguality could potentially rest on these task-specific fine-tuning. With GELECTRA, a German based language model, the researchers demonstrated superior performance when pre-training on German data on hate speech classification and NER compared to models like multilingual BERT (Chan et al., 2020). Other examples for language-specific BERT-based models include french versions FlauBERT and CamemBERT (Le et al., 2020; Martin et al., 2020), Spanish BETO (Cañete et al., 2023) or Dutch RobBERT (Delobelle et al., 2020).

Almost a language-specific dataset, the remarkable generative capacity of GPT-3 (Brown et al., 2020) created great interest in both research and general media. A distinguishing feature of GPT-3 lies in the expansive and diverse dataset it has been trained on. Though primarily an English language model, noteworthy is the fact that it has been trained on data containing multiple languages, intentionally unfiltered. Despite the relatively small portion that non-English languages constitute in the dataset GPT-3 has been surprisingly adept in processing these languages. Conventionally, such limited data would be insufficient for attaining high-quality performance in low-resource languages. This intuition is backed by the vast amount of data needed for training language-specific models.

Transfer learning techniques allow models to leverage pre-trained representations on large monolingual corpora before fine-tuning on parallel data. Prominent language models, including GPT-3 (Brown et al., 2020), display exceptional few-shot and zero-shot learning capabilities in a multitude of languages beyond English. Arunkumar et al. (2023) explored the multilingual capabilities within predominantly English-trained large-scale language models. These models, despite their primary training in English, demonstrate a remarkable ability to understand and process multiple languages. This is attributed to the expansive and diverse nature of the data used for training, which includes instances of various languages. The research's findings are pivotal in the field of language modeling, indicating that even models primarily trained in one language can develop a level of understanding in multiple languages, although the degree of this understanding can vary based on several factors, including the representation of the language in the training data and the linguistic similarities between the languages.

[2] https://bard.google.com/

2.2 Text Classification

2.2.1 Challenges and Wins

Text classification is a fundamental task in natural language processing (NLP) where text data is categorized into predefined classes or categories. This process involves analyzing the content of text and assigning it to one or more classes based on its content. It's a widely applicable technique used in various domains, ranging from sentiment analysis, where the emotional tone of a text is determined (Taboada, 2016), to topic detection, which involves identifying the main themes or subjects of a text (Allan, 2002).

Although text classification benefits from automation and sophisticated algorithms, it faces significant challenges. The nature of text data, such as social media content or online reviews, is inherently unstructured, lacking a standard format and necessitating extensive processing for effective use. This data often includes extraneous elements like words or symbols that add no value, but make the classification task more complex. Additionally, the issue of class imbalance is prevalent, where some classes are disproportionately represented in datasets. This imbalance can lead to biased model predictions favoring the more dominant classes (Ali et al., 2015).

Even the nature of language itself complicates text classification. Sarcasm, idioms, and slang often require contextual understanding and nuances that are difficult for a computational model to grasp. For example, the word 'hot' can refer to a spicy dish or a high temperature, and discerning the correct meaning can be difficult for an algorithm without using context clues. Ideas and thoughts are not only expressed by means of characters and words. Emojis, for instance, add another layer of complexity to classification of textual data, as their visual form is used to express emotional state and augments the textual content with nuanced meanings (Bai et al., 2019).

Yet, overcoming these challenges holds significant benefits. Detailed and comprehensive text classification can facilitate more accurate sentiment analysis, making it a more trusted tool for data-driven decision making. It can also lead to more effective spam detection algorithms, preventing the propagation of unwanted content (J. Li et al., 2013). Well-defined text classification systems can also improve automated customer service, helping businesses understand and respond to their customers more effectively (Olujimi and Ade-Ibijola, 2023), underlining the importance of advancements in this area.

2.2.2 Overview Techniques

The early stages of text classification approaches witnessed the predominance of rule-based methods, characterized by their straightforward yet rigid *if-then* rules. For instance, a rule may state that "if a text contains words *shipping* or *semiconductor*, then map it to the category *trade*". In 2002, H. Li and Yamanishi introduced an approach for text classification using stochastic decision lists, exemplifying the rule-based method's advantages in readability and refinability, especially when dealing with texts belonging to specific categories or shorter in length. While being simple and interpretable, such predefined rules can be restrictive and fail to capture the complexity and nuanced meaning inherent in human language. Advancing this concept, X. Zhang et al. (2008) utilized a granular

computing approach in rule-based text classification, which demonstrated its effectiveness in certain scenarios, contrasting it with methods based on statistical theory. Later, in 2016, Kamaruddin et al. explored the challenges of associative classification in text due to the high dimensionality of data, proposing a modified Multi-Class Association Rule Method to enhance classification accuracy while managing the complexity inherent in the vast amount of generated rules.

Simple machine learning models like Naive Bayes, Support Vector Machines (SVM), and Logistic Regression have also been widely used (J. Chen et al., 2009; M. Liu et al., 2016; Sun et al., 2009). These methods start with feature extraction processes like TF-IDF or Bag-of-Words, transforming unstructured text into numerical vectors. For example, Naive Bayes uses probability theory to predict the class of a given text based on the features. Support Vector Machines work by finding hyperplanes that classify the data into different categories. Logistic Regression, on the other hand, uses the logistic function to predict probabilities of different classes. Asha et al. (2023) explored the effectiveness of Naive Bayes, Support Vector Machine, and other algorithms on Twitter data. Their study demonstrated that Naive Bayes outperformed other models, achieving higher accuracy in sentiment classification. However, the research had limitations, particularly in the context of word ambiguity and multi-polarity in sentiment analysis, which could affect the robustness of their findings. Similarly, Ho et al. (2019) proposed a two-stage data analytic approach for sentiment analysis on tweets using machine learning algorithms like Logistic Regression, Naive Bayes, and SVM, combined with combinatorial fusion. This approach aimed to address the challenge of imbalanced data content in social media. While this method showed promise, its effectiveness was contingent on the performance ratio of the algorithms used, indicating potential limitations in generalization.

These models, while being more flexible than rule-based systems, often struggle with the high-dimensionality of textual data. Furthermore they have difficulties to capture the dependencies among words, and lack the ability to comprehend semantic details. Advancements in computational power and data availability marked a transition towards modern, more sophisticated approaches to text classification. They involve the use of (contextual) word embeddings. Kilimci and Akyokuş (2019) addressed the challenge of Turkish document (text) classification by employing word embedding models like Word2Vec, GloVe, and FastText, along with deep learning architectures such as Recurrent Neural Networks (Schmidt, 2019), Long Short-Term Memory Units (Hochreiter and Schmidhuber, 1997), and Convolutional Neural Networks (Albawi et al., 2017). Their work highlighted the enhancement in text classification performance when combining deep learning algorithms with word embedding models. However, the study's limitation was its focus on Turkish language texts, potentially limiting the generalization of the findings to other languages or diverse datasets. Nassif and Fahkr (2020) investigated the combination of statistical and deep learning techniques, including HMM (Rabiner and Juang, 1986) and LSTM, with word embedding models like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) for supervised document classification. Their research demonstrated the importance of both the meaning and order of words in topic modeling, a perspective often overlooked in previous studies. However, their findings were constrained by the specific datasets they used, and their approach may require adaptations for different types of text data or classifications tasks. Around 2019, revolutionary pre-trained language models like BERT (Devlin et al., 2019) and RoBERTa (Y. Liu et al., 2019) emerged, considerably improving word representation through embeddings. Subsequent model iterations

brought refinements; however, they shared a common requirement for extensive labeled datasets of effectively tackle the downstream tasks. In instances of labeled data scarcity or complexity in real-world applications, more intricate methods become vital. Recent research in prompt-based learning seeks to more effectively leverage the inherent knowledge of pre-trained language models. This method involves reformulating the task into a prompt or instruction, with the anticipation that the desired outcome will appear in the language model's response (P. Liu et al., 2023). This technique capitalizes on the large language model's capacity to internalize the provided contextual information. Another advanced method involves employing models specifically trained on Natural Language Inference (NLI). Here, the model evaluates to what degree a premise, or the text in question, supports a hypothesis that corresponds to a certain label or class. The premise-hypothesis pair with the maximal entailment score is then designated as the result, referring to the text-label classification pairing (Yin et al., 2020).

2.2.3 Multilingual Text Classification

Building upon the foundational concepts of text classification introduced earlier, this area represents a significant leap in complexity and application, as it involves the analysis and categorization of text across multiple languages. Unlike monolingual text classification, multilingual text classification confronts additional challenges such as language variability, script differences, and cultural nuances.

Amine and Mimoun (2007), presents a method that utilizes WordNet for Multilingual Text Categorization, capitalizing on its linguistic database to classify documents in various languages under a common classification tree. This approach, while innovative for the time, may face limitations in languages not extensively covered by WordNet or in the context of linguistic nuances not captured by a standard database. On the other hand, Prof. Praveen Dhyani et al. (2015) explores the use of character n-gram frequencies for language identification and text categorization. This method, effective across various languages, might struggle with languages having less distinct n-gram profiles and can be challenged by texts containing multiple languages (code-switching).

Transfer learning and cross-lingual embeddings (Devlin et al., 2019) (Radford and Narasimhan, 2018) from language models equip NLP applications with significant word representations, enabling understanding and handling text classification across various languages, especially in low-resource settings (Conneau, Khandelwal, et al., 2020). Khalil et al. (2019), explores methods for multilingual intent classification in situations where resources are limited. This paper underscores the significance of recent contextual text representation advances and their potential in cross-lingual transfer, a crucial aspect in industrial applications where language resources may not be abundant. The research suggests a blend of techniques, including the use of machine-translated text, to achieve practical results. However, they acknowledge the limitations in the application of these methods, necessitating a combination of strategies for optimal outcomes. Research in the domains of NER, document classification (Keung, Lu, Szarvas, et al., 2020) and dependency parsing (Conneau and Lample, 2019) has observed similarly strong cross-lingual behavior in classification tasks, thus highlighting the benefits conferred by contextual embeddings from pre-trained large language models.

Cross-lingual text classification (CLTC) shares certain similarities with the challenge of multilingual text classification. The objective involves categorizing documents across various languages, utilizing a consistent taxonomy composed of pre-established categories. M'hamdi et al. (2019) emphasizes in their research the effectiveness of bilingual and multilingual word embeddings, especially in neural architectures for CLTC, offering significant advantages for low-resourced languages. However, their conclusion also points out that the performance of such embeddings can vary significantly based on the resources available for the target languages, highlighting a gap in the current capabilities of these methods

Prompt-based learning emerges as a promising approach for (multilingual) text classification, utilizing pre-trained language models and tailored prompts to achieve classification across various languages. Polyglot Prompt, by Fu et al. (2022), proposes an architectural framework for multitask multilingual learning. This framework, exploits prompting methods to develop a unified semantic space, accommodating different languages and tasks through multilingual prompt engineering. A notable result of this study is the successful demonstration of prompt-based learning's efficacy across multiple task and languages, illustrating its versatility. However, the authors acknowledge the complexity in balancing performance across this diversity of tasks and languages, underscoring the intricacies of finding universally effective prompts. Y. Chen et al. (2022) introduces an approach for prompt-based multilingual relation classification. The study's method efficiently constructs prompts from relation triples, requiring minimal translation for class labels. This approach proved superior in both fully supervised and few-shot scenarios, outperforming competitive baselines. However, the study confronts limitations in the dependency on the quality of prompts and the challenge of creating effective prompts for languages with limited resources, highlighting areas for future enhancement.

2.2.4 Few-Shot Classification

In the realm of machine learning, models can assimilate to new tasks or generate accurate predictions with a limited set of labeled examples (Brown et al., 2020). This capability of recognizing fundamental patterns within text is termed as few-shot learning. This approach is inspired by human cognitive abilities where one can often learn new concepts with only a few demonstrations. This learning paradigm is well suited for real-world applications, especially in domains where collecting large datasets is impractical or impossible due to constraints such as privacy, rarity of instances, or lack of linguistic expertise. For instance, in computer aided medical diagnostics, rare diseases may have only a handful of documented cases, yet a diagnostic model is expected to recognize and classify these cases accurately (W. Wang et al., 2023). The application of few-shot learning becomes handy in text classification as we now can achieve classification for new instances to rare classes with just a few classification examples. Initially, few-shot text classification with language models primarily involved leveraging pre-trained models and fine-tune them on a small number of examples per class. This approach, derived from transfer learning, adapts a model trained on a large dataset to perform a specific task with limited data. Experiments on few-shot classification typically relied on the k-shot, N-way framework, where tasks consist of N class labels, each with k instances for training. Such tasks are often extracted from a single, extensive dataset, ensuring task relevance to one another (M. Yu et al., 2018). (Howard and Ruder, 2018) demonstrate with their transfer learning method, called 'ULMFiT', strong few-shot classification performance for classes with only 100

Input: The happy cat sat on the mat.
Sentiment: Good

Input: It was a dark and stormy night.
Sentiment: Bad

Input: The fox jumps over the brown dog.
Sentiment: Normal

Input: I'm happy. I finally adopted the cat.
Sentiment: *[output]*

Figure 2.2: Formulation of a few-shot sentiment classification configuration. It exemplifies a prompt, augmented with few demonstrations of classification completion. A language model can process this input prompt and generate a desired output. "[output]" designates the start where the model's completion follows. The model derives the underlying task by its capability to rapidly adapt to new tasks, hence few-shot transfer learning. To be noted, the prompt does not incorporate an explicit task description.

labeled examples. Pre-trained language models, such as BERT (Devlin et al., 2019) or RoBERTa (Y. Liu et al., 2019), generally demonstrate strong performance in few-shot text classification, due to their capability transferring knowledge coherently across tasks. Classic fine-tuning on large dataset provides strong performance on the downstream task but require acquisition of the task-specific dataset and is computationally expensive. In few-shot scenarios, fine-tuning demonstrates moderate performance while maintaining low computation demands. Leveraging the extensive knowledge embedded within large pre-trained models by prompting methods has become feasible with the scaling of model sizes, particularly in the context of generative models like GPT-3 (Brown et al., 2020). In the realm of text classification, these prompting methods entail formulating tasks in natural language and soliciting the model to perform classification. Augmenting the prompt with a select number of demonstrations, a technique known as prompt augmentation, aligns with the few-shot learning paradigm and can be considered a variant of few-shot text classification (Brown et al., 2020). Rooted in the principles of few-shot learning, the inclusion of completion examples in prompt augmentation aids the model in classifying unseen instances or adapting to novel tasks. In Figure 2.2, an illustration of a example few-shot classification setting in which a few labeled examples contextualize the model, and showcase how the task could be solved. GPT-3 showcases strong few-shot performance across diverse NLP datasets, including translation, question-answering, and cloze test, as well as tasks requiring on-the-fly reasoning or domain adaptation (Brown et al., 2020).

Task descriptions can also enhance performance in few-shot setting when combined with supervised learning. Schick and Schütze (2021) introduces Pattern-Exploiting Training (PET), a semi-supervised training procedure. This method reformulates input examples as cloze-style phrases, which help language models understand a given task. The approach of PET is innovative in combining task descriptions in natural language with semi-supervised learning, providing a way to

Input: I'm happy. I finally adopted the cat.
Sentiment: *[output]*

Figure 2.3: Formulation of a zero-shot sentiment classification configuration. It exemplifies a zero-shot prompt, which will be the input to a language model. "[output]" designates the start where the model's completion follows. Demonstrations illustrating the resolution of the task are not supplied, thereby characterizing the approach as zero-shot. To be noted, the prompt does not incorporate an explicit task description. However, incorporating such a description does not conflict with the zero-shot setting.

leverage pre-trained language models more effectively. The study shows that PET outperforms both supervised training and other strong semi-supervised approaches by a significant margin across various tasks and languages.

2.2.5 Zero-Shot Classification

" the probability for *contradiction* is likely to be low, indicating a mismatch. Conversely, presenting the statement "The text is about [animals]"elicits a high probability in *entailment*, signifying alignment between the input and the hypothesis. Yin et al. (2020) authors propose unifying the zero-shot text classification of diverse aspects (topic, emotion, situation) within a textual entailment formulation. This approach studies the classification through the lens of entailment, evaluating whether a piece of text entails a certain label description. While this approach provides a comprehensive benchmark for future research, it faces limitations in its heavy reliance on the diversity of the datasets and potential challenges in accurately capturing and generalizing across varied and novel classification tasks. R. Zhong et al. (2021) presents an approach that revolves around meta-tuning, which involves fine-tuning pre-trained language models on a zero-shot learning objective using an aggregated collection of diverse datasets transformed into a unified question-answering format. This method aims to align the language model's training more closely with zero-shot tasks. Such a capabilities proves particularly advantageous within a multilingual framework. The zero-shot cross-lingual transfer capabilities enable users to solve multilingual tasks across a broad set of languages by interacting with the models in a natural way.

2.2.6 Sentiment Analysis

The significance of Sentiment Analysis lies in its wide-ranging applications, from monitoring social media sentiment (Stieglitz and Dang-Xuan, 2013), analyzing customer feedback (Ordenes et al., 2014), to gauging public opinion on various issues (Almazrouei et al., 2023). For businesses, it plays a crucial role in brand monitoring and market research, allowing companies to understand consumer reactions to products or services (Ziegler and Skubacz, 2012). In the realm of public services, it can be used to assess public sentiment on policy decisions or social events (Verma, 2022).

It involves the automatic identification and categorization of opinions or emotions within text data, aiming to discern the writer's or speaker's sentiment towards particular topics or the overall tonality of the document. This process ranges from detecting basic sentiments like *positive*, *negative*, or *neutral* to more complex emotions such as *happiness*, *anger*, or *disappointment*. Sentiment Analysis is distinct from, yet related to, other NLP tasks like text classification, where the goal is to categorize text into predefined categories. While text classification may involve identifying the topic or theme of a text, Sentiment Analysis specifically focuses on the emotional tone behind the words (Taboada, 2016).

