# Master's Thesis
# Exploring the Power of Large Language Models: News Intention Detection using Adaptive Learning Prompting


**Master's Thesis**

**Alberto Caballero Hinojosa**

Master Thesis for the

Master in Language Technologies

Universidad Nacional de Educacion a Distancia

Directed by

**Prof. Dr. D. Álvaro Rodrigo Yuste**

**Prof. Dr. D. Roberto Centeno Sánchez**

September 2023

# Abstract

In today's interconnected global landscape, political news wields significant influence in shaping voter perceptions. Political news stands as a primary source of information for citizens in democratic societies. Understanding the intentions underlying news articles is indispensable for ensuring transparency in political discourse. It empowers the public to make informed decisions and to hold media platforms accountable for the information they disseminate, whether directly or indirectly.

Precise identification of news intentions holds the potential to safeguard democratic processes, forming the core focus of this Master's thesis. Accurate identification of news intentions assumes a central role in Natural Language Processing (NLP) for effective information analysis and categorization. This research project delves into the innovative concept of "Adaptive Learning Prompting". It harnesses the formidable capabilities of Large Language Models (LLMs) for efficient news intention classification through iterative prompt engineering, with a specific emphasis on resource-constrained scenarios.

Employing a dynamic evaluation methodology conducted in iterative batches, this research meticulously tracks model performance. Through trend analysis, the evolution of the proposed model architecture emerges as statistically significant, effectively demonstrating the adaptability inherent in this proposed approach.

In summary, this Master's thesis serves as a testament to the potential of LLMs in adaptive news intention identification. Particularly noteworthy is the "Adaptive Learning Prompting" strategy. While specific performance metrics may exhibit variations, the adaptive approach underscores the feasibility of iterative feedback prompting for operational enhancements, especially in resource-scarce scenarios. These findings transcend the realm of news intention analysis, offering broader applications in NLP, while shedding light on the reasoning capabilities intrinsic to LLMs, thereby amplifying their utility.

# Index

# List of Figures

# List of Tables

# Chapter 1

# Introduction

News articles are not mere conveyors of facts, they may carry intentions meticulously crafted by their authors, whether to inform, persuade, advocate, or provoke. These intentions can be as diverse as the topics they cover and the ideologies they represent. The ability to accurately identify these intentions is a critical aspect of media literacy and information quality assessment, especially in our interconnected and digital society.

In an era defined by the relentless overflow of news and information through digital channels and social media, the importance of news intention identification cannot be overstated. The proliferation of fake news, sensationalism, and the deliberate manipulation of news content underscore the urgency of developing effective techniques in this domain.In today's information-rich landscape, assessing the credibility and impact of news articles goes beyond simply verifying their factual accuracy. While ensuring the accuracy of news content remains a fundamental aspect of journalism and information dissemination, it has become increasingly evident that understanding the intentions driving the creation of news articles is equally crucial (Bermes, 2021).

News articles can be crafted with various aims, ranging from providing objective information to persuading or manipulating readers. These intentions can impact the tone, framing, and selection of facts within a news piece, ultimately shaping how the information is interpreted and received by the audience.

By considering the intentions behind news articles, we can gain insights into potential biases, agendas, or hidden narratives that may be at play. This broader perspective allows us to critically evaluate news sources, discern potential misinformation or disinformation campaigns, and better understand the broader context in which news is produced and consumed.

In summary, while factual accuracy remains essential, recognizing and analyzing the intentions behind news articles provide a more comprehensive understanding of their potential impact on public perception and the overall health of our information ecosystem. This shift in focus underscores the evolving complexities of media literacy and the need for a more nuanced approach to news evaluation and consumption. Furthermore, news intention identification contributes to the broader landscape of information verification within Natural Language Processing (NLP). It aligns with the overarching goal of distinguishing between authentic and misleading content, a challenge that extends beyond news articles to encompass various forms of textual information, including social media posts, reviews, and opinions (Shu et al., 2017).This foundation sets the stage for a comprehensive examination of the technologies that form the core of the "Adaptive Learning Prompting" strategy within the domain of news intention identification. Importantly, these developments align with the ideals of democracy and freedom as advocated in ancient Athens, where the informed citizenry's critical thinking and reasoning abilities were considered essential for governance and the well-being and development of the Polis.

In this modern context this Master's thesis delves deep into the realm of Natural Language Processing, focusing on the classification of news intentions. This research harnesses the latest advancements in Large Language Models (LLMs) to address the challenges posed by the contemporary information landscape, especially in low-resource scenarios.

This work introduces a novel approach, "Adaptive Learning Prompting", that leverages the formidable capabilities of Large Language Models. This approach not only enhances classification accuracy but also provides cost-effective and operationally ready systems through iterative prompt engineering, enabling the system to learn from misclassified inputs and providing clearer guidance for labeling future items.

Another contribution of this work to the study of political news intention identification is the creation of a curated and specific dataset that includes human-labeled news intentions. To our knowledge, this dataset represents the first of its kind and aspires to facilitate further research in this direction. This dataset derived from the FakeNewsNet (Shu et al., 2018) is explored, and two distinct types of models are developed to assess

the efficiency of different approaches.

On one hand, a fine-tuned pre-trained BERT model is created as the project's baseline, representing the traditional approach to multiclass classification problems (Han et al., 2021). On the other hand, a series of progressive adaptations in prompt engineering are introduced, utilizing various prompting strategies to evaluate the performance of this approach in the specific task (Balkus and Yan, 2022).

In conclusion, this Master's thesis sets up to explore the potential of LLMs in adaptive news intention identification through the innovative "Adaptive Learning Prompting" strategy for news intention identification. These findings underscore the importance of contextually informational inputs in improving LLMs classification performance, even in resource-constrained environments. Importantly, this approach demonstrates that iterative feedback can be achieved without extensive fine-tuning, rendering it operationally viable from the early stages of implementation.

These insights extend beyond news intention analysis, opening doors to further exploration in various realms of NLP applications, particularly those facing low-resource challenges. Furthermore, this study suggests that Generative Pretrained Transformers (GPT's) capabilities extend beyond text generation, demonstrating evidence of generating behavioral patterns resembling reasoning. The capacity for learning and improving from past errors further amplifies these reasoning abilities.

All the code generated through this whole work is available in its Github repository [1]

[1] https://github.com/alberto2020china/thesis/

## 1.1. Hypotheses and Research Questions

In this section, we proceed to formulate the hypotheses and research questions, which will serve as the guiding principles steering the course of this research work. These hypotheses and questions are instrumental in providing a structured framework for investigation, enabling us to explore, analyze, and draw meaningful insights from this study in LLMs and how they can be applied successfully to real case scenarios.

### 1.1.1. Hypotheses

These hypotheses revolve around the transformative potential of Large Language Models and their potential impact in the field of news intention identification. We also delve into the realms of iterative prompt engineering and adaptive strategies, hypothesizing their potential to elevate the effectiveness of news intention identification.

#### Hypothesis 1

LLMs can significantly improve the accuracy of news intention identification.
We hypothesize that LLMs, with their remarkable language understanding capabilities, can be effectively employed to improve the accuracy of news intention identification. By leveraging their contextual awareness and reasoning abilities, we anticipate that these models will outperform traditional methods in this task, especially in situations of data-scarcity.

#### Hypothesis 2

Iterative prompt engineering can enhance the precision and recall of news intention identification.
We hypothesize that the iterative refinement of prompts, guided by performance feedback, will result in increased precision and recall in news intention identification. This iterative approach enables the model to adapt and refine its understanding of the task, ultimately leading to more accurate classifications.

### Hypothesis 3

The "Adaptive Learning Prompting" strategy will significantly improve news intention identification performance.
Our hypothesis posits that the "Adaptive Learning Prompting" strategy, with its dynamic integration of misclassified instances into the prompt, will lead to an increase in overall precision, recall, and F1-score. This adaptive approach is expected to demonstrate the potential for continuous enhancement in news intention identification.

## 1.1.2. Research Questions

As we delve deeper into our research, a set of fundamental questions emerges to guide our exploration. These research questions direct our efforts towards a deeper understanding of the complex task of news intention identification. Each question represents a distinct facet of our study, designed to further our understanding about the utilization of LLMs, the role of iterative prompt engineering, and the promise of adaptive strategies in this area.

### Research Question 1

How can Large Language Models be effectively leveraged for news intention identification?.

This fundamental question drives our exploration into the realm of NLP and the potential of LLMs to enhance the accuracy of news intention identification. We seek to uncover novel strategies and techniques for harnessing the power of these models in this specific domain.

### Research Question 2

What role does iterative prompt engineering play in improving news intention identification accuracy?.

Our second research question focuses on the iterative refinement of prompts and how it can steer LLMs toward more precise news intention recognition. We aim to understand

the impact of iterative feedback in fine-tuning LLMs for this task.

### Research Question 3

Can news intention identification benefit from adaptive strategies, such as the "Adaptive Learning Prompting" proposed in this thesis?.

The third question probes the potential of adaptive techniques, particularly the "Adaptive Learning Prompting" strategy introduced in this research. We seek to evaluate whether dynamic and contextually nuanced cues can significantly enhance the classification of news intentions.

## 1.2. Thesis Overview

In this section, we provide an overview of the structure and content of the thesis, offering a glimpse into the chapters and their contributions.

**Chapter 2. Previous Work**
In this section we delve into the existing body of knowledge and research that forms the foundation of this work. We explore the evolution of NLP technologies, with a particular focus on Large Language Models and their role in shaping the possibilities of news intention identification. This chapter provides insights into the state of the art and sets the stage for the contributions introduced by this thesis.

**Chapter 3. Methodology**
This chapter introduces the dataset employed as well as the models tested and the methodology used to compare and extract insights into their performance.

**Chapter 4. Results**
In this chapter, we present the results of our research. We begin by examining metrics and evaluation in detail, focusing on the significance of the F1 score. Both static and dynamic evaluation results for each model are presented.

**Chapter 5. Discussion**
Our analysis extends to a comprehensive exploration of the results. We compare the performance of the different models, identifying trends and interpreting metrics to gain insights into the strengths and weaknesses of each approach, returning to the original hypothesis and research questions to determine their validity.

**Chapter 6: Conclusion and Further Work**

The chapter moves forward by discussing the broader implications and contributions of our research. We examine how the "Adaptive Learning Prompting" strategy impacts news intention identification and discuss the operational viability and reduction of fine-tuning. We underscore the significance of contextually nuanced cues and iterative prompt engineering in improving news intention identification.

Additionally, we highlight potential directions for further research and the broader applications of our approach in the realm of NLP.

To conclude this introduction, we can refer to the wisdom of ancient Greece, which reminds us that Democracy is the foundation of society, and an informed citizenry is its strength. Just as the Athenians valued informed participation in governance, today, we must leverage technology to discern the intentions behind news content and uphold democratic principles.

# Chapter 2

# Previous Work

This chapter reviews the work laid by previous research in the field of Natural Language Processing. Herein, we explore the multifaceted landscape of information verification in NLP, dissecting the methodologies and approaches that have paved the way for this thesis. Furthermore, we unravel the transformative influence of Large Language Models and their role in the realm of NLP.

## 2.1. Information Verification in NLP

The landscape of Natural Language Processing has witnessed a burgeoning interest in addressing the intricate challenge of information verification. This endeavor assumes paramount importance in light of the escalating spread of misinformation and disinformation across digital platforms. As individuals navigate a complex web of textual content, the critical need to distinguish between authentic, reliable information and deceptive, misleading narratives becomes increasingly pronounced (Oshikawa et al., 2018).

In this context, the task of information verification emerges as a capital aspect in safeguarding the integrity of communication channels. Information verification encompasses multifaceted aspects, ranging from fact-checking to discerning the underlying intentions and motives behind textual content. The broader goal is to equip individuals, organizations as well as automated systems with the ability to ascertain the trustworthiness and credibility of textual information (Guo et al., 2022).

### 2.1.1. Approaches to Fake News Detection

The detection of fake news has been a foundational area of research, serving as a precursor to news intention identification. Traditional approaches, such as fact-checking and source analysis, have long been instrumental (Zubiaga et al., 2018). Fact-checking organization like Snopes[1]and PolitiFact[2] have played a crucial role in meticulously evaluating the accuracy of news articles through in-depth investigations. Additionally, source analysis involves assessing the credibility of news outlets and authors, aiding in determining the reliability of the information presented (Opdahl et al., 2023).

Linguistic pattern recognition has also been instrumental in identifying potentially fake news. This approach focuses on identifying grammatical inconsistencies, sensational language, and the overuse of emotionally charged expressions as potential indicators of misinformation (Conroy et al., 2015). Leveraging linguistic analysis, these methods aim to uncover deceptive narratives.

