Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science & Marketing Analytics

# Prompt-Based Sentiment Analysis of Sustainability Discourse on Social Media Using Generative AI

Bas van Roozendaal (536165)

| | |
|---|---|
| Supervisor: | Bas Donkers |
| Second assessor: | Rommert Dekker |
| Date final version: | 6th July 2025 |

**Abstract**

This work investigates the application of prompt-based Generative AI (GenAI) models for sentiment analysis of sustainability-related discourse on Twitter. Here, sentiment analysis is operationalised as classifying the stance expressed in tweets, used in this work as a proxy for emotional sentiment, due to the lack of data labelled with sentiment. Building on recent advancements in natural language processing, this work evaluates whether large language models (LLMs) such as GPT-4o and DeepSeek-Chat can effectively classify stance (used as a proxy for sentiment) using various prompting techniques. A modular prompt engineering framework is developed to systematically test the influence of prompt structure on classification performance. Two publicly available datasets, consisting of climate change- or sustainability-related tweets of the period 2015-2018, are used to benchmark GenAI models against traditional rule-based (VADER) and transformer-based (Twitter-RoBERTa) classifiers.

The results indicate that, while fine-tuned transformer models outperform GenAI in overall accuracy, the best-performing prompts enable LLMs to achieve competitive results without additional training or fine-tuning. Prompt design significantly affects model outputs: prompts that incorporate label definitions and explicit instructions enhance classification accuracy, whereas misleading or ambiguous phrasing degrades performance. Topic-stratified evaluations further reveal that model performance varies across sustainability themes. LLMs achieve the highest accuracy on tangible environmental topics and lower performance on politically framed content.

The findings underscore the importance of prompt robustness when deploying GenAI models for social media analysis and highlight their potential as accessible, flexible tools for stance-based sentiment classification in data sparse environments, with methods transferable to future settings involving affective sentiment labels. This work contributes to the theoretical and methodological literature on prompt-based NLP and supports future applications of GenAI in sustainability communication research.

# Contents

# Chapter 1

# Introduction

## 1.1 Background & Motivation

### 1.1.1 Rise of Social-Media Sustainability Discourse

In recent years, sustainability has evolved from a specialist concern to a central topic in public discourse. Social media platforms, particularly X (formerly known as Twitter) [1], have become dynamic arenas for sharing opinions, frustrations, and support around environmental initiatives, corporate responsibility, and climate policy. This shift has profound implications: Not only does it democratise access to the sustainability conversation, but it also enables real-time public sentiment monitoring to shape the strategies of businesses, governments, and NGOs.

The study by Lineman et al. (2015) illustrates how search volume and sentiment around terms like "climate change" and "global warming" correlate with publicity and emotive framing on digital platforms. Public awareness and support for sustainability topics is thus largely mediated by online exposure, and more specifically, by the emotional valence with which they are discussed (Lineman et al., 2015). Twitter, due to the brevity and interactivity of its messages, is particularly suited for capturing rapid fluctuations in sentiment, especially around polarising topics like environmental policy or greenwashing claims.

Beyond individual engagement, social media now serves as a real-time indicator of societal attitudes. Researchers and institutions have increasingly employed social listening and sentiment-labelled data are used to approximate sentiment-related insights from these vast, unstructured conversations (Ballestar et al., 2020). In the domain of sustainability, this presents an opportunity to better understand public perceptions of corporate ESG initiatives, regulatory proposals, or environmental activism.

**Clarification:** Importantly, due to limited availability of high quality data with ground-truth sentiment labels, the ground-truth labels in the main dataset used in this work denote stance (i.e., the position expressed toward climate change) rather than emotional sentiment. Although this work performs the task of "sentiment classification", the labels used in the data represent opinionated stance toward climate-related issues. This conceptual mismatch has implications for both model evaluation and result interpretation.

[1] Since the datasets analysed in this work originate from the period when the platform was still known as *Twitter*, this term will be used throughout the remainder of this work for clarity and consistency.

### 1.1.2 Emergence of Generative AI for Text Analytics

Concurrently, the rise of Generative AI, particularly Large Language Models (LLMs) such as ChatGPT and DeepSeek, have transformed the landscape of text analytics. Traditionally, sentiment analysis relied on lexicon-based approaches (e.g., VADER) or fine-tuned Deep-Learning models (e.g., BERT), which required labelled datasets and often struggled with nuanced or context-dependent expressions common in social media texts (Chakriswaran et al., 2019; Krugmann & Hartmann, 2024).

Recent advances in prompt engineering allow zero-shot or few-shot applications of LLMs to complex NLP tasks, enabling high performance without domain-specific training data (Krugmann & Hartmann, 2024). Notably, LLMs offer flexible capabilities for text classification, including the ability to identify the stance expressed toward specific topics such as sustainability and climate action, when guided through carefully designed prompts. [2] This opens new methodological possibilities for sustainability researchers and practitioners seeking to interpret public discourse in more nuanced and scalable ways.

Despite this promise, the application of prompt-based LLMs to sustainability-focused sentiment analysis remains in its early stages. There is no cohesive framework for using GenAI through the science of prompt engineering to classify sentiment, operationalised here as stance, from climate-related tweets. This work aims to fill that gap. In this work, two state-of-the-art LLMs, GPT-4o and DeepSeek-Chat, are evaluated for their ability to classify stance (used here as a proxy for sentiment) through prompt engineering.

## 1.2 Research Gap & Problem Statement

While traditional models such as VADER and BERT have been widely used for sentiment analysis of social media texts, they often lack contextual understanding, especially in domain-specific areas like sustainability. In this work, sentiment is operationalised as the *stance expressed toward sustainability and climate-related topics*, rather than general emotional tone or mood. This introduces a conceptual gap between the affective nature of sentiment and the positional nature of stance. Prompt-based Generative AI models show promise in addressing these shortcomings, yet there is no established framework for their application in sustainability-related sentiment analysis.

Specifically, no study has systematically explored how prompt engineering can be used to classify the sentiment (stance) of sustainability-related tweets, compare such methods with traditional models, and assess the accuracy and robustness of outputs across prompt designs. Topic modelling is addressed through Latent Dirichlet Allocation (LDA), with subsequent analysis of model accuracy per topic, enabling insights into how well LLM performance generalises across different sustainability-related themes.

[2] However, their effectiveness in this task is partially contingent on how well prompts are capable of inferring stance under the label of "sentiment".

## 1.3 Research Objectives & Questions

The overarching aim of this work is to develop and evaluate a framework that uses prompt-engineered Generative AI to classify sentiment from sustainability-related tweets.

The research is guided by the following main research question and derived subquestions:

> *How can Generative AI, through prompt engineering, be effectively applied to classify sentiment, and to analyze how sentiment varies across different sustainability-related topics, and what managerial insights can be gained from its application?*

### 1.3.1 Sub-Questions

1. How can prompt engineering be applied to classify the sentiment of tweets using Generative AI?
2. How do Generative AI models compare to traditional sentiment analysis methods (e.g., VADER, BERT) in terms of accuracy and efficiency, given the stance-based nature of the ground-truth labels?
3. To what extent are the outputs of Generative AI models robust across different prompt formulations?
4. How can topic modelling (using Latent Dirichlet Allocation) be applied to identify the main topics in sustainability-related tweets, and how do Generative AI models perform across these topics?

## 1.4 Academic Relevance

This work contributes to three interlinked streams within the academic literature: First, sentiment analysis in sustainability contexts, second, methodological applications of prompt-engineered LLMs, and third, evaluation of prompt robustness. It addresses gaps identified in recent literature, including the need for more context-sensitive, adaptable models in sustainability analytics (Anderson et al., 2024; M. Liu et al., 2023), a systematic way to evaluate prompt variations, and the absence of systematic benchmarking for prompt-based GenAI models in this domain. Furthermore, it responds to calls for more transparent and practically useful NLP approaches capable of offering actionable insights in complex socio-environmental contexts (Chakriswaran et al., 2019). Moreover, it acknowledges the limitation that the main dataset used reflects stance rather than emotional sentiment, offering a transferable methodology that can be adapted to future sentiment-labelled corpora.

## 1.5 Managerial Relevance

From a managerial perspective, this work provides a blueprint for evaluating and applying advanced AI techniques to monitor and understand public discourse on sustainability. Companies, NGOs, and policymakers increasingly seek to align communication strategies with stakeholder

expectations and to potentially select reliable AI tools for this purpose. The methodology presented here focuses on systematically evaluating model performance, enabling more informed managerial decisions. It supports:

- Evaluation of model performance across different sustainability-related topics, providing managers with insights into which topics can be reliably monitored using AI-based sentiment analysis.
- Identification of strengths and limitations of different Generative AI models (e.g., GPT-4o, DeepSeek-Chat) in classifying sentiment within the sustainability domain, supporting evidence-based selection of tools for practical applications.
- Assessment of prompt robustness, informing trade-offs between prompt variants and accuracy in real-world social listening contexts.

These insights are particularly relevant for organizations that aim to implement AI-based monitoring of public discourse on sustainability topics at scale. By highlighting which prompts and models perform best for specific types of sustainability-related content and topics, this work enables managers to select and deploy AI tools that match their monitoring objectives and risk management needs.

Managers should note, however, that the sentiment labels used in this work represent stance toward sustainability issues, which may not accurately reflect emotional valence or consumer sentiment, but rather indicate attitudinal orientation. Results and interpretations should be contextualised accordingly.

## 1.6 Thesis Structure

The remainder of this work is structured as follows:

- **Chapter 2 - Literature Review:** synthesises existing work on sentiment analysis and topic modelling in social media, the rise of prompt-based generative models, and theoretical concepts related to sustainability communication, trust, and explainability.
- **Chapter 3 – Conceptual Framework:** presents the conceptual model, defines key constructs, and formulates propositions where applicable.
- **Chapter 4 – Data:** describes the datasets, data collection procedures and preprocessing steps, ethical considerations, and basic descriptive statistics.
- **Chapter 5 – Methodology:** details the research design, models, prompt engineering approach, baseline methods, LDA, and evaluation criteria.
- **Chapter 6 – Results:** reports the findings of the prompt engineering approach, sentiment classification, model robustness and performance evaluation, and analysis of model performance across sustainability-related topics based on LDA.
- **Chapter 7 – Discussion:** interprets the results in light of research questions and prior literature, and outlines theoretical, managerial, and methodological implications.

- **Chapter 8 – Conclusion:** summarises the key insights and answers the main research question.