In the early stages, but also current, approaches analyzed documents and text by rule-based techniques (Hutto and Gilbert, 2014; Tan et al., 2015; Zahoor and Rohilla, 2020). Lexicons, filled with words and corresponding sentiment scores, facilitate the assessment of emotional valence through straightforward mathematical summation. Despite the ease of configuration, these systems lack robustness and adaptability. The requirement for periodic updates constrains lexicon-based approaches, rendering them inflexible and challenging to scale. Furthermore, complexities introduced by negation and code-switching demand additional customized rules, further complicating the application of rules in sentiment analysis.

In the realm of NLP, sentiment analysis has experienced a paradigm shift from lexicon-based methods to those that leverage supervised machine learning techniques. This evolution reflects the broader trend within NLP towards embracing more sophisticated approaches. These techniques for handling with sentiment analysis involve training algorithms on pre-labeled datasets to classify text into some specified opinion or sentiment label. This approach leverages algorithms like Naïve Bayes, Support Vector Machines, and Decision Trees. A comparative study by Bhavitha et al. (2017) reveals that supervised learning techniques, especially when compared to unsupervised methods, show higher accuracy. These techniques rely heavily on pre-labeled datasets to train models that can classify sentiments in sentences or documents. However, they often face limitations such as the challenge of adapting to different contexts or domain-specific language. Duong and Nguyen-Thi (2021) highlighted the challenges of building large, pre-labeled datasets for supervised learning, noting that it can be tedious, expensive, and time-consuming, with difficulties in handling unseen data.

Fully supervised approaches have limitations, such as the need for extensive and diverse training data to accurately capture sentiment, and challenges in interpreting context and sarcasm, which can lead to misclassification of sentiments. Numerous research studies have demonstrated the enhancements by (contextual) word embeddings and deep neural networks, to a variety of NLP tasks, so to sentiment analysis (Devlin et al., 2019; Brown et al., 2020; Wei et al., 2022). Within the existing corpus of research on sentiment analysis, large language models have emerged as a preeminent instrument. Frameworks built upon the strategy of pre-training and fine-tuning have shown proficiency in grappling with the nuances and complexities inherent in sentiment analysis (Hoang et al., 2019; Praveen and Vajrobol, 2023). With XLM-T Barbieri et al. (2022) released a multilingual RoBERTa-based model, fine-tuned on a linguistically diverse corpus of tweets. Thereby creating a baseline model for multilingual Twitter sentiment analysis. This approach capitalizes on Twitter's unique linguistic properties, enhancing the model's relevance to social media contexts. Additionally, they released sentiment analysis datasets for eight languages, offering valuable resources for cross-linguistic sentiment analysis studies. However, while effective for Twitter data, this model's broad

linguistic scope may limit its precision in capturing language-specific nuances. Also, as technology progresses, larger models with more parameters have emerged, potentially surpassing the capabilities of XLM-T in handling complex linguistic data and providing more nuanced understanding across diverse languages.

While "pre-train and fine-tune" techniques demonstrate robust performance, their drawbacks are evident in scenarios with limited data availability. Under such conditions, the aforementioned methods encounter limitations, rendering them less effective in realizing their full capabilities. In low-resource contexts, which include not only the intricacies of real-world data but also multilingual environments, there exists a demand for strategies that capitalize on the inherent knowledge of language models. Tasks such as assigning an emotional tone to social media posts within custom categories or interpreting tweets with a mix of African and English dialects underscore the critical requirement for systems that display comprehensive linguistic understanding across a broad spectrum of languages. Leveraging the capabilities of OpenAI's GPT for sentiment analysis was investigated by Kheiri and Karimi (2023). The system, called SentimentGPT, combines several strategies for exploiting GPT in advanced sentiment analysis. This includes predictions from a fine-tuned GPT instance, but also prompting methods and training classifiers on GPT-encoded embeddings. Their strategy demonstrates robust capabilities in analyzing sentiment within the English language. However, this focus on English represents a limitation, as researchers have confined their examination exclusively to content in this language. The instructions utilized were designed solely in English, prompting an inquiry into the effectiveness of multilingual prompt engineering.

Zhou et al. (2023) introduced an automatic approach, APE, for generating natural language instructions. This method capitalizes on a language model's ability to produce coherent natural language instructions. Employing an in-context learning scenario, it facilitates the creation of instructions for various NLP tasks, such as sentiment analysis. These instructions are human-readable and effectively prompt a model to analysis the text and response with a sentiment class. Much like SentimentGPT, this research concentrates on sentiment analysis in a monolingual (English) context. The efficacy of multilingual instructions over English instructions in enhancing (multilingual) sentiment analysis remains an open area of inquiry. Research by Fu et al. (2022) explored the potential of integrating various tasks from different languages into a unified framework without the need for language or task-specific modules. Their evaluation indicates that language models for multilingual sentiment analysis, when tested on the Multilingual Amazon Reviews Corpus (Keung, Lu, and Bhardwaj, 2019), benefits from multilingual multitask training. They have also contributed valuable insights on the topic of multilingual prompt engineering. It emerges that cross-lingual prompt templates, where the template is in English and the test input is in target language, surpass those formulated specifically in the target language, albeit by a slim margin. While the authors manually created the cross-lingual prompt templates, the methods for creating the in-lingual variants remains undocumented. This omission raises the question of whether an automated method could yield equivalent or superior cross-lingual or in-lingual templates.

2.3 Prompting Methods

Over an extended period, the field of NLP has been dominated by supervised learning paradigms (Amine and Mimoun, 2007) (Asha et al., 2023). More recently, frameworks that engage in pre-training followed by fine-tuning have become instrumental in addressing a variety of challenges within this discipline. P. Liu et al. (2023) summarized the shift in paradigms from manually *feature engineering* (Lafferty et al., 2001) through *architecture engineering* where neural networks were exploited to learn salient features (J. Wang et al., 2016) towards a *pre-train and fine-tune* paradigm. Adapting language models via fine-tuning on downstream tasks has markedly influenced almost any field in NLP (Conneau, Khandelwal, et al., 2020; Devlin et al., 2019).

As language models scale in parameter size by pre-training on enormous raw text corpora, leveraging their deep natural language understanding becomes feasible without tailoring task-specific systems (Brown et al., 2020). This approach deviates from the conventional supervised learning framework and relies on the capability of large language models to directly model coherent text. It entails the use of templates in the form of textual string. This template act as a frame for transforming the initial input into a structured prompt that incorporates slots to be completed by the language model. Templates can also incorporate task description or instructions, describing a desired behavior. There are no restrictions for the form of the template. For instance, in sentiment analysis a review can be *Ï like the functionalities of the new device"* and for analyzing the sentiment it can be transformed into *"Classify the sentiment of given sentence. Sentence: I like the functionalities of the new device. Sentiment: "*.

However, this transformation is fundamentally not tied to any template. Input can also be presented without any transformation. Employing a suitable language model, the example prompt is expected to elicit it's coherent language modeling ability and generate the desired output effectively. In contrast, without any description, the likelihood of eliciting the correct sentiment classification directly from the context cannot be assured.

Addressing complex tasks through posing the problem as a question or instruction reflects a method akin to human problem-solving strategies. The process may appear straightforward, yet the transformation of input or context into an optimal format is not trivial. Language models frequently get labeled by the term *black-box* due to the opaque nature of their input-output transformations. Nonetheless, models generation is contextualized by input provided, thereby emphasizing the importance of creating a coherent input. Prompt engineering is the act of analyzing how different prompt formulations influence the models generation, along with the deployment of strategic search to find the optimal prompt for a designated task.

2.3.1 Prompt Engineering

Prompt engineering constitutes a crucial process within the domain of prompting methods, entailing the systematic exploration, optimization, and crafting of prompts to adeptly guide the model's response towards the desired task. The structure of the prompt, regardless of its simplicity or complexity, significantly influences the task's alignment with the capabilities of the model. Formulating prompts in an interpretable natural language represents the most intuitive, human-like method.

(P. Liu et al., 2023) mention two distinguish designs: Either a cloze-style form, where the slot for the text to be generated by the model, is within the prompt (e.g. ”l like this book. This implies a [slot] sentiment”) or a formulation in which the slot is placed at the end (e.g. ”This book is awesome! The sentiment is [slot]”).

The creation of these prompts involves two primary methodologies: manual engineering and automated approaches. Crafting templates manually represents a intuitive but mentally demanding task, often susceptible to errors. Schick and Schütze (2021) defined templates in a few-shot learning setting on text classification. These researchers employed cloze-style templates, probing the knowledge of language models. Their evaluation demonstrated good performance on text classification. However, the necessity to manually generate templates for each specific task presented a limitation, as this process requires the crafting of novel templates for each new task encountered.

Automated Discrete Prompts Creating prompts manually is mentally taxing, error-prune and quite possibly sub-optimal. In the case of multiple languages, it also requires linguistic expertise. Presented with an inadequate prompt, the model may likely fail to grasp the underlying task. Researchers have since proposed to delegate the engineering process to the entity responsible for processing the prompt — a language model. Harnessing the potential of language models to generate coherent text, an automated method supported by the deep language understanding from large language models shows promise. AutoPrompt represents an automatic approach, automating the generation of prompts to elicit knowledge from language models without fine-tuning. Developed by Shin et al. (2020), the technique utilizes gradient-guided search to find prompt templates, more precisely to find a set of optimal trigger tokens, aiming to enhance a model’s performance in tasks such as sentiment analysis. A key advantage of AutoPrompt is its ability to produce more effective prompts than manual creation, particularly in complex NLP tasks. However, one limitation is its potential variability in performance across different tasks or datasets.

Jiang et al. (2020) investigated the automatic engineering of prompts. The study introduced methods to automatically generate optimized prompts for querying language models, enhancing the accuracy of extracting factual knowledge. It highlights the potential limitations of manually created prompts and demonstrates that their automated alternatives can yield a more accurate estimation of a language model’s knowledge. T. Gao et al. (2021) utilized a text-to-text transformer-based model (T5) for automatically generating a template. The model received training sentences featuring annotated positions (slots) to guide the token insertion process. This approach yields templates that were intuitively reasonable in natural language form; however, upon closer inspection, these formulations were often brief, containing only a few words or tokens and occasionally displaying irregularities, as noted by the authors.

Research by Zhou et al. (2023) explored the capabilities of language models in generating and selecting natural language instructions, thereby acting as effective prompt engineers. The authors propose their method called Automatic Prompt Engineer (APE), that optimizes instructions by searching through a pool of candidate proposed by OpenAIs GPT model to maximize the accuracy on downstream tasks. Different to the approach aforementioned, they do not generate template tokens but task the model to generate a whole sentence, formulating as an instruction and describing the task which has to be accomplished. Their experiments across many NLP tasks demonstrate that the instructions generated by APE outperform those generated by human annotators in most cases.

While there approach demonstrate an effective automatic approach for generating natural language instructions, the researchers limited themself to investigation of refining English instructions. This is a shortcut considering the immense body of multilingual data available and a similar need for such approaches. The researchers noted the remarkable proficiency of large language models in following instructions, a proficiency not solely restricted to the English language context. The ability to learn across languages provides a foundation that shows large language models are able to understand languages that are considered low-resource. Strategies for the automatic generation of prompts across languages could represent a significant step forward in the advancement of robust multilingual NLP applications.

Automated Continuous Prompts To this point, all the discussed methods, whether manually or automated, have centered on generating input that constitutes natural language text, rendering it interpretable by humans. However, since language models are unable to process this format directly and necessitate that the words be embedded within a specific vector space, certain studies have explored the efficacy of prompting language models directly within this embedding space. This introduces the concept of continuous (also known as *soft*) prompts, whose token composition is not confined to a discrete representation of words structured by a language but can instead be initialized by arbitrary vectors in $\mathbb{R}^d$.

Building on the foundation of soft prompting, X. L. Li and Liang (2021) introduced Prefix-Tuning, an approach designed for natural language generation tasks. This technique involves prefixing a sequence of continuous, task-specific vectors, to the input of a language model. These prefixes, which function like sequences of *virtual* tokens direct the model's attention during generation. A significant advantage of this approach is its efficiency: unlike full fine-tuning, which requires updating and storing a separate model for each task, prefix-tuning only optimizes and stores the small prefix, greatly reducing overhead and allowing a single model to support multiple tasks simultaneously. However, a potential limitation of prefix-tuning is its reliance on the quality and design of these prefixes, as the effectiveness of the model's output heavily depends on how well these prefixes are optimized for their respective tasks. Leveraging well-crafted discrete prompt as a starting point for the search of continuous representations also constituted the baseline for some studies (Z. Zhong et al., 2021) (Qin and Eisner, 2021). One significant advantage of vector representations over discrete representations lies in the capacity for vector representations to be learned and adapted to specific contexts. (Goswami et al., 2023) proposed a gated method in which language models are able to dynamically switch between general-domain soft prompts and domain-specific ones. These vectors are either randomly initialized or represent domain-specific keywords and contribute with semantic information of the special domain. The prompts are then optimized by training on sentence classification task and demonstrate good performance on low-resource clinical datasets. As hypothesized, domain-specific soft prompts effectively retrieve domain-specific knowledge from language models.

Soft prompts presents an additional, promising strategy to the methods by which humans can prompt large language models. Prompt engineering seeks to identify the optimal prompt that enables the task to be solved efficiently; therefore, there is no necessity to confine prompt tokens to a specific (e.g. discrete) representation. Instead, one can tune continuous parameters based on task-specific data. However, when words are no longer represented in natural language, they become unintelligible to humans and their semantics are difficult to study. Even when initializing the embeddings around

discrete word tokens, arbitrarily updates during learning can drift them apart from real words. Khashabi et al. (2022) uncovered a noteworthy phenomenon in their research, revealing that the process of mapping continuous prompts into their discrete counterparts can result in text that is arbitrary or even contradictory. This issue underscores the difficulty in comprehending continuous prompts and their potential for generalization across different tasks.

2.3.2 In-context Learning

The advent of generative transformer models, showcased remarkable adaptability in scenarios with limited or no prior task-specific training. This adaptability is conceived through what is known as in-context learning (Brown et al., 2020), where a model, rather than being explicitly fine-tuned, leverages patterns it has internalized from the breadth of its pre-training data to make predictions about new, unseen tasks. (Brown et al., 2020) suggest that language models exemplify meta-learning capabilities, as evidenced by the superior performance in one- and few-shot scenarios compared to zero-shot settings. They described a in-context learning scenario as providing the model with demonstrations at inference time, showcasing input-output transformation. Figure 2.2 illustrates an example for sentiment analysis. The prompt is augmented with demonstrations of completion, in this concrete case the sentiment label. Numerous experiments have highlighted the capability of large language models in learning from demonstrations, including multilingual relation classification, question answering and machine translation (S. Agrawal et al., 2023; T. Gao et al., 2021; H. Zhang et al., 2022). P. Agrawal et al. (2023) underscore the advantages of few-shot multilingual in-context learning within the domain of machine translation, introducing a parameter-efficient approach that leverages prompt tuning for the automatic generation of multilingual question-answering datasets. The enhancement of a QA model through fine-tuning, as delineated in their study, suggests that the synthetic dataset generation coupled with subsequent fine-tuning of the QA model can surpass natural language prompting in effectiveness. Nonetheless, P. Agrawal et al. also acknowledge the computational demands of prompt tuning, given that it necessitates optimization on a language model. Additionally, the parameters adjusted during this process tied to the specific language model, thereby reducing the versatility of the method.

Lin et al. (2022) conducted analyses on the efficacy of incorporating in-language and cross-lingual demonstrations within natural language prompts for enhancing performance on the cross-lingual sentence understanding task XNLI (Conneau, Rinott, et al., 2018). They examined the potential enrichment of embedding high-resource language demonstrations, such as those from English, into prompts. Comparisons were drawn between this approach and an in-language setting. The evaluation yields insights that medium and low-resource settings can benefit from cross-lingual demonstrations. Nonetheless, in several instances, the cross-lingual prompts demonstrated inferior performance when compared to the in-language ones. Prompt template were generated through a combination of manual creation by native speakers of the respective target languages and translation from English. The observation that low-resource languages can benefit from cross-lingual in-context learning, points to a potential improvement of approaches for automatic prompt generation through in-context learning from non-English languages.

Nevertheless, prompt engineering and learning from demonstrations represent a subset of the prompting methods available. As the input in an hypothetical prompt-based system play an important role, so does the choice of a language model. Furthermore, methods like Prompt Answer Engineering (Zeng et al., 2023) or more advanced 'Instruction Tuning' (Wei et al., 2022) can significantly elevate language model performance.

2.3.3 Large Language Models

Selecting suitable large language models emerges as an equal important component in prompting methods as strategies finding the optimal prompt. Previous paragraphs underscored the existing methods that address the nature of inputs to the model. Their composition and the processes by which they are crafted. Given that this input will be transformed and process by some language model, the design and training procedure of these models can significantly influence the handling of the inputs. The choice of an appropriate model depends on both the nature of the input and overall objective. This decision warrants consideration in the development of systems based on prompt-based learning. In consideration of the task, P. Liu et al. (2023) suggests for text generation, such as machine translation, autoregressive models. The auto-regressive nature of transformers, such as GPT-2 or 3, BERT or FLAN-T5 (Brown et al., 2020; Chung et al., 2022; Devlin et al., 2019; Radford, J. Wu, et al., 2019), condition the next token generation based on the previous ones. This sequential processing enables them to capture the dependencies and subtleties inherent in the text. This makes them ideal for tasks where output is highly context-dependent. As previously mentioned in the introduction of Section 2.3.1, the formulation of templates can also be in a cloze-style format. This closely resembles the masked-language modeling objective that underlies the pre-training of models such as BERT or T5 (Devlin et al., 2019; Raffel et al., 2020). This is why masked language models are a good choice in such situations.