Machine learning-based approaches have played a prominent role in fake news detection. Feature engineering, which involves extracting pertinent information from news articles, has been a common practice. These features encompass textual content, metadata, and user-generated data (Castillo et al., 2011).

Supervised learning algorithms, including decision trees, random forests, and support vector machines, have been applied to classify news articles as fake or genuine, though the scarcity of labeled training data has posed a significant challenge (Shu et al., 2018).

Unsupervised learning techniques, such as clustering and anomaly detection, have also been employed to identify unusual patterns or outliers within news content as potential markers of fake news (Yang et al., 2016).

Deep learning techniques have revolutionized the field of fake news detection. Recurrent Neural Networks (RNNs) have been utilized to model sequential dependencies in textual data, capturing temporal information effectively |(Mikolov et al., 2013).

Convolutional Neural Networks (CNNs) are employed to detect local patterns and structural information within news articles (Ma et al., 2017). Attention mechanisms further enhance model performance by focusing on critical phrases or sentences within the text (Qazi et al., 2020).

[1]Snopes. (n.d.). https://www.snopes.com
[2]PolitiFact. (n.d.). https://www.politifact.com

### 2.1.2. Existing Methods for News Intention Identification

Parallel to fake news detection, the emerging field of news intention identification focuses on understanding the motives and biases underlying news articles. This nuanced endeavor delves into comprehending the purpose behind news narratives, unraveling the potential impacts they might have on readers'perspectives and actions.

There has been attempts to use sentiment analysis to mine opinions (Liu, 2015), emotions and latent topics. While others have conducted a survey of techniques, datasets, and evaluation measures for stance detection (Mohammad et al., 2016). Some researchers have developed probabilistic topic models to identify the underlying themes in a document (Blei, 2012).

The advent of Large Language Models has entailed a transformative dimension to news intention identification. By harnessing the inherent context-capturing capabilities of LLMs, In this work (Radford et al., 2018), Radford was able to dive deeper into the subtleties of textual content to unveil underlying intentions. These models exhibit the capacity to comprehend linguistic intricacies, tone, and connections within the text, thereby facilitating a more comprehensive understanding of news narratives.

More recently some teams have proposed a novel approach to intent analysis of social media information using uncertainty-aware reward-based deep reinforcement learning. The proposed approach first uses a generative pre-trained transformer model to extract features from the text (Guo et al., 2023). These features are then used to train a deep reinforcement learning agent to learn to predict the intent of the text. The agent is rewarded for making accurate predictions and penalized for making inaccurate predictions. The uncertainty of the agent's predictions is also taken into account, so that the agent is more likely to explore promising actions even when it is uncertain about the correct intent.

In summary, the study of information verification in NLP emerges as a critical undertaking in an era characterized by information abundance and complexity. The need to discern authenticity from deception and uncover intentions from narratives has encouraged researchers to explore innovative techniques and strategies. As this thesis delves into the Adaptive Learning Prompting strategy for news intention identification it takes a significant stride toward enhancing the veracity and depth of information verification in the age of digital communication.

## 2.2. Overview of Large Language Models (LLMs)

The evolution of Natural Language Processing has witnessed a revolutionary stride with the advent of Large Language Models (Radford et al., 2018). These game changer creations have redefined the boundaries of machine learning, showcasing their transformative ability to comprehend and generate human language with unparalleled precision and fluidity. In this section we proceed to explore the concept of LLMs. At the core of LLMs lies an complex architecture that enables them to decipher the intricacies of language. These models harness the power of deep learning and sophisticated neural networks to internalize the vast intricacies of language patterns, syntax, and semantics. This goes beyond mere rule-based approaches, embracing the nuances and context are the very core of the nature of human language.

One of the most remarkable achievements of LLMs is their capacity to generate text that mirrors the fluidity and coherence of human language. Through exposure to massive datasets, LLMs learn to emulate the stylistic variations, vocabulary diversity, and semantic subtleties that constitute effective communication. The outcome is text generation that seamlessly flows, resonates naturally, and is often indistinguishable from human-authored content. A defining feature of LLMs is their pretraining phase, where models are exposed to colossal volumes of textual data. This immersion equips LLMs with a foundational understanding of linguistic structures, cultural context, and idiomatic expressions.

Consequently, these models emerge from pretraining with an innate grasp of grammar, semantics, and even societal nuances, a solid foundation upon which subsequent fine-tuning is built. The influence of LLMs spans across a spectrum of applications. From sentiment analysis and language translation to question-answering systems and content summarization, LLMs demonstrate their adaptability and prowess. Their ability to unravel context, decode idiomatic expressions, and navigate linguistic ambiguity contributes to their transformative impact across diverse NLP tasks (Radford et al., 2019).

### 2.2.1. Transformers Architecture

At the core of LLMs lies the innovative Transformer architecture (Vaswani et al., 2017). the Transformer architecture revolutionized NLP by introducing a novel mechanism for handling sequential data, such as text. Unlike traditional recurrent neural networks or convolutional neural networks, which rely on sequential processing, Transformers employ attention mechanisms to capture relationships between words in a non-sequential manner.

The Transformer architecture facilitates parallelization, enabling the model to process input data more efficiently. This parallelization, coupled with self-attention mechanisms, allows LLMs to consider the context of every word in a sentence simultaneously.

This context-awareness is a crucial factor in understanding the nuances and complexities of natural language, making LLMs particularly adept at tasks like language translation, sentiment analysis, and, significantly and as we will explore in this thesis, news intention identification.

At the heart of the Transformer architecture is the concept of self-attention. This mechanism allows the model to consider the relationships between all words in a sentence simultaneously, rather than processing them sequentially as in RNNs.

This parallelization significantly accelerates the training and inference processes, making Transformers highly efficient.

To grasp the significance of Transformers in NLP, it's crucial to explore their core components. A clearer understanding of the Transformer's architecture can be gained by referring to Figure 2.1, which showcases a typical architecture, highlighting its various components.

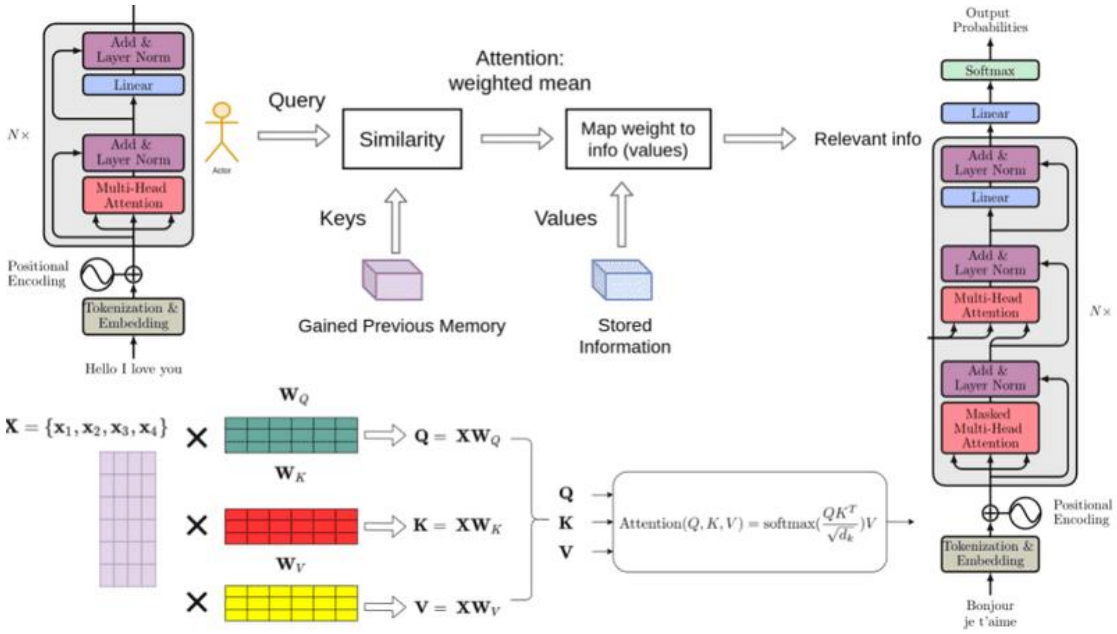


Figure 2.1: Transformers Inner Architecture

**Multi-Head Self-Attention:**

This mechanism enables each word in a sentence to focus on different parts of the input, capturing complex relationships and dependencies.

**Positional Encoding:**

Transformers lack inherent information about the order of words in a sequence. Positional encodings are added to the input embeddings to provide the model with information about the positions of words.

**Feedforward Neural Networks:**

After attention mechanisms, feedforward neural networks are applied to each word's representation, introducing non-linearity and further enhancing the model's expressiveness.

**Layer Normalization and Residual Connections:**

These techniques help stabilize training and mitigate vanishing gradient issues, allowing for the training of deep models.

The impact of Transformers is most evident in the development of pre-trained language models. Models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and GPTs (Radford et al., 2018) have achieved groundbreaking results across a wide range of NLP tasks. BERT, for instance, employs a bidirectional training approach that enables it to understand the context of a word by considering both its left and right surroundings.

This contextual understanding has revolutionized tasks like text classification, named entity recognition, and sentiment analysis. GPTs, on the other hand, is a generative model capable of producing coherent and contextually relevant text. It has demonstrated remarkable performance in tasks involving text generation, language translation, and even question-answering.

The following part of this section will delve into the specific ways in which Transformers, and a greater extent Large Language Models, have been leveraged for specific tasks, namely news intention identification.

**Task-Specific Dataset:**

A dataset relevant to the task is prepared, containing labeled examples. For sentiment analysis, this dataset might consist of text samples with corresponding sentiment labels as in our particular case study, the different intentions of news.

**Loss Function and Optimizer:**

A loss function that measures the model's performance on the task and an optimizer to update the model's parameters are chosen.

**Fine-Tuning Process:**

During fine-tuning, the model is exposed to the task-specific dataset, and its parameters are adjusted based on the loss computed after each iteration (batch). The model's original weights are retained, and only the final layers are updated to adapt to the task.

**Transfer Learning:**

Fine-tuning leverages the knowledge gained from pre-training, allowing the model to generalize well even on smaller task-specific datasets.

**Regularization and Hyperparameter Tuning:**

Regularization techniques, such as dropout, and hyperparameter tuning ensure the model's performance doesn't overfit the task-specific data.

In summary, LLMs owe much of their versatility to their ability to be fine-tuned for specific tasks. The process of prompt engineering further enhances their adaptability by providing task-related context and guidance. This combined approach empowers LLMs to deliver remarkable performance across a range of NLP tasks, including the innovative "Adaptive Learning Prompting" strategy explored in this thesis.

### 2.2.2. Harnessing the power of Large Language Models (LLMs): Fine-Tuning and Prompt Engineering

Large Language Models have brought about a paradigm shift in Natural Language Processing, enabling remarkable language understanding and generation capabilities. LLMs are trained on massive datasets of text and code, and they learn to represent the statistical relationships between words and phrases. This allows them to perform a variety of tasks, such as text translation, question answering and summarization.

Among the prominent LLMs, OpenAI's GPT series (Brown et al., 2020) and Bidirectional Encoder Representations from Transformers (BERT)(Devlin et al., 2018) have demonstrated exceptional performance. These models have been fine-tuned on a variety of tasks, and they have achieved state-of-the-art results.

Fine-tuning is the process of retraining an LLM on a new dataset of data that is relevant to the task at hand. This can be done by adjusting the weights of the model's parameters so that it learns to perform the task more accurately. Fine-tuning can be a time-consuming and computationally expensive process, but it can significantly improve the performance of an LLM on a specific task.

Prompting is a contemporary approach to harnessing and leveraging LLMs. In this method, users provide concise textual prompts that instruct the model on the desired task. For instance, a prompt like Compose a Greek drama set in Ancient Athens, directs the model to generate a script inspired by the training data of classical theatrical works. Prompting is a less computationally expensive approach than fine-tuning, and it can be used to generate text on a variety of tasks, even tasks that the model has not been explicitly trained on.

Both fine-tuning and prompting have their own advantages and disadvantages. While Fine-tuning is more likely to produce accurate results it is also more time-consuming and computationally expensive. Prompting is less accurate, but it is faster and easier to use (Wang et al., 2022)(Shi and Lipani, 2023).

Some of the main differences between these two approaches are:

**Fine-tuning:**

Requires a large dataset of labeled data for the specific task. Can be more difficult to implement, as it requires the user to have extensive knowledge of machine learning and coding. Fine-tuning can be more expensive, as it requires more computational resources. Prompting is less expensive, as it can be done with fewer computational resources depending on whether we are using open or close sourced models.

**Prompting:**

Does not require labeled data, but it may require more experimentation to find effective prompts. It is easier to implement, as it does not require any machine learning knowledge.