# Chapter 2

# Literature Review

## 2.1 Text Analytics on Social Media in the Sustainability Domain

### 2.1.1 Lexicon-Based Approaches for Sentiment Analysis

Lexicon-based sentiment analysis techniques rely on predefined dictionaries that associate words or phrases with specific sentiment values (e.g., positive, negative, neutral). These methods are especially appealing in contexts with limited labelled data, offering interpretable, rule-based outputs suitable for rapid analysis of large-scale social media content. This is especially relevant when sentiment annotations are unavailable, and stance-labelled datasets must serve as imperfect proxies for affective sentiment, as in this work. Such substitution requires caution, as affective sentiment and opinionated stance reflect distinct communicative intent.

One of the most widely used lexicon-based tools for social media is 'VADER' (Valence Aware Dictionary for Sentiment Reasoning), developed specifically to handle the idiosyncrasies of online communication, including emojis, capitalisation, and slang (Hutto & Gilbert, 2014). VADER performs well on microblogging platforms such as Twitter, where message length and informal tone complicate traditional NLP models. It produces compound and component sentiment scores and is often used as a benchmark for lightweight, domain-agnostic sentiment classification.

'SO-CAL' (Semantic Orientation CALculator), introduced by Taboada et al. (2011), advanced the field by integrating grammatical structures, part-of-speech information, and context-modifiers such as negations, intensifiers, and irrealis markers. This made it particularly robust for nuanced text and cross-domain sentiment detection, outperforming several standard lexicons across classification benchmarks.

Several empirical studies have applied these tools in sustainability contexts. For instance, Ballestar et al. (2020) analysed 15,000 sustainability-related tweets using VADER and social network analysis, revealing that 85% of tweets exhibited neutral or positive sentiment and that online discourse was fragmented yet aligned with academic framing. Similarly, Lineman et al. (2015) combined VADER with Google Trends data to compare the public emotional response to 'climate change' versus 'global warming', finding that the latter evoked more negative sentiment and political polarisation.

Radi and Shokouhyar (2021) and Neri et al. (2012) further demonstrated the utility of

lexicon-based methods for brand and media analysis. Radi evaluated 106,350 tweets to compare consumer perceptions of two mobile phone brands along environmental, social, and economic dimensions, revealing distinct emotional profiles aligned with environmental reputations. Neri et al., in an earlier Facebook study, used semantic parsing combined with supervised methods to validate sentiment differences between competing Italian news outlets.

While lexicon-based methods are transparent and computationally efficient, they are limited in capturing sarcasm, ambiguity, or complex sentiment patterns. As highlighted by Anderson et al. (2024), such tools often underperform on domain-specific sustainability texts compared to machine-/deep-learning alternatives, particularly when applied to stance-labelled data used as a stand-in for sentiment, where subtle attitudinal cues may not align neatly with affective lexicon scoring. Nonetheless, lexicon approaches remain useful as baselines or hybrid components, especially when interpretability and speed are prioritised.

### 2.1.2 Machine-/Deep-Learning Models for Sentiment Analysis

Machine and deep learning methods have significantly advanced sentiment analysis by enabling context-aware representations, feature learning, and flexible adaptation to domain-specific data. In contrast to lexicon-based approaches, these models can learn complex sentiment cues, handle sarcasm and negation more effectively, and be fine-tuned for particular use cases. Through fine-tuning, they also show greater flexibility in modelling stance, which often requires interpreting ideological or attitudinal positions rather than emotional tone.

Early surveys such as Pang and Lee (2008) and Yue et al. (2018) laid the groundwork by cataloging the shift from rule-based techniques to machine learning classifiers, including Naive Bayes, Support Vector Machines (SVM), and decision trees. As these approaches matured, deep learning architectures, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) models, began to dominate performance benchmarks (L. Zhang et al., 2018).

The introduction of transformer-based models, especially BERT (Bidirectional Encoder Representations from Transformers), marked a further leap. Devlin et al. (2019) demonstrated that pre-training on large text corpora followed by task-specific fine-tuning yields state-of-the-art results across eleven NLP benchmarks. This has since become a standard methodology for sentiment analysis across domains.

Benchmark studies have affirmed the superiority of deep learning in sustainability contexts. Anderson et al. (2024) compared VADER, TextBlob, SVM, Random Forest, BERT, GPT-2, and LSTM on over 38,000 sustainability-related tweets, concluding that Random Forest and BERT performed best in terms of accuracy and contextual understanding, which is essential when ground-truth labels capture stance rather than sentiment and models must infer intent beyond emotional tone, as in climate-related discourse. Chakraborty et al. (2020) applied LSTM and a fuzzy-rule system to analyse COVID-19 tweets, revealing that negative or neutral tweets were disproportionately amplified through retweets and that deep learning models reached up to 81% accuracy.

The emergence of Generative AI has further influenced this space. Krugmann and Hartmann (2024) benchmarked GPT-3.5, GPT-4, and LLaMA 2 on twenty datasets, finding that GPT-

4 achieved top-tier binary sentiment classification accuracy, while LLaMA 2 offered superior interpretability. These findings highlight the growing relevance of large language models (LLMs) in sentiment tasks even without supervised fine-tuning.

Recent applications illustrate the adaptability of deep learning models to ESG and climate domains. M. Liu et al. (2023) used a hybrid of topic modelling and sentiment classification to analyse ESG-related Weibo posts in China, finding mostly positive public sentiment but persistent concerns about greenwashing. Ghahramani et al. (2021) employed LDA and sentiment scoring on TripAdvisor reviews of Dublin's urban green spaces to assess latent sentiment structures, revealing actionable insights for urban planning.

Overall, machine and deep learning methods offer notable advantages in predictive performance and contextual depth. However, they often require substantial labelled data and computational resources, and their black-box nature can pose interpretability challenges, prompting interest in more explainable and human-aligned approaches.

### 2.1.3 Methods for Topic modelling

Topic modelling plays a critical role in uncovering latent themes within large corpora of social media text, especially in contexts where public discourse is diffuse and rapidly evolving. These methods are particularly valuable in sustainability research, where understanding dominant concerns, perceptions, and narratives can inform both academic inquiry and policy interventions.

The foundational model in this area is Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003). LDA models a document as a mixture of topics, with each topic represented as a distribution over words. Its generative framework enables scalable unsupervised learning of thematic structures in text corpora. The method has been widely adopted and adapted across domains due to its simplicity and robustness.

Building on LDA, extensive reviews such as Jelodar et al. (2017) have surveyed model variants, applications, and unresolved challenges, highlighting the model's limitations in handling short, noisy text like tweets. These challenges include sparse word co-occurrence, semantic drift, and low topic coherence in microblog data.

To address these issues, more recent contributions have benchmarked and refined topic modelling techniques. Egger and Yu (2022) compared LDA, Non-Negative Matrix Factorisation (NMF), Top2Vec, and BERTopic using 31,800 COVID-related tweets. They found that BERTopic, which integrates class-based TF-IDF and contextual embeddings, produced the most coherent and interpretable topics, particularly suitable for social science applications. Similarly, Grootendorst (2022) introduced BERTopic as a neural topic modelling framework capable of capturing contextual nuance via transformer embeddings and dynamic clustering.

The integration of topic modelling into applied sustainability analytics has also been demonstrated. Reyes-Menendez et al. (2018) combined topic modelling and sentiment analysis to map Twitter discourse on #WorldEnvironmentDay to Sustainable Development Goals (SDGs), identifying climate change, water issues, and pollution as dominant themes. Likewise, M. Liu et al. (2023) applied topic modelling to ESG-related tweets from China, uncovering public focus on transparency, ratings, and the credibility of environmental claims.

These developments suggest a methodological evolution from probabilistic models like LDA

to hybrid and embedding-based approaches that better handle short-text corpora. Nonetheless, classic models retain relevance due to their interpretability and computational efficiency, making them valuable baselines or components in ensemble frameworks.

## 2.2 Large Language Models & Prompt Engineering

### 2.2.1 Zero-/Few-Shot Capabilities

Large Language Models (LLMs) have revolutionised natural language processing by enabling high-quality task performance without explicit fine-tuning. This capability, known as zero- or few-shot learning, enables users to formulate tasks as prompts, which the model processes by leveraging its pre-trained statistical associations, allowing for flexible and rapid deployment across domains.

The pioneering work by Brown et al. (2020) introduced GPT-3 and demonstrated that sufficiently large transformer-based models could generalise to a wide array of NLP tasks through in-context learning. Without any task-specific training, GPT-3 achieved strong performance on summarisation, translation, and sentiment classification, thereby establishing the viability of in-context learning as a viable paradigm.

Subsequent work has refined and extended these ideas. Schick and Schütze (2020) proposed Pattern-Exploiting Training (PET) and iterative PET (iPET), which allowed small masked language models to match or outperform GPT-3 in few-shot settings. These techniques also enhanced efficiency and interpretability, reducing computational costs while preserving model performance. Z. Zhao et al. (2021) introduced contextual calibration to correct output bias in few-shot learning, showing that content-free prompts could improve accuracy and reduce label imbalance effects.

In the sentiment analysis domain, Krugmann and Hartmann (2024) benchmarked LLMs like GPT-3.5, GPT-4, and LLaMA 2 against traditional models (e.g., RoBERTa, SiEBERT) across 20 datasets. GPT-4 achieved state-of-the-art performance on binary sentiment classification, suggesting that prompt-based methods can rival or exceed fine-tuned models even on specialised sentiment tasks. This is especially promising in contexts like sustainability, where available datasets often label stance rather than true sentiment.

Synthesising these contributions, LLMs exhibit strong generalisation abilities in sentiment classification and other NLP tasks, offering an attractive alternative where labelled data are scarce or deployment agility is essential.

### 2.2.2 Robustness and Interpretability Studies

Despite their promise, prompt-based LLMs raise questions regarding robustness, semantic understanding, and explainability. A key issue is that LLMs often exhibit inconsistent behaviour across prompt formulations, undermining their reliability in high-stakes settings.

Webson and Pavlick (2022) critically examined prompt semantics and found that LLMs sometimes perform equally well with instructive, irrelevant, or misleading prompts. This indicates that performance is more influenced by target word choices than true semantic comprehension. Similarly, W. Zhang et al. (2023) evaluated LLMs across 13 sentiment analysis tasks and showed

that, while models perform well in zero-/few-shot scenarios, they lag behind fine-tuned models on aspect-based sentiment analysis (ABSA). The authors introduced SENTIEVAL as a unified benchmark to more realistically assess LLM capabilities.