Many NLP tasks can easily be described via natural language instructions. Consequently, the task for large language models consists in following these instructions proficiently, thereby responding with the desired output. Ouyang et al. (2022) introduced InstructGPT, a fine-tuned version of GPT-3, which is designed to enhance the model's responsiveness to utterances framed as instructions. Their application of reinforcement learning from human feedback with prompt-response pair that are favored by humans, effectively aligns the model towards an improved instruction following. Similarly, Wei et al. (2022) evaluated the effectiveness of fine-tuning language models on tasks formulated as instructions – instruction tuning. Their released model called FLAN, demonstrates strong zero-shot learning for translation, closed-book question answering or natural language inference. Further analysis underscores that training on instructions boosts the zero-shot performance on unseen tasks. Leveraging this fine-tuning process, Chung et al. (2022) emphasized the expansion of training datasets and broadened the array of tasks employed for refining a FLAN-based model. The incorporation of a substantial array of natural instructions and reasoning tasks was prominent in their strategy. This enhancement in the model's architecture facilitated an elevation in its proficiency to follow instructions and execute reasoning tasks.

The Falcon series model, as introduced by Almazrouei et al. (2023), emerges as a competitor, exhibiting performance similar to other models such as PaLM (Chowdhery et al., 2022) or the GPT series (Brown et al., 2020; Radford, J. Wu, et al., 2019). This achievement stems from well-defined

filtering strategies that led to the creation of a high-quality web-based dataset. For pre-training, the dataset comprised only a minimal portion of multilingual data, due to the challenges posed by multilinguality. The 'curse of multilinguality' (Conneau, Khandelwal, et al., 2020) refers to the phenomenon where multilingualism compromises the quality of English understanding, coupled with a scarcity of high-quality multilingual data sources. Nonetheless, multilingual data constituted 10% of the pre-training corpus. They released two variants of the Falcon model: a base version and a subsequent iteration refined through instruction tuning, which is anticipated to exhibit robust instruction-following capabilities.

Chapter 3

Methods

This research investigates the effectiveness of integrating demonstrations within prompt templates to properly contextualize large language models for generating or rephrasing high-quality instructions in the target language.

This chapter presents the proposed approach for automatic multilingual prompt engineering. The methodology encompasses a detailed explanation of the prompt templates used to infer instructions, the process of generating instructions using large language models, the selection of the most suitable instructions, and the subsequent rephrasing of these instructions. Finally, the experimental setup for conducting the experiments is described in detail.

3.1 Automatic Multilingual Prompt Engineering

Building upon recent advancements made with Automatic Prompt Engineering (APE) by Zhou et al. (2023), this research investigates a similar approach but for non-English languages, which is designated as Automatic Multilingual Prompt Engineering (AMPE). The central hurdle in this exploration is tailoring APE strategies to a multilingual context, with a particular focus on examining the efficacy of target language demonstrations to enhance the generation or rephrasing of instructions in the target language. This section aims to provide a detailed overview of the methods and system deployed to investigate the research questions and proposes a multilingual approach for the automated generation of instructions in diverse target languages. Figure 3.1 visualizes the approach of automatically generating target language instructions.

3.2 Instruction Generation

The process of generating instructions in a target language commences with the selection or creation of a reference natural language instruction, initially in English. This reference instruction is critical as it acts as a reference or bridge. The instruction is specified by task demonstrations. These demonstrations are formulated as input-output pairs. An optimal instruction is defined by the fact that the desired output is obtained when it is executed on the input. Therefore, the semantics of the instruction implicitly or explicitly describe the task at hand. It can be created manually or automated, with tools like APE (Zhou et al., 2023) offering assistance. Achieving a well-defined English instruction is pivotal, as its quality directly influences the outcome.

This instruction is element of a template aimed at producing instructions in the target language. For instance as illustrated by the Transfer Template Figure 3.2. Furthermore, a template encompasses a frame or meta-prompt. This frame contextualize and semantically position the prompt within the context of the input-output demonstrations and overarching goal of transferring the instruction to the target language. The study examines various templates, differing based on their inclusion of demonstrations. For those incorporating demonstrations to support in-context learning, a randomly chosen set of n demonstrations melds into the prompt template. These demonstrations, or input-output pairs, delineate the task at hand, for instance, sentiment analysis. By leveraging large language models with such structured prompts, the aim is to direct these models towards the generation of quality instructions in the target language. The approach employs a language model to produce an initial batch of candidate instructions in the target language. This same model also appraises these candidates, evaluating them against a reserved test set. Subsequently, this process results in a set of target language instructions that are ranked according to their performance.

The proposed method is similar to, yet diverges from, traditional in-context learning. Typically, demonstrations serve as a form of guidance, prompting the large language model to deduce the implicit task they illustrate. When prompted with a series of demonstration pairs $\{i_k, o_k\}$, the language model is expected to match the subsequent input i_{k+1} with the correct output o_{k+1}. The divergence lies in the fact that, within the scope of this research, the language model is not tasked with continuing the input sequence. Instead, it aims to generate a natural language instruction that effectively describes and characterizes the task, which, in turn, achieves the completion of the input i_{k+1} (Honovich et al., 2023).

Zhou et al. (2023) introduced a novel method for the automated generation and selection of instructions linked to given tasks and specified through input-output pairs during generation. Their methodology treats the generation as an optimization problem, aiming to identify an instruction p generated by a language model M that optimizes the score for a given sample across a task set of input-output demonstrations $\{i_k, o_k\}$. The approach follows the few-shot in-context learning paradigm (Brown et al., 2020). In this context, few-shot implies that the prompt for generation is augmented with a limited number of demonstrations. These demonstrations outline a specific task or instruction necessary for generating the output from the given input. Nonetheless, their approach had a limitation: it was solely focused on generating instructions in English, thereby only addressing instruction generation for tasks presented in English.

Generation of target language instructions via in-context learning In-context learning involves the capability of large language models to swiftly adjust to new tasks through a limited (or great) number of demonstrations (Brown et al., 2020). However, it is not solely the demonstrations that can effectively guide the model towards completing a task. The overall context within a prompt has an important role as well (P. Agrawal et al., 2023). The efficacy of augmenting prompt templates with demonstrations for improving the process of target language instruction generation is examined by comparing two methods. Both integrate an instruction for a sentiment analysis task in English as reference. The frame enhances the template with additional context. (Honovich et al., 2023) refer to this frame as meta-prompt. Initial APE (Zhou et al., 2023) experiments facilitated the generation of English instructions by incorporating input-output demonstrations in English within the prompt. Initial testing using the AMPE approach highlighted limitations in producing instructions in the target language through simple augmentation of the prompt with demonstrations.

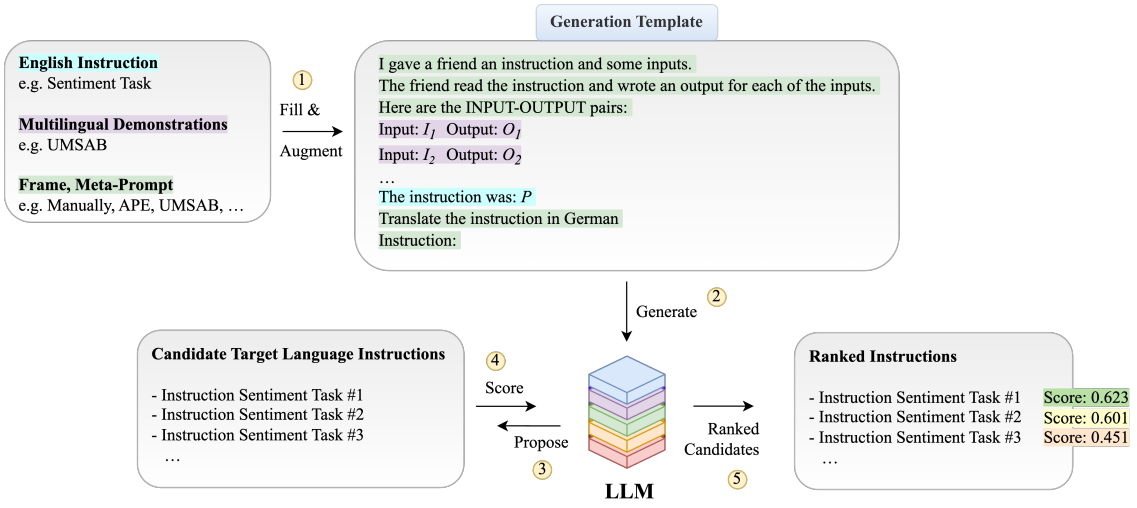


Figure 3.1: The approach, Automatic Multilingual Prompt Engineering (AMPE), produces instructions in the target language by leveraging a large language model conditioned with prompt templates that include, optionally exclude, demonstrations. The generation template, used for inference of target language instructions, draws inspiration from (Zhou et al., 2023). Demonstrations, specifying the task (e.g. sentiment), are optional and can be in English or in the target language. The same large language models is used for inference and scoring. The instruction with the best performance score is selected.

The assumption was that the language model used could not be adequately contextualized to achieve two objectives simultaneously: inferring the task and transferring the language. Consequently, an English instruction for the task was incorporated into the template. This inclusion was aimed at enhancing the model's understanding and its ability to generate instructions in the target language. Given that a language model's language understanding is crucial for success, future research is encouraged to explore more advanced models. These models might more effectively deduce latent meanings. Section Section 3.2.1 described the templates utilized in detail.

Target language in-context learning This research aims to generate target language instructions p for a range of target languages, adapting the optimization problem to a multilingual setting and addressing the research questions. In the domain of sentiment analysis, this involves specifying the task with a set of target language input-output demonstrations. A concrete example could be "Input: The food tastes amazing¡' and the corresponding sentiment as output "Output: Positive". A small sample n of this set is incorporated into the prompt template. For comparative analysis, this experiment also generates instructions from templates containing only English input-output demonstrations. The hypothesis under investigation posits that instructions generated with demonstrations in the target language will outperform those generated by prompt templates with English-only demonstrations, leading to improved instruction generation in the target languages. Diversifying the languages within the prompt templates used for contextualizing the language model at inference is one of the adaption and differences to the original APE approach. The following section elaborates the prompt templates employed to shape the prompts.

Transfer Template

Instruction: {instruction}.
Translate the instruction in {target_lang}.
Instruction: [slot].

Transfer-Inclusion Template

I gave a friend an instruction and some inputs.
The friend read the instruction and wrote an output for each of the inputs.
Here are the INPUT-OUTPUT pairs:

INPUT: {input} OUTPUT: {output}
INPUT: {input} OUTPUT: {output}
. . .

The instruction was: {instruction}
Translate the instruction in {target_lang}
Instruction: [slot]

Figure 3.2: (Left) Prompt template for generating instructions across languages. The template comprises three segments: Firstly, the input phrase; secondly, an meta-prompt, which specifies a translation objective into an assigned target language; and lastly, the output section, marking the beginning of the models response with "[slot]". (Right) This template comprises two more segments: Firstly, a frame or additional meta-prompt, introducing more context; secondly a section designated for the inclusion of demonstrations.

3.2.1 Prompt Templates

The structure and content of prompts or instructions stand as a crucial element in directing model behavior and gauging task performance. Hence, the design or engineering of prompt templates for generating them hold equal importance. This subsection outlines the templates utilized throughout this research, offering insights into the motivation and methods behind their selection. Templates not only serve as the body but also critically influence the quality and direction of the generated output. Appendix A lists the collection of prompts and templates employed for generating and evaluating instructions.

To determine if demonstrations provide useful context and enhance the effectiveness of target language instruction generation, the experiments considered various templates. As seen on the left side of Figure 3.2, one of the templates used to generate instructions. The template depicts a translation setting, with the objective to transfer the reference instruction into the target language. The shape of this template is manually crafted. The complexity of this task remains low and its formulation is intuitive. Additionally, these templates can be easily used or adapted to other NLP tasks (Honovich et al., 2023; Zhou et al., 2023). The prompt template (Transfer-Inclusion) on the right side in Figure 3.2 follows the same objective but it incorporates a frame or meta-prompt and section for demonstrations, following the analysis by Honovich et al. (2023).

Despite their simplicity, both the "Transfer" templates, formulated for the process of language transfer and instruction generation, are expected to yield a limited variety of instructions due to their basic context, which focuses solely on translation. Bearing this in mind, another template

Semantically-Similar Template

Instruction: {instruction}.
Write a semantically similar instruction in {target_lang}.
Instruction: [slot]

Semantically-Similar-Demo Template

I gave a friend an instruction and some inputs.
The friend read the instruction and wrote an output for each of the inputs.
Here are the INPUT-OUTPUT pairs:

INPUT: {input} OUTPUT: {output}
INPUT: {input} OUTPUT: {output}
...

The instruction was: {instruction}
Write a semantically similar instruction in {target_lang}
Instruction: [slot]

Figure 3.3: Prompt templates for generating semantically similar instructions to match the meaning of the initial instruction. The shape of the templates follows the same logic as the "Transfer" templates but with an adjustment of the meta prompt (visualized in bold).

– "Semantically-Similar" – was additionally created. This prompt template aims to prompt the discovery of semantically similar sets of instructions instead of merely translating the original instruction. The expectation is to achieve a broader and more varied distribution of generated target language instructions. Figure 3.3 illustrates the two approaches – one including demonstrations and the other not – in the context of the first research question

Demonstrations Demonstration learning, also known as in-context learning, represents a paradigm shift in the approach to training large language models. It fundamentally involves the inclusion of example-based context directly within the input prompt, guiding the model to perform a specific task. The templates that adhere to this paradigm, incorporates labeled demonstrations to explicitly guide language models. Leveraging language models via in-context learning from demonstrations depends on two key aspects: demonstration selection and ordering (P. Liu et al., 2023). Finding the optimal number of demonstrations is not straightforward and may involve complex optimization methods (Deng et al., 2022). Additionally, the context window imposes a limitation on the number. This refers to the maximum input that a language model can process in a single run. The experiments in this research thus follows a simple approach and selects one demonstration for each label category of the sentiment task. This method intentionally incorporates all labels to prevent bias and ensure every label is covered. As for sample ordering, it's crucial to present these examples in one sequence that enhances the learning process (Lu et al., 2022), without overwhelming the model with to much context.

3.2.2 Instruction Selection

The method for selecting the appropriate target language instructions, generated by the large language model, adheres to the same framework outlined by Honovich et al. (2023). During the process of generating candidate instructions the execution accuracy (Honovich et al., 2023) is calculated. Within the scope of the experiments for the sentiment analysis task, this entails a direct comparison for string matching between the model's response and the correct label.

To make the process of selecting the promising instructions more computationally efficient, Zhou et al. (2023) proposed an iterative evaluation strategy. This features an adaptive filtering mechanism where the accuracy of instructions is evaluated on subsets of the dataset. Instructions that achieve accuracy above a threshold progress, after which a new, non-overlapping subset is taken for further evaluation. The procedure was modified to incrementally increase the test subset with each iteration. Iterative evaluations continue until only a small set of optimal instructions remain. Zhou et al. (2023) suggests that increases computational efficiency and decreases the resources required. In fact, such a scalable approach ensures that early evaluations quickly eliminate less effective instructions, while those advancing through earlier rounds are subject to more rigorous testing.

The process concludes by selecting the best instruction from the smaller optimal set after conducting evaluations on the entire held-out test set. Calculation of the F_1 score serves as the metric for final evaluation. While the accuracy metric stands adequate for datasets with balanced class distributions, the F_1 score proves to be more appropriate for those with imbalanced distributions. Hence, to take account for potentially imbalanced datasets, the F_1 score is used.

3.2.3 Rephrasing the Instructions

This section presents the strategy of "Instruction Rephrasing" as an advanced technique aimed at improving the process of instruction generation in target language. The experiments investigate whether in-context learning with demonstrations can enhance this process. The process of rephrasing instructions is displayed in Figure 3.5

Although efforts focus on sampling high-quality initial target language instruction candidates, the templates mentioned before may not yield an optimal initial set of instructions. To address this issue and investigate the efficacy of suboptimal instructions as demonstrations on rephrasing optimal instructions, the "Rephrase" template was developed. The design of the template follows an underlying logic seen in the approach by Zhou et al. (2023). Nonetheless, the prompt template underwent some modifications: Firstly, the formulation of a more descriptive meta-prompt; Secondly a section for the inclusion of demonstrations. Such enhancements aim to enrich the contextual body and thereby augment the process of refining the target language instruction. The hypothesis posits that suboptimally performing instructions in the target language can effectively facilitate the guidance of both search and refinement processes when used as demonstrations. Previous research indicates that cross-lingual prompting results in robust performance, particularly in languages that are resource-scarce (P. Agrawal et al., 2023). It is therefore expected to improve the refinement process by incorporating these target language demonstrations in the English prompt template, as illustrated in Figure 3.4

Rephrase-Demo Template

Task Description: Generate a more explicit and in-depth variant of the provided initial instruction, ensuring the core meaning is conveyed in {target_lang}.
Initial Instruction: {initial_instruction}.
The following examples are suboptimal and should be avoided: {suboptimal_examples}.
Variant: [slot]

Figure 3.4: Prompt template for rephrasing instructions. The template comprises four segments: Firstly, the meta-prompt or task description, describing the rephrase objective; secondly, a section for the initial instruction to be rephrased; a section dedicated for the inclusion of suboptimal demonstrations; and lastly, the output section, marking the beginning of the models response.