### Fine-Tuning of LLMs for Specific Tasks

The process of fine-tuning Large Language Models is a pivotal methodology for adapting these pre-trained models to perform specialized tasks, enhancing their utility across various domains. Fine-tuning involves taking a pre-trained LLM, which has already captured a wealth of general language knowledge and refining it to excel at a specific task. This approach facilitates the transfer of the extensive linguistic understanding encoded in the LLMs parameters to domain-specific applications (Iman et al., 2023).

The fine-tuning process unfolds through a sequence of essential steps, designed to align the LLMs capabilities with the targeted task's requirements. The journey commences with the selection of an appropriate pre-trained model, serving as the foundation on which task-specific expertise will be built. The chosen LLM, having learned linguistic nuances and contextual intricacies from a broad range of text sources, sets the stage for focused adaptation. The next stage involves training the pre-trained LLM on task-specific data, which is meticulously annotated with relevant labels (Devlin et al., 2018). The annotations serve as the guiding compass, steering the model toward learning the correlations between input text and desired outputs. During this training phase, the

model's parameters are fine-tuned, allowing it to gradually grasp the intricacies and idiosyncrasies inherent to the targeted task. This nuanced learning process equips the LLM to generate task-specific responses with remarkable precision.

However, despite its effectiveness, the fine-tuning process is not without limitations. One key challenge lies in its resource-intensive nature, demanding substantial computational power and labeled task-specific data. Moreover, fine-tuned models may exhibit overfitting, performing exceedingly well on training data but struggling to generalize to new, unseen examples. These limitations are particularly pronounced in scenarios where data or resources are scarce, rendering downstream fine-tuning infeasible. Given these constraints, especially in data-scarce situations, researchers and practitioners have tried to explore alternative strategies that balance model performance, efficiency, and generalization.

This thesis investigates such an alternative strategy, prompt engineering, in the context of news intention identification, aiming to address the challenges posed by fine-tuning while fostering adaptability and efficiency.

### 2.2.3. Generative Pretrained Transformers (GPTs)

At the heart of the transformative advancements within Large Language Models lies the emergence of Generative Pretrained Transformers. These models encapsulate in many ways the cutting-edge of Natural Language Processing by seamlessly integrating generative capabilities with context-aware text generation. By delving into GPT's underlying architectural foundation and its implications for reasoning, we can discern the symbiotic relationship between Transformers, reasoning capabilities, and the objectives of the present research (Azaria and Mitchell, 2023).

GPTs models, as transformative offspring of transformers, embody the epitome of NLP innovation. Transformers, with their self-attention mechanisms and multi-head architectures, revolutionized sequence-to-sequence tasks by capturing contextual dependencies with unprecedented efficacy. GPTs, in its generative essence, inherits this foundation, enabling it to generate coherent and contextually relevant text that resonates naturally.

While GPT's generative prowess is evident, it raises intriguing questions about its reasoning abilities. Reasoning, as a cognitive function, encompasses more than pattern recognition, it involves extrapolation, inference, and the capacity to draw conclusions from acquired information. GPT's foundation in pretraining, while providing substantial lin-

guistic understanding, may not inherently confer it with explicit reasoning capabilities. This urges us to explore the nexus between generative language models, such as GPT, and the extent to which they embody rudimentary reasoning mechanisms.

This exploration finds resonance in the context of the research on "Adaptive Learning Prompting" for news intention identification. This research quest mandates the discernment of nuanced linguistic inputs within news articles, often requiring a form of inferential reasoning. While GPTs may not fully replicate human-like reasoning, its ability to weave contextually coherent narratives positions it as an important element for bridging the gap between language generation and some sort of reasoning.

### 2.2.4. Role of Prompting in LLMs

Pre-training endows Large Language Models with a fundamental understanding of language. However, the pivotal role of prompting cannot be understated, as it guides LLMs'behavior toward specific tasks. A prompt essentially serves as textual input that instructs or contextualizes the LLM for a particular task. In the field of news intention identification, the utilization of prompts becomes a potent mechanism to drive and guide the model's attention (Liu et al., 2023).

Prompt engineering is a strategic process that involves designing prompts that effectively guide the LLM's generation towards the desired outputs. Through the creation of well-crafted prompts and an iterative learning feedback loop, LLMs can be finely tuned to excel at distinct tasks, including the intricate domain of news intention identification (Brown et al., 2020).

#### Techniques for Prompt Engineering

Prompt engineering has emerged as a highly effective and innovative strategy within the realm of Natural Language Processing, providing an alternative avenue to overcome the challenges associated with resource-intensive fine-tuning. This approach not only mitigates the need for extensive fine-tuning but also enhances the flexibility and versatility of Large Language Models (Schick and Schutze, 2021). The concept revolves around crafting specialized prompts that guide the behavior of the model, channeling its language generation capabilities toward specific objectives.

The advancement of prompt engineering techniques has ushered in a new era of fine-tuning alternatives, offering precise control over the output of Large Language Models in a wide array of NLP tasks. These techniques serve as a detour to overcome the challenges of traditional fine-tuning, enabling practitioners to achieve task-specific outcomes with remarkable efficiency and adaptability.

**Instructional Prompts**

The introduction of explicit instructions within prompts presents a powerful avenue for guiding LLM behavior (Schick and Schutze, 2021). This technique equips practitioners with the ability to shape the model's output by providing specific task-related directions. These instructions act as a compass, steering the model's linguistic prowess toward generating responses aligned with the desired objectives.

**Contrastive Prompts**

Drawing inspiration from Schick's work, the application of contrastive prompts takes advantage of the model's discriminative capabilities. By presenting multiple choices and prompting the model to select the correct one, practitioners encourage the model to engage in a process of selection, promoting precise understanding of task nuances and requirements. This technique taps into the model's decision-making abilities, leveraging its capacity to discern subtle differences and arrive at optimal choices.

**Cloze-style Prompts**

Rooted in traditional language completion exercises, cloze-style prompts offer a novel approach to eliciting contextually appropriate responses (Radford et al., 2018). By masking specific portions of the input and prompting the model to predict the missing content, this technique compels the LLM to internalize contextual dependencies. Through cloze-style prompts, the model delves into the intricate fabric of the provided context, unraveling the threads of meaning and generating responses that seamlessly fit the narrative.

**Generative Prompts**

As showcased by (Radford et al., 2018), the introduction of generative prompts encourages the LLM to venture beyond predefined responses and explore creative avenues, allowing the model to generate outputs that transcend mere adherence to rules.

**Prompt Engineering through Reinforcement Learning**

The integration of reinforcement learning principles into prompt engineering brings about a dynamic approach (Schick and Schutze, 2021). By subjecting prompts to iterative optimization guided by reinforcement signals, practitioners nudge the model toward generating desired outcomes. This technique embraces an adaptive learning process, iteratively shaping the model's behavior to align with predefined objectives.

The primary strategy emphasized in this work draws inspiration from this latest technique. This innovative approach, known as Prompt Engineering through Reinforcement Learning, serves as the guiding principle for enhancing the performance of Large Language Models in various tasks, including news intention identification.

By harnessing the power of Reinforcement Learning-based Prompt Engineering, this strategy enables practitioners to iteratively optimize prompts to guide the behavior of LLMs. This technique capitalizes on reinforcement signals to dynamically adjust prompts, steering the model's output generation toward desired outcomes. This approach stands in contrast to traditional fine-tuning methods, which can be resource-intensive and may lead to overfitting (Devlin et al., 2018).

The adoption of this cutting-edge strategy allows for the efficient adaptation of LLMs to specific tasks, even when data or resources are limited. It empowers the model to excel in tasks that demand contextually accurate responses, such as news intention identification (Brown et al., 2020). Through continuous refinement of prompts based on reinforcement signals, the model's performance is improved iteratively, offering a dynamic and effective alternative to traditional fine-tuning.

### 2.2.5. Reasoning Capabilities of LLMs

One of the defining features of LLMs is their remarkable reasoning capabilities (Radford et al., 2018). Unlike earlier NLP models, LLMs can perform intricate tasks that involve understanding context, making inferences, and generating coherent responses. These capabilities are achieved through pre-training on massive datasets containing vast amounts of text from the internet, which allows LLMs to learn grammar, semantics and world knowledge (Yu et al., 2023). In the context of news intention identification, LLMs excel in capturing the subtle cues and patterns in news articles that reveal the intentions behind the content. They can recognize the language used in opinion pieces, the factual reporting in news articles, and the emotional appeal in commonly found in advertisements. Leveraging these reasoning abilities, LLMs hold immense promise for automating the identification of news intentions.

The preceding section underscores the critical importance of information verification in the context of Natural Language Processing. It showcases the evolution of techniques, from traditional approaches like fact-checking and source analysis to the emergence of cutting-edge Large Language Models. These models, with their inherent understanding of linguistic nuances, hold immense potential for enhancing the identification of news intentions. However, while previous work has paved the way, there remain notable gaps and limitations. The primary challenge lies in the efficient adaptation of LLMs to the specific task of news intention identification, particularly in scenarios with limited labeled data. In this respect, this thesis intends to explore possible solutions to these limitations.

# Chapter 3

# Methodology

The methodology section introduces the most relevant methodological aspects of this work, crucial to the execution of any research endeavor in the realm of Natural Language Processing. This journey encompasses three pivotal dimensions: the dataset employed, the models implemented and the evaluation framework used to compare the different models and approaches.

## 3.1. Dataset

In the realm of Natural Language Processing, datasets serve as the very foundation upon which research can be built and contrasted. The FakeNewsNet dataset (Shu et al., 2018) stands out as a comprehensive collection of news articles that exemplifies the interplay between authentic narratives and deceptive content. This dataset comprises a diverse range of articles sourced from various platforms, mirroring the multifaceted nature of information dissemination in this digital age.

One notable contribution of this Master's thesis to the study of political news intention identification is the incorporation of human-based news intention labels into the original FakeNewsNet dataset. To our knowledge, this dataset represents the first of its kind and aspires to facilitate further research in this direction. This addition significantly enriches the dataset, allowing for a deeper exploration of the complexities of news intentions. Furthermore, the FakeNewsNet dataset encompasses a rich variety of variables, including user and news dissemination variables. This multifaceted nature of the dataset provides

researchers with a robust foundation for exploring various approaches to information verification and news intention identification. It is our expectation that this dataset will serve as a valuable resource for advancing the field and shedding light on the evolving landscape of information verification, with a special focus on the use of Large Language Models for news classification.

This contribution sets the stage for our own research, aiming to bridge gaps and overcome limitations in the existing literature concerning news intention identification within the broader field of Information Verification. Regrettably, the initial set of news articles available for scrutiny has undergone a reduction due to the removal of pertinent links, sources, and platforms. This phenomenon is evident in the transition from an original corpus of 336 fake and 447 real news articles, as initially sourced from Politifact, the section in the dataset focusing on political news, this is the subset used in this thesis. Unfortunately, this count has dwindled to 133 fake news articles and 118 real news articles as the webscraping was conducted on July 2023, those from which the original news content could still be extracted. This selective pruning process has yielded a diminished dataset of 251 news articles, forming the foundation of this work.

The reduction in available news articles highlights a common challenge faced in real-world scenarios, where access to extensive and well-curated datasets is often limited. This scenario underscores the critical need for developing robust and adaptable tools for news intention identification, particularly in low-resource situations. In many practical applications, such as monitoring news content on emerging platforms or in resource-constrained environments, researchers and practitioners encounter similar challenges related to data scarcity.

Studying and developing tools tailored to these low-resource scenarios is of paramount importance, as it reflects the real-world conditions where these techniques are most urgently needed. The ability to achieve accurate and efficient news intention identification in such scenarios holds the potential to mitigate the spread of misinformation, improve information quality assessment, and enhance media literacy. Thus, this research project not only addresses a pressing need but also serves as a valuable step towards empowering stakeholders in various fields to navigate the information landscape effectively, even when resources are limited.

### 3.1.1. Labeling Process

The annotation of news intention within the dataset was undertaken to discern distinct types of intent underlying news articles. These intents, as classified by their impact on political narratives, were defined in the labels provided in Table 3.1. These labels were derived from the analysis of political narratives in prior research (Shenhav, 2006).