Broader interpretability concerns are highlighted by Miller (2019), who argued that effective explanations must account for the contrastive, selective, and social nature of human reasoning. This aligns with findings from Minh et al. (2022), who categorised explainability methods into pre-model, interpretable model, and post-model approaches, and emphasised trade-offs between fidelity and comprehensibility. Antoniadi et al. (2021) further reinforced the importance of user-centred design in explainable AI, calling attention to the lack of standard evaluation frameworks in real-world decision support systems.

While LLMs like LLaMA 2 have shown promising interpretability traits in sustainability contexts (Krugmann & Hartmann, 2024), current systems still lack consistency and transparency. In this work, the focus lies primarily on evaluating robustness to prompt variations, rather than on generating or evaluating model explanations. In conclusion, research increasingly converges on the need for robust prompt design strategies, improved evaluation methods, and hybrid systems that combine the adaptability of LLMs with the interpretability of symbolic reasoning.

## 2.3 Sustainability Communication & Public Perception Theories

Understanding how sustainability issues are perceived and discussed publicly is essential for effective communication strategies, particularly in the context of social media discourse. Research in this area draws heavily from environmental psychology, communication science, and behavioural economics to identify the cognitive and emotional mechanisms that shape public engagement with environmental topics.

Corner et al. (2014) argue that individual human values, especially self-transcendent ones such as altruism and universalism, strongly influence public engagement with climate change. Their work highlights that message framing aligned with these values can enhance the persuasiveness and emotional resonance of climate communication efforts. Similarly, Weber (2010) proposed a dual-process framework that integrates affective, analytical, and rule-based systems of decision-making to explain the widespread public underreaction to climate risks. The model suggests that climate threats are often perceived as abstract, distant, or temporally delayed, an effect known as psychological distance, which is hypothesised to dampen emotional response and weaken motivation to act.

At the strategic level, Nisbet (2009) emphasise the importance of message framing to connect climate change issues with audience values and identities. By broadening the discourse to include economic, public health, and moral frames, beyond the traditional scientific frame, communicators can expand the relevance of sustainability messages to diverse publics. This is echoed by Moser (2010), who provide a comprehensive review of climate communication strategies, noting the key barriers to effective messaging, such as invisibility of impacts, uncertainty, and emotional distancing. They propose message design principles and audience segmentation strategies for achieving long-term engagement.

Complementing these theoretical frameworks, empirical studies have examined actual patterns of public discourse. For instance, Lineman et al. (2015) compared public emotional responses to the terms "climate change" and "global warming" using Twitter and Google Trends data. Their findings indicate that "global warming" elicits more negative sentiment and political polarisation, hypothesising that the framing of sustainability topics significantly shapes public affect and engagement.

Together, these contributions demonstrate that sustainability communication is not merely a function of message accuracy but also of emotional, cultural, and identity-based alignment. For computational models aiming to analyze such communication, this means that effective sentiment classification requires more than just surface-level lexical analysis. AI systems must be able to account for the framing effects and value-laden cues that shape how messages are received and interpreted by the public.

## 2.4 Trust, Greenwashing, and Explainable AI

As sustainability discourse becomes increasingly mediated by digital technologies and automated analytics, concerns around trust, greenwashing, and the interpretability of AI systems have gained prominence. These dimensions are interlinked: effective sustainability communication requires credibility, which is shaped not only by the transparency and integrity of the message itself, but also by the perceived reliability of the analytical tools that extract and present insights. When AI systems are used to classify sentiment or detect topics, their opacity can raise doubts about the validity of conclusions, potentially undermining public trust.

Greenwashing, the practice of misleading stakeholders about environmental performance, has been systematically studied across disciplines. Delmas and Burbano (2011) proposed a multi-level framework that explains greenwashing based on regulatory, organisational, and psychological factors. Their model highlights how firms balance external pressure with internal capacity and reputational risk. Lyon and Montgomery (2015) expanded this view by developing a comprehensive taxonomy of greenwashing practices, such as selective disclosure and symbolic management, while also identifying gaps in empirical measurement and enforcement.

Building on these foundations, de Freitas Netto et al. (2020) conducted a systematic review of 67 greenwashing studies, classifying deceptive practices into claim-based and executional forms. The authors emphasise that greenwashing occurs at both product and firm levels, often exploiting consumer trust in sustainability labels and communication. These insights are critical for NLP researchers seeking to detect deceptive environmental claims and for training AI models to distinguish between genuine and symbolic sustainability efforts. While this work does not focus on detecting greenwashing directly, it remains an important backdrop for understanding public sentiment and trust dynamics in sustainability communication.

While the previous studies address public trust in the content of sustainability communication, a related, but distinct, concern is trust in the AI systems used to analyse such content. Explainability in AI models has emerged as a crucial factor for sustaining trust among analysts, practitioners, and end users.

Miller (2019) argued that explanations in AI should draw from social science theories, noting that users prefer contrastive and context-sensitive explanations that align with how humans

reason. In parallel, Minh et al. (2022) provided a comprehensive taxonomy of explainable AI (XAI) techniques, identifying trade-offs between accuracy and interpretability, and highlighting the need for rigorous evaluation frameworks.

The intersection of greenwashing and XAI is especially pertinent in sentiment and topic analysis of social media. As shown by Leippold et al. (2023), annotated datasets can reveal patterns of environmental claim inflation in corporate communication, offering empirical support for model training and evaluation. Moreover, transparency in model decisions is essential for sustaining trust in AI tools among analysts and decision makers, while public trust in sustainability discourse depends on the credibility of both the content and the analytical methods applied.

Together, these studies underscore that technical performance alone is insufficient. Trust in AI-enabled sustainability analysis must be grounded in both ethical content evaluation and transparent, interpretable model behaviour.

## 2.5 Synthesis and Identification of Research Gap

The literature reviewed thus far provides a comprehensive landscape of methods and theories relevant to sentiment and topic analysis in sustainability discourse. Lexicon-based approaches remain valuable for their interpretability and efficiency, especially in exploratory analyses (Hutto & Gilbert, 2014; Taboada et al., 2011), yet they lack the contextual depth necessary for more nuanced or dynamic datasets (Anderson et al., 2024). Machine and deep learning models, including BERT and LSTM-based classifiers, have demonstrated superior predictive performance (Anderson et al., 2024; Chakraborty et al., 2020; Devlin et al., 2019), but require labelled data and often operate as black boxes. Meanwhile, advances in topic modelling, from classical LDA (Blei et al., 2003) to embedding-based models like BERTopic (Grootendorst, 2022), have enabled scalable, interpretable insights into public discourse themes. In this work, the task of sentiment classification is conducted using stance labels, which serve as a proxy for emotional sentiment. The goal is to evaluate model performance in interpreting expressed positions toward sustainability themes, acknowledging that these labels reflect opinionated stance rather than affective tone.

In parallel, large language models (LLMs) and prompt-based learning have introduced flexible, data-efficient alternatives. Studies have shown that LLMs such as GPT-4 can match or surpass fine-tuned models in zero- and few-shot sentiment classification (Brown et al., 2020; Krugmann & Hartmann, 2024), though concerns remain about prompt sensitivity and semantic robustness (Webson & Pavlick, 2022; W. Zhang et al., 2023). A growing body of work now addresses interpretability and trust in AI, linking explainability with greenwashing detection and ethical sustainability communication (de Freitas Netto et al., 2020; Leippold et al., 2023; Miller, 2019).

Despite these contributions, several gaps persist. First, most LLM studies focus on generic or commercial domains, with limited application to sustainability-specific social media content. Second, while benchmark comparisons of traditional and generative models exist (Anderson et al., 2024; Krugmann & Hartmann, 2024), there is no cohesive framework that systematically applies prompt-engineered LLMs to sentiment classification within climate discourse, and eval-

uates how model performance varies across sustainability-related topics. Third, robustness of prompt-based outputs, especially across varying prompt formulations, remains underexplored in applied environmental contexts. Finally, few studies link these technical insights to managerial applications, such as supporting the selection and deployment of AI tools for reliable sentiment monitoring of online sustainability discourse.

This work addresses these gaps by investigating how generative AI, through prompt engineering, can be applied to classify stance, as an operational proxy for sentiment, and extract topics from sustainability-related tweets. It will benchmark performance against traditional methods, assess robustness across prompt designs, and analyze model performance across topics identified through Latent Dirichlet Allocation (LDA), providing actionable insights for sustainability communication strategy. Findings are interpreted in light of the stance-vs-sentiment label limitation, and results are framed as indicative of attitudinal alignment rather than emotional expression. This distinction frames the interpretation of results throughout this work and adds necessary nuance to all generalisations about sentiment, framing them explicitly as reflections of stance.

# Chapter 3

# Conceptual Framework

## 3.1 Conceptual Model Narrative

This work proposes a conceptual framework that outlines how generative AI can be applied to derive meaningful insights from sustainability-related tweets. Given the data constraints, these insights rely on stance-labelled tweets serving as a proxy for emotional sentiment. The model integrates four interrelated constructs: *Sentiment Classifications*, *Topic Prevalence*, *Prompt Robustness*, and *Perceived Reliability of Model Outputs*. These constructs span both the technical and social dimensions of AI-assisted sustainability analysis.

At the foundation of the model is the premise that public conversations on social media convey identifiable patterns of stance (used in this work as a proxy variable for emotional sentiment due to data limitations) and key topics. Generative AI, through prompt-based analysis, can extract this information efficiently. However, the quality of insights is not solely determined by output content. It is also shaped by the reliability of the prompting procedure and the extent to which users perceive the outputs as credible and reliable.

The conceptual model proposes that:

- (Stance-based) **Sentiment Classifications** and **Topic Prevalence** are direct analytical outputs derived from applying LLMs and LDA to sustainability-related tweets.
- The perceived quality and credibility of these outputs are mediated by one key technical factor: **Prompt Robustness** (i.e., the consistency of outputs across prompt variations).
- **Prompt Robustness** influences a downstream construct, **Perceived Reliability of Model Outputs**, which reflects the extent to which stakeholders, researchers, policymakers, or communication professionals are likely to rely on AI-derived outputs for interpretation and decision-making.

The conceptual relationships are summarised in the diagram below.