The process begins with the ranked target language instruction set, obtained by prompting the language model with a generation template (e.g. Transfer Template). Zhou et al. (2023) adopt a simplified method, simply prompting the model to generate variations of the given instruction. This strategy does not take into account the multilingual context where preserving the instruction's original language is critical. Consequently, some adjustments were implemented for improvement. The first pertains to the redesign of the meta-prompt.

Secondly, the rephrase template with the respective parameters as illustrated in Figure 3.4. The rephrasing method is enhanced by considering the potential learning value of including suboptimal instructions for few-shot in-context learning. Language models are known for their ability to interpret instructions and adapt to novel contexts (Brown et al., 2020). This is substantially influenced by the model's ability to discern the prompt's context as well as the quality of that context. Therefore, its hypothesized that by contextualizing the prompt on both optimal and suboptimal instruction demonstrations, the quality of the context can be augmented, consequently prompting the model to rephrase a more accurate and descriptive target language instructions out of the optimal instruction.

Leveraging the language model for rephrasing, the initial run outputs a rephrased variant. This variant is evaluated against a held-out validation set. An iterative approach allows the further refinement of instructions over iterations. During the steps, instructions are ranked by their effectiveness, with those falling below a predefined threshold (e.g., an F_1 score of 0.3) being eliminated to conserve computational resources and on the assumption that such instructions contribute little to the rephrasing process. Superior instructions are identified by the highest F_1 scores. Those scoring above the threshold yet below the best are classified as suboptimal. Every iteration concludes with a new rephrased instruction. The 'survival' threshold is adjusted to the mean F_1 score of the remaining instruction set. The hypothesis is that this iterative refinement will eliminate weak instructions while enhancing the rephrasing process through examples pooled from a set of ever-improving instruction set. .

This approach draws inspiration from the *Iterative Monte Carlo Search* method proposed by Zhou et al. (2023), which emphasizes resampling instructions semantically close to the approach's best-found instruction. Within the idea of instruction generation, this involves using a large language model to rephrase the best instruction, based on the premise that resampling around it is more

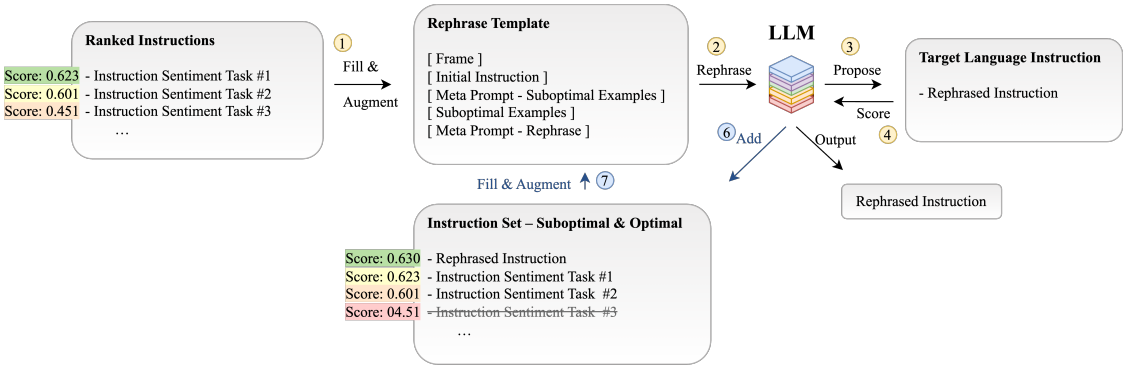


Figure 3.5: The (iterative) instruction rephrasing approach, rephrases the best-scored target language instruction. The prompt template undergoes augmentation, incorporating demonstrations and the meta-prompt to steer the model towards successful instruction rephrasing. Demonstrations are characterized by suboptimal instructions, based on their performance. The "Rephrase" template can include or exclude such demonstrations. After the initial run, a rephrased instruction is generated. Through iterative methods, there is an opportunity for the continuous refinement of instructions across multiple iterations (visualized by blue arrows). Within these iterations, instructions that perform poorly are systematically eliminated.

likely to yield good variants (Zhou et al., 2023). Recognizing the limitation of initial instruction sets, which may suffer from a lack of diversity or fail to include optimal candidates, this research investigates a similar strategic rephrasing approach. While Zhou et al. (2023) observed only a nominal benefit from iterative rephrasing, Fernando et al. (2023) highlighted the effectiveness of iterative refinement methods and the dynamic *mutation* of prompts.

3.2.4 Data Collection

The selection of appropriate datasets is guided by the specific objectives of the research. In the case of multilingual prompt engineering, the datasets must be representative of diverse languages. This diversity is essential for testing the efficacy of in-context learning in transferring instructions from English to other languages. The chosen datasets should encompass a great number of data points to allow for a comprehensive evaluation and analysis of the instruction generation process.

UMSAB The Unified Multilingual Sentiment Analysis Benchmark (UMSAB) (Barbieri et al., 2022) aims to create a balanced multilingual dataset for sentiment analysis (SA), where eight diverse languages are equally represented. These languages include typologically distant ones, ensuring a wide linguistic range. To maintain balance, the dataset size is limited to the smallest dataset, which is Hindi, consisting of 3,033 tweets. Each language dataset is pruned to this size, resulting in 1,839 training tweets (15% reserved for validation) and 870 for testing, totaling 24,262 tweets. The dataset ensures a balanced distribution across positive, negative, and neutral labels, maintaining original training/test splits. Some languages in the dataset represent groups of similar languages or dialects.

Outside the scope of the experiments, important data sets for multilingual sentiment analysis are presented below.

XED The XED Dataset (Öhman et al., 2020) is a significant multilingual resource designed for sentiment analysis and emotion detection. It contains human-annotated Finnish (25,000) and English sentences (30,000), along with annotations for 30 additional languages. This dataset utilizes Plutchik's core emotions (Plutchik, 2001), including a neutral category, to create a multilabel multiclass format. The total number of emotion annotations is 24,164 (excluding neutral) across 17,520 unique sentences. The majority of English subtitles in the dataset were assigned one emotion label, with some having multiple labels.

AfriSenti Is a valuable sentiment analysis dataset focused on African languages, encompassing over 110,000 annotated tweets in 14 languages such as Amharic, Algerian Arabic, Hausa, and Swahili. The dataset includes languages from various families and employs different writing systems including Latin, Arabic, and Ethiopic scripts. Data collection involved location-based and vocabulary-based methods, using tools like sentiment lexicons and language detection to gather relevant tweets. AfriSenti's diverse language coverage makes it a valuable resource for NLP research in African languages, particularly for those that are under-represented. (Muhammad et al., 2023)

3.2.5 Selection of Language Models

The selection of large language models for this study was guided by several considerations to ensure the effectiveness of multilingual prompt engineering. Firstly, instruction-tuned, multilingual models were prioritized due to their innate capacity to understand and follow the instructions in the prompt but also produce text across a diverse range of languages, particularly those included in the experiments. Additionally, the necessity for open-source models was emphasized to guarantee the reproducibility of the experiments, allowing for broader verification and application of the findings. Each of the models is publicly accessible over the internet. To generate meaningful instructions, the models required significant computational power, as evidenced by a great number of parameters. This attribute not only facilitates a deep understanding of natural languages but also enables the generation of coherent text. Among the chosen models for experimentation were notable Text-to-Text Generation Models such as Flan-T5 (Chung et al., 2022) and mT5 (Xue et al., 2021), alongside Generative Models like Falcon (Almazrouei et al., 2023) or Llama (Touvron et al., 2023). Table 3.1 contains brief information on the models.

Table 3.1: Overview of the large language models utilized in the experiments.

Model Name	Characteristics	Parameter Size
Falcon-7B-Instruct	Generative, Multilingual, Instruction-Tuned	7B
Flan-T5 XXL	Text-2-Text, Multilingual, Instruction-Tuned	13B
mT5-XXL	Text-2-Text, Multilingual, TEST	13B
Llama-2-7B	Generative, Multilingual, Instruction-Tuned	7B

Table 4.1: Example of target language instructions generated by the defined method. “Inclusion“ describes a template that includes demonstrations. The utilized large language model was Falcon-7b-Instruct

Method	English Instruction (Bridge)	Target Language Instruction
Semantically-Similar-Inclusion	Write ’2’ if the input is a positive tweet, ’0’ if the input is a negative tweet and ’1’ if the input is a neutral tweet.	2 schreiben, wenn positiv, 0 schreiben, wenn negativ, 1 schreiben, wenn neutral.
Transfer-Inclusion	Write ’2’ if the input is a positive tweet, ’0’ if the input is a negative tweet and ’1’ if the input is a neutral tweet.	Schreibe ’2’ wenn die Eingabe positiv ist, schreibe ’0’ wenn die Eingabe positiv ist, schreibe wenn die Eingabe neutral ist.

Chapter 4

Results

4.1 Effectiveness of Demonstrations on Target Language Instruction Generation

The research on optimizing target language instruction generation for various languages has led to an investigation centered around the effectiveness of including English demonstrations within prompt templates for instruction generation. This section reports the potential of such demonstrations on the generation process. For the purpose of conducting an comparison across multiple languages, the study employs the standardized UMSAB (Barbieri et al., 2022) dataset, which encompasses sentiment analysis data in eight distinct languages. Instructions are produced for each language in UMSAB by employing distinctive prompt templates. These templates are elaborated upon in Section 3.2.1. Subsequently, the inferred instructions for the target language undergo evaluation against the held-out test set designated for that specific target language. To set benchmarks, the research includes an evaluation of a manually created English instruction. This particular evaluation is carried out across all language test sets within UMSAB.

4.1.1 Comparative Analysis: Inclusion versus Exclusion of Demonstrations

Table 4.2 showcases the findings from the INCLUSION-EXCLUSION-DEMONSTRATIONS experiments. These experiments contrast the effectiveness of different templates. The first set uses templates that include English demonstrations (Inclusion) within the templates to generate target language instruction. The second set employs templates devoid of any demonstrations (Exclusion) within the template. In contrast to his, a human crafted English instruction serves as a *benchmark*. This comparison analyses their efficacy in producing instructions in the target language, utilizing large language models (presented in the model column). The study specifically examines the impact of incorporating English demonstrations within the template on the generation of target language instructions. Instructions were generated using the designated prompt template (method). The final column, labeled "Macro F_1-Score,"presents the scores for each language included in the UMSAB dataset. Filled versions of these prompt templates are provided in Appendix A for reference.

Building on the premise that prompting a language model on templates incorporating English demonstrations enhances its capacity to generate instructions, it is hypothesized that templates containing such demonstrations should generate instructions with improved performance. This improvement is expected to be evident through higher F_1 scores across various languages. Indeed, the experiments reveal that upon integrating demonstrations, the prompt templates enabled the Falcon-7B-instruct language model to generate instructions with a slightly better zero-shot performance compared to those produced without demonstrations. As underlined in the Table 4.2, it can be observed that the inclusion of English demonstrations yields instructions that perform better than those generated by templates without these demonstrations. Yet, the extent of this enhancement is not uniform across different templates. Specifically, the instructions generated by the 'Transfer' templates outperformed the instructions generated by the 'Semantically-Similar' templates. The most effective instructions emerged from the 'Transfer' templates augmented with demonstrations, even though the gain over the original English instruction is modest for the languages being evaluated. The dataset in *German* experienced the most significant performance enhancement when employing a target language instruction created via the "Transfer" template. This outperformed the results of utilizing an English instruction crafted by humans. The manually crafted benchmark English instruction exhibited marginally better performance than the instructions derived from 'Semantically-Similar' templates. Yet, this English instruction either matched or fell short of the performance of target language instructions generated through the "Transfer" templates.

4.1.2 Comparative Analysis: Inclusion of English or Target Language In-Context Demonstrations

Previous experiments have demonstrated that the incorporation of demonstrations generally leads to improved outcomes, as evidenced by instructions achieving higher F1-scores than those without integrated demonstrations. Subsequent investigations aim to ascertain the impact of integrating demonstrations in the target language on the generation of instructions within that same target language, in contrast to solely utilizing English demonstrations.

Table 4.2: Results for the INCLUSION-EXCLUSION-DEMONSTRATIONS experiments on the UMSAB sentiment task dataset. Compared are different prompt templates (method) used to generate target language instructions. This method comparison showcases the effectiveness of the integration of English demonstrations on instruction generation. "Transfer" indicates a template with the objective of translating the instruction, while "Semantically-Similar" describes the search for instructions with similar semantic in the target language. Shown are the zero-shot test F_1-Score of those instructions on the sentiment task.

Model	Method	Macro F_1 Score							
		ar	**de**	**en**	**es**	**fr**	**hi**	**it**	**pt**
Falcon-7B-Instruct	Transfer-Exclusion†	.60	.65	.63	.64	.56	.60	.61	.61
	Transfer-Inclusion⋆	.62	.65	.67	.64	.55	.63	.64	.62
	Semantically-Similar-Exclusion†	.52	.58	.60	.50	.52	.59	.51	.52
	Semantically-Similar-Inclusion⋆	.53	.58	.60	.57	.51	.57	.60	.55
	Human Crafted English Instruction	.60	.61	.65	.61	.57	.61	.61	.61

†: 'Exclusion' denotes templates that exclude demonstrations
⋆: 'Inclusion' denotes templates that include demonstrations
Note: Languages codes and their names are listed in Appendix B

The research outcomes of the ENGLISH-VS-TARGET-LANGUAGE-DEMONSTRATIONS experiments based on the UMSAB dataset are elaborated in Table 4.3. Falcon-7B-Instruct, the selected model for generation and inference, was examined in distinct configurations: one that encapsulated demonstrations in English (Inclusion-EL), and another which incorporated demonstrations in the target language (Inclusion-TL). These configurations were applied to both templates "Transfer" and "Semantically-Similar". Like in the experiments before, a human crafted English instruction serves as benchmark.

An examination of the "Transfer-Inclusion-EL" setup shows distinct scores for different target languages in the range of .55 to .67. When shifting demonstration inclusion to target languages, the " Transfer-Inclusion-TL" configuration reported a divergence in scores across languages with Hindi and Italian improving to .65 and .67 respectively. The "Semantically-Similar-Inclusion-EL" and "Semantically-Similar-Inclusion-TL" methods assess the application of prompt templates that closely align with demonstrations. In these settings, the model obtained a range of scores varying from .51 to .60, indicating varying degrees of effectiveness across multiple languages. Lastly, the human engineered English instruction was examined as another baseline. The measures of its efficacy fluctuated between .59 and .65, pinpointing strengths and weaknesses in comparison to the generated instructions by the aforementioned methods.

To respond quantitatively, the inclusion of target language demonstrations led to minor improvements in performance. Specifically, the "Transfer-Inclusion-TL" method showed a marginal improvement in instructions with German and Portuguese achieving Macro F1 scores of .67 and .66 respectively when contrasted against their English-centered counterparts. However, it is crucial to underline that the inclusion of target-language demonstrations within model generated instructions did not

Table 4.3: Results for the English-Vs-Target-Language-Demonstrations experiments on the UMSAB dataset. This analysis compares prompt templates to evaluate their efficacy in generating instructions in the target language. Through this comparison, the effectiveness of incorporating English or target language demonstrations on instruction generation is highlighted. "EL" in the method name designates the inclusion of English demonstrations, while "TL" in the name refers to the inclusion of target language demonstrations.

Model	**Method**	**Macro F_1 Score**							
		ar	**de**	**en**	**es**	**fr**	**hi**	**it**	**pt**
Falcon-7B-Instruct	Transfer-Inclusion-EL*	.62	.65	.67	.64	.55	.63	.64	.62
	Transfer-Inclusion-TL*	.61	.67	.67	.64	.54	.65	.67	.66
	Semantically-Similar-Inclusion-EL*	.53	.58	.60	.57	.51	.57	.60	.55
	Semantically-Similar-Inclusion-TL*	.55	.59	.60	.58	.53	.57	.60	.55
	Human Crafted English Instruction	.60	.61	.65	.61	.57	.61	.61	.61

*: 'Inclusion' denotes templates that include in-context demonstrations

significantly amplify the overall performance in zero-shot performance. Any increments were slight, with some languages indicating no change, or at best, a minor enhancement. The relayed results offer compelling insights for the research question focused on the efficacy of incorporating English or target language demonstrations in the process of instruction generation.

Figure 4.1 provides a detailed comparison of the impact various model architectures have on the quality of instruction generation. The process involved each large language model being prompted with an identical prompt template, labeled "Transfer-Inclusion-TL", to create instructions in the target language. The analysis indicates that large language models fine-tuned with instruction data consistently yield better zero-shot performance and more uniform results compared to those without instruction tuning (mt5-XXL). However, differences in performance among the instruction-tuned models were slight for the various languages evaluated in this study. This comparison highlights the beneficial impact of incorporating target-language demonstrations into the instruction generation process across several model architectures.