The process encapsulates the intricate dynamics of news categorization, where the interplay of intent and effect shapes the final annotation. These specific labels were chosen as they align with the predominant categorization of political narratives, providing valuable context for the reader to understand their origin and relevance.

| **Type** | **Description** | **Label** |
|---|---|---|
| Reputation Impact | Influence the reputation of a political actor or institution | A |
| Associative Manipulation | Manipulate voting intentions or discourage participation | B |
| Electoral Legitimacy Concerns | Undermine the legitimacy of the electoral process | C |
| Fear and Uncertainty Propagation | Generate uncertainty, fear, or incite violence | D |

Table 3.1: News Intention Classification

Guided by the provided label definitions, each news article together with its headline was meticulously analyzed by two separate annotators, primary and secondary. Notably, these annotations were conducted at different time points. This dual-annotator approach adds a layer of complexity, capturing diverse perspectives on the same content. The labeling process entailed categorizing shuffled news articles without supplying any background information, such as their type, source, or potential biases, throughout the process.

Significantly, during the annotation process, a notable and substantial divergence of 36 % in labels emerged between the two annotators. This stark contrast underscores the inherent subjectivity involved in categorizing nuanced intent within news articles.

It serves as a crucial reference point for assessing the performance limitations of various models in capturing the intricacies of natural language. Furthermore, it highlights the greater challenge that artificial systems face in this regard, considering that even a human-based annotation system cannot provide perfect alignment. illuminate the challenges inherent in dealing with such complex and subjective categorizations.

Ultimately, only the labels of the primary annotator, intention-main, were utilized due to these discrepancies. Nonetheless, these variations in annotation clearly illuminate the challenges inherent in dealing with such complex and subjective categorizations.

### 3.1.2. Descriptive Statistics of the Dataset

In this segment of our study, we embark on the exploration of the dataset, aiming to attain a profound comprehension of its composition and inherent attributes. Our focus centers on two pivotal variables that better reflect the dataset intricacies and its connections with the field of information verification:

**Type**

This variable dichotomizes the news content into two distinct categories, fake or real as labeled by the compilers of the dataset. By discerning the authenticity of the news, we establish a foundational distinction between fabricated narratives and genuine journalistic content. This variable was originally provided in the original dataset.

**Intention-main**

The label variable assigns distinct intention categories to each news article. This categorization spans the labels mentioned in the previous section. This approach allows for a nuanced examination of the dataset, with the main appendix corresponding to the labels assigned by the primary annotator.

In Figure 3.1, we can see the distribution of fake and real news in the dataset. The distribution is similar, with a slight majority of fake news.

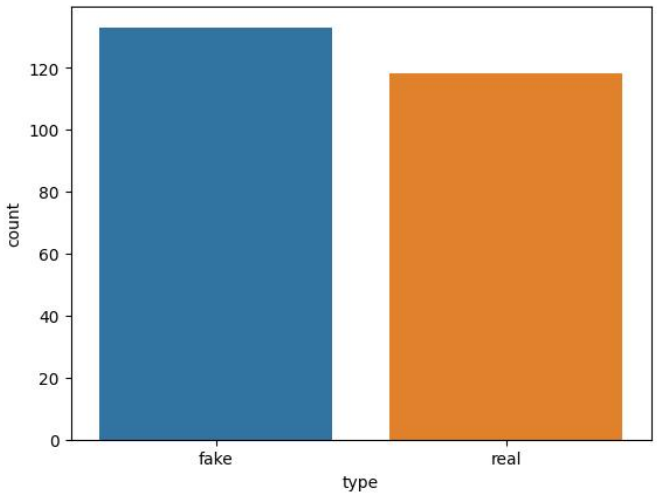


Figure 3.1: Fake Vs Real News Distribution

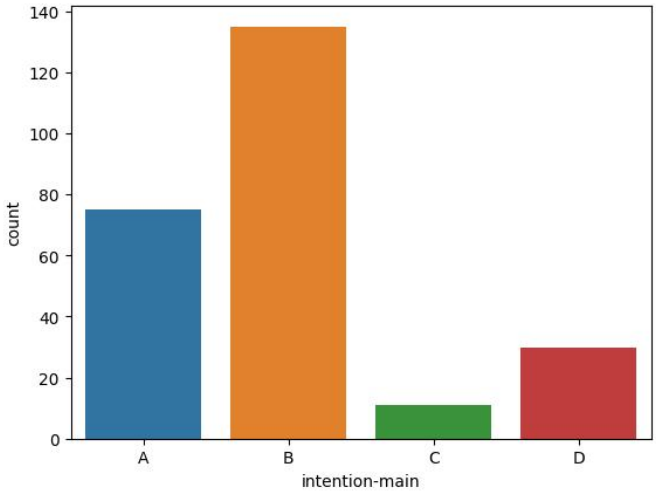


Figure 3.2: News Intention Distribution

In Figure 3.2 we can see how label **B** (Associative Manipulation) is the most prevalent, accounting for 52,4 % of the news articles in the dataset. In contrast, intention type **C** (Electoral Legitimacy Concerns) is notably less common. This imbalanced distribution of intentions introduces a further challenge into the task, particularly in countering the inclination of models to favor label **B** due to its higher occurrence rate. Later on during the analysis, we will explore how this imbalance might affect the performance of the models.

### Relationship between type and intention

The chi-squared test serves to evaluate the independence between categorical variables (Zibran, 2007). In this context, it probes the existence of a notable association between the news type (fake or real) and the intention label attributed to the news articles. The remarkably low p-value underscores substantial evidence against the null hypothesis of independence, signifying a robust association between the two variables.

The chi-squared test yielded a calculated value of 12.96 for the cross-analysis of the variables type and intention-main accompanied by a p-value of approximately 0.0047, below 0.05. This statistical analysis carries significant implications for comprehending the interrelation between the categorization of news articles as fake or real and the intention labels assigned to them. These findings lead us to conclude that a discernible

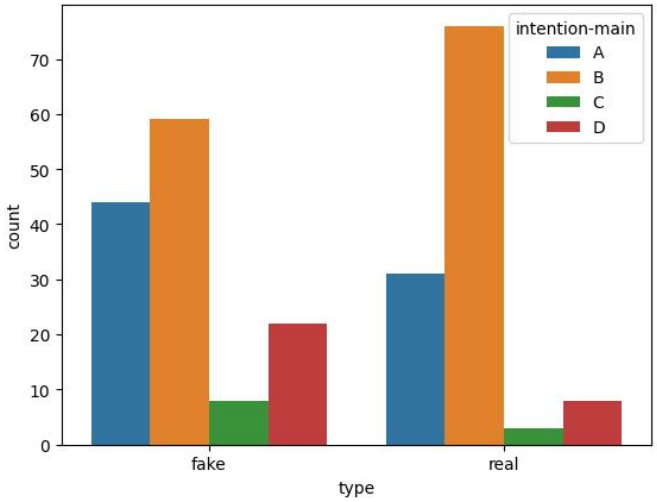


Figure 3.3: Type and Intention Co-distribution

relationship exists between the intention conveyed by news articles and their veracity. Despite the limited scope of the employed dataset, our statistical analysis allows us to confidently assert the presence of a statistically significant correlation between these variables in the dataset subject of this work as we can see in Figure 3.3 where we can observe how **B** (Associative Manipulation) is relatively more common among real news.

Furthermore, Table 3.2 provides a concise overview of the conditional probabilities associated with news articles being classified as either fake or real, contingent on their respective intention labels. Notably, with the exception of label **B** which exhibits a stronger tendency to be categorized as real, all the other labels demonstrate a pronounced inclination towards being associated with fake news. This insight highlights the significance of intention labels **A**,**C** and **D** in identifying news articles with a higher likelihood of being deceptive, further emphasizing their relevance in discerning misinformation within the dataset.

Table 3.2: Probability of Authenticity Assessment Given the Intention

| Intention Label | Probability of Being Fake | Probability of Being Real |
|---|---|---|
| A | 0.428 | 0.572 |
| B | 0.389 | 0.611 |
| C | 0.746 | 0.254 |
| D | 0.733 | 0.267 |

### Cluster and Topic Modeling for Data Understanding

In this section, we delve into the application of cluster and topic modeling techniques to enhance our understanding of the dataset.

Initially, a preprocessing phase is applied to the news articles. To manage the complexity of the data, a limitation is imposed on the length of tokenized sequences (500). This approach aligns with the common practice of sequence truncation in NLP tasks, ensuring computational efficiency without sacrificing essential information.

Leveraging the power of Large Language Models, we employ BERT embeddings to encode the semantic information within the articles (Devlin et al., 2018). BERT contextual understanding of language allows us to capture nuanced relationships between words, enhancing the quality of embeddings.

The dimensionality reduction technique UMAP is then employed (McInnes et al., 2018). UMAP ability to preserve local and global structures makes it an effective choice for visualizing high-dimensional data. This technique facilitates the identification of clusters, a crucial step in uncovering patterns and themes.

Furthermore, HDBSCAN, a density-based clustering algorithm, is utilized. Its capability to identify clusters of varying shapes and sizes ensures a comprehensive categorization of the articles, this clustering process identifies two clusters within the dataset.

In the realm of NLP, topic modeling is a well-established practice for uncovering latent themes within textual data. By using the TF-IDF method, we can identify keywords that hold significant importance within each cluster. This approach draws from the core principles of term frequency and inverse document frequency to highlight terms that are discriminative within specific clusters (Grootendorst, 2022).

The integration of cluster and topic modeling techniques, along with the incorporation of news type and intention, has unveiled complex thematic structures within the dataset. This insight offers a deeper understanding of prevalent themes and narratives. These clusters pinpoint distinct content categories, each characterized by a unique assembly of terms and semantic connections.

This comprehensive analysis provides valuable insights into the diverse range of news intentions encompassed by the dataset presented in Table 3.3. Additionally, it sheds light on the interrelation between this clustering approach and the key variables of the study,

enhancing our comprehension of the dataset's underlying dynamics.

Table 3.3: Cluster Definitions

| Cluster | Size | Topic Modeling Interpretation | By Type | By Intention |
|---|---|---|---|---|
| 0 | 114 | Misinformation and Conspiracy Theories | Fake: 84 Real: 30 | A: 42 B: 51 C: 8 D: 13 |
| 1 | 137 | Healthcare and Policy | Fake: 49 Real: 88 | A: 33 B: 84 C: 3 D: 17 |

Similarly to the analysis conducted in the previous section (Relationship between Type and Intention), we can now turn our attention to studying the correlation between the clusters to which news articles belong and their veracity and intention. This exploration will further unveil the intricate relationships between news clusters and these crucial attributes within the dataset.

**Cluster vs. Type:**

The chi-square test of independence between the Cluster variable and the Type variable reveals significant insights. The very small p-value, close to 0, indicates a highly significant association between clusters and news types (Fake or Real). This suggests that the distribution of news types is not independent of the clusters; certain clusters are strongly associated with specific types of news. The chi-square statistic of approximately 34.41 further supports this finding, indicating that the observed differences between the clusters and types are not likely due to random chance.

**Cluster vs. Intention:**

Similarly, the chi-square test of independence between the Cluster variable and the Intention variable provides valuable insights. The p-value of approximately 0.0192 is below the significance level of 0.05, indicating a statistically significant association between clusters and intentions (**A**, **B**, **C**, **D**). This suggests that certain clusters are related to specific intentions, although the association might be relatively weaker compared to the Cluster vs. Type test. The chi-square statistic of approximately 9.93 supports this interpretation, that the relationship is not likely due to random chance.

The cluster analysis has uncovered distinct thematic patterns within the dataset. Cluster 0 predominantly involves the propagation of misinformation and conspiracy theories, primarily within the realm of fake news. The content's intentions are spread across different

categories, indicating a diverse range of narratives aimed at generating sensationalized and potentially misleading stories.

On the other hand, Cluster 1 is centered around healthcare policy, reform, and benefits, and is mainly composed of real news articles. The intention behind these articles is predominantly associative manipulation (**B**), suggesting a focus on policy-related discussions and potentially more balanced narratives. The presence of terms related to healthcare policy and reform underscores the cluster's focus on addressing healthcare challenges and exploring potential policy changes.

Overall, these findings validate the significant associations observed in the chi-square tests and provide deeper insights into the distinctive thematic content of each cluster. The cluster analysis enhances our understanding of prevalent themes and narratives in the dataset.

### 3.1.3. Data Set Division

We divided our dataset into distinct portions to facilitate model training and evaluation. Specifically, we utilized a training dataset consisting of 151 news articles, employing a stratified method to ensure representation from all the labels from our original dataset of 251 items.

Furthermore, we allocated the remaining 100 articles from our original dataset for the final test set evaluation. This partitioning strategy enables us to train and validate different models using a representative sample of the data

## 3.2. Models

In our quest for accurate news intention identification, we have developed a series of models to assess the validity of the questions and hypotheses at the core of this study. All this is done while taking into account the constraints of operating in a low-resource environment. This section delves into each of these models, providing detailed insights into their design, configuration, and the unique paradigms they embody specifically, fine-tuning and prompt engineering.