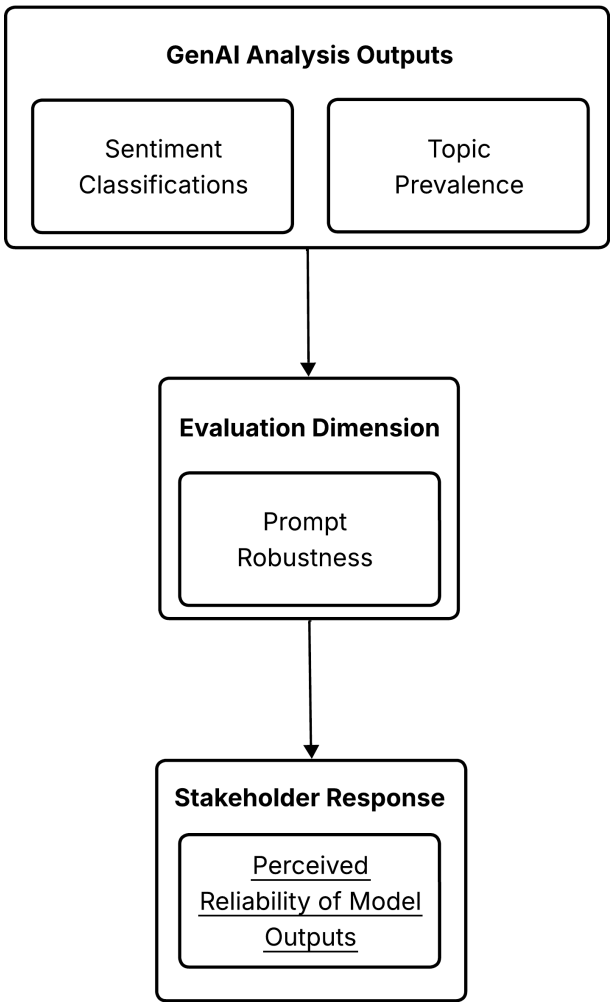


Figure 3.1: Conceptual framework showing how GenAI outputs and evaluation dimensions relate to the perceived reliability of model outputs.

This flow captures how generative AI outputs (sentiment, topics) must meet standards of robustness in order to establish credibility and build reliability in the output of the model. This framing also reflects the exploratory nature of this work, where sentiment classification is operationalised via stance-labelling due to the lack of affective sentiment ground-truth data.

## 3.2 Definitions of Constructs

| Construct | Definition |
|---|---|
| **Sentiment Classifications** | The stance expressed (positive, neutral, or negative) toward sustainability and climate-related topics in tweets (used in this work as a proxy for sentiment due to the unavailability of affectively annotated data) as classified by a language model. |
| **Topic Prevalence** | The relative prominence or recurrence of particular themes or subjects (e.g., climate policy, greenwashing, activism) in online sustainability discourse. |
| **Prompt Robustness** | The stability and consistency of outputs generated by large language models when the prompt wording is varied slightly but semantically equivalent. |
| **Perceived Reliability of Model Outputs** | The extent to which stakeholders are willing to rely on and act upon the outputs (e.g., sentiment classifications) generated by a generative AI system. |

Table 3.1: Construct definitions used in the conceptual framework.

## 3.3 Implications for Data and Methods

The conceptual framework shapes the research design and methods in several ways:

1. **Construct operationalisation:**
    - *Sentiment Classifications* are obtained by using both GenAI (GPT-4o, DeepSeek-Chat) and baseline models (VADER, Twitter-RoBERTa (Base and Fine-tuned models) on tweet-level data. These classifications reflect stance, not direct sentiment, and are treated as approximate indicators of public emotional positioning.
    - *Topic Prevalence* is estimated via LDA. Topic prevalence and per-topic sentiment classifications are analyzed based on GenAI and LDA output
    - *Prompt Robustness* is quantified by comparing sentiment classifications across prompt variants.
    - *Perceived Reliability of Model Outputs* is not directly measured through surveys but is inferred from the consistency and stability of model outputs under prompt variation and across topics.
2. **Evaluation strategy:** The framework justifies a multi-layered evaluation, combining quantitative accuracy with qualitative robustness. Prompt variants are tested to examine sensitivity and reliability.
3. **Relevance to managerial insights (Stakeholders):** By linking model design (e.g., prompt formulation) to perceived reliability, the framework ensures that the findings of

this work are not only technically valid but also practically meaningful for policy-makers seeking to apply GenAI in sustainability or sentiment analysis contexts.

In summary, the conceptual framework bridges text analytics, user-centered evaluation, and sustainability communication. It provides a theoretical and operational guide for subsequent chapters, ensuring alignment between research objectives, model outputs and evaluations, and stakeholder relevance. This includes transparency about the use of stance as a stand-in for sentiment, ensuring all findings are interpreted in light of this limitation.

The framework presented in this chapter guides the evaluation of model performance and robustness in the subsequent analysis. No formal hypotheses are tested, as this work is exploratory in nature.

# Chapter 4

# Data

## 4.1 Data Sources

### 4.1.1 Twitter Datasets

This work integrates two publicly available datasets containing Twitter posts related to environmental sustainability and climate change. Both datasets are used in prior text analytics research related to sustainability (Anderson et al., 2024). Together they provide a comprehensive and diverse representation of online sustainability discourse.

- **Dataset A – Climate Sentiment in Twitter Dataset (Guzman, 2020):** Originally scraped for academic research, this dataset contains 396 climate-related tweets labelled using TextBlob for polarity and subjectivity scores. Although the sentiment labels are automatically generated, the dataset offers broad topical coverage of environmental events and reactions from the public.
- **Dataset B – Twitter Climate Change Dataset (Qian, 2019):** This large dataset includes 43,943 climate-related tweets manually labelled into five categories related to climate change (pro-climate (positive), neutral, anti-climate (negative), news, and irrelevant). It provides a useful resource for sentiment classification in the area of sustainability, due to the quality and high level of accuracy of the data due to the process of manual labelling.

One additional dataset (Leippold et al., 2023) was initially considered but excluded, as it did not consist of tweets, but rather corporate disclosures such as sustainability reports and earnings calls. As the focus of this work is on public discourse rather than corporate self-representation, this dataset was deemed out of scope.

Together, these two datasets enable benchmarking across both traditional and prompt-based sentiment classification methods and allow exploration of topic prevalence within the discourse. Similar to the multi-source approach adopted in Anderson et al. (2024), this approach supports more generalisable findings.

### 4.1.2 Collection Procedure and Inclusion Criteria

The final dataset used in this work is a merged version of two publicly available datasets described above (Dataset A and Dataset B). No new data is collected manually. Instead, the datasets are obtained from Kaggle and Hugging Face repositories and preprocessed to ensure consistency in structure and content.

An overview of the dataset merging and cleaning process is shown in Figure 4.1. To maintain comparability and analytical coherence, several inclusion and transformation criteria were applied during dataset consolidation:

1. **Retention of Relevant Content:** Only the tweet text and corresponding sentiment label are retained. All other metadata, such as tweet ID, date, username, and URLs, are excluded.

2. **Sentiment Label Harmonisation:** Labels are mapped onto a consistent three-label sentiment (stance) scale to allow unified analysis:

   - **Negative**: Includes tweets labelled as "anti-climate" in Dataset B and tweets with a negative polarity in Dataset A.
   - **Neutral**: Includes tweets labelled as "neutral" in Dataset B and tweets with zero polarity in Dataset A.
   - **Positive**: Includes tweets labelled as "pro-climate" in Dataset B and tweets with a positive polarity in Dataset A.

3. **Quality Filtering:** Duplicates, and tweets with malformed or empty content, are removed.

4. **Removal of Retweets:** Retweets are removed to avoid duplication and ensure that the dataset reflects original user-generated content. Tweets beginning with the "RT" pattern are excluded

The resulting dataset offers a clean, structured, and diverse set of sustainability-related tweets and sentiment labels suitable for both traditional and Generative AI-based sentiment classification methods.

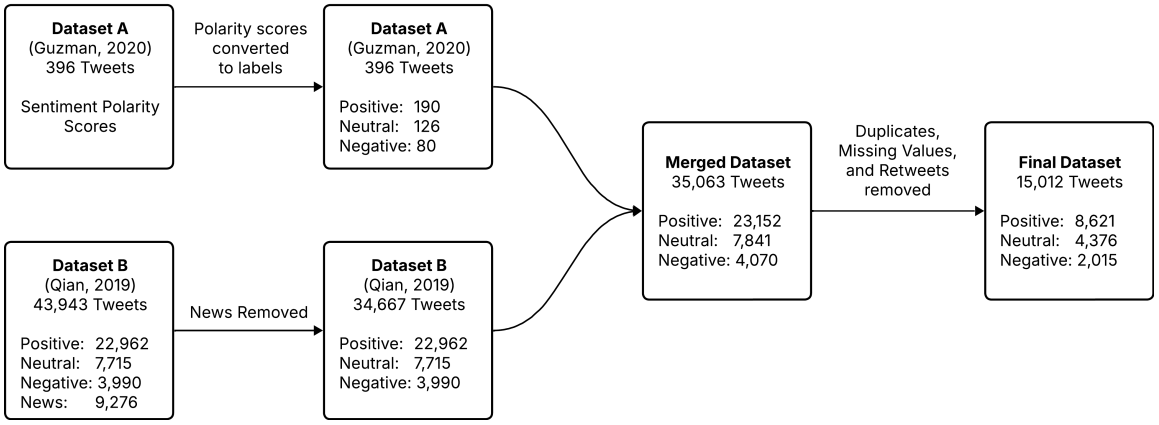


Figure 4.1: Overview of dataset harmonisation and cleaning process. Sentiment labels are standardised, and duplicates, missing values, and retweets are removed.

## 4.2 Data Cleaning and Pre-Processing

To prepare the data for sentiment classification and topic extraction, a tailored data cleaning and pre-processing pipeline is applied. The pre-processing methodology is adapted to suit the requirements of two different modelling approaches: First, traditional NLP models, including rule-based classifiers (VADER), supervised transformers (Twitter-RoBERTa), and topic models (LDA). Second, generative large language models (LLMs), such as GPT-4o and DeepSeek-Chat, which rely on natural, unaltered language structure for optimal performance.

### 4.2.1 Preprocessing Pipeline for Traditional Models

Traditional benchmark models require structured and noise-reduced input. For these methods, the following standardised pre-processing steps are applied:

1. **Language Filtering:** Non-English tweets are removed using Python language detection tools.
2. **Lowercasing:** All text is converted to lowercase to normalise the representation of tokens.
3. **Noise Removal:**
   - URLs, hyperlinks, and email addresses are removed.
   - User mentions (e.g., `@username`) and HTML artefacts (e.g., `&`) are removed.
   - Special characters are removed, excluding hashtags and emojis.
   - Punctuation is removed for VADER only.
4. **Tokenisation and Optional Lemmatisation:** Tweets are tokenised using spaCy. For wordclouds and topic modelling (e.g., LDA), optional lemmatisation is performed to group inflected word forms.
5. **Stopword Handling:** Stopwords are retained for VADER and Twitter-RoBERTa models to preserve context, but are removed for LDA topic modelling to enhance topic coherence.