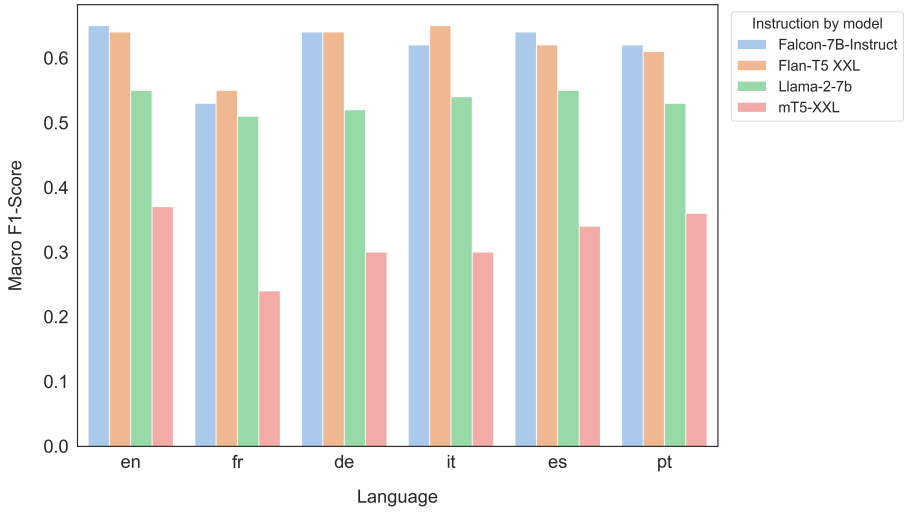
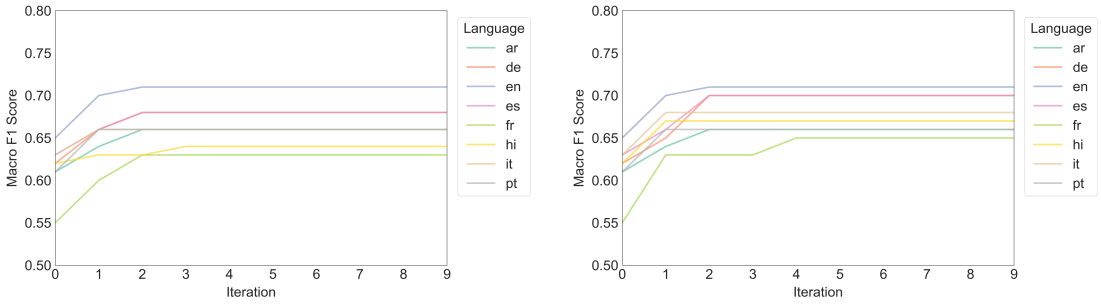


Figure 4.1: Results of the zero-shot performance of instructions on the UMSAB sentiment task. Target language instructions were generated by different language model architectures utilizing the "Transfer-Inclusion-Target-Language" prompt template. To be noted, the languages Arabic and Hindi were excluded from this comparison due to the lack of support for these languages in the Flan-T5 models.



(a) Rephrasing without suboptimal instruction demonstrations

(b) Rephrasing with suboptimal instruction demonstrations

Figure 4.2: Comparison of strategies on iterative rephrasing the best selected instruction for the displayed target languages. Shown are macro F_1 scores of some target language instruction iteratively rephrased and evaluated on a held-out validation set in UMSAB. (a) represents the Rephrase-Without-Demos strategy, which doesn't incorporate suboptimal instruction demonstrations within the rephrasing prompt template. Conversely, (b) employs the Rephrase-With-Demos approach, integrating less effective and suboptimal instruction demonstrations within the template. The experiments utilized generative model Falcon-7B-Instruct.

4.2 Effectiveness of Demonstrations on Instruction Rephrasing

To investigate the impact of suboptimal instructions as demonstrations in the rephrasing process, two comparative analyses were executed, centering on the quantifiable effects of such demonstrations on the rephrased instructions evaluated across all languages in the UMSAB sentiment task. The essence of this experiments aims to determine whether the integration of demonstrations in prompt templates supports the process of instruction rephrasing when contrasted with prompt templates that do not contain such demonstrations. The underlying assumption posits that including demonstrations of suboptimal instructions within prompt templates can serve as a valuable guidance mechanism in prompting methods for language models. Through exposure to these demonstrations, its expected to augment the model's capability in generating refined instructions.

The results, shown in Figure 4.2 reveals findings regarding the rephrasing of instructions. Notably, when examining the trajectory of the F_1 scores for the rephrased instructions across multiple iterations, a consistent trend emerged: As seen in Figure 4.2 (b), the integration of suboptimal instructions as in-context demonstrations uniformly improves the zero-shot performance of the rephrased instructions. This was contrasted with the templates devoid of such demonstrations, as seen in Figure 4.2 (a), which demonstrated a similar pattern but a smaller improvement boost across the languages.

A notable pattern was identified within the two strategies under examination. The initial iterations demonstrate a consistent enhancement of instructions via iterative rephrasing. However, this improvement reaches a plateau after only a few iterations. A limited number of iterations are adequate to attain the optimum rephrased instructions using either strategy. Integrating suboptimal

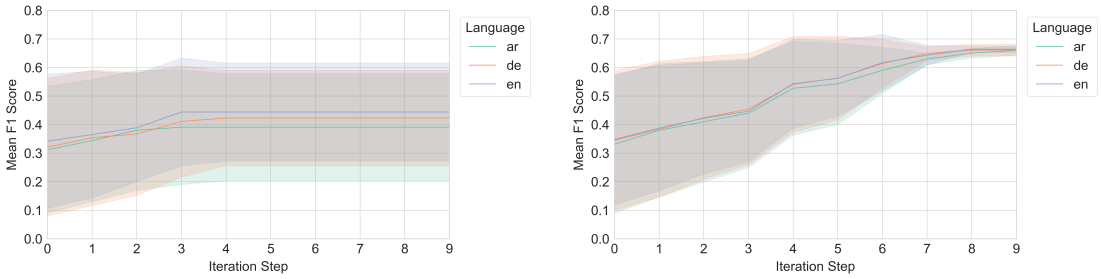
instructions as demonstrations within templates marginally enhances the instruction rephrasing. This improvement is almost comparable to the zero-shot performance observed in templates without these demonstrations.

Figure 4.2 illustrates minimal variance in the enhancement of instruction rephrasing, as shown by the zero-shot performance across the two examined strategies. Additionally, an analysis of the iterative progression and overall rephrase instruction set, depicted in Figure 4.3, indicates an upward trend in mean performance associated with the instruction set for each strategy as the number of iterations increased. Nevertheless, the exclusion(a), which omits suboptimal instructions demonstrations, reaches a performance peak and after few iterations. Subsequent iterations of rephrasing do not see variations in the mean or the standard deviation, suggesting a stagnation of the discovery of new instructions. On the contrary, the inclusion (b) demonstrates continuous refinement of the instruction set, culminating in an optimized set of instructions with better zero-shot performance. For clarity in visualization, the illustration exclusively utilizes German, English, and Arabic.

Addressing the research question of whether the approach of rephrasing target language instructions with prompt templates, incorporating suboptimal instructions demonstrations, yields greater efficacy than those without, the data from Figure 4.4 provides additional insightful observations. The rephrasing strategies consistently yielded higher average macro F1 scores. Specifically, the strategy without rephrasing "Transfer-Inclusion-TL" did not achieve high macro F1 scores as the iterative rephrasing methods. In line with the initial hypothesis, rephrasing of instructions can further enhance the quality of instructions, thus leading to an optimized automatic instruction generation process. This process can be even further enriched by incorporating demonstrations that exhibit examples of suboptimal instructions. This is showcased by the left chart in Figure 4.4.

In summary, the results showed interesting differences in the instruction generation by various strategies. English demonstrations within prompt templates yielded better performing instructions than templates that do not make use of such demonstrations. The zero-shot performance of those generated instructions were higher in comparison to a human crafted English instruction, nevertheless the differences a marginal higher. The evaluation of instruction generation with templates that include target language demonstrations showed marginal improvements or same performance compared to templates with English demonstrations.

Following experiments revealed that rephrasing instructions into the target language enhances the quality of the instruction set. When instructions are rephrased in the absence of demonstrations using suboptimal instructions, the enhancements are negligible. Incorporating suboptimal instructions as negative demonstrations during the rephrasing cycle leads to minor improvements in the quality of instruction modification. Nonetheless, in both methods, enhancements plateau after a few iterations.



(a) Rephrasing without suboptimal instructions demonstrations

(b) Rephrasing with suboptimal instructions demonstrations

Figure 4.3: Zero-shot performance of the instructions as the iterations goes up. Reported are the mean F_1 scores of the target language instruction set and standard deviation, for the UMSAB sentiment task. (Left) iterative rephrasing without demonstrations within the template. (Right) rephrasing with demonstrations within the template. The iterative rephrasing process adhering to in-context learning, generates the slightly better instructions, and ensures a superior set of instructions through iterative refinement.

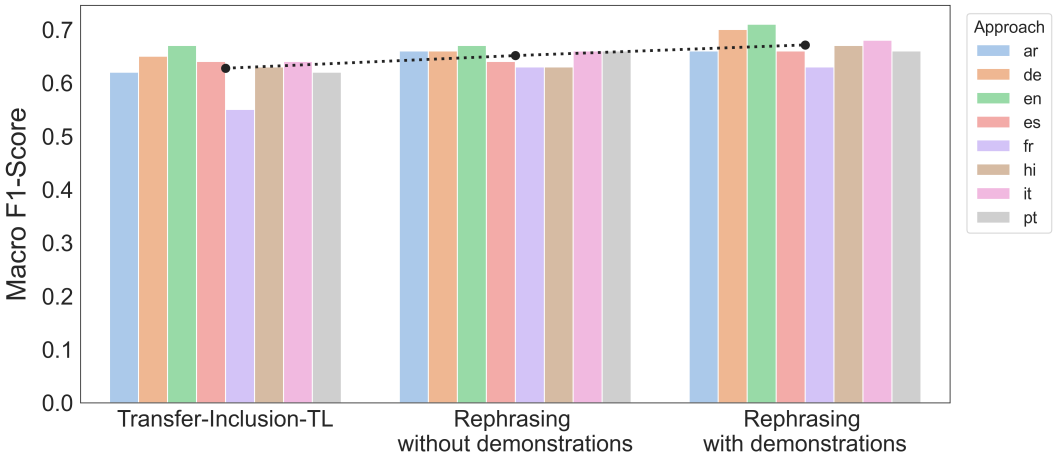


Figure 4.4: Comparison of different approaches for generating target language instructions to solve the UMSAB sentiment task. Shown are the macro F_1 scores for each approach on all languages in UMSAB. Iterative rephrasing with demonstrations enhances the average zero-shot performance across languages, as indicated by the trend of higher macro F_1 scores in this approach.

Chapter 5

Discussion

The results of this study highlight an advance in the field of multilingual prompt engineering. They show conclusively that in-context learning, when augmented with curated demonstrations, plays a fundamental role in improving the generation of accurate instructions for sentiment analysis in a variety of languages. This finding not only extends the utility of large language models in processing multilingual sentiment information, but also lays the foundation for more comprehensive and effective sentiment analysis methods that can cover a wide range of languages with accuracy and efficiency.

Table 5.1 provides a summary of results from all experiments on target language instruction generation. The initial experiments explored the impact of including or excluding demonstrations, either in the specific target language or in English. These experiments uncovered a positive effect. The evaluation revealed that including demonstrations not only enables the generation of target language instructions with performance comparable to the baseline English instructions but also results in instructions in certain target languages exhibiting slightly superior performance. Prior research has already established the understanding that in-context learning with demonstrations enable large language model to infer the latent task described by a few examples (S. Agrawal et al., 2023; Brown et al., 2020). For the specific undertaking of instruction generation, Zhou et al. (2023) and Honovich et al. (2023) revealed that these models can produce English instructions based on the context provided by a few task demonstrations. The findings presented in Table 5.1 demonstrate that this method is also effective in a multilingual context. It is evident that in-context learning with demonstrations adds valuable information in the process.

Subsequent experiments focused on assessing the impact of incorporating demonstrations in the target language within prompt templates on instruction generation. The hypothesis posited that in-context learning, enriched with demonstrations from the same target language intended for instruction generation, would provide additional context compared to demonstrations in English. P. Agrawal et al. (2023) highlights in their research on multilingual QA a cross-lingual prompting method. Upon examining Table 5.1, it becomes evident that incorporating demonstrations in the target language does not significantly enhance instruction generation as initially anticipated. Generating instructions through prompt templates with demonstrations in either English or the target language yields variants with comparably effective performance. This finding indicates that large language models exhibit no preference for target language demonstrations during the instruction generation process for that specific language. S. Agrawal et al. (2023) demonstrated through their research that demonstrations which implicitly describe the task and structure provides complementary benefits. The data presented in Table 5.1 supports this claim. However, it seems that demonstrations in the target language do not additionally amplify these benefits, particularly in the context of generating instructions for multilingual sentiment analysis.

Table 5.1: Presents the outcomes from experiments conducted on generating instructions in all target languages. The performance metrics of instructions produced via each suggested prompt template (method) are summarized herein. The notation "EL" indicates that English instructions were incorporated within the template, whereas "TL" signifies the incorporation of instructions in the target language. The rephrasing of instructions was performed either with or without the augmentation by suboptimal instructions as negative demonstrations. For reference, an instruction in English is used as a benchmark. The experiments utilized generative model Falcon-7B-Instruct.

Method	**Macro F_1 Score**							
	ar	**de**	**en**	**es**	**fr**	**hi**	**it**	**pt**
Transfer-Exclusion†	.60	.65	.63	.64	.56	.60	.61	.61
Transfer-Inclusion-EL*	.62	.65	.67	.64	.55	.63	.64	.62
Transfer-Inclusion-TL*	.61	.67	.67	.64	.54	.65	.67	.66
Rephrasing-Exclusion†	.66	.66	.67	.64	.63	.63	.66	.66
Rephrasing-Inclusion*	.66	.69	.69	.66	.63	.67	.68	.66
Human Crafted English Instruction	.60	.61	.65	.61	.57	.61	.61	.61

†: 'Exclusion' denotes templates that exclude in-context demonstrations
*: 'Inclusion' denotes templates that include in-context demonstrations

The experiments carried out before, but also previous research (S. Agrawal et al., 2023; Brown et al., 2020; Zhou et al., 2023), highlighted the complementary advantage of in-context learning with demonstrations. The application of this learning paradigm to the rephrasing of target language instructions has not been explored previously. In the research's experiments, demonstrations include instructions in the target language that yield poor or suboptimal results. In examining the impact of in-context learning alongside such demonstrations on the rephrasing process of target language instructions, findings summarized in Table 5.1 and detailed in Section 4.2 reveal the following insights: The initial rephrasing process of target language instructions results in a marginal enhancement of the instructions' performance. However, incorporating suboptimal instructions as negative demonstrations in the prompt template does not significantly improve the rephrasing process beyond the methods that produce instructions without such rephrasing. This outcome aligns with the findings of (Zhou et al., 2023), who noted that their iterative search algorithm yielded negligible improvements in performance. This study corroborates the same observation within the multilingual context of sentiment analysis.

It is posited that while large language models can generate optimized instructions for a target language, they encounter challenges in capturing every nuanced aspect of the prompt used to infer them. When suboptimal instructions are used as demonstrations, the utilized model likely struggles to extract the semantic and syntactic elements that define a suboptimal instruction. Furthermore, these models face significant hurdles in executing complex abstract reasoning (Huang and Chang, 2023) and in eliciting creative responses. Future research is encouraged to explore the effectiveness of this rephrasing strategy when implemented with more advanced models.

5.1 Limitations

The research leveraged language models with parameter sizes ranging from 1 to 7 billion (detailed in Section 3.2.5). While these models are significant in capability, it is acknowledged that larger models exist (e.g. GPT-4), boasting advanced learning and generative capacities that could potentially offer more nuanced insights into instruction generation and sentiment analysis in multiple languages. This limitation suggests that the findings might vary when applied or tested with these more sophisticated models, indicating a potential area for further exploration.

A notable challenge in this research stems from the prevalent English-centric training data that many large language models are based on. Despite their multilingual processing abilities, the predominant English data bias in these models could hinder their efficiency in handling and generating non-English texts (Mao et al., 2023; Papadimitriou et al., 2023). Highly advanced models such as GPT-4 demonstrate strong cross-lingual transfer learning capabilities but often are not publicly available. This scenario underlines a significant limitation, as the rare availability of efficient multilingual open-source models poses challenges in conducting robust and unbiased sentiment analysis across a diverse linguistic spectrum. Such limitations underscore the need for more inclusive language model training practices that better represent the global linguistic diversity. On the other hand, there is also the curse of multilingualism, which highlights the challenges of multilingual NLP.

Regarding the methods, incorporating English instructions into the prompt templates was a conscious choice. Initial experiments uncovered challenges for the models in deriving instructions in the target language based on the APE methodology (Zhou et al., 2023). The process of eliciting the model to internalize demonstrations and subsequently generate target language instructions for the task described by the demonstrations, proved to be complex. However, this contrasts with the APE approach, where models infer tasks autonomously from demonstrations without any reference point. This delineation points to a limitation in directly assessing the models' capacity for instruction generation without English guidance. Nevertheless, its assumed that prompt templates formulated in the specific target language mitigate the problem, as the context then will be aligned with the target language.

Lastly, the transformer models employed in this study are subject to the inherent limitation of conditioning only on a finite amount of input data due to their bounded context window. This constraint restricts the size and variety of demonstrations that can be processed, potentially limiting the depth and richness of the in-context learning and, by extension, the efficacy of instruction generation for sentiment analysis across languages. This technical limitation underscores a critical challenge in exploring the full capabilities of transformers in the context of prompt engineering.

Chapter 6

Conclusion and Future Work

In conclusion, this thesis has made significant strides in the realm of multilingual prompt engineering, demonstrating that the integration of in-context learning, coupled with the inclusion of demonstrations, enhances the generation of target language instructions for sentiment analysis. These findings answer the research questions affirmatively, revealing an improvement in instruction generation when demonstrations are embedded within prompt templates. These demonstrations can be formulated in the target language or in English. This enhancement in generating targeted language instructions underscores the potential of large language models in effectively tackling multilingual sentiment analysis. The proposed iterative approach to rephrase optimal target language instructions found with the previous methods has the potential to further improve the search. The addition of demonstrations, in this case the collected suboptimal instructions, also brings complementary benefits.

However, the exploration was conducted using language models with parameter sizes of only a few billions. Given the rapid advancement in the field, future studies could benefit from employing larger, advanced models to uncover deeper insights and potentially achieve even greater enhancements in instruction generation for multilingual sentiment analysis. As sentiment analysis is only one facet of many NLP tasks, the strategy used in this study is certainly applicable to other challenges.