The foundation of our exploration lies in a Baseline model, which serves as the reference point for our entire study, especially within a context with resource limitations. The baseline model acts as the cornerstone of our investigation, and its core involves fine-tuning a pre-trained BERT model to specialize its parameters and features for the specific task of multiclass classification. This crucial step establishes the groundwork for our subsequent advancements.

The basic Zero-Shot In-Context-Learning (ICL) model introduces a novel approach. It employs label definitions as prompts, even operational under low-resource scenarios where no examples are available.

The Few-Shot in-Context-Learning model builds upon this methodology by incorporating both label definitions and examples.

The most significant leap occurs with the introduction of the "Adaptive Learning Prompting" strategy. This approach dynamically reintroduces previously misclassified instances into the original Zero-Shot model prompt, now correctly labeled.

In our pursuit of accurate news intention identification, these models provide a comprehensive exploration of different prompting strategies, from Zero-Shot to Few-Shot and "Adaptive Learning Prompting". Each of these models represents distinct paradigms of prompt engineering within the framework of Large Language Models, all while considering the dynamic and resource-constrained nature of Natural Language Processing under specific circumstances.

These models not only embody various prompt designs but also reflect the adaptability required to navigate low-resource scenarios effectively. Through this progression, we gain insights into the strategies that can guide LLMs in comprehending and categorizing news intentions, shedding light on the nuances of language understanding and classification in the realm of political news while efficiently managing limited resources (Brown et al., 2020).

### 3.2.1. Baseline Model: Fine-tuned Pretrained BERT Model

The foundation of our exploration into news intention identification begins with the Baseline Model. This model employs a fine-tuned pretrained BERT ("bert-base-uncased") model for the task of multiclass classification. The model's architecture leverages the powerful pre-trained language understanding capabilities of BERT, enabling it to learn intricate patterns and relationships within the textual data. To rigorously assess the performance of this baseline and ensure its robustness within the constraints of limited resources, we implement a 10-fold cross-validation strategy on the training dataset.

This approach partitions the 151 training articles into ten distinct subsets, with nine folds used for training and one for validation in each iteration.

The cross-validation process iterates ten times, providing us with a comprehensive understanding of the model's behavior across various subsets of the training data, all within the low-resource context. This technique aids in minimizing overfitting and generalizing the model's capabilities to unseen data instances (Rodriguez et al., 2009).

#### Fine-Tuning with 10-fold Cross Validation

Fine-Tuning with 10-fold Cross Validation In the pursuit of robust model evaluation, the baseline model undergoes training and validation using a 10-fold cross-validation methodology. This technique partitions the dataset into ten distinct subsets, or folds, ensuring that each fold serves as both a training and validation set in rotation. During each iteration, nine folds are utilized for training, while the remaining fold is held out for validation.

This process is iterated ten times, with each fold taking a turn as the validation set. The rationale behind 10-fold cross-validation is to evaluate the model's performance across various subsets of the data, mitigating the potential for bias introduced by a single training-validation split, we can have a clearer grasp of this technique by looking at Figure 3.4 to see how the original data is divided and iterated through to obtain a more robust estimation of the model performance.

In this study the baseline model is aligned with the traditional approach of fine-tuning a pre-trained BERT model for news intention identification. The training and testing of this model entails several stages.

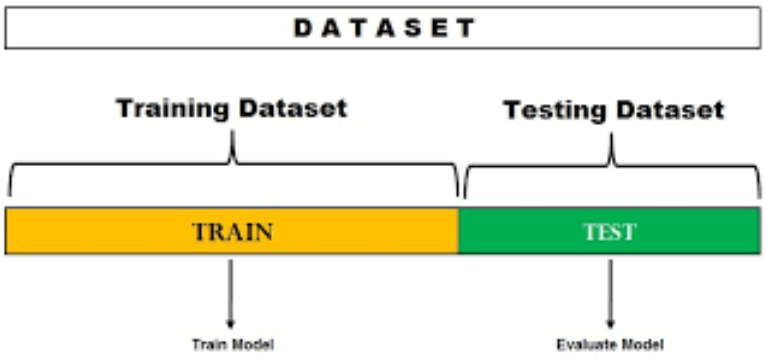


Figure 3.4: Cross-Validation Methodology Summary

1. **Tokenization and Data Transformation**
   It employs the "bert-base-uncased" pre-trained BERT tokenizer to perform tokenization and transform raw textual data into a format compatible with BERT's input requirements. This format includes various encoding components such as input IDs, attention masks, and token type IDs. The transformation process ensures consistent padding, truncation, and adherence to the maximum sequence length.

2. **Data Partitioning**
   The dataset is divided into distinct subsets for training, validation, and testing. This partitioning ensures that separate datasets are available for training the model, validating its performance during training, and assessing its final performance as we have seen in section 3.1.3 (Data Set Division).

3. **Model Initialization and Hyperparameters**
   The baseline model is instantiated using the pre-trained BERT model "bert-base-uncased" configured to address the specific multilabel classification task. The number of unique labels is determined using the label encoder. The AdamW optimizer is chosen with a learning rate of 2e-5. The loss function, CrossEntropyLoss, is employed due to the softmax activation used in multiclass classification tasks.

4. **Training Loop**
   The training process spans a predefined number of epochs (5 in this case) and is organized into batches of size 8. The model is iteratively trained using these batches. For each batch, the optimizer's gradients are reset, and the model's predictions are compared against the actual labels. The CrossEntropyLoss computes the loss, which is then back-propagated to update the model's parameters.

5. **Validation**
   After each epoch, the performance of the model is evaluated on the validation dataset. The model generates predictions on the validation inputs while in evaluation mode. The validation loss is computed and monitored, providing insights into the model's convergence and performance on unseen data.

6. **Testing and Metric Calculation**
   Following the completion of training, performance is assessed on the held-out testing dataset. The model's logits are transformed into class probabilities using the softmax activation function. Each instance's predicted label is determined by selecting the class with the highest probability.

7. **Results Aggregation**
   Testing metrics are aggregated over multiple iterations (10 in this case) to ensure statistical robustness. Predicted labels and true labels are collected across these iterations for further analysis.

#### Developing an Ensemble Voting Model as the Baseline Model

To further elevate the performance and resilience of our baseline model, we employed ensemble learning. Recognizing the merits of combining predictions from multiple models, we develop an Ensemble Voting Model using the ten iterations of our fine-tuned BERT model obtained through 10-fold cross-validation.

This ensemble aggregates predictions from these iterations, effectively synthesizing their outputs to yield a collective verdict. This approach seeks to capitalize on the diversification of individual models and harnesses their collective intelligence to make more informed decisions.

The Ensemble Voting Model follows a majority voting scheme, wherein the label predicted by the majority of iterations is selected as the final prediction. This mechanism not only aids in noise reduction but also harnesses the synergy among diverse models to enhance classification accuracy (Rojarath et al., 2016).

This ensemble-based approach offers the potential to showcase the efficacy of leveraging multiple instances of our baseline model to improve classification accuracy, offering a more robust and stronger baseline comparison.

### 3.2.2. Large Language Models (LLMs) for Prompted Models

The subsequent phase of our exploration delves into the 3 models representing innovative paradigms of prompt engineering within the realm of LLMs. These models push the boundaries of news intention identification by leveraging the capacities of LLMs.

#### Selection of LLM for Experimentation and parameters specification

In our experimentation, we employed "gpt-3.5-turbo model" Large Language Model provided by OpenAI [1] via their API interface. This model served as the cornerstone for designing and executing the experiments detailed in this study, enabling rapid and streamlined investigations.

To foster responses that closely adhere to the provided news context while maintaining coherence and clarity, we adjusted the configuration parameters of the "gpt-3.5-turbo model". Specifically, we set the temperature parameter to 0.0, resulting in responses that exhibit heightened focus and determinism. By mimicking the news context in this manner, we ensure that the generated content aligns closely with the intention identification task.

To prevent excessive verbosity in the generated responses, we retained the default value for the maximum number of tokens parameter. This safeguard ensures that the generated outputs remain concise and informative without venturing into unnecessary elaboration. The top p parameter, which governs nucleus sampling remained unaltered. By preserving the default setting, we ensured that the diversity of word sampling was appropriately balanced while still maintaining a strong connection to the input context. Furthermore, the presence penalty parameter was also set to 0.0, emphasizing the importance of utilizing the context provided in the input. By doing so, the model is encouraged to seamlessly incorporate the news context into its responses, thereby enhancing the overall coherence of generated content. Adhering to the natural flow of conversation, we maintained the n parameter at its default value. This choice ensured that a single completion was generated for each input, promoting coherent and contextually relevant responses.

[1] https://openai.com/

### Modular Design and Introduction of Prompted Models

The process of introducing the prompted models, each representing distinct stages of prompt engineering within the LLMs framework, involves a modular design approach enabled by the LangChain library[2]. LangChain stands as a framework tailored to facilitate the development of applications harnessing the capabilities of language models.

This framework not only provides abstractions for effective interactions with language models but also offers an array of implementations for these abstractions, promoting modularity and ease of use. LangChain's core principles align seamlessly with the objectives of this thesis, which seeks to create models that are both data-aware and capable of engaging with their environment in an agentic manner.

The core of LangChain's utility is found in its two key value propositions:

1. **Components**
   LangChain fosters a component-based approach for seamless interaction with language models. These components serve as modular building blocks that can be employed independently or as integral parts of LangChain-powered applications. Regardless of whether one is utilizing the broader LangChain framework or not, these components offer a versatile and user-friendly means of interfacing with language models. This characteristic is especially relevant in our context, as we aim to construct models that can be fine-tuned, customized, and tailored to specific needs while maintaining simplicity and flexibility.

2. **Off-the-shelf Chains**
   Beyond individual components, LangChain provides pre-configured chains that streamline the accomplishment of higher-level tasks. These off-the-shelf chains facilitate rapid prototyping and deployment by offering structured assemblages of components. Such an approach accelerates the development of applications and empowers users to delve into tasks without becoming mired in intricate technical details.

By leveraging LangChain's modularity and its off-the-shelf chains each model's architecture is constructed as a chain of components, enabling the systematic manipulation of input, interaction with the LLM, and output generation.

[2] https://www.langchain.com/

### Explanation of Each Model's Architecture and Prompting Strategies

This section provides a comprehensive overview of the architectural designs and prompting strategies employed by the three proposed models. These models serve as integral components of our research, each contributing a distinct approach to the task of news intention identification. Below, we delve into the specifics of each model's architecture and its associated prompting strategy:

## Zero-Shot Model

Zero-Shots grounded in the in-context learning paradigm and harnesses the advanced capabilities of the "gpt-3.5-turbo" language model. In this approach, the model is presented with news inputs and tasked with the classification of these inputs into predefined categories. The unique architecture of the prompt allows the model to examine the descriptions and characteristics associated with the provided categories. Subsequently, it leverages this information to deduce the underlying intention of the news content (Min et al., 2021). The primary objective of this model is to explore the capacity of LLMs to classify news articles solely based on the definitions provided for the labels. This approach challenges the model to make sense of news content by relying solely on its understanding of the label descriptions, showcasing the model's ability to comprehend and categorize information in a contextually informed manner.

## Few-Shot Model

This model employs a few-shot learning strategy to enhance contextual comprehension. The architecture involves presenting examples from various categories alongside the news input. This approach enables the model to develop a broader understanding of intention identification by learning from a limited set of examples not originally present in the dataset. Through this methodology, we aim to demonstrate the model's ability to generalize from a small set of examples, showcasing its context-awareness and adaptability (Parnami and Lee, 2022). We achieve this by providing a set of examples (8 in total, with 2 for each label) into the prompt.

### Adaptive-Learning Model

This final model represents an approach that evolves through adaptive learning. Similar to the Few-Shot Model, the architecture involves presenting the news input and relevant examples. However, "Adaptive Learning Prompting" starts with no examples, notably, this model introduces a mechanism wherein examples emerge from previously misclassified news items. These misclassified news items are reintroduced into the prompt with corrected labels, acting as exemplars for the model's learning.

This dynamic adaptation empowers the model to iteratively refine its understanding over time. The model's architecture, in conjunction with LangChain's capabilities, imbues it with the ability to adapt to subtle intention categories, resulting in progressively refined estimations.

Through these explanations, we offer a more detailed introduction to the three models, enabling a comprehensive understanding of their architectures and underlying strategies. This knowledge serves as a foundation for the subsequent evaluation, analysis, and comparison of these models'performances in the context of identifying news intentions.

## 3.3. Evaluation Metrics and Techniques

In this section we introduced the framework utilized to measure and evaluate the effectiveness of the diverse models and methodologies introduced within this study. The evaluation phase assumes a vital role in quantifying the efficiency and robustness of these models, aiding researchers in informed decision-making regarding their applicability and possible improvements. This is especially significant within the context of news intention identification, where the task involves multi-label classification spanning various categories. The presence of label imbalance accentuates the critical nature of choosing an appropriate evaluation metric.