These steps reflect common practices in short-text sentiment analysis and topic modelling (Anderson et al., 2024; Pang & Lee, 2008), ensuring clean, consistent input for quantitative model evaluation.

### 4.2.2 Preprocessing Pipeline for Generative AI Models

In contrast, generative LLMs, such as GPT-4o and DeepSeek-Chat, are provided with lightly processed text to preserve linguistic nuance, contextual cues, and expressive variation. Over-cleaning may reduce performance by removing elements that contribute to the model's contextual understanding.

For LLM inputs, the following minimal cleaning steps are applied:

- **Encoding Normalisation:** Tweets are UTF-8 encoded and stripped of corrupted characters.
- **Whitespace Trimming:** Leading/trailing whitespace and line breaks are removed.
- **Language Filtering:** Non-English tweets are excluded, consistent with the traditional pipeline.
- **Metadata Removal:** Only the raw tweet text is passed to the model; all metadata (e.g., username, timestamp, tweet ID) is excluded.
- **Noise Removal:**
  - URLs, hyperlinks, and email addresses are removed.
  - User mentions (e.g., `@username`) are replaced by "`@user`" and HTML codes (e.g., `&`) are removed.

All punctuation, emojis, capitalisation, hashtags, and emotive formatting are retained. These features are valuable for LLMs, which rely on token context to infer sentiment, emotion, irony, and relevance.

### 4.2.3 Summary

This dual-pipeline approach allows for an accurate comparison between traditional and generative models, each using optimised input formats. Table 4.1 summarises the differences in preprocessing pipelines.

Table 4.1: Comparison of Pre-processing for Traditional Models and LLMs

| **Step** | **Traditional Models** | **LLM Models** |
|---|---|---|
| Lowercasing | Yes | No |
| URL/user mention removal | Yes | Yes |
| Punctuation removal | VADER only | No |
| Lemmatisation | LDA only | No |
| Emoji/hashtag retention | Yes | Yes |
| Tokenisation | Yes | Handled by API |
| Duplicate filtering | Yes | Yes |
| HTML removal | Yes | Yes |
| Stopword Handling | LDA only | No |

Before applying these preprocessing steps, the dataset is split into training and test sets using an 80/20 stratified split based on sentiment labels. A validation set, consisting of 1000 tweets stratified by sentiment, is sampled, and subsequently removed, from the test set for prompt engineering purposes. Stratification ensures that each sentiment class is proportionally represented in all subsets, enabling fair performance comparison across classification models. All cleaning and transformation steps are applied separately to the training, test, and validation sets to prevent data leakage and ensure validity of the evaluation results.

All pre-processing is implemented using Python (pandas, spaCy, regex) and executed prior to model input. The final cleaned datasets provide a robust foundation for the comparative analyses of sentiment accuracy, topic modelling, and model robustness in Chapters 5 and 6. It is important to note that these sentiment labels, while harmonised, reflect stance expressions rather than affective sentiment.

## 4.3 Descriptive Statistics

This section summarises the main characteristics of the dataset after preprocessing and splitting. Table 4.2 presents descriptive statistics for the train, test and validation subsets, along with the overall totals. Sentiment balance and tweet properties are preserved across splits due to stratified sampling.

Table 4.2: Descriptive Statistics for Train, Test, and Full Datasets. Rounded to two decimals.

| Metric | Train | Test | Validation | Total |
|---|---|---|---|---|
| Tweet Count | 12,009 | 2,003 | 1,000 | 15,012 |
| Positive Tweets | 6,896 | 1,150 | 575 | 8,621 |
| Neutral Tweets | 3,501 | 584 | 291 | 4,376 |
| Negative Tweets | 1,612 | 269 | 134 | 2,015 |
| Avg. Length (chars) | 115.03 | 115.47 | 115.63 | 115.13 |
| Avg. Length (tokens) | 16.74 | 16.64 | 16.68 | 16.62 |
| Hashtag Usage (%) | 16.74 | 18.32 | 16.7 | 16.95 |
| Emoji Usage (%) | 0.86 | 0.85 | 0.8 | 0.85 |

Figure 4.2a and Figure 4.2b show word clouds of the most common lemmatised tokens in positive and negative tweets, respectively. These highlight key themes and lexical differences across sentiment classes.


(a) Positive Tweets

(b) Negative Tweets

Figure 4.2: Word clouds of the most frequent lemmatised words in positive and negative tweets. Common stopwords and domain-generic terms are removed to improve clarity.

## 4.4 Ethical Considerations and GDPR Compliance

This work adheres to ethical standards for secondary data use. All tweets were obtained from open-access academic datasets, and no new data scraping was performed. The datasets are anonymised at the tweet level and do not contain personal identifiers beyond publicly available handles, which were removed during pre-processing.

As outlined in GDPR Article 6(1)(e) and Recital 159, research using publicly accessible data for scientific purposes is permissible provided that privacy safeguards are in place. The work complies with these standards by:

- Using only publicly available and pre-existing datasets.
- Removing usernames, location tags, and other identifiers.
- Avoiding profiling, automated decision-making, or inferences about individuals.
- Reporting results at the aggregate (topic/sentiment) level.

As no direct interaction with subjects occurs, and no personal or sensitive data is retained, a formal ethical review is not required. The handling and storage of all files follow Erasmus University's guidelines on secure research data management.

# Chapter 5

# Methodology

## 5.1 Introduction

This chapter outlines the methodological framework used to evaluate sentiment classification performance using stance-labelled data as a proxy for affective sentiment in sustainability-related Twitter discourse. The design integrates two parallel pipelines: one using prompt-based Generative AI models, and the other applying traditional sentiment analysis methods as benchmarks. Throughout this chapter, the term 'sentiment classification' refers to the classification of stance labels, which reflect a position toward climate action and are treated here as an approximate indicator of underlying sentiment.

The chapter begins by describing the overall research design and the dataset structure. It then details the modular prompt engineering process used to construct and select effective prompts for Generative AI models, followed by an assessment of prompt robustness. Next, it introduces the sentiment classification methods used in both pipelines and presents the evaluation metrics applied to assess model performance. Finally, the chapter describes the use of topic modelling to support stratified evaluation and concludes with a summary of the methodological approach.

## 5.2 Overall Research Design

This work employs a comparative, multi-method design to evaluate the effectiveness and robustness of Generative AI in classifying sentiment within sustainability-related discourse on Twitter, benchmarked against traditional sentiment analysis models. The methodological framework comprises two parallel pipelines: a Generative AI pipeline based on prompt engineering, and a traditional pipeline using lexicon- and transformer-based models. The aim is to assess both classification performance and robustness across varying conditions, while enabling topic-specific analysis.

The dataset used in this work consists of 15,012 sustainability-related tweets, split into a stratified 80/20 train-test partition to ensure balanced sentiment representation. A validation set, consisting of 1000 tweets stratified by sentiment, is sampled, and subsequently removed, from the test set for prompt engineering purposes. The prompt engineering process is performed on the 1000-tweet validation set and all model evaluations are performed on the 2,003 tweets in the

test set. Each tweet is annotated with one of three stance-based sentiment proxy labels (positive, neutral, negative). This use of stance labels in place of true affective sentiment introduces limitations that are addressed in the evaluation and interpretation of results. All tweets are preprocessed according to the dual-pipeline approach outlined in Chapter 4.

The Generative AI pipeline involves prompt-based sentiment classification using two large language models: GPT-4o (`gpt-4o-2024-08-06`), accessed via the OpenAI API, and DeepSeek-Chat (`DeepSeek-V3-0324`), accessed via the DeepSeek API. Prompts are designed and evaluated through a modular engineering approach, which is further discussed in Subsection 5.3 A tuning subset of 1,000 tweets, stratified by sentiment, is randomly sampled from the test set to test all possible combinations of prompt modules. For each model, the prompt achieving the highest classification accuracy on this subset is selected as the final prompt for deployment on the full test set.

To assess robustness, two controlled variations of the selected prompt are tested independently: one with all punctuation removed, and another incorporating a deliberately misleading ("false") example. Classification performance under each variation is compared to the baseline to evaluate sensitivity to prompt structure. Given that the ground-truth labels reflect stance rather than direct sentiment, testing robustness also assesses how well prompts help models interpret implicit attitude under ambiguous or indirect expressions.

Traditional sentiment classification is conducted using three models: VADER (a rule-based model with fixed sentiment thresholds), Twitter-RoBERTa Base, and a fine-tuned Twitter-RoBERTa variant. All traditional models operate on preprocessed textual inputs prepared through a dedicated cleaning pipeline, ensuring compatibility and fairness in evaluation.

To complement sentiment classification, Latent Dirichlet Allocation (LDA) is applied to the full dataset to identify latent thematic structures. The LDA model is trained on lemmatised text with stopwords removed (including domain-redundant terms such as "climate" and "change"). The number of topics is determined using coherence scores. Each tweet is assigned to a dominant topic based on the highest topic probability, enabling stratified evaluation of model performance across topical segments.

This work employs three evaluation metrics: Accuracy, macro-averaged F1-score, and confusion matrices. These metrics are computed on the full test set as well as within each identified topic, allowing both general and topic-specific performance comparisons. A stance labelled ground-truth test set, used as a proxy for sentiment, is used as the reference standard for all quantitative evaluations. This framing is particularly important for interpreting performance differences across models trained on affective sentiment versus stance-based ground truth.

Together, this dual-pipeline architecture supports a comprehensive analysis of model capabilities, comparative strengths, and reliability under varied prompt and content conditions. This approach also facilitates managerial insights into the feasibility of using Generative AI for monitoring public sentiment on sustainability-related issues.

## 5.3 Prompt Engineering Process

### 5.3.1 Modular Prompt Design and Prompt Selection

To systematically design effective prompts for sentiment classification, this work employs a modular prompt engineering strategy. All prompts are based on a shared template: a clear task instruction followed by a tweet to be classified and a placeholder for the model's output. This base structure is extended with additional components that reflect different types of contextual cues and reasoning instructions. The goal is to evaluate how various combinations of prompt components influence model accuracy and robustness.