Looking ahead, this body of work proposes avenues for further studies and practical application, specifically through the integration of task-specific contexts and adopting a “Prompt Evolution“ (Fernando et al., 2023) approach. Iteratively refining the context of prompt templates and infusing them with relevant information may yield improvements in the process. These efforts not only promise refinement but also aim to push the boundaries of multilingual NLP in an increasingly multilingual world.

Appendix A

Prompt Templates

Table A.1: Prompt templates used in the experiments, which are text strings that incorporate instructions or placeholders for context and demonstrations. These templates incorporate input variables denoted by curly brackets, such as *{input}*.

Name	Template
Transfer-Exclusion	Instruction: {instructions}\n Translate the instruction in \{target_lang\}\n Instruction:
Transfer-Inclusion	I gave a friend an instruction and some inputs.\n The friend read the instruction and wrote an output for each of the inputs.\n Here are the INPUT-OUTPUT pairs:\n INPUT: {input} OUTPUT: {output}\n INPUT: {input} OUTPUT: {output}\n INPUT: {input} OUTPUT: {output} \n The instruction was: {instruction}\n Translate the instruction in {target_lang}\n Instruction: [OUTPUT]
Semantically-Similar-Exclusion	Instruction: {instructions}\n Write a semantically similar instruction in {target_lang}\n Instruction:

Continued on next page

Table A.1: Prompt templates used in the experiments, which are text strings that incorporate instructions or placeholders for context and demonstrations. These templates incorporate input variables denoted by curly brackets, such as *{input}*. (Continued)

Semantically-Similar-Inclusion	I gave a friend an instruction and some inputs.\n The friend read the instruction and wrote an output for each of the inputs.\n Here are the INPUT-OUTPUT pairs:\n Instruction: {instructions}\n INPUT: {input} OUTPUT: {output}\n INPUT: {input} OUTPUT: {output}\n INPUT: {input} OUTPUT: {output} \n The instruction was: {instruction} Write a semantically similar instruction in {target_lang}\n Instruction:
English Instruction	Write '2' if the input is a positive tweet, '0' if the input is a negative tweet and '1' if the input is a neutral tweet.
Rephrasing-Without-Demonstrations	Task Description: Generate a more explicit and in-depth variant of the provided initial instruction, ensuring the core meaning is conveyed in {target_lang}.\n Instruction: {instruction}\n Variation:
Rephrasing-With-Demonstrations	Task Description: Generate a more explicit and in-depth variant of the provided initial instruction, ensuring the core meaning is conveyed in {target_lang}.\n Initial Instruction: {initial_instruction}\n The following examples are suboptimal and should be avoided: {suboptimal_examples}\n Variant:
Zero-shot Evaluation	{instruction}\n INPUT: {input}\n OUTPUT:

Table A.2: Prompts that emerge when the specified prompt templates are populated with task data from the UMSAB dataset.

Prompt
Transfer Inclusion Template:
I gave a friend an instruction and some inputs.\n The friend read the instruction and wrote an output for each of the inputs.\n Here are the INPUT-OUTPUT pairs:\n INPUT: Arschlöcher!!! MOBIL http OUTPUT: 0 \n INPUT: @user ja, so siehts aus o,o wieso? OUTPUT: 1 \n INPUT: @user Aw, na gut, dann schlaf fein, klein Delalein :) OUTPUT: 2 \n The instruction was: ”Write '2' if the input is a positive tweet, '0' if the input is a negative tweet and '1' if the input is a neutral tweet.”\n Translate the instruction in German\n Instruction:
Rephrase Inclusion Template:
Task Description: Generate a more explicit and in-depth variant of the provided initial instruction, ensuring the core meaning is conveyed in German.\n Initial Instruction: ”Schreibe 0 wenn negativ, schreibe 1 wenn neutral und schreibe 2 wenn positiv”\n The following examples are suboptimal and should be avoided:\n ”0, 1, 2”,\n ”schreibe 0, schreibe 1, schreibe 2”,\n Variant:

Appendix B

Language Codes

Table B.1: Codes for the Representation of Names of Languages

ISO 639-1 Code	English name of Language
ar	Arabic
de	German
en	English
es	Spanish
fr	French
hi	Hindi
pt	Portuguese

Bibliography

Agrawal, P., C. Alberti, F. Huot, J. Maynez, J. Ma, S. Ruder, K. Ganchev, D. Das, M. Lapata (Dec. 2023). "QAmeleon: Multilingual QA with Only 5 Examples". In: *Transactions of the Association for Computational Linguistics* 11, pp. 1754–1771. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00625. URL: https://doi.org/10.1162/tacl_a_00625 (visited on 01/28/2024) (cit. on pp. 15, 17, 36, 40, 44, 59).

Agrawal, S., C. Zhou, M. Lewis, L. Zettlemoyer, M. Ghazvininejad (July 2023). "In-context Examples Selection for Machine Translation". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, N. Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 8857–8873. DOI: 10.18653/v1/2023.findings-acl.564. URL: https://aclanthology.org/2023.findings-acl.564 (visited on 01/14/2024) (cit. on pp. 36, 59, 60).

Ahmad, G. I., J. Singla, A. Ali, A. A. Reshi, A. A. Salameh (2022). "Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus - A Comprehensive Review". en. In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 13.2. Number: 2 Publisher: The Science and Information (SAI) Organization Limited. ISSN: 2156-5570. DOI: 10.14569/IJACSA.2022.0130254. URL: https://thesai.org/Publications/ViewPaper?Volume=13&Issue=2&Code=IJACSA&SerialNo=54 (visited on 01/21/2024) (cit. on p. 20).

Ahuja, K., H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, S. Sitaram (Dec. 2023). "MEGA: Multilingual Evaluation of Generative AI". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, K. Bali. Singapore: Association for Computational Linguistics, pp. 4232–4267. DOI: 10.18653/v1/2023.emnlp-main.258. URL: https://aclanthology.org/2023.emnlp-main.258 (cit. on pp. 15, 16).

Albawi, S., T. A. Mohammed, S. Al-Zawi (Aug. 2017). "Understanding of a convolutional neural network". In: *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186. URL: https://ieeexplore.ieee.org/document/8308186 (visited on 01/28/2024) (cit. on p. 26).

Ali, A., S. Shamsuddin, A. Ralescu (2015). "Classification with class imbalance problem: A review". In: URL: https://www.semanticscholar.org/paper/Classification-with-class-imbalance-problem%3A-A-Ali-Shamsuddin/1e4870524f8de44d4f18c8f9f80eb797dfd25c89 (visited on 01/28/2024) (cit. on p. 25).

Allan, J. (2002). "Introduction to Topic Detection and Tracking". en. In: *Topic Detection and Tracking: Event-based Information Organization*. Ed. by J. Allan. The Information Retrieval Series. Boston, MA: Springer US, pp. 1–16. ISBN: 978-1-4615-0933-2. DOI: 10.1007/978-1-4615-0933-2_1. URL: https://doi.org/10.1007/978-1-4615-0933-2_1 (visited on 01/28/2024) (cit. on p. 25).

Almazrouei, E., H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, G. Penedo (Nov. 2023). *The Falcon Series of Open Language Models*. arXiv:2311.16867 [cs]. DOI: 10.48550/arXiv.2311.16867. URL: http://arxiv.org/abs/2311.16867 (visited on 01/27/2024) (cit. on pp. 30, 37, 47).

Amine, B. M., M. Mimoun (May 2007). "WordNet based Cross-Language Text Categorization". In: *2007 IEEE/ACS International Conference on Computer Systems and Applications*. ISSN: 2161-5330, pp. 848–855. DOI: 10.1109/AICCSA.2007.370731. URL: https://ieeexplore.ieee.org/document/4231059 (visited on 01/23/2024) (cit. on pp. 27, 33).

Arunkumar, A., S. Sharma, R. Agrawal, S. Chandrasekaran, C. Bryan (Apr. 2023). *LINGO : Visually Debiasing Natural Language Instructions to Support Task Diversity*. arXiv:2304.06184 [cs]. DOI: 10.48550/arXiv.2304.06184. URL: http://arxiv.org/abs/2304.06184 (visited on 04/22/2023) (cit. on p. 24).

Asha, P., K. S. Amirtha Varsini, S. Vidhya, M. S. K. Reddy, J. A. Mayan, L. K. Joshila Grace (July 2023). "Analysis of Twitter Sentiments using Machine Learning Algorithms". In: *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 782–786. DOI: 10.1109/ICESC57686.2023.10193310. URL: https://ieeexplore.ieee.org/document/10193310 (visited on 01/23/2024) (cit. on pp. 26, 33).

Bai, Q., Q. Dan, Z. Mu, M. Yang (2019). "A Systematic Review of Emoji: Current Research and Future Perspectives". In: *Frontiers in Psychology* 10. ISSN: 1664-1078. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02221 (visited on 01/23/2024) (cit. on p. 25).

Bang, Y., S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung (Feb. 2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. arXiv:2302.04023 [cs]. DOI: 10.48550/arXiv.2302.04023. URL: http://arxiv.org/abs/2302.04023 (visited on 04/22/2023) (cit. on p. 12).

Barbieri, F., L. Espinosa Anke, J. Camacho-Collados (June 2022). "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27 (visited on 07/31/2023) (cit. on pp. 23, 31, 46, 49).

Barnes, J., R. Klinger, S. Schulte im Walde (July 2018). "Bilingual Sentiment Embeddings: Joint Projection of Sentiment Across Languages". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych, Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 2483–2493. DOI: 10.18653/v1/P18-1231. URL: https://aclanthology.org/P18-1231 (visited on 01/23/2024) (cit. on p. 21).

Bhaumik, A., A. Bernhardt, G. Katsios, N. Sa, T. Strzalkowski (July 2023). "Adapting Emotion Detection to Analyze Influence Campaigns on Social Media". In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Ed. by J. Barnes, O. De Clercq, R. Klinger. Toronto, Canada: Association for Computational Linguistics, pp. 441–451. DOI: 10.18653/v1/2023.wassa-1.38. URL: https://aclanthology.org/2023.wassa-1.38 (visited on 01/28/2024) (cit. on p. 11).

Bhavitha, B. K., A. P. Rodrigues, N. N. Chiplunkar (Mar. 2017). "Comparative study of machine learning techniques in sentimental analysis". In: *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 216–221. DOI: 10.1109/ICICCT.2017.7975191. URL: https://ieeexplore.ieee.org/document/7975191 (visited on 01/24/2024) (cit. on p. 31).

Blanco, G., A. Lourenço (May 2022). "Optimism and pessimism analysis using deep learning on COVID-19 related twitter conversations". In: *Information Processing & Management* 59.3, p. 102918. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2022.102918. URL: https://www.sciencedirect.com/science/article/pii/S0306457322000437 (visited on 01/28/2024) (cit. on p. 11).

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (visited on 07/28/2023) (cit. on pp. 12, 13, 17, 24, 28, 29, 31, 33, 36, 37, 40, 45, 59, 60).

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez (Aug. 2023). *Spanish Pre-trained BERT Model and Evaluation Data*. arXiv:2308.02976 [cs]. DOI: 10.48550/arXiv.2308.02976. URL: http://arxiv.org/abs/2308.02976 (visited on 01/28/2024) (cit. on p. 24).

Cao, S., N. Kitaev, D. Klein (Feb. 2020). *Multilingual Alignment of Contextual Word Representations*. arXiv:2002.03518 [cs]. DOI: 10.48550/arXiv.2002.03518. URL: http://arxiv.org/abs/2002.03518 (visited on 01/17/2024) (cit. on p. 19).

Chan, B., S. Schweter, T. Möller (Dec. 2020). "German's Next Language Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by D. Scott, N. Bel, C. Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6788–6796. DOI: 10.18653/v1/2020.coling-main.598. URL: https://aclanthology.org/2020.coling-main.598 (visited on 01/20/2024) (cit. on p. 24).

Chen, J., H. Huang, S. Tian, Y. Qu (Apr. 2009). "Feature selection for text classification with Naïve Bayes". In: *Expert Systems with Applications: An International Journal* 36.3, pp. 5432–5435. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2008.06.054. URL: https://doi.org/10.1016/j.eswa.2008.06.054 (visited on 01/28/2024) (cit. on p. 26).

Chen, X., A. H. Awadallah, H. Hassan, W. Wang, C. Cardie (July 2019). "Multi-Source Cross-Lingual Model Transfer: Learning What to Share". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, L. Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3098–3112. DOI: 10.18653/v1/P19-1299. URL: https://aclanthology.org/P19-1299 (visited on 01/19/2024) (cit. on p. 22).

Chen, Y., D. Harbecke, L. Hennig (Dec. 2022). "Multilingual Relation Classification via Efficient and Effective Prompting". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 1059–1075. DOI: 10.18653/v1/2022.emnlp-main.69. URL: https://aclanthology.org/2022.emnlp-main.69 (visited on 01/23/2024) (cit. on p. 28).

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel (Oct. 2022). *PaLM: Scaling Language Modeling with Pathways*. arXiv:2204.02311 [cs]. DOI: 10.48550/arXiv.2204.02311. URL: http://arxiv.org/abs/2204.02311 (visited on 06/26/2023) (cit. on p. 37).

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei (Dec. 2022). *Scaling Instruction-Finetuned Language Models*. arXiv:2210.11416 [cs]. URL: http://arxiv.org/abs/2210.11416 (visited on 07/30/2023) (cit. on pp. 37, 47).

Clark, J. H., E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, J. Palomaki (July 2020). "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages". In: *Transactions of the Association for Computational Linguistics* 8, pp. 454–470. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00317. URL: https://doi.org/10.1162/tacl_a_00317 (visited on 01/18/2024) (cit. on p. 21).

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, J. Tetreault. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747 (visited on 01/19/2024) (cit. on pp. 19, 22–24, 27, 33, 38).

Conneau, A., G. Lample (2019). "Cross-lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html (visited on 01/20/2024) (cit. on pp. 23, 27).

Conneau, A., R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, V. Stoyanov (Oct. 2018). "XNLI: Evaluating Cross-lingual Sentence Representations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. URL: https://aclanthology.org/D18-1269 (visited on 01/27/2024) (cit. on p. 36).

Delobelle, P., T. Winters, B. Berendt (Nov. 2020). "RobBERT: a Dutch RoBERTa-based Language Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by T. Cohn, Y. He, Y. Liu. Online: Association for Computational Linguistics, pp. 3255–3265. DOI: 10.18653/v1/2020.findings-emnlp.292. URL: https://aclanthology.org/2020.findings-emnlp.292 (visited on 01/28/2024) (cit. on p. 24).

Deng, M., J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, Z. Hu (Dec. 2022). "RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3369–3391. DOI: 10.18653/v1/2022.emnlp-main.222. URL: https://aclanthology.org/2022.emnlp-main.222 (visited on 10/03/2023) (cit. on p. 43).

Devlin, J., M.-W. Chang, K. Lee, K. Toutanova (June 2019). "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (visited on 07/28/2023) (cit. on pp. 12, 22, 23, 26, 27, 29, 31, 33, 37).

Duong, H.-T., T.-A. Nguyen-Thi (Jan. 2021). "A review: preprocessing techniques and data augmentation for sentiment analysis". In: *Computational Social Networks* 8.1, p. 1. ISSN: 2197-4314. DOI: 10.1186/s40649-020-00080-x. URL: https://doi.org/10.1186/s40649-020-00080-x (visited on 01/24/2024) (cit. on p. 31).

El Bolock, A., I. Khairy, Y. Abdelrahman, N. T. Vu, C. Herbert, S. Abdennadher (2020). "Who, When and Why: The 3 Ws of Code-Switching". en. In: *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness. The PAAMS Collection*. Ed. by F. De La Prieta, P. Mathieu, J. A. Rincón Arango, A. El Bolock, E. Del Val, J. Jordán Prunera, J. Carneiro, R. Fuentes, F. Lopes, V. Julian. Communications in Computer and Information Science. Cham: Springer International Publishing, pp. 83–94. ISBN: 978-3-030-51999-5. DOI: 10.1007/978-3-030-51999-5_7 (cit. on pp. 19, 20).

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, W. Wang (May 2022). "Language-agnostic BERT Sentence Embedding". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 878–891. DOI: 10.18653/v1/2022.acl-long.62. URL: https://aclanthology.org/2022.acl-long.62 (visited on 01/19/2024) (cit. on p. 22).

Fernando, C., D. Banarse, H. Michalewski, S. Osindero, T. Rocktäschel (Sept. 2023). *Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution*. arXiv:2309.16797 [cs]. DOI: 10.48550/arXiv.2309.16797. URL: http://arxiv.org/abs/2309.16797 (visited on 10/19/2023) (cit. on pp. 46, 63).

Fu, J., S.-K. Ng, P. Liu (Dec. 2022). "Polyglot Prompt: Multilingual Multitask Prompt Training". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9919–9935. DOI: 10.18653/v1/2022.emnlp-main.674. URL: https://aclanthology.org/2022.emnlp-main.674 (visited on 10/17/2023) (cit. on pp. 15, 28, 32).

Fung, P., T. Schultz (May 2008). "Multilingual spoken language processing". In: *IEEE Signal Processing Magazine* 25.3. Conference Name: IEEE Signal Processing Magazine, pp. 89–97. ISSN: 1558-0792. DOI: 10.1109/MSP.2008.918417. URL: https://ieeexplore.ieee.org/abstract/document/4490205 (visited on 01/28/2024) (cit. on p. 22).