### 3.3.1. Introduction to Evaluation Metrics: Static Evaluation

In a multilabel classification setting, where news articles can belong to multiple intention categories, conventional evaluation metrics such as accuracy may not sufficiently capture the intricacies of model performance. Due to the inherent imbalance in label distribution,

where some categories may have more instances than others, the risk of biased assessment arises. For instance, if category **B** (Associative Manipulation) appears frequently, a model biased towards predicting this label might yield high accuracy, masking its shortcomings in predicting the rarer types.

To counter this challenge, the reference metric for evaluating the models'performance in news intention identification becomes the F1-score. The F1-score, a harmonic mean of precision and recall, is an effective measure in scenarios with imbalanced label distributions. It considers both the true positive rate (recall) and the precision (accuracy of positive predictions), providing a balanced assessment of the model's ability to correctly identify both the majority and minority classes.

In the static evaluation, we measured the F1-score for each model based on a single test set of 100 news articles. This snapshot of performance provides insights into the initial predictive capabilities of our models.

### 3.3.2. Dynamic Batch-Based Evaluation Methodology

To comprehensively evaluate the performance of our intention detection models, particularly focusing on the Adaptive-Learning model, a batch-based evaluation methodology was devised. This approach allows us to systematically assess how the model's performance evolves as it is provided with more references of correctly classified instances. We seek to explore how the model's performance can be incrementally improved by introducing additional examples of accurately labeled news articles. The methodology involves the following steps.

1. **Evaluation Setup**
   We initially evaluate the performance of the model using the test set (100), resulting in a baseline evaluation score.

2. **Incremental Introduction of Correctly Classified Instances**
   In subsequent evaluations, we introduce batches (10) of news articles, re-feeding those items misclassified back into the model through the prompt. However, this time, these instances are accurately labeled as examples for the model to learn from. This simulates a scenario where the model receives constructive feedback in the form of correctly classified instances, enabling it to refine its decision-making process.

3. **Batch-Based Evaluation**
   The evaluation is performed in batches. The evaluation metrics, including precision, recall, and F1-score, are computed for each batch separately. This approach helps us analyze the model's response to the additional correctly labeled instances in a granular manner.

4. **Observing Performance Improvement**
   By scrutinizing the batch-wise evaluation metrics, we gain insight into how the model's performance develops when presented with increasing instances of accurately classified examples. This nudges the system toward a more discerning decision-making process, anchored in references derived from instances that the system initially mislabeled. In essence, we guide the system by introducing highly valuable examples that steer the LLM toward a label that, in the absence of further insights, it would otherwise misclassify. In this context, we look for trends indicating whether the model's precision, recall, and F1-score are consistently improving with each batch.

The presented evaluation method provides valuable insights into how the Adaptive-Learning model's performance benefits from an iterative learning process. As more correctly labeled examples are introduced, we can gauge whether the model's predictive capabilities are strengthened, leading to higher precision, recall, and F1-score values. This approach addresses the challenge of limited dataset availability by making the most of the available data and tracking the model's progress as it adapts and refines its decision-making strategies based on newly introduced references.

To summarize, the adoption of the F1-score as the primary evaluation metric, alongside the utilization of a dynamic batch-based evaluation methodology, lays the foundation for the comprehensive assessment of our models'performance in the realm of news intention identification. This approach, which accounts for both precision and recall, adeptly tackles the challenges posed by label imbalance, ensuring an equitable and accurate evaluation of the models effectiveness across diverse intention categories. In the following section, we delve into the results and engage in a detailed discussion to gain deeper insights into our models'performance and their implications for news intention identification.

# Chapter 4

# Results

In this chapter we present the results and the subsequent discussion arising from the experiments conducted with the different news intention identification models. This chapter serves as a comprehensive exploration of our models'performance, providing insights into their efficacy laying the ground for a more in-depth and informative discussion about the effectiveness of the "Adaptive Learning Prompting" strategy.

The results we present encompass both static and dynamic batch-based evaluations for each of our models. Static evaluations provide a snapshot of performance, while batch-based evaluations unveil the adaptability and the cumulative impact of iterative learning. Through this dual evaluation approach, we assess its abilities to not only make accurate predictions but also to refine their decision-making processes with experience.

The comparative analysis that follows serves as the heart of this chapter. We conduct an in-depth examination of the proposed models, shedding light on their respective strengths and weaknesses. By dissecting their performances across intention categories and evaluating their consistency through batch-based analysis.

## 4.1. Model Performance Overview

In this section, we delve into the evaluation metrics we employed, with a particular focus on the F1-score, and present the results of both static and batch-based evaluations.

In the subsequent sections, we present the results of these evaluations for each model, enabling us to better comprehend their initial performance and how they refine their decision-making with iterative learning. Through this detailed analysis, we aim to provide a more comprehensive view of model capabilities, setting the stage for a comparative assessment of their strengths and weaknesses in the following section.

## 4.2. Static Model Analysis

In this section, we delve into a comprehensive comparative analysis of the performance of the Models in the context of news intention identification. We assess their respective strengths and weaknesses, aiming to gain deeper insights into their effectiveness, particularly in the context of the "Adaptive Learning Prompting".

### 4.2.1. Baseline Model

Table 4.1 summarizes the performance metrics of a classification model in a multilabel scenario. It includes precision, recall, and F1-score for each label. Support indicates the number of instances for each label. The table also shows overall accuracy, as well as average metrics considering all labels with equal and weighted representation.

Table 4.1: Classification Report Baseline Model

| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| A | 0.4 | 0.2 | 0.27 | 30 |
| B | 0.53 | 0.87 | 0.66 | 52 |
| C | 0 | 0 | 0 | 6 |
| D | 0 | 0 | 0 | 12 |
| **accuracy** | 0.51 | | | 100 |
| **macro avg** | 0.23 | 0.27 | 0.23 | 100 |
| **weighted avg** | 0.4 | 0.51 | 0.41 | 100 |

The Baseline Model performance highlights the critical influence of class imbalance within the dataset. This model exhibits a notable inclination towards adhering to the majority

classes, such as category **B**, while largely neglecting the less frequent ones, categories **C** and **D**. This behavior is a direct consequence of the substantial class imbalance that exists in the dataset.

On the positive side, this model demonstrates commendable precision (53 %) and outstanding recall (87 %) for Category **B**. This implies that the model effectively identifies a substantial portion of category **B** articles, striking a balance between precision and recall with an F1-score of 66 %. In essence, when it identifies a category **B** article, it is highly likely to be correct. However, the primary limitation of the model lies in its inability to effectively predict articles in the less frequent classes, categories **C** and **D**. These categories are severely underrepresented in the dataset, resulting in precision, recall, and F1-score values of 0 %. this fine-tune model tends to overlook these minority classes, primarily due to their limited presence in the training data. This issue underscores the significant impact of class imbalance on its performance.

Furthermore, while this baseline achieves a precision of 40 % for category **A**, its recall is merely 20 %. This indicates that the model correctly identifies some category **A** articles but misses a substantial portion of them. This conservative approach minimizes false positives but leads to an increased rate of false negatives. These implications are relevant to the study of news intentions in a multiclass setting. It emphasizes the critical need to address class imbalance within the dataset. The model's tendency to adhere to majority classes while neglecting less frequent ones could be problematic in scenarios where those less frequently represented labels might be of great importance.

### 4.2.2. Zero-Shot Model

In Table 4.2 we can have a deeper insight into the performance of the first prompted model, Zero-Shot is this case.

Table 4.2: Classification Report Zero-Shot Model

| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| A | 0.33 | 0.77 | 0.46 | 30 |
| B | 0.73 | 0.21 | 0.33 | 52 |
| C | 0.44 | 0.67 | 0.53 | 6 |
| D | 0.29 | 0.17 | 0.21 | 12 |
| **accuracy** | 0.40 | | | 100 |
| **macro avg** | 0.45 | 0.45 | 0.38 | 100 |
| **weighted avg** | 0.54 | 0.40 | 0.37 | 100 |

This model exhibits a performance profile that reflects the impact of class imbalance while demonstrating improvements in identifying less frequent categories, specifically categories **C** and **D**. Like the previous model, the Zero-Shot model struggles with the challenge of class imbalance, a prevalent issue in multiclass classification problems. This imbalance significantly influences the model's behavior, leading it to favor the majority class **B** while encountering difficulties with the less frequent ones. Zero-Shot prompting improves upon precision when it comes to category **B**, achieving a 73 %. This high precision implies a high likelihood of correctly identifying articles in this category.

However, this precision comes at the expense of recall, which is notably lower at 21 %. In essence, while this model accurately identifies a substantial portion of label **B** articles, it also misses a significant number of them. This trade-off between precision and recall underscores the model's struggle to achieve a balance between minimizing false negatives and false positives.

One notable strength of this model compared to the baseline is its marked improvement in handling categories **C** and **D**. For **C**, it attains a substantially improved recall of 67 %, paired with a precision of 44 %. This means that, while there are still some false positives, the model successfully identifies a noteworthy portion of type **C** articles. Similarly, for category **D**, the model exhibits a higher recall (17 %) and a comparable precision (29 %), resulting in an F1-score of 21 %. These performance metrics signify that despite its overall limitations, demonstrates an enhanced ability to predict articles belonging to categories **C** and **D**. Zero-Shot Model faces the classic precision-recall trade-off, a common challenge in classification tasks. It excels in recall for category **A**, achieving a commendable 77 %, indicating effective identification. However, it struggles with precision, which is at 33 %. This trade-off highlights the inherent difficulty of simultaneously minimizing false positives while maximizing true positives in a multiclass setting.

This model ability to improve upon the identification of types **C** and **D** is significant, particularly in scenarios where these categories hold substantial weight or importance. While Zero-Shot may appear to lag behind in general performance metrics, its enhanced capability in recognizing less frequent categories positions it as a preferred choice when accurate identification of such articles is of relatively high importance.

### 4.2.3. Few-Shot Model

In Table 4.3 we can study the results of the second prompted model, using a set of examples to guide the LLM in the decision-making process.

Table 4.3: Classification Report Few-Shot Model

| Label | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| A | 0.37 | 0.87 | 0.51 | 30 |
| B | 0.69 | 0.17 | 0.28 | 52 |
| C | 0.5 | 0.67 | 0.57 | 6 |
| D | 0.38 | 0.25 | 0.3 | 12 |
| **accuracy** | 0.42 | | | 100 |
| **macro avg** | 0.48 | 0.49 | 0.42 | 100 |
| **weighted avg** | 0.54 | 0.42 | 0.37 | 100 |

Similar to the previous models, the Few-Shot approach faces some when dealing with the challenge of class imbalance, which has a substantial impact on its behavior. The model tends to favor majority classes, such as **A**, while facing challenges with less frequent ones. This model demonstrates strengths and trade-offs in terms of precision and recall. It excels in precision for category **B**, achieving a notable 69 %. This high precision implies a high likelihood of correctly identifying articles in this category, minimizing false positives. However, this precision comes at the expense of recall, which is notably lower at 17 %. In essence, while few-shot accurately identifies a portion of category **B** articles, it misses a significant number of them, resulting in a trade-off between precision and recall.

One notable improvement in this model's performance compared to the previous models is its ability to handle **C** and **D** more effectively. For category **C**, it achieves a substantially improved recall of 67 %, paired with a precision of 50 %. This indicates a balanced performance in identifying type **C** articles with fewer false positives. Similarly, for category **D**, the model exhibits a higher recall (25 %) and a comparable precision (38 %), resulting in an F1-score of 30 %. These performance metrics mean that, while still facing class imbalance challenges, demonstrates an enhanced ability to predict articles belonging to the lest frequent categories **C** and **D**.

This set up continues to face the classic precision-recall trade-off, a common challenge in classification tasks. It excels in recall for category A, achieving a commendable 87 %, indicating effective identification. However, it struggles with precision, which is at 37 %. This trade-off highlights the inherent difficulty of simultaneously minimizing false positives while maximizing true positives in a multiclass setting. With respect to the

previous models, this model improves upon the identification of categories **C** and **D** is significant, particularly in scenarios where these categories hold substantial weight or importance. While Few-Shot may have trade-offs in precision and recall, its enhanced capability in recognizing less frequent categories positions it as a valuable choice when accurate identification of such articles is of paramount importance.

### 4.2.4. Adaptive-Learning Model

In Table 4.4 we can finally observe the results from the Adaptive-Learning strategy.