The base prompt provides the task specification and embeds the tweet text in a standardised format:

```
Classify the sentiment of the following tweet as 'positive',
'neutral', or 'negative'.
Tweet: "{tweet_text}"
Sentiment:
```

Building on this foundation, six modular components are defined and combined in all 64 possible ways ($2^6$) with the base prompt. The full text of each component is as follows:

- Role:

```
You are an impartial social-media analyst.
```

- Domain:

```
Tweets discuss climate change, climate action, or sustainability.
```

- Label Explanation:

```
 - positive: if the tweet supports climate action and sustainability,
expresses concern about climate change, or affirms its reality.
- negative: if the tweet denies, mocks, criticises, or downplays
climate change, climate action, or sustainability.
- neutral: if the tweet does not clearly express a stance or is purely
factual.
```

- Few-Shot Examples:

```
 Here are some examples:
Tweet:  "Climate change is real and we need to act now to save the
planet."
Sentiment:  positive
Tweet:  "Global warming is a hoax created to control us."
Sentiment:  negative
Tweet:  "The IPCC released its latest climate report today."
Sentiment:  neutral
```

- Sarcasm Cue:

```
Tweets may contain sarcasm.
```

- Self-Check Instruction:

```
Verify label before replying.
```

These specific components and the base prompt are chosen to reflect theoretically grounded and practically relevant dimensions of prompt construction for sentiment analysis in the sustainability domain. The base prompt establishes a clear, minimal task formulation that serves as a consistent reference point across all variations. Each modular component targets a distinct cognitive or contextual aid: the Role cue frames the model's perspective, the Domain cue orients it toward the correct context, the Label Explanation clarifies the meaning of the label, the Few-Shot Examples demonstrate task format and logic, the Sarcasm Cue primes the model for nuanced tone interpretation, and the Self-Check Instruction encourages response verification. Together, these elements cover a spectrum from task comprehension to reasoning reliability, enabling a systematic evaluation of how prompt content influences model behavior and performance in a complex real-world classification task.

To analyse the contribution of each component, average accuracy scores are compared between prompts that include the component and those that do not. A two-sample t-test is conducted to assess whether the observed differences are statistically significant. In addition, a delta ranking is computed for each component, measuring the average change in accuracy when the component is included versus excluded. Finally, to evaluate the significance of each prompt component, a linear regression is performed with accuracy as the dependent variable and the components as independent variables.

The best-performing prompt for each model is selected based on highest overall accuracy and used to classify the full test set.

### 5.3.2 Prompt Robustness Evaluation

Following prompt selection, robustness is evaluated by applying controlled variations to the best-performing prompt for each model. Two perturbations are introduced independently to

test how resilient the model's performance is to structural changes.

The first variation removes all punctuation from the prompt, including commas, full stops, colons, empty lines and spaces. This tests the model's sensitivity to structure and formatting.

The second variation introduces a deliberately misleading example into the prompt:

```
Tweet: "Global warming is a hoax created to control us."
Sentiment: positive
```

This contradicts the correct sentiment label and is designed to assess whether the internal logic of the prompt can handle a false example.

These two specific variations are chosen to evaluate distinct dimensions of prompt robustness. The removal of punctuation isolates the role of syntactic and visual structure in model comprehension, testing whether language models rely on formatting cues to anchor their reasoning. In contrast, the false example challenges the semantic coherence of the prompt itself, examining whether the model blindly replicates patterns or exhibits internal consistency by rejecting misleading demonstrations. Together, these perturbations simulate real-world risks such as noisy input formatting and adversarial examples, providing a targeted stress test of prompt resilience under imperfect conditions.

Each variant is applied separately to the selected prompt and re-evaluated on the full test set of 2,003 tweets. Performance is then compared to the original prompt baseline using accuracy and Macro F1-scores. While no formal stability threshold is imposed, qualitative interpretation of robustness is reserved for discussion in the Chapter 7. The full text for the base prompt, prompt components, best-performing prompts and robustness prompt variants can be found in Appendix A.

## 5.4 LLM Setup and Execution

All prompts are evaluated using two state-of-the-art large language models: GPT-4o (`gpt-4o-2024-08-06`), accessed via the OpenAI API, and DeepSeek-Chat (`DeepSeek-V3-0324`), accessed via the DeepSeek API. Both models are interfaced through chat-based endpoints and are queried using structured prompts constructed as described in Section 5.3.

To standardize the interaction, a unified Python wrapper is used to submit each prompt to the respective API and retrieve the model response. The generation temperature is fixed at 0.0 across all experiments to promote deterministic behavior and minimize output randomness. Other decoding parameters are left at their default values to retain model-native behavior.

Because model outputs are not constrained to a specific format, a post-processing step is applied to extract the predicted sentiment label. This involves lowercasing the output and scanning for the first occurrence of one of the three valid sentiment labels: `positive`, `neutral`, or `negative`. If no valid label is detected, the response is marked as `unknown`. No reruns or manual corrections are applied. This ensures consistency and transparency in evaluation, especially when dealing with malformed or ambiguous outputs.

This implementation allows the selected prompts to operate in either zero-shot or few-shot configurations, depending on the components included. Final prompts, selected based on validation accuracy (see Section 5.3), are used for all evaluations on the full 2,003-tweet test set in the subsequent sentiment classification phase.

## 5.5 Sentiment Classification Methods

### 5.5.1 Prompt-Based Classification with LLMs

Sentiment classification, based on stance-oriented labels that serve as an imperfect proxy for emotional sentiment, is conducted by applying the finalized prompt to the full test set of 2,003 tweets. The predicted sentiment labels (positive, neutral, or negative) are extracted using the post-processing method described in the previous section. These predictions form the basis for evaluating overall classification performance and for topic-specific sentiment analysis in the next sections.

### 5.5.2 Traditional Models (VADER, Twitter-RoBERTa Base, Twitter-RoBERTa Fine-Tuned)

The traditional sentiment analysis pipeline benchmarks the performance of the Generative AI models against established lexicon-based and transformer-based classifiers. Three models are used:

1. **VADER** (Hutto & Gilbert, 2014): A rule-based sentiment classifier optimised for social media text. It outputs a compound sentiment score for each input tweet, which is mapped to sentiment classes using standard thresholds: A compound score $> 0.05$ is labelled as `positive`, a compound score $< -0.05$ as `negative`, and scores in between as `neutral`.

2. **Twitter-RoBERTa Base** (Barbieri et al., 2020): A transformer model pre-trained on 58 million tweets using the RoBERTa architecture. The version used is `cardiffnlp/twitter-roberta-base-sentiment`, loaded directly from the HuggingFace model hub.

3. **Twitter-RoBERTa Fine-Tuned:** The same pretrained model is further fine-tuned on the train set used in this work. Fine-tuning is performed for up to five epochs using HuggingFace's Trainer class with early stopping based on validation accuracy (patience = 2). The model is trained end-to-end, meaning all parameters, including those in the transformer backbone, are updated. To mitigate overfitting on the relatively small training set, several regularisation techniques are applied: a low learning rate of $2 \times 10^{-5}$, weight decay of 0.01, and dynamic padding via a DataCollatorWithPadding for efficient batching. Performance is evaluated using weighted F1-score and accuracy.

All three models operate on the preprocessed text as described in Chapter 4.

Tokenisation for the Twitter-RoBERTa models is handled using the associated tokenizer from the `transformers` library. Model predictions are obtained by applying the softmax function to

the output logits and selecting the class with the highest probability. These predictions are compared against the ground-truth labels in the test set for model evaluation.

This traditional pipeline serves as a baseline for evaluating the effectiveness of prompt-based Generative AI methods and provides complementary insights into model reliability, interpretability, and topic sensitivity.

## 5.6 Model Evaluation Metrics

To assess model performance, this work uses three metrics: accuracy, macro-averaged F1-score, and confusion matrices. These metrics collectively capture overall classification quality, class-level balance between precision and recall, and the distribution of model errors. (Sokolova & Lapalme, 2009) All evaluation metrics are computed using ground-truth labels that reflect stance, which serves as a proxy for affective sentiment in this work due to data limitations.

*Accuracy* is defined as the proportion of correct predictions out of all predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

where $TP$ and $TN$ denote true positives and true negatives, and $FP$ and $FN$ denote false positives and false negatives, respectively. In the multi-class setting, accuracy is computed as the number of correctly predicted labels divided by the total number of predictions.

Although not reported directly, *precision* and *recall* are used to compute the F1-score for each class $c$:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{5.2}$$

The *F1-score* for class $c$ is the harmonic mean of its precision and recall:

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{5.3}$$

The *macro-averaged F1-score* is then calculated by averaging the F1-scores across all $K$ classes:

$$\text{Macro-F1} = \frac{1}{K} \sum_{c=1}^{K} \text{F1}_c \tag{5.4}$$

Confusion matrices are used to visualise the distribution of predicted versus actual sentiment labels and to identify systematic patterns of misclassification.

All metrics are computed for each model under evaluation. For the Generative AI models, GPT-4o and DeepSeek-Chat, metrics are reported both for the best-performing prompt and for the two robustness variants (punctuation removed, false example injected). For the traditional models (VADER, Twitter-RoBERTa Base, and Twitter-RoBERTa Fine-Tuned), metrics are reported for a single run on the test set. All metrics are computed on the 2,003-tweet test set, which contains ground-truth stance labels as described in Chapter 4. These ground-truth labels represent stance rather than affective sentiment, and interpretations of model performance should be understood in this context.

Evaluation is conducted on two levels: the entire test set and within each topic segment as

assigned by the LDA model. This approach enables performance comparison across thematic subsets of the data, offering insights into model behaviour across different sustainability-related topics.

## 5.7 Topic Modelling and Topic-Based Evaluation

### 5.7.1 Latent Dirichlet Allocation

To uncover latent themes in sustainability-related tweets, this work applies Latent Dirichlet Allocation (LDA) (Jelodar et al., 2017) using the `gensim` Python library. The model is trained on the full dataset (train, test, and validation) of lemmatised tweets, with domain-specific terms and general stopwords removed. In particular, the terms *climate* and *change* are excluded to reduce semantic redundancy. Tokens appearing in fewer than five tweets or more than 40% of tweets are also filtered out.

To determine the number of topics $k$, LDA models are fit across a range of values from 2 to 20, in increments of 2. For each value of $k$, the model's performance is evaluated using two metrics:

- **Topic coherence** ($c_v$) (Röder et al., 2015): a measure of semantic similarity among the top words in each topic, based on their co-occurrence in a sliding window.
- **Topic diversity**: the proportion of unique words among the top-$n$ words across all topics, defined as:

$$\text{Topic Diversity} = \frac{|\bigcup_{k=1}^{K} \text{TopN}_k|}{K \times n} \tag{5.5}$$

  where $\text{TopN}_k$ is the set of top-$n$ words for topic $k$, and $K$ is the total number of topics.

While the coherence and diversity curves provide guidance, the final value of $k$ is manually selected to balance topic interpretability, coherence, and diversity. Individual topic coherence scores are also reviewed to identify and discard topics with nonsensical or low-quality word distributions.