Gao, S., X.-C. Wen, C. Gao, W. Wang, M. R. Lyu (Apr. 2023). *Constructing Effective In-Context Demonstration for Code Intelligence Tasks: An Empirical Study*. arXiv:2304.07575 [cs]. DOI: 10.48550/arXiv.2304.07575. URL: http://arxiv.org/abs/2304.07575 (visited on 06/19/2023) (cit. on p. 17).

Gao, T., A. Fisch, D. Chen (Aug. 2021). "Making Pre-trained Language Models Better Few-shot Learners". 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)*. Online: Association for Computational Linguistics, pp. 3816–3830. DOI: 10.18653/v1/2021.acl-long.295. URL: https://aclanthology.org/2021.acl-long.295 (visited on 06/15/2023) (cit. on pp. 14, 34, 36).

Gary F., S., D. M. Eberhard, F. Charles D. (2023). *What are the top 200 most spoken languages?* en. URL: https://www.ethnologue.com/insights/ethnologue200/ (visited on 01/17/2024) (cit. on pp. 19, 21).

Golbeck, J., C. Robles, K. Turner (May 2011). "Predicting personality with social media". In: *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '11. New York, NY, USA: Association for Computing Machinery, pp. 253–262. ISBN: 978-1-4503-0268-5. DOI: 10.1145/1979742.1979614. URL: https://doi.org/10.1145/1979742.1979614 (visited on 01/28/2024) (cit. on p. 12).

Goswami, K., L. Lange, J. Araki, H. Adel (May 2023). "SwitchPrompt: Learning Domain-Specific Gated Soft Prompts for Classification in Low-Resource Domains". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by A. Vlachos, I. Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2689–2695. DOI: 10.18653/v1/2023.eacl-main.197. URL: https://aclanthology.org/2023.eacl-main.197 (visited on 01/26/2024) (cit. on p. 35).

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi (Aug. 2018). "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5, 93:1–93:42. ISSN: 0360-0300. DOI: 10.1145/3236009. URL: https://dl.acm.org/doi/10.1145/3236009 (visited on 01/28/2024) (cit. on p. 12).

Haviv, A., J. Berant, A. Globerson (Apr. 2021). "BERTese: Learning to Speak to BERT". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by P. Merlo, J. Tiedemann, R. Tsarfaty. Online: Association for Computational Linguistics, pp. 3618–3623. DOI: 10.18653/v1/2021.eacl-main.316. URL: https://aclanthology.org/2021.eacl-main.316 (visited on 12/29/2023) (cit. on p. 14).

Ho, J., D. Ondusko, B. Roy, D. F. Hsu (Aug. 2019). "Sentiment Analysis on Tweets Using Machine Learning and Combinatorial Fusion". In: *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 1066–1071. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00191. URL: https://ieeexplore.ieee.org/document/8890520 (visited on 01/23/2024) (cit. on p. 26).

Hoang, M., O. A. Bihorac, J. Rouces (Sept. 2019). "Aspect-Based Sentiment Analysis using BERT". In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, pp. 187–196. URL: https://aclanthology.org/W19-6120 (visited on 07/28/2023) (cit. on pp. 12, 31).

Hochreiter, S., J. Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (visited on 01/28/2024) (cit. on p. 26).

Honovich, O., U. Shaham, S. R. Bowman, O. Levy (July 2023). "Instruction Induction: From Few Examples to Natural Language Task Descriptions". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 1935–1952. DOI: 10.18653/v1/2023.acl-long.108. URL: https://aclanthology.org/2023.acl-long.108 (visited on 09/03/2023) (cit. on pp. 40, 42, 44, 59).

Howard, J., S. Ruder (July 2018). "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych, Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://aclanthology.org/P18-1031 (visited on 01/24/2024) (cit. on p. 28).

Huang, J., K. C.-C. Chang (July 2023). "Towards Reasoning in Large Language Models: A Survey". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, N. Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1049–1065. DOI: 10.18653/v1/2023.findings-acl.67. URL: https://aclanthology.org/2023.findings-acl.67 (visited on 01/30/2024) (cit. on p. 60).

Hutto, C., E. Gilbert (May 2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1. Number: 1, pp. 216–225. ISSN: 2334-0770. DOI: 10.1609/icwsm.v8i1.14550. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (visited on 01/24/2024) (cit. on pp. 11, 31).

Jiang, Z., F. F. Xu, J. Araki, G. Neubig (2020). "How Can We Know What Language Models Know?" In: *Transactions of the Association for Computational Linguistics* 8. Ed. by M. Johnson, B. Roark, A. Nenkova. Place: Cambridge, MA Publisher: MIT Press, pp. 423–438. DOI: 10.1162/tacl_a_00324. URL: https://aclanthology.org/2020.tacl-1.28 (visited on 01/25/2024) (cit. on pp. 14, 34).

Joshi, P., S. Santy, A. Budhiraja, K. Bali, M. Choudhury (July 2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, J. Tetreault. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. URL: https://aclanthology.org/2020.acl-main.560 (visited on 01/21/2024) (cit. on p. 20).

Kamaruddin, S. S., Y. Yusof, H. Husni, M. H. A. Refai (Aug. 2016). "TEXT CLASSIFICATION USING MODIFIED MULTI CLASS ASSOCIATION RULE". en. In: *Jurnal Teknologi* 78.8-2. Number: 8-2. ISSN: 2180-3722. DOI: 10.11113/jt.v78.9553. URL: https://journals.utm.my/jurnalteknologi/article/view/9553 (visited on 01/23/2024) (cit. on p. 26).

Keung, P., Y. Lu, V. Bhardwaj (Nov. 2019). "Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, X. Wan. Hong Kong, China: Association for Computational Linguistics, pp. 1355–1360. DOI: 10.18653/v1/D19-1138. URL: https://aclanthology.org/D19-1138 (visited on 01/23/2024) (cit. on p. 32).

Keung, P., Y. Lu, G. Szarvas, N. A. Smith (Nov. 2020). "The Multilingual Amazon Reviews Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, Y. Liu. Online: Association for Computational Linguistics, pp. 4563–4568. DOI: 10.18653/v1/2020.emnlp-main.369. URL: https://aclanthology.org/2020.emnlp-main.369 (visited on 01/23/2024) (cit. on pp. 11, 27).

Khalil, T., K. Kie\lczewski, G. C. Chouliaras, A. Keldibek, M. Versteegh (Nov. 2019). "Cross-lingual intent classification in a low resource industrial setting". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, X. Wan. Hong Kong, China: Association for Computational Linguistics, pp. 6419–6424. DOI: 10.18653/v1/D19-1676. URL: https://aclanthology.org/D19-1676 (visited on 01/23/2024) (cit. on p. 27).

Khashabi, D., X. Lyu, S. Min, L. Qin, K. Richardson, S. Welleck, H. Hajishirzi, T. Khot, A. Sabharwal, S. Singh, Y. Choi (July 2022). "Prompt Waywardness: The Curious Case of Discretized Interpretation of Continuous Prompts". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 3631–3643. DOI: 10.18653/v1/2022.naacl-main.266. URL: https://aclanthology.org/2022.naacl-main.266 (visited on 01/26/2024) (cit. on p. 36).

Kheiri, K., H. Karimi (July 2023). *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning*. arXiv:2307.10234 [cs]. DOI: 10.48550/arXiv.2307.10234. URL: http://arxiv.org/abs/2307.10234 (visited on 01/21/2024) (cit. on p. 32).

Kilimci, Z. H., S. Akyokuş (Sept. 2019). "The Evaluation of Word Embedding Models and Deep Learning Algorithms for Turkish Text Classification". In: *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pp. 548–553. DOI: 10.1109/UBMK.2019.8907027. URL: https://ieeexplore.ieee.org/document/8907027 (visited on 01/23/2024) (cit. on p. 26).

Kim, Y. (Oct. 2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, W. Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. URL: https://aclanthology.org/D14-1181 (visited on 01/28/2024) (cit. on p. 12).

Koehn, P. (Sept. 2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11 (visited on 01/18/2024) (cit. on p. 21).

Koppel, M., J. Schler (May 2006). "The Importance of Neutral Examples for Learning Sentiment." In: *Computational Intelligence* 22, pp. 100–109. DOI: 10.1111/j.1467-8640.2006.00276.x (cit. on p. 11).

Lafferty, J. D., A. McCallum, F. C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1-55860-778-1 (cit. on p. 33).

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab (May 2020). "FlauBERT: Unsupervised Language Model Pre-training for French". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis. Marseille, France: European Language Resources Association, pp. 2479–2490. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.302 (visited on 01/28/2024) (cit. on p. 24).

Li, H., K. Yamanishi (Nov. 1999). "Text classification using ESC-based stochastic decision lists". In: *Proceedings of the eighth international conference on Information and knowledge management*. CIKM '99. New York, NY, USA: Association for Computing Machinery, pp. 122–130. ISBN: 978-1-58113-146-8. DOI: 10.1145/319950.319966. URL: https://dl.acm.org/doi/10.1145/319950.319966 (visited on 01/23/2024) (cit. on p. 25).

Li, J., C. Cardie, S. Li (Aug. 2013). "TopicSpam: a Topic-Model based approach for spam detection". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by H. Schuetze, P. Fung, M. Poesio. Sofia, Bulgaria: Association for Computational Linguistics, pp. 217–221. URL: https://aclanthology.org/P13-2039 (visited on 01/28/2024) (cit. on p. 25).

Li, X. L., P. Liang (Aug. 2021). "Prefix-Tuning: Optimizing Continuous Prompts for Generation". 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)*. Ed. by C. Zong, F. Xia, W. Li, R. Navigli. Online: Association for Computational Linguistics, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353. URL: https://aclanthology.org/2021.acl-long.353 (visited on 12/28/2023) (cit. on pp. 14, 35).

Lin, X. V., T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, X. Li (Dec. 2022). "Few-shot Learning with Multilingual Generative Language Models". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9019–9052. DOI: 10.18653/v1/2022.emnlp-main.616. URL: https://aclanthology.org/2022.emnlp-main.616 (visited on 01/20/2024) (cit. on pp. 15, 36).

Liu, M., G. Haffari, W. Buntine (Dec. 2016). "Learning cascaded latent variable models for biomedical text classification". In: *Proceedings of the Australasian Language Technology Association Workshop 2016*. Ed. by T. Cohn. Melbourne, Australia, pp. 128–132. URL: https://aclanthology.org/U16-1014 (visited on 01/28/2024) (cit. on p. 26).

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig (Jan. 2023). "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Computing Surveys* 55.9, 195:1–195:35. ISSN: 0360-0300. DOI: 10.1145/3560815. URL: https://dl.acm.org/doi/10.1145/3560815 (visited on 04/22/2023) (cit. on pp. 13, 27, 33, 34, 37, 43).

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov (July 2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692 [cs]. DOI: 10.48550/arXiv.1907.11692. URL: http://arxiv.org/abs/1907.11692 (visited on 01/20/2024) (cit. on pp. 23, 26, 29).

Lu, Y., M. Bartolo, A. Moore, S. Riedel, P. Stenetorp (May 2022). "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL: https://aclanthology.org/2022.acl-long.556 (visited on 07/21/2023) (cit. on p. 43).

M'hamdi, M., R. West, A. Hossmann, M. Baeriswyl, C. Musat (Mar. 2019). "Expanding the Text Classification Toolbox with Cross-Lingual Embeddings". In: *ArXiv*. URL: https://www.semanticscholar.org/paper/Expanding-the-Text-Classification-Toolbox-with-M'hamdi-West/6a51d898ebe339bcc259acecee9880de6aa182dd (visited on 01/23/2024) (cit. on p. 28).

Mao, R., Q. Liu, K. He, W. Li, E. Cambria (July 2023). "The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection". In: *IEEE Transactions on Affective Computing* 14.3. Conference Name: IEEE Transactions on Affective Computing, pp. 1743–1753. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2022.3204972. URL: https://ieeexplore.ieee.org/document/9881877 (visited on 01/27/2024) (cit. on p. 61).

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, B. Sagot (July 2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, J. Tetreault. Online: Association for Computational Linguistics, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645. URL: https://aclanthology.org/2020.acl-main.645 (visited on 01/28/2024) (cit. on p. 24).

Mikolov, T., Q. V. Le, I. Sutskever (Sept. 2013). *Exploiting Similarities among Languages for Machine Translation*. arXiv:1309.4168 [cs]. DOI: 10.48550/arXiv.1309.4168. URL: http://arxiv.org/abs/1309.4168 (visited on 01/19/2024) (cit. on pp. 22, 26).

Muhammad, S. H., I. Abdulmumin, S. M. Yimam, D. I. Adelani, I. S. Ahmad, N. Ousidhoum, A. A. Ayele, S. Mohammad, M. Beloucif, S. Ruder (July 2023). "SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval)". In: *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 2319–2337. URL: https://aclanthology.org/2023.semeval-1.315 (visited on 07/30/2023) (cit. on p. 47).

Nandi, S. (Jan. 2022). "Bilingualism & Multilingualism: A New Perspective to Language Learning". en. In: *International Journal of English Learning & Teaching Skills* 4.2, pp. 1–8. ISSN: 2639-7412. DOI: 10.15864/ijelts.4211. URL: https://www.ingentaconnect.com/content/10.15864/ijelts.4211 (visited on 01/21/2024) (cit. on p. 19).

Nassif, R., M. W. Fahkr (Feb. 2020). "Supervised Topic Modeling Using Word Embedding with Machine Learning Techniques". In: *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, pp. 1–6. DOI: 10.1109/AECT47998.2020.9194177. URL: https://ieeexplore.ieee.org/document/9194177 (visited on 01/23/2024) (cit. on p. 26).

Öhman, E., M. Pàmies, K. Kajava, J. Tiedemann (Dec. 2020). "XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6542–6552. DOI: 10.18653/v1/2020.coling-main.575. URL: https://aclanthology.org/2020.coling-main.575 (visited on 07/30/2023) (cit. on p. 47).

Olujimi, P. A., A. Ade-Ibijola (May 2023). "NLP techniques for automating responses to customer queries: a systematic review". en. In: *Discover Artificial Intelligence* 3.1, p. 20. ISSN: 2731-0809. DOI: 10.1007/s44163-023-00065-5. URL: https://doi.org/10.1007/s44163-023-00065-5 (visited on 01/28/2024) (cit. on p. 25).

OpenAI et al. (Dec. 2023). *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. DOI: 10.48550/arXiv.2303.08774. URL: http://arxiv.org/abs/2303.08774 (visited on 01/28/2024) (cit. on p. 24).

Ordenes, F. V., B. Theodoulidis, J. Burton, T. Gruber, M. Zaki (Aug. 2014). "Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach". en. In: *Journal of Service Research* 17.3. Publisher: SAGE Publications Inc, pp. 278–295. ISSN: 1094-6705. DOI: 10.1177/1094670514524625. URL: https://doi.org/10.1177/1094670514524625 (visited on 01/28/2024) (cit. on p. 30).

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe (Dec. 2022). "Training language models to follow instructions with human feedback". en. In: *Advances in Neural Information Processing Systems* 35, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html (visited on 01/27/2024) (cit. on pp. 13, 37).

Pang, B., L. Lee (Jan. 2008). "Opinion Mining and Sentiment Analysis". In: *Foundations and Trends in Information Retrieval* 2, pp. 1–135. DOI: 10.1561/1500000011 (cit. on pp. 11, 12).

Papadimitriou, I., K. Lopez, D. Jurafsky (May 2023). "Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by A. Vlachos, I. Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1194–1200. DOI: 10.18653/v1/2023.findings-eacl.89. URL: https://aclanthology.org/2023.findings-eacl.89 (visited on 01/21/2024) (cit. on pp. 20, 61).

Pennington, J., R. Socher, C. Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, W. Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162 (visited on 01/28/2024) (cit. on p. 26).

Pires, T., E. Schlinger, D. Garrette (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, L. Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493 (visited on 01/20/2024) (cit. on p. 23).

Plutchik, R. (2001). "The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice". In: *American Scientist* 89.4. Publisher: Sigma Xi, The Scientific Research Society, pp. 344–350. ISSN: 0003-0996. URL: https://www.jstor.org/stable/27857503 (visited on 01/29/2024) (cit. on p. 47).

Poplack, S. (Dec. 2001). "Code Switching: Linguistic". In: *International Encyclopedia of the Social & Behavioral Sciences*. Journal Abbreviation: International Encyclopedia of the Social & Behavioral Sciences, pp. 2062–2065. ISBN: 978-0-08-043076-8. DOI: 10.1016/B0-08-043076-7/03031-X (cit. on p. 20).

Praveen, S. V., V. Vajrobol (Aug. 2023). "Understanding the Perceptions of Healthcare Researchers Regarding ChatGPT: A Study Based on Bidirectional Encoder Representation from Transformers (BERT) Sentiment Analysis and Topic Modeling". en. In: *Annals of Biomedical Engineering* 51.8, pp. 1654–1656. ISSN: 1573-9686. DOI: 10.1007/s10439-023-03222-0. URL: https://doi.org/10.1007/s10439-023-03222-0 (visited on 01/24/2024) (cit. on p. 31).

Prof. Praveen Dhyani, Sonam Mittal, B K Birla Institute of Engineering Technology (Mar. 2015). "Multilingual Text Classification". en. In: *International Journal of Engineering Research and* V4.03, IJERTV4IS030032. ISSN: 2278-0181. DOI: 10.17577/IJERTV4IS030032. URL: http://www.ijert.org/view-pdf/12550/multilingual-text-classification (visited on 01/23/2024) (cit. on p. 27).

Prokhorova, A. (2020). "Multilingual Communicative Competence of Future Engineers: Essence, Structure, Content". en. In: *Integrating Engineering Education and Humanities for Global Intercultural Perspectives*. Ed. by Z. Anikina. Lecture Notes in Networks and Systems. Cham: Springer International Publishing, pp. 11–20. ISBN: 978-3-030-47415-7. DOI: 10.1007/978-3-030-47415-7_2 (cit. on p. 20).