Table 4.4: Classification Report Adaptive-Learning Model

| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| A | 0.49 | 0.57 | 0.52 | 30 |
| B | 0.61 | 0.54 | 0.57 | 52 |
| C | 0.5 | 0.5 | 0.5 | 6 |
| D | 0.46 | 0.5 | 0.48 | 12 |
| **accuracy** | 0.54 | | | 100 |
| **macro avg** | 0.51 | 0.53 | 0.52 | 100 |
| **weighted avg** | 0.55 | 0.54 | 0.54 | 100 |

The adaptive learning model strikes a balance between precision and recall in some categories while facing trade-offs in others. Notably, it demonstrates balanced precision and recall for category **A**, with precision at 49 % and recall at 57 %. This indicates effective identification of category **A** articles with a relatively low rate of false positives. Similarly, for type **B**, the model maintains a balance with a precision of 61 % and a recall of 54 %, resulting in an F1-score of 57 %.

Types **C** and **D** continue to be challenging for this model, as they are less frequent and underrepresented in the dataset. For category **C**, the model achieves a balanced precision and recall, both at 50 %, with an F1-score of 50 %. This suggests that while the model can identify category **C** articles, there are some false positives. For category **D**, the model faces a similar situation, with a precision and recall of 46 % and 50 %, respectively, resulting in an F1-score of 48 %. These results indicate that this model struggles to predict articles in these less frequent categories.

This Adaptive-Learning model, like its predecessors, grapples with the classic precision-recall trade-off, a common challenge in multiclass classification tasks. It excels in recall for categories **A** and **B**, signifying effective identification. However, it faces challenges in precision, especially for category **C** and **D**, due to the difficulty of minimizing false

positives while maximizing true positives in a multiclass setting.

This model achieves an overall accuracy of 54 %, which reflects its ability to correctly classify articles into their respective categories. However, it's essential to consider the trade-offs and challenges it encounters, particularly with less frequent categories.

### 4.2.5. Cross-Model Comparison

In this section, we perform a cross-model comparison to provide an overall assessment of all four models. These models exhibit distinct characteristics in terms of their architecture and approach to news intention identification. The comparative analysis takes into consideration the significant challenges posed by class imbalance within the dataset and the models'respective abilities to address these challenges.

The Baseline model displays notable strengths and limitations. It excels in identifying articles belonging to the majority class, category **B**, with a precision of 53 % and a remarkable recall of 87 %. This indicates that when it classifies an article as category **B**, it is highly likely to be correct, striking a balance between precision and recall.

However, the main limitation of this model lies in its inability to effectively predict articles in the less frequent classes, categories **C** and **D**, where precision, recall, and F1-score values are consistently at 0 %. This behavior stems from the severe under representation of these categories in the dataset. It underscores the significant impact of class imbalance on its performance. Furthermore, it achieves a precision of 40 % for category **A**, with a recall of merely 20 %. This conservative approach minimizes false positives but leads to an increased rate of false negatives.

Prompted models, constructed using prompt engineering techniques, exhibit a more balanced approach to class imbalance while presenting their own trade-offs. Zero-Shot shows improvements in handling types **C** and **D**. It achieves a substantially better recall of 67 % for **C** and an F1-score of 21 % for **D**. Despite this, it faces trade-offs between precision and recall, particularly for category **B**.

Few-Shot, similarly, showcases enhancements in precision for category **B** and a balanced performance for categories **C** and **D**. It attains a recall of 67 % for type **C** and an F1-score of 30 % for **D**. The model continues to grapple with the precision-recall trade-off, emphasizing the challenge of minimizing false positives while maximizing true positives.

Finally the Adaptive-Learning model maintains a balance between precision and recall for categories **A** and **B**, indicating effective identification. Like the other prompted models, it faces challenges with **C** and **D** news, achieving balanced precision and recall but exhibiting limitations in precision-recall trade-offs.

In summary, the Baseline model performance highlights the influence of class imbalance and its inclination towards majority classes. This could be problematic if categories **C** and **D** carry significant weight or importance. On the other hand models built upon prompt engineering, demonstrate improved capabilities in handling less frequent categories. While they still face precision-recall trade-offs, their enhanced performance in these categories positions them as preferable choices when the accurate identification of such articles is crucial.

## 4.3. Dynamic Model Analysis.

This section is dedicated to studying the dynamic behavior of the 4 different models. It aims to understand how the performance of the news intention identification models changes over time or as they adapt to varying conditions, especially in the case of testing the iterative "Adaptive Learning Prompting" strategy, comparing to the performance and evolution of the other models. It is particularly important when dealing with tasks like news intention identification in low-resource scenarios where the limited amount of training data could represent a major obstacle for model development.

### 4.3.1. Batch-based Comparison

In Figure 4.1, a visual representation is provided, illustrating the regression lines of various models and how they evolve across different batches. This visualization offers insight into the modelsínherent ability to adapt to their informational environment.

We examine the performance of different models across batches for news intention identification. Several noteworthy observations can be made from this analysis: The Baseline demonstrates a relatively stable performance across batches. This stability is evident from its flatter trend line, indicating that its F1-scores remain consistent as batches progress.

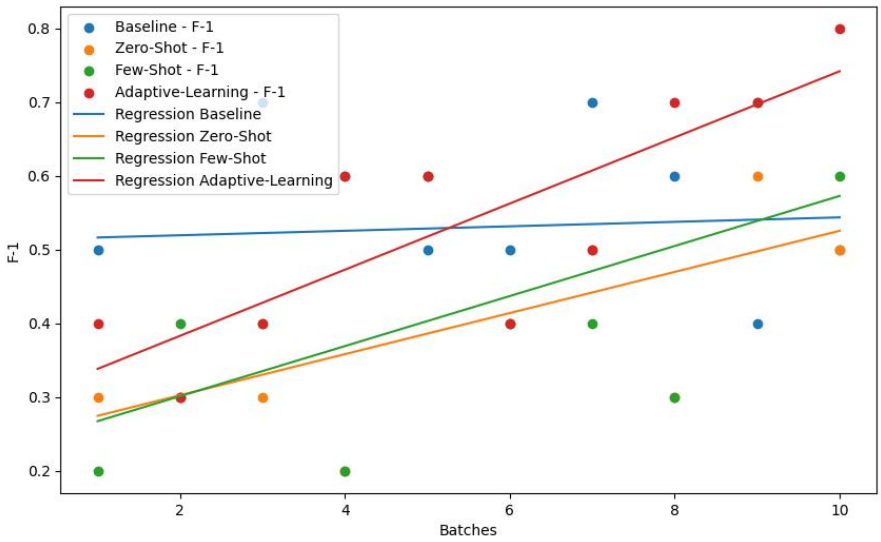


Figure 4.1: Model Performance Across Batches: F1-Scores

In contrast, the prompt-engineered models exhibit more pronounced trends in their performance over batches. These models display varying F1-scores as batches evolve, suggesting that their performance is dynamic and adaptable.

Among the prompt engineered models, the Adaptive-Learning Model stands out with the most prominent trend. The steeper trend line of this last model compared to the other 2 prompted models indicates a more substantial change in performance as batches advance. This suggests that the Adaptive-Learning model is particularly responsive to iterative learning and demonstrates a more pronounced ability to adapt.

To delve deeper into these observed trends and provide statistical support for our findings, we conduct a slope test analysis. The primary objective of this analysis is to assess whether there is statistically significant evidence to confirm these claims.

### 4.3.2. Slope-Testing

Table 4.5 presents the outcomes of a slope test analysis conducted on four different models in the context of news intention identification. This analysis aims to determine whether there is a statistically significant trend in the evolution of the models'performance over batches.

Table 4.5: Performance Trend Analysis Results

| **Model** | **Linear Regression Slope** | **P-value** |
|---|---|---|
| Baseline | 0.003 | 0.8405 |
| Zero-Shot | 0.0279 | 0.0685 |
| Few-Shot | 0.0339 | 0.0615 |
| Adaptive-Learning | 0.448 | 0.0033 |

The Baseline exhibits a very shallow slope of 0.003 in the linear regression analysis, indicating a nearly flat trend in its performance over batches. The corresponding p-value of 0.8405 is much greater than the conventional significance level of 0.05. This suggests that there is no statistically significant trend in its performance across batches.

Zero-Shot shows a slightly steeper slope of 0.0279. However, the p-value of 0.0685 is still higher than 0.05, indicating that the observed trend is not statistically significant. Similar to the Baseline, this model also demonstrates a relatively stable performance over batches, with minor variations that are not statistically significant. The Few-Shot approach exhibits a slightly steeper slope than Zero-Shot, with a value of 0.0339. However the p-value of 0.0615 is greater than 0.05, suggesting that there is no statistically significant trend in its performance. its performance remains relatively stable across batches, with minor fluctuations that do not reach statistical significance.

In contrast to the previous models, the Adaptive-Learning model displays a significantly steeper slope of 0.448. Importantly, the p-value associated is 0.0033, which is less than the 0.05 threshold. This indicates that there is a statistically significant increasing trend in the performance over batches. In other words, it shows clear and significant improvement in its performance as batches progress.

The results from this analysis highlight different performance dynamics among the four models. The baseline, Zero-Shot, and Few-Shot demonstrate relatively stable performance over batches, with no statistically significant trends. On the other hand, the Adaptive-Learning model stands out as having a statistically significant increasing trend, indicating its ability to adapt and improve its performance with iterative learning.

These findings provide valuable insights into the adaptability and learning capabilities of the models, contributing to a more comprehensive understanding of their performance dynamics in the context of news intention identification. They also support the earlier observations made in Figure 4.1 regarding the varying performance trends across the models

## 4.4. Overall Analysis

In this final section of Chapter 4, we provide a high-level analysis of the main results obtained from the extensive evaluation and analysis, which explored the performance and adaptability of different AI-NLP models in the context of news intention identification. These findings constitute a crucial part that sheds light on the strengths and limitations of the various models and settings under consideration in this thesis.

The evaluation of the four models highlighted their performance disparities. The Baseline Model, leveraging fine-tuned BERT, performed relatively well in identifying majority-class articles but faced significant challenges with less frequent categories due to class imbalance. In contrast, prompt-engineered models, especially the Adaptive-Learning model, demonstrated improved adaptability to these less frequent categories, although they grappled with trade-offs between precision and recall.

The dynamic analysis conducted across batches unveiled noteworthy insights. The Baseline maintained a stable performance with no discernible trend. Conversely, Zero-Shot and Few-Shot exhibited minor fluctuations that lacked statistical significance. However, it was the Adaptive-Learning prompting that stood out, showcasing a statistically significant increasing trend. This trend underscored its ability to adapt and enhance its performance iteratively. This observation holds great relevance as it highlights its capability to dynamically adjust to specific contexts. This adaptive process operates akin to a feedback loop, allowing the model to rectify and improve its classification capabilities, better aligning with the evolving requirements of newly presented tasks and informational contexts.

This process effectively guides the model's learning, resulting in substantial improvements in overall precision, recall, and F1-score. During our batch-wise analysis, we observe a noteworthy and statistically significant enhancement in model performance, improving progressively across batches, with the last two batches achieving accuracy rates of 70 % and 80 %, respectively. These results approach the expected performance ceiling of the dataset, even in low-resource scenarios where there is a 36 % discrepancy in labeling by different annotators. This underscores the effectiveness of the model's learning process over successive iterations.

# Chapter 5

# Discussion

In this chapter of the thesis, we conclude our exploration of "Adaptive Learning Prompting" for news intention identification. We delve into the implications and contributions of our research, discussing its significance, operational viability, and potential for broader applications in the field of Natural Language Processing.

## 5.1. Research Questions Revisited

In this section, we revisit the research questions posed at the outset of this study and reflect on their relevance and alignment with the findings and insights gained throughout the research journey. The research questions have served as guiding questions, leading our analysis of the "Adaptive Learning Prompting". As we near the culmination of this work, it is imperative to assess whether these questions have been adequately addressed and whether any adjustments or refinements are necessary to capture the essence of this research work.

### Research Question 1

LLMs can significantly improve the accuracy of news intention identification.

### Results

The experiments using Large Language Models demonstrated improvements in the accuracy of news intention identification. Specifically, the Adaptive-Learning strategy, which leveraged LLMs'language understanding capabilities, outperformed traditional methods in this particular context.

### Response

The results support Hypothesis 1. LLMs, with their contextual awareness and reasoning abilities, significantly enhanced the accuracy of news intention identification, especially in situations with limited data and with less frequent categories.

### Research Question 2

Iterative prompt engineering can enhance the precision and recall of news intention identification.

### Results

The research employed iterative prompt engineering, refining prompts based on performance feedback. This iterative approach led to increased precision and recall in news intention identification, as evidenced by the improved evaluation metrics.

### Response

Hypothesis 2 is supported by the results. Iterative prompt engineering was effective in enhancing both precision and recall, indicating that the model adapted and refined its understanding of the task.