Each topic is examined by inspecting its top eight terms, extracted from the model's learned topic-word distributions. These keyword lists are used to verify semantic coherence and distinctiveness. Topic labels are inferred from these top terms and used in subsequent topic-stratified evaluation.

### 5.7.2 Topic Assignment for Model Comparison

After training, the LDA model is applied to the preprocessed test set. Each tweet is assigned a probability distribution over the $k$ topics, and a single dominant topic is selected based on the highest posterior probability. This one-topic-per-tweet assignment allows for analysis of model performance within and between each distinct topic.

All evaluation metrics, accuracy, macro-averaged F1-score, and confusion matrices, are computed both on the full test set and separately within each topic-defined subset. This enables topic-level diagnostic evaluation, revealing how classification accuracy and model performance

vary across different areas of sustainability discourse. However, since sentiment labels reflect stance, topic-level results are interpreted as topic-wise patterns in expressed positions toward sustainability, not necessarily emotional sentiment.

## 5.8 Summary of Methodological Approach

This chapter has outlined the full methodological framework used to evaluate sentiment classification performance in sustainability-related Twitter discourse. A dual-pipeline structure is implemented, comparing prompt-based Generative AI models to traditional sentiment analysis methods. Prompt design follows a modular approach, with 64 prompt variants tested and the best-performing version selected for each model. Robustness is assessed through controlled prompt perturbations. The traditional pipeline includes VADER, a pre-trained Twitter-RoBERTa model, and a fine-tuned variant trained on this work's training data.

Model performance is evaluated using accuracy, macro-averaged F1-score, and confusion matrices. Additionally, Latent Dirichlet Allocation is used to extract topics from the tweet corpus, allowing for stratified performance comparison across topic-defined subsets of the test set. Together, these methodological components support a rigorous and interpretable comparison of model behaviour across models, prompt designs, and topics. Although the current analysis relies on stance-labelled data, the methodological design and evaluation framework are transferable to future datasets with affective sentiment annotations.

# Chapter 6

# Results

## 6.1 Introduction

This chapter presents the empirical results of this work. It begins with an evaluation of the prompt engineering process, showing how the modular prompt design impacts classification performance and how the best prompt is selected for each LLM. It then compares the overall performance of all models, examines the robustness of the selected LLM prompts under controlled perturbations, and evaluates model performance across topic-defined segments. Throughout this chapter, it is important to recall that the labels reflect stance toward climate change rather than emotional sentiment. This distinction shapes all interpretations of model outputs.

## 6.2 Prompt Engineering Results

This section presents the results of the modular prompt engineering process used to guide classification of stance (used as a proxy for sentiment) by Large Language Models (LLMs). A total of 64 prompt variants are tested for each model, GPT-4o and DeepSeek-Chat, on a validation subset of 1,000 labelled tweets. Each variant represents a unique combination of six modular prompt components. The evaluation focuses on overall accuracy, the contribution of individual components, and the selection of the best-performing prompt for each model.

### 6.2.1 Accuracy per Prompt Variant

The performance of all prompt variants is visualised in Figure 6.1, which displays the classification accuracy of each combination for both models. For DeepSeek-Chat, the accuracy ranges from 0.379 (lowest-performing variant) to 0.640 (best-performing variant), while GPT-4o varies from 0.259 to 0.605 across the same range of prompts.

Overall, prompts enriched with contextual guidance and reasoning instructions substantially outperform the base prompt, which contains only a minimal task description. In particular, combinations including the *Label Explanation* component consistently produce higher accuracy. This can most likely be attributed to the fact that ground-truth labels reflect stance instead of sentiment. The addition of the *Role* and *Domain* components also yields performance gains, particularly for GPT-4o. The complete set of accuracy scores per prompt variant is provided in

Appendix B, Table B.1.

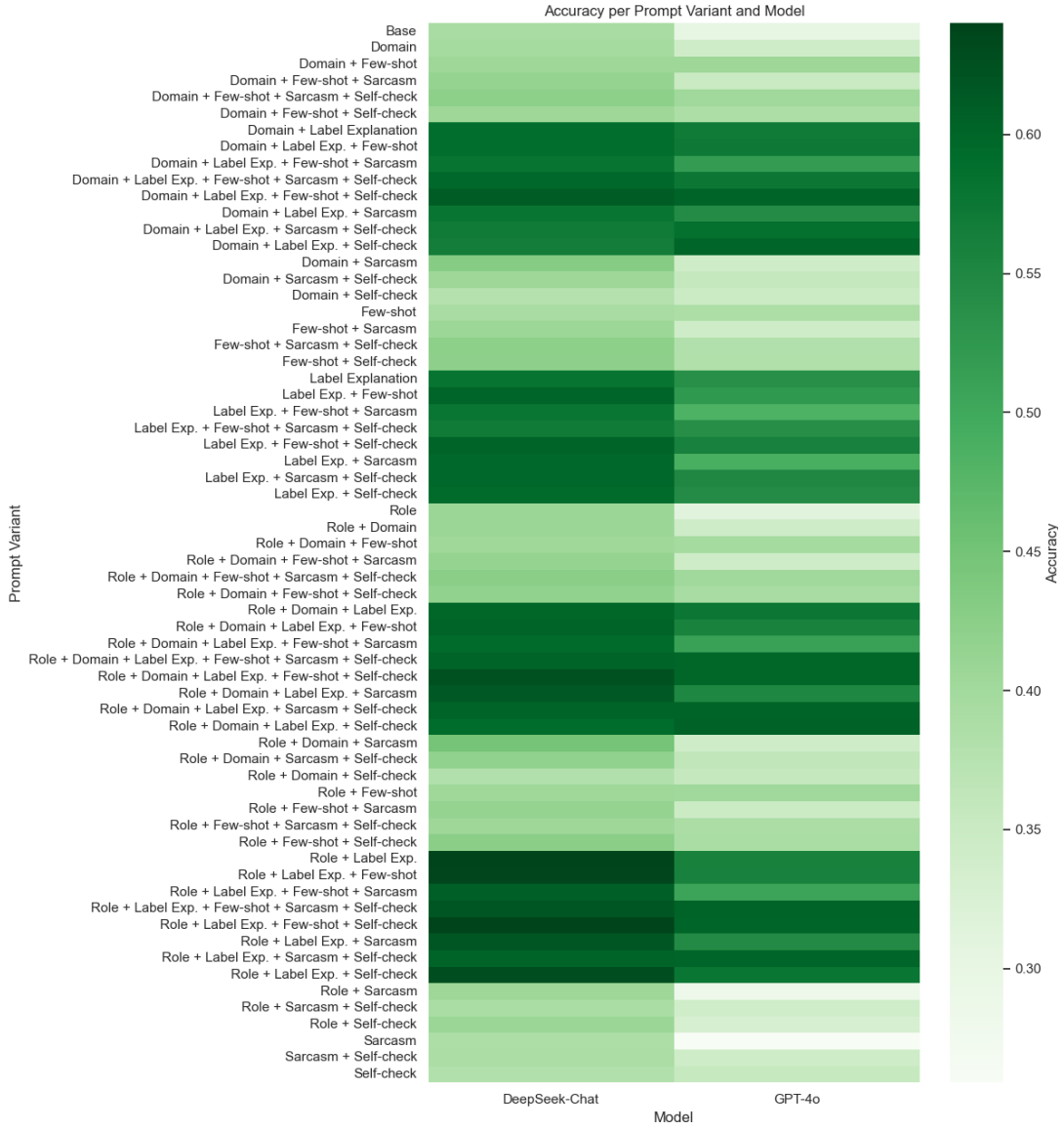


Figure 6.1: Accuracy per Prompt Variant and Model

### 6.2.2 Component-Level Effects

To evaluate the effect of individual prompt modules, mean accuracy is compared between prompts that include a given component and those that do not. The results, shown in Apendix B, Figure B.1, are aggregated across all prompt variants and both LLMs. Actual values are provided in Appendix B, Table B.2.

Among the six components, *Label Explanation* has the largest positive impact, increasing mean accuracy by 20.4 percentage points for GPT-4o and 19.4 for DeepSeek-Chat. These effects are statistically significant at the $p < .001$ level (see Appendix B, Table B.3). The *Role Instruction* and *Self-Check Step* also show consistent, although smaller, positive effects across both models. Conversely, the *Sarcasm Cue* shows minimal or negative influence, particularly for GPT-4o, suggesting that such instructions may confuse the model.

The delta ranking analysis (Appendix B, Table B.4) confirms this pattern. *Label Explanation* ranks highest in average performance gain for both LLMs, followed by *Role Instruction*. These results underscore the importance of including task-specific label guidance when prompting LLMs for sentiment classification.

Table 6.1: Regression Results for GPT-4o and DeepSeek-Chat

| | **GPT-4o** | **DeepSeek-Chat** |
|---|---|---|
| Constant | 0.3204*** | 0.3962*** |
| | (0.0080) | (0.0052) |
| Role | 0.0130** | 0.0170*** |
| | (0.0061) | (0.0039) |
| Domain | 0.0223*** | -0.0024 |
| | (0.0061) | (0.0039) |
| Label explanation | 0.2040*** | 0.1938*** |
| | (0.0061) | (0.0039) |
| Few-shot | 0.0205*** | 0.0082** |
| | (0.0061) | (0.0039) |
| Sarcasm | -0.0181*** | 0.0003 |
| | (0.0061) | (0.0039) |
| Self-Check | 0.0360*** | -0.0010 |
| | (0.0061) | (0.0039) |
| R-squared | 0.9550 | 0.9773 |
| Adjusted R-squared | 0.9502 | 0.9749 |
| Observations | 64 | 64 |

*Note.* Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

The regression results in Table 6.1 confirm the findings from the component-level comparisons: Label Explanation has by far the strongest and most significant effect on accuracy for both models. Other components, such as Role Instruction and Few-Shot Examples, also show positive and statistically significant contributions. Interestingly, several components that are significant for GPT-4o, such as Domain, Sarcasm, and Self-Check, do not reach significance for DeepSeek-Chat, suggesting that GPT-4o is more sensitive to prompt structure. The high R-squared values indicate that the component indicators explain most of the variance in prompt performance. Notably, the standard errors are identical across predictors within each model, this is a consequence of the balanced design: each component is included in exactly half of the 64 prompts, and the predictors are orthogonal by construction. This setup leads to homoskedastic and uncorrelated regressors, resulting in equal standard errors.

### 6.2.3 Best Prompt Selection

The final prompt for each LLM is selected based solely on accuracy on the validation set. For GPT-4o, the best-performing prompt combines four modules:

**Role + Domain + Label Explanation + Self-Check** (Accuracy: 0.605)

For DeepSeek-Chat, a shorter configuration yields the highest score:

**Role + Label Explanation** (Accuracy: 0.640)

These optimal prompts are subsequently used for all test set predictions and robustness evaluations. Notably, while GPT-4o benefits from multiple guiding components, DeepSeek-Chat achieves better results with a more concise prompt, suggesting model-specific sensitivities to prompt complexity. Differences in performance may also relate to how well each model interprets stance-oriented language when prompted under sentiment framing.

## 6.3 Overall Performance Comparison

This section presents the comparative evaluation of all five sentiment classification models tested in this work: GPT-4o, DeepSeek-Chat, VADER, Twitter-RoBERTa Base, and a fine-tuned version of Twitter-RoBERTa. Performance is assessed on the full 2,003-tweet test set using both accuracy and macro-averaged F1-score. In addition, confusion matrices are analysed to reveal common misclassification patterns and class-specific performance dynamics.

### 6.3.1 Accuracy and Macro-F1 Overview

Table 6.2 summarises the classification performance of all models. Among the two prompt-based LLMs, DeepSeek-Chat outperforms GPT-4o with an accuracy of 0.603 and a macro-F1 of 0.512, compared to 0.553 and 0.504, respectively. However, both models lag behind the fine-tuned Twitter-RoBERTa model, which achieves the highest overall scores (accuracy = 0.759, macro-F1 = 0.702).

Unsupervised and non-fine-tuned baselines underperform considerably. VADER achieves an accuracy of 0.352, while the "zero-shot" Twitter-RoBERTa Base model scores just 0.311, highlighting the limitations of these models in handling domain-specific sentiment. This gap is likely amplified by the mismatch between the stance-based ground-truth and the sentiment orientation of these baseline models.

Table 6.2: Comparison of Sentiment Classification Models

| Model | Accuracy | Macro-F1 |
|---|---|---|
| VADER | 0.352 | 0.327 |
| Twitter-RoBERTa Base | 0.311 | 0.301 |
| Twitter-RoBERTa Fine-Tuned | 0.759 | 0.702 |
| GPT-4o | 0.553 | 0.504 |
| DeepSeek-Chat | 0.603 | 0.512 |

### 6.3.2 Confusion Matrices

Tables 6.3 through 6.7 display the confusion matrices for each model. These reveal distinct patterns of classification behaviour.

VADER and RoBERTa Base show severe over-classification into the *negative* and *neutral* classes. For instance, VADER misclassifies 488 positive tweets as negative, while Twitter-RoBERTa Base misclassifies 538 and 519 positive tweets as negative and neutral, respectively.

These errors indicate that non-adaptive models tend to overlook subtle stance cues indicating support for climate action, particularly when such cues do not carry strong affective signals.

In contrast, the prompt-based models, GPT-4o and DeepSeek-Chat, show improved balance between classes but still over-classify the negative class. Both LLMs frequently misclassify positive or neutral tweets as negative, with GPT-4o assigning 621 (300 positive, 321 neutral) such errors and DeepSeek-Chat 548 (233 positive, 315 neutral). This suggests that, despite their contextual reasoning capabilities, LLMs may exhibit a bias toward identifying critical or sceptical sentiment, potentially reflecting the framing of sustainability discourse in the data or an overemphasis on negative linguistic cues. This may reflect a conflation of anti-climate stance with negative sentiment, revealing a limitation of sentiment-oriented tools when applied to attitudinal data.

The fine-tuned Twitter-RoBERTa model exhibits the most balanced performance, accurately classifying the majority of examples across all sentiment classes. It shows strong precision on the positive class and reduced spillover into incorrect labels, reflecting the benefits of task-specific tuning. Importantly, this model is fine-tuned on the same dataset with stance-based labels, meaning it learns to classify attitudinal position rather than emotional sentiment.

Table 6.3: Confusion Matrix: `VADER`

| | Pred: Negative | Pred: Neutral | Pred: Positive |
|---|---|---|---|
| True: Negative | 133 | 42 | 94 |
| True: Neutral | 159 | 148 | 277 |
| True: Positive | 488 | 238 | 424 |

Table 6.4: Confusion Matrix: `Twitter-RoBERTa Base`

| | Pred: Negative | Pred: Neutral | Pred: Positive |
|---|---|---|---|
| True: Negative | 179 | 84 | 6 |
| True: Neutral | 191 | 350 | 43 |
| True: Positive | 538 | 519 | 93 |

Table 6.5: Confusion Matrix: `Twitter-RoBERTa Fine-Tuned`

| | Pred: Negative | Pred: Neutral | Pred: Positive |
|---|---|---|---|
| True: Negative | 161 | 43 | 65 |
| True: Neutral | 45 | 334 | 205 |
| True: Positive | 34 | 91 | 1025 |

Table 6.6: Confusion Matrix: `GPT-4o`

| | Pred: Negative | Pred: Neutral | Pred: Positive |
|---|---|---|---|
| True: Negative | 250 | 10 | 9 |
| True: Neutral | 321 | 168 | 95 |
| True: Positive | 300 | 161 | 689 |

Table 6.7: Confusion Matrix: `DeepSeek-Chat`

| | Pred: Negative | Pred: Neutral | Pred: Positive |
|---|---|---|---|
| True: Negative | 245 | 11 | 13 |
| True: Neutral | 315 | 112 | 157 |
| True: Positive | 233 | 66 | 851 |

## 6.4 Prompt Robustness Analysis

This section evaluates the sensitivity of prompt-based LLM performance to structural and semantic changes in the prompt design. Robustness is assessed in two ways. First, models are re-evaluated under controlled prompt variations, removal of punctuation, and inclusion of a misleading example. Second, qualitative inspection of misclassified tweets is used to identify recurring failure modes.

### 6.4.1 Accuracy Drop from Baseline

Table 6.8 presents the accuracy and macro-F1 scores of both LLMs under three prompt conditions: the selected baseline prompt, a version with all punctuation removed, and a version containing a deliberate false example. Inference times are also included for reference. The exact prompts can be found in Appendix A.

Table 6.8: Robustness Comparison Across Prompt Variants and Models

| Model | Accuracy | Macro-F1 |
|---|---|---|
| GPT-4o (Baseline) | 0.553 | 0.504 |
| GPT-4o (No Punctuation) | 0.555 | 0.511 |
| GPT-4o (False Example) | 0.495 | 0.450 |
| DeepSeek-Chat (Baseline) | 0.603 | 0.512 |
| DeepSeek-Chat (No Punctuation) | 0.555 | 0.475 |
| DeepSeek-Chat (False Example) | 0.513 | 0.454 |

The results indicate that GPT-4o is largely robust to the removal of punctuation, surprisingly with a small improvement in both accuracy and macro-F1. In contrast, DeepSeek-Chat experiences a more substantial decline under the same condition. Both models, however, show

significant performance drops when a false example is inserted into the prompt, especially in macro-F1, suggesting vulnerability to semantically misleading inputs. This is particularly relevant when prompts are designed for sentiment classification, but labels encode attitudinal stance, which can be semantically subtle or indirect.

### 6.4.2 Misclassified Observations

To better understand model limitations, Table 6.9 presents a selection of six test-set tweets that are misclassified by GPT-4o and/or DeepSeek-Chat. These examples highlight key robustness challenges when classifying stance toward climate change under a sentiment framing.

Table 6.9: Examples of Misclassified Tweets and Model Predictions

| ID | Tweet Text | True Label | GPT-4o | DeepSeek-Chat |
|---|---|---|---|---|
| 1 | *Where is global warming when you want it?* | Neutral | Negative | Negative |
| 2 | *@user @user next, learn how to grow cocoa and wait for global warming to induce the correct climate* | Positive | Negative | Negative |
| 3 | *Just think – Trump still maintains that global warming is a hoax. (Maybe when Mar-a-Lago is drowning...)* | Positive | Negative | Positive |
| 4 | *next you're gonna blame oil companies for global warming! 'yes because they are to blame' user* | Positive | Negative | Positive |
| 5 | *Another school shooting today. Can we talk about climate change NOW?* | Neutral | Negative | Positive |
| 6 | *@user @user this is what you tools sound like when you say global warming isn't real.* | Positive | Positive | Negative |

Three recurring failure modes emerge from these examples:

1. Sarcasm and Irony: Tweets with sarcastic or ironic phrasing (e.g., Example 1) are interpreted as negative sentiment, even when their stance toward climate change is neutral or supportive. Similarly, GPT-4o misclassifies Example 3, which ironically critiques climate denial, due to its emotionally ambiguous tone.

2. Contextual Ambiguity: Examples 4, 5, and 6 illustrate the difficulty of interpreting attitudinal stance when tweets contain rhetorical questions, quoted speech, or emotionally charged context. The models diverge in their predictions, likely because the stance signal is embedded in indirect language and due to the models being prompted for sentiment instead of stance.

3. Lexical Triggers vs. Intent: In Example 2, the presence of phrases like "global warming" and "induce the correct climate" may be interpreted as negative sentiment, despite the tweet expressing a positive stance. This suggests that the models sometimes rely too heavily on sentiment-associated keywords rather than inferring attitudinal alignment.

These observations underscore that even strong prompt-based models can struggle with nuanced stance detection, particularly when using sentiment-oriented prompts and evaluation frameworks. This limitation is especially evident on social media platforms like Twitter, where indirect, sarcastic, or emotionally complex expressions often encode stance without overt affect.

## 6.5 Topic-Specific Evaluation

This section evaluates model performance across different semantic segments of the tweet corpus, defined by latent topics extracted via Latent Dirichlet Allocation (LDA). The analysis provides insight into which types of sustainability-related content are easier or harder for different models to classify correctly.

### 6.5.1 LDA Topics

The optimal number of topics is selected using a combination of topic coherence and topic diversity metrics, as shown in Appendix B, figures B.1 and B.2. Based on this analysis, the model is fit using $k = 8$ topics. A relatively low number of topics is preferred in this work to ensure interpretability and alignment with the evaluation scope. Only five of the eight topics are retained for downstream evaluation based on interpretability, diversity, and coherence. The discarded topics are low in semantic structure or dominated by noisy or generic language.

Table 6.10 summarises the five retained topics, including their top terms, coherence scores, and assigned labels. Topics are manually interpreted and labelled to reflect their dominant themes.

Table 6.10: Summary of Topics Extracted via LDA ($k = 8$)

| Topic | Coherence | Top Terms |
|---|---|---|
| Topic