Qian, Q., M. Huang, J. Lei, X. Zhu (July 2017). "Linguistically Regularized LSTM for Sentiment Classification". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by R. Barzilay, M.-Y. Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 1679–1689. DOI: 10.18653/v1/P17-1154. URL: https://aclanthology.org/P17-1154 (visited on 01/28/2024) (cit. on p. 12).

Qin, G., J. Eisner (June 2021). "Learning How to Ask: Querying LMs with Mixtures of Soft Prompts". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou. Online: Association for Computational Linguistics, pp. 5203–5212. DOI: 10.18653/v1/2021.naacl-main.410. URL: https://aclanthology.org/2021.naacl-main.410 (visited on 01/26/2024) (cit. on p. 35).

Rabiner, L., B. Juang (Jan. 1986). "An introduction to hidden Markov models". In: *IEEE ASSP Magazine* 3.1. Conference Name: IEEE ASSP Magazine, pp. 4–16. ISSN: 1558-1284. DOI: 10.1109/MASSP.1986.1165342. URL: https://ieeexplore.ieee.org/abstract/document/1165342?casa_token=UKVAbHzDG4oAAAAA:l4T7VNN7apBt6riOy_O3HDDQuhCRPNOoP0kKD3LtUqLf7DimEUWwakIYvd9IBX6TwKM3GO7XXDg (visited on 01/28/2024) (cit. on p. 26).

Radford, A., K. Narasimhan (2018). "Improving Language Understanding by Generative Pre-Training". In: URL: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035 (visited on 01/24/2024) (cit. on p. 27).

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In: URL: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe (visited on 01/24/2024) (cit. on pp. 24, 37).

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu (Jan. 2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1, 140:5485–140:5551. ISSN: 1532-4435 (cit. on p. 37).

Ramesh, G., S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, M. S. Khapra (2022). "Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages". In: *Transactions of the Association for Computational Linguistics* 10. Ed. by B. Roark, A. Nenkova. Place: Cambridge, MA Publisher: MIT Press, pp. 145–162. DOI: 10.1162/tacl_a_00452. URL: https://aclanthology.org/2022.tacl-1.9 (visited on 01/17/2024) (cit. on pp. 19, 21).

Rao, T., S. Srivastava (2012). "Analyzing stock market movements using Twitter sentiment analysis". en. In: Washington, DC, USA: IEEE Computer Society, pp. 119–123. ISBN: 978-0-7695-4799-2. DOI: 10.1109/ASONAM.2012.30. URL: http://dx.doi.org/10.1109/ASONAM.2012.30 (visited on 01/28/2024) (cit. on p. 11).

Razumovskaia, E., G. Glavas, O. Majewska, E. M. Ponti, A. Korhonen, I. Vulic (July 2022). "Crossing the Conversational Chasm: A Primer on Natural Language Processing for Multilingual Task-Oriented Dialogue Systems". In: *Journal of Artificial Intelligence Research* 74, pp. 1351–1402. ISSN: 1076-9757. DOI: 10.1613/jair.1.13083. URL: http://www.jair.org/index.php/jair/article/view/13083 (visited on 01/21/2024) (cit. on p. 20).

Sanh, V., A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, A. M. Rush (Oct. 2021). "Multitask Prompted Training Enables Zero-Shot Task Generalization". en. In: URL: https://openreview.net/forum?id=9Vrb9D0WI4 (visited on 12/28/2023) (cit. on p. 13).

Schick, T., H. Schütze (Apr. 2021). "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 255–269. DOI: 10.18653/v1/2021.eacl-main.20. URL: https://aclanthology.org/2021.eacl-main.20 (visited on 10/05/2023) (cit. on pp. 29, 34).

Schmidt, R. M. (Nov. 2019). *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. arXiv:1912.05911 [cs, stat]. DOI: 10.48550/arXiv.1912.05911. URL: http://arxiv.org/abs/1912.05911 (visited on 01/28/2024) (cit. on p. 26).

Schuster, T., O. Ram, R. Barzilay, A. Globerson (June 2019). "Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing". 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)*. Ed. by J. Burstein, C. Doran, T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1599–1613. DOI: 10.18653/v1/N19-1162. URL: https://aclanthology.org/N19-1162 (visited on 01/17/2024) (cit. on p. 19).

Shin, T., Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh (Nov. 2020). "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4222–4235. DOI: 10.18653/v1/2020.emnlp-main.346. URL: https://aclanthology.org/2020.emnlp-main.346 (visited on 06/10/2023) (cit. on pp. 13, 14, 34).

Smith, S. L., D. H. P. Turban, S. Hamblin, N. Y. Hammerla (Nov. 2016). "Offline bilingual word vectors, orthogonal transformations and the inverted softmax". en. In: URL: https://openreview.net/forum?id=r1Aab85gg (visited on 01/19/2024) (cit. on p. 22).

Stieglitz, S., L. Dang-Xuan (Apr. 2013). "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior". In: *Journal of Management Information Systems* 29.4. Publisher: Routledge _eprint: https://doi.org/10.2753/MIS0742-1222290408, pp. 217–248. ISSN: 0742-1222. DOI: 10.2753/MIS0742-1222290408. URL: https://doi.org/10.2753/MIS0742-1222290408 (visited on 01/28/2024) (cit. on pp. 11, 30).

Sun, A., E.-P. Lim, Y. Liu (Dec. 2009). "On strategies for imbalanced text classification using SVM: A comparative study". In: *Decision Support Systems*. Information product markets 48.1, pp. 191–201. ISSN: 0167-9236. DOI: 10.1016/j.dss.2009.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0167923609001754 (visited on 01/28/2024) (cit. on pp. 12, 26).

Taboada, M. (2016). "Sentiment Analysis: An Overview from Linguistics". In: *Annual Review of Linguistics* 2.1. _eprint: https://doi.org/10.1146/annurev-linguistics-011415-040518, pp. 325–347. DOI: 10.1146/annurev-linguistics-011415-040518. URL: https://doi.org/10.1146/annurev-linguistics-011415-040518 (visited on 01/28/2024) (cit. on pp. 25, 31).

Talat, Z., A. Névéol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev, S. Sharma, A. Subramonian, J. Tae, S. Tan, D. Tunuguntla, O. Van Der Wal (May 2022). "You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings". In: *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Ed. by A. Fan, S. Ilic, T. Wolf, M. Gallé. virtual+Dublin: Association for Computational Linguistics, pp. 26–41. DOI: 10.18653/v1/2022.bigscience-1.3. URL: https://aclanthology.org/2022.bigscience-1.3 (visited on 01/21/2024) (cit. on p. 20).

Tan, L. I., W. S. Phang, K. O. Chin, A. Patricia (Oct. 2015). "Rule-Based Sentiment Analysis for Financial News". In: *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1601–1606. DOI: 10.1109/SMC.2015.283. URL: https://ieeexplore.ieee.org/abstract/document/7379415?casa_token=8hJaF_o1K2oAAAAA:pPbHfkO5TegafezoJGXzChfgcHFRyslcJY--YfBENkeSZLDt9Hk637eRXHj2zdzOAgNrjAN9O_4 (visited on 01/24/2024) (cit. on p. 31).

Tang, D., F. Wei, B. Qin, N. Yang, T. Liu, M. Zhou (Feb. 2016). "Sentiment Embeddings with Applications to Sentiment Analysis". In: *IEEE Transactions on Knowledge and Data Engineering* 28.2. Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 496–509. ISSN: 1558-2191. DOI: 10.1109/TKDE.2015.2489653. URL: https://ieeexplore.ieee.org/document/7296633 (visited on 01/28/2024) (cit. on p. 12).

Tiedemann, J. (Nov. 2020). "The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT". In: *Proceedings of the Fifth Conference on Machine Translation*. Ed. by L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri. Online: Association for Computational Linguistics, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139 (visited on 01/18/2024) (cit. on p. 21).

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample (Feb. 2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971 [cs]. DOI: 10.48550/arXiv.2302.13971. URL: http://arxiv.org/abs/2302.13971 (visited on 07/30/2023) (cit. on p. 47).

Tsvetkov, Y., V. Prabhakaran, R. Voigt (June 2018). "Socially Responsible NLP". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Ed. by M. Bansal, R. Passonneau. New Orleans, Louisiana: Association for Computational Linguistics, pp. 24–26. DOI: 10.18653/v1/N18-6005. URL: https://aclanthology.org/N18-6005 (visited on 01/21/2024) (cit. on p. 20).

Tu, L., J. Qu, S. Yavuz, S. Joty, W. Liu, C. Xiong, Y. Zhou (Apr. 2023). *Efficiently Aligned Cross-Lingual Transfer Learning for Conversational Tasks using Prompt-Tuning*. arXiv:2304.01295 [cs]. DOI: 10.48550/arXiv.2304.01295. URL: http://arxiv.org/abs/2304.01295 (visited on 04/22/2023) (cit. on p. 15).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 01/21/2024) (cit. on p. 12).

Verma, S. (July 2022). "Sentiment analysis of public services for smart society: Literature review and future research directions". In: *Government Information Quarterly* 39.3, p. 101708. ISSN: 0740-624X. DOI: 10.1016/j.giq.2022.101708. URL: https://www.sciencedirect.com/science/article/pii/S0740624X22000417 (visited on 01/28/2024) (cit. on p. 30).

W3Techs (Jan. 2024). *Usage Statistics and Market Share of Content Languages for Websites, January 2024*. URL: https://w3techs.com/technologies/overview/content_language (visited on 01/17/2024) (cit. on pp. 19, 21).

Wang, J., L.-C. Yu, K. R. Lai, X. Zhang (Aug. 2016). "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by K. Erk, N. A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 225–230. DOI: 10.18653/v1/P16-2037. URL: https://aclanthology.org/P16-2037 (visited on 01/28/2024) (cit. on p. 33).

Wang, W., Y. Li, K. Lu, J. Zhang, P. Chen, K. Yan, B. Wang (2023). "Medical tumor image classification based on Few-shot learning". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1–11. ISSN: 1557-9964. DOI: 10.1109/TCBB.2023.3282226. URL: https://ieeexplore.ieee.org/document/10147294 (visited on 01/24/2024) (cit. on p. 28).

Wei, J., M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le (Feb. 2022). *Finetuned Language Models Are Zero-Shot Learners*. arXiv:2109.01652 [cs]. DOI: 10.48550/arXiv.2109.01652. URL: http://arxiv.org/abs/2109.01652 (visited on 12/28/2023) (cit. on pp. 13, 31, 37).

Weiss, K., T. M. Khoshgoftaar, D. Wang (May 2016). "A survey of transfer learning". In: *Journal of Big Data* 3.1, p. 9. ISSN: 2196-1115. DOI: 10.1186/s40537-016-0043-6. URL: https://doi.org/10.1186/s40537-016-0043-6 (visited on 01/19/2024) (cit. on p. 22).

Winata, G. I., A. Madotto, Z. Lin, R. Liu, J. Yosinski, P. Fung (Nov. 2021). "Language Models are Few-shot Multilingual Learners". In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Ed. by D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, G. G. Sahin. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1–15. DOI: 10.18653/v1/2021.mrl-1.1. URL: https://aclanthology.org/2021.mrl-1.1 (visited on 01/24/2024) (cit. on p. 15).

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6 (visited on 07/30/2023) (cit. on p. 24).

Workshop, B. et al. (June 2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv:2211.05100 [cs]. DOI: 10.48550/arXiv.2211.05100. URL: http://arxiv.org/abs/2211.05100 (visited on 07/13/2023) (cit. on p. 23).

Wu, S., M. Dredze (Nov. 2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, X. Wan. Hong Kong, China: Association for Computational Linguistics, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077 (visited on 01/20/2024) (cit. on p. 23).

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41 (visited on 07/13/2023) (cit. on p. 47).

Yang, Y., D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil (July 2020). "Multilingual Universal Sentence Encoder for Semantic Retrieval". In: *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics: System Demonstrations. Ed. by A. Celikyilmaz, T.-H. Wen. Online: Association for Computational Linguistics, pp. 87–94. DOI: 10.18653/v1/2020.acl-demos.12. URL: https://aclanthology.org/2020.acl-demos.12 (visited on 01/19/2024) (cit. on p. 22).

Yin, W., N. F. Rajani, D. Radev, R. Socher, C. Xiong (Nov. 2020). "Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, Y. Liu. Online: Association for Computational Linguistics, pp. 8229–8239. DOI: 10.18653/v1/2020.emnlp-main.660. URL: https://aclanthology.org/2020.emnlp-main.660 (visited on 01/23/2024) (cit. on pp. 27, 30).

Yu, M., X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, B. Zhou (June 2018). "Diverse Few-Shot Text Classification with Multiple Metrics". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by M. Walker, H. Ji, A. Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1206–1215. DOI: 10.18653/v1/N18-1109. URL: https://aclanthology.org/N18-1109 (visited on 01/24/2024) (cit. on p. 28).

Yu, X., T. Chatterjee, A. Asai, J. Hu, E. Choi (Dec. 2022). "Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Y. Goldberg, Z. Kozareva, Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3725–3743. DOI: 10.18653/v1/2022.findings-emnlp.273. URL: https://aclanthology.org/2022.findings-emnlp.273 (visited on 01/18/2024) (cit. on p. 20).

Zahoor, S., R. Rohilla (June 2020). "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study". In: *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 537–542. DOI: 10.1109/ICRITO48877.2020.9197910. URL: https://ieeexplore.ieee.org/abstract/document/9197910 (visited on 01/24/2024) (cit. on p. 31).

Zeng, G., Y. Zhang, Y. Zhou, B. Fang, G. Zhao, X. Wei, W. Wang (Oct. 2023). "Filling in the Blank: Rationale-Augmented Prompt Tuning for TextVQA". In: *Proceedings of the 31st ACM International Conference on Multimedia*. MM '23. New York, NY, USA: Association for Computing Machinery, pp. 1261–1272. DOI: 10.1145/3581783.3612520. URL: https://dl.acm.org/doi/10.1145/3581783.3612520 (visited on 01/25/2024) (cit. on p. 37).

Zeroual, I., A. Lakhouaja (Jan. 2020). "MulTed: a multilingual aligned and tagged parallel corpus". In: *Applied Computing and Informatics* 18.1/2. Publisher: Emerald Publishing Limited, pp. 61–73. ISSN: 2210-8327. DOI: 10.1016/j.aci.2018.12.003. URL: https://doi.org/10.1016/j.aci.2018.12.003 (visited on 01/17/2024) (cit. on p. 19).

Zhang, H., X. Zhang, H. Huang, L. Yu (Dec. 2022). "Prompt-Based Meta-Learning For Few-shot Text Classification". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 1342–1357. DOI: 10.18653/v1/2022.emnlp-main.87. URL: https://aclanthology.org/2022.emnlp-main.87 (visited on 01/19/2024) (cit. on p. 36).

Zhang, X., Y. Yin, X. Meng, H. Zhao (Oct. 2008). "Text Classification Based on Rule Mining by Granule Network Constructing". In: *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*. Vol. 2, pp. 514–518. DOI: 10.1109/FSKD.2008.444. URL: https://ieeexplore.ieee.org/document/4666170 (visited on 01/23/2024) (cit. on p. 25).

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen (June 2023). *A Survey of Large Language Models*. arXiv:2303.18223 [cs]. DOI: 10.48550/arXiv.2303.18223. URL: http://arxiv.org/abs/2303.18223 (visited on 07/30/2023) (cit. on p. 12).

Zhao, X., S. Ouyang, Z. Yu, M. Wu, L. Li (July 2023). "Pre-trained Language Models Can be Fully Zero-Shot Learners". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 15590–15606. URL: https://aclanthology.org/2023.acl-long.869 (visited on 07/28/2023) (cit. on p. 12).

Zhong, R., K. Lee, Z. Zhang, D. Klein (Nov. 2021). "Adapting Language Models for Zero-shot Learning by Meta-tuning on Dataset and Prompt Collections". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2856–2878. DOI: 10.18653/v1/2021.findings-emnlp.244. URL: https://aclanthology.org/2021.findings-emnlp.244 (visited on 01/24/2024) (cit. on p. 30).

Zhong, Z., D. Friedman, D. Chen (June 2021). "Factual Probing Is [MASK]: Learning vs. Learning to Recall". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou. Online: Association for Computational Linguistics, pp. 5017–5033. DOI: 10.18653/v1/2021.naacl-main.398. URL: https://aclanthology.org/2021.naacl-main.398 (visited on 01/26/2024) (cit. on p. 35).

Zhou, Y., A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba (Mar. 2023). *Large Language Models Are Human-Level Prompt Engineers*. arXiv:2211.01910 [cs]. DOI: 10.48550/arXiv.2211.01910. URL: http://arxiv.org/abs/2211.01910 (visited on 01/29/2024) (cit. on pp. 13–17, 32, 34, 39–42, 44–46, 59–61).

Zhu, W., H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, L. Li (May 2023). *Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis*. arXiv:2304.04675 [cs]. DOI: 10.48550/arXiv.2304.04675. URL: http://arxiv.org/abs/2304.04675 (visited on 07/28/2023) (cit. on p. 12).

Ziegler, C.-N., M. Skubacz (2012). "Towards Automated Reputation and Brand Monitoring on the Web". en. In: *Mining for Strategic Competitive Intelligence: Foundations and Applications*. Ed. by C.-N. Ziegler. Studies in Computational Intelligence. Berlin, Heidelberg: Springer, pp. 109–119. ISBN: 978-3-642-27714-6. DOI: 10.1007/978-3-642-27714-6_6. URL: https://doi.org/10.1007/978-3-642-27714-6_6 (visited on 01/28/2024) (cit. on p. 30).