### Research Question 3

"Adaptive Learning Prompting" will significantly improve news intention identification performance.

### Results

"Adaptive Learning Prompting", which dynamically adjusted prompts and incorporated misclassified instances, resulted in a increase in overall precision, recall, and F1-score. This adaptive approach continuously improved news intention identification performance.

### Response

Hypothesis 3 is supported by the results on this dataset. The "Adaptive Learning Prompting" significantly enhanced the overall performance of news intention identification, demonstrating its potential for continuous improvement in the task.

In summary, the results of the experiments described in Chapter 4 align closely with the hypotheses. LLMs, when combined with the Adaptive-Learning strategy and iterative prompt engineering, substantially improved the accuracy, precision, and recall of news intention identification. These findings provide empirical evidence of the effectiveness of the proposed strategies in enhancing the performance of Artificial Intelligence models in this context, particularly in low-resource scenarios where these systems can be implemented from the very onset, having the learning capabilities to improve their performance over time.

## 5.2. Impact of Adaptive Prompt Refinement

In this part, we provide a detailed examination of how the "Adaptive Learning Prompting" directly influences the process of news intention identification. We offer insights into how this innovative approach dynamically adjusts and refines the prompts provided to Large Language Models. Through this exploration, we shed light on the mechanism behind the notable performance improvements observed, showcasing its effectiveness in fine-tuning LLMs for specific tasks.

A capital aspect of our research is the operational viability of the Adaptive-Learning strategy. We discuss how this approach reduces the necessity for extensive fine-tuning of LLMs. By adaptively refining prompts, we achieve remarkable performance enhancements without the resource-intensive process of retraining. This discussion underscores the practicality of our approach, especially in resource-constrained scenarios.

Another significant aspect we address is the accelerated deployment and applicability of the "Adaptive Learning Prompting", even in the early stages of implementing LLM-based models. We emphasize how this strategy can streamline the development process and reduce time-to-deployment, making it a valuable asset for organizations seeking efficient solutions for news intention identification.

## 5.3. Study Limitations

It is imperative to acknowledge the inherent limitations of our study. While our research shows promise by highlighting the reasoning potential of Large Language Models and the applicability of the Adaptive Learning strategy, even in scenarios with minimal or no prior training data, it is essential to emphasize that the findings presented in this thesis are context-specific.

The scope of our results is limited to the dataset we utilized, namely the available news data within the Fakenews Politifact dataset during the time of this research. It is crucial to recognize that the nuances and characteristics of this particular dataset may have influenced our outcomes. Therefore, the direct transferability of our findings to different datasets and broader contexts remains an area for further exploration. To achieve a comprehensive understanding of the capabilities of LLMs and the effectiveness of the specific strategy proposed in this work, future research must encompass more extensive datasets and diverse contextual scenarios. Only through such comprehensive investigations can we aspire to draw broader conclusions about the general applicability and potential of LLMs.

## 5.4. Potential Future Directions

One of the most promising frontiers in Natural Language Processing today lies in the development of open-source Large Language Models[1]. While, for practical reasons, our research employed closed-source LLMs provided by OpenAI, recent strides in open-source LLMs herald a new era of possibilities. In this emerging landscape, it is possible to envision a future where both open and closed-source LLMs coexist harmoniously, forming a diverse and thriving ecosystem.

At the time of writing this thesis, implementing systems akin to the one described herein, or even conversational AI systems, demanded substantial hardware investments and sometimes exhibited less-than-optimal performance. However, the rapid pace of advancements and innovations in the field allows us to envision a not-so-distant future. In this future, efficient and fully functional LLM-based Natural Language Processing systems can be designed, operated, and deployed using standard laptop computers. This democratization of LLM technology points towards a new era in AI, poised to reshape the trajectory of humanity in the years ahead.

### 5.4.1. Potential Uses of LLM-Based NLP Systems

Beyond the academic realm, LLM-based NLP systems hold immense potential for real-world applications. These applications include enhanced content creation, where LLMs assist content creators in generating high-quality written material, from news articles to creative stories, saving time and enhancing productivity. Additionally, advanced chatbots powered by LLMs can provide personalized and efficient customer support, handling inquiries and resolving issues with human-like understanding. LLMs also play a pivotal role in breaking down language barriers through accurate and context-aware translations, benefiting international communication and global businesses. In the medical field, LLMs aid professionals in diagnosing diseases, analyzing patient data, and contributing to medical research by processing vast amounts of scientific literature. Legal professionals benefit from LLMs by conducting legal research, drafting documents, and analyzing case law, streamlining legal processes. In education, LLMs personalize educational content, offering tailored lessons and assessments to students, making learning more engaging and effective.

[1] https://huggingface.co/spaces/HuggingFaceH4/open$_l$*lm*$_l$*eaderboard*

Lastly, LLMs assist in content moderation on online platforms, helping identify and filter out harmful or inappropriate content. These diverse applications illustrate the transformative potential of LLM-based NLP systems across various industries and sectors, transcending traditional academic boundaries and becoming integral components of our daily lives and industries.

### 5.4.2. The Role of Langchain in Modular AI Systems

Another pivotal element employed in our experiments throughout this thesis has been Langchain[2]. Its modular structure opens up a world of possibilities for designing and implementing AI systems. Through these modular architectures, AI researchers gain precise control over information flow within these systems, establishing checks and balances.

This not only simplifies system construction but also elevates the development of Explainable AI (XAI) systems (Datta and Dickerson, 2023). By visualizing the entire flow of information and decisions made by constituent elements, Langchain empowers AI professionals and system users in effectively controlling and auditing these systems.

## 5.5. Envisioning the Future

Although we currently find ourselves in the nascent stages of this revolution, the results and experiments conducted in this work offer a glimpse into a future where modular LLM-based AI systems take center stage, serving as a seamless human-computer interface, they enhance efficiency while granting users unparalleled control and transparency over outcomes and results.

This transformative potential promises to shape the future of AI and NLP in profound ways, transcending the boundaries of academia, affecting all industries and becoming an integral part of our daily lives.

[2] https://python.langchain.com/docs/get_started/introduction.html

# Chapter 6

# Conclusions and Further Work

In this Master's thesis, we have discussed the importance of news intention detection to enhance the formation of the public's opinion, particularly in light of the current circumstances characterized by information overflow and attempts by various entities to distort public opinion. This distortion occurs not only through the creation and dissemination of fake news but also through subtle manipulation of the intentions embedded within legitimate news.

This relatively new vision is hindered by the ongoing insufficiency of efforts to address this issue, coupled with a shortage of datasets for training and comparing various potential architectures.

In this particular context we have introduced the transformative power of Large Language Models within the realm of Natural Language Processing in particular and of Artificial Intelligence in general. Our journey began presenting the "Adaptive Learning Prompting" that leverages LLMs to perform input classification tasks through iterative prompt engineering, even with no training data available, capital aspect to consider in the still recent news intention identification field,

This approach emerged as a linchpin, enabling the adaptation of LLMs for specific tasks with high efficiency. It not only showcased substantial performance enhancements but also proved operationally viable, reducing the need for resource-intensive fine-tuning. We demonstrated its accelerated deployment and applicability, even in the early stages of implementing LLM-based models. This promising feature streamlines model development, reducing time-to-deployment, and providing valuable solutions for news intention

identification, even with no data or in resource-constrained scenarios.

However, it is imperative to acknowledge the context-specific nature of our findings, rooted in the dataset we employed during this research. To unleash the full potential of LLMs and the "Adaptive Learning Prompting", broader investigations across diverse datasets and contexts are crucial. Only through such comprehensive research can we chart a course toward generalizable conclusions and widespread applicability.

Moreover, we are still scratching the surface of the immense potential of LLM-based AI systems for real-world applications that extend beyond academia. From content creation to customer support, language translation to medical diagnosis, legal assistance to education, and content moderation, LLMs offer solutions that can revolutionize industries and enhance human experiences.

Our journey also introduced the possibilities of modular architectures (LangChain), offering a higher degree of control over information flow within AI systems. This modular architecture not only simplifies system construction but also catalyzes the development of Explainable AI systems, empowering AI professionals and system users alike.

As we look to the future, we envision a landscape where LLMs are at the heart of AI systems, serving as seamless human-computer interfaces. This vision promises enhanced efficiency, greater user control, and transparency over outcomes and results. It is a future that transcends academic boundaries and emerges as an integral part of our daily lives and industries.

In closing, this research underscores the significance of contextually nuanced cues and iterative prompt engineering, offering a glimpse into the potential that LLMs hold. In alignment with our earlier discussions on the importance of information validation and news intention detection, we must emphasize how these advancements align with the core principles of democratic societies, echoing the values that inspired early democracy in ancient Athens. The ability to discern truth from falsehood and to understand the intentions behind news articles bolsters the democratic values of informed decision-making and an engaged citizenry. However, the possibilities presented by the LLM-based modular systems proposed in this thesis are equally significant, carrying the potential to democratize the entire field of AI. These systems offer more than just improved performance, they pave the way for a deeper transparency and interactivity between AI systems and users. This transparency empowers users to understand the decision-making processes of AI, fostering trust and accountability.

# References

## Bibliografía


[Azaria and Mitchell, 2023] Azaria, A. and Mitchell, T. (2023). The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

[Balkus and Yan, 2022] Balkus, S. V. and Yan, D. (2022). Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, pages 1–30.

[Bermes, 2021] Bermes, A. (2021). Information overload and fake news sharing: A transactional stress perspective exploring the mitigating role of consumersâ resilience during covid-19. *Journal of Retailing and Consumer Services*, 61.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Castillo et al., 2011] Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th international conference on World wide web*, pages 675–684.

[Conroy et al., 2015] Conroy, N. K., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.

[Datta and Dickerson, 2023] Datta, T. and Dickerson, J. P. (2023). Who's thinking? a push for human-centered evaluation of llms using the xai playbook. *arXiv preprint arXiv:2303.06223*.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*.

[Grootendorst, 2022] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[Guo et al., 2022] Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

[Guo et al., 2023] Guo, Z., Zhang, Q., An, X., Zhang, Q., JÃ¸sang, A., Kaplan, L. M., and Cho, J. H. (2023). Uncertainty-aware reward-based deep reinforcement learning for intent analysis of social media information. *arXiv preprint arXiv:2302.10195*.

[Han et al., 2021] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., and Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

[Iman et al., 2023] Iman, M., Arabnia, H. R., and Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40.

[Liu, 2015] Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions.

[Liu et al., 2023] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

[Ma et al., 2017] Ma, J., Gao, W., and Wong, K. F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. *Association for Computational Linguistics*.

[McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Min et al., 2021] Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. (2021). Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

[Mohammad et al., 2016] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of International Workshop on Semantic Evaluation, Semeval-2016*, pages 31–41.

[Opdahl et al., 2023] Opdahl, A. L., Tessem, B., Dang-Nguyen, D.-T., Motta, E., Setty, V., Throndsen, E., Tverberg, A., and Trattner, C. (2023). Trustworthy journalism through ai. *Data Knowledge Engineering*, 146:102182.

[Oshikawa et al., 2018] Oshikawa, R., Qian, J., and Wang, W. Y. (2018). A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

[Parnami and Lee, 2022] Parnami, A. and Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*.

[Qazi et al., 2020] Qazi, M., Khan, M. U., and Ali, M. (2020). Detection of fake news using transformer model. In *2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–6.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Rodriguez et al., 2009] Rodriguez, J. D., Perez, A., and Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575.

[Rojarath et al., 2016] Rojarath, A., Songpan, W., and Pong-inwong, C. (2016). Improved ensemble learning for classification techniques based on majority voting. In *2016 7th IEEE international conference on software engineering and service science (ICSESS)*, pages 107–110.

[Schick and Schutze, 2021] Schick, T. and Schutze, H. (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*, page 255â269.

[Shenhav, 2006] Shenhav, S. R. (2006). Political narratives and political reality. *International Political Science Review / Revue Internationale de Science Politique*, 27(3):245â262.

[Shi and Lipani, 2023] Shi, Z. and Lipani, A. (2023). Don't stop pretraining? make prompt-based fine-tuning powerful learner. *arXiv preprint arXiv:2305.01711*.

[Shu et al., 2018] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

[Shu et al., 2017] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, volume 30.

[Wang et al., 2022] Wang, C., Yang, Y., Gao, C., Peng, Y., Zhang, H., and Lyu, M. (2022). No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 382–394.

[Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

[Yu et al., 2023] Yu, F., Zhang, H., and Wang, B. (2023). Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*.

[Zibran, 2007] Zibran, M. F. (2007). Chi-squared test of independence. *Department of Computer Science, University of Calgary, Alberta, Canada*, 1(1):1–7.

[Zubiaga et al., 2018] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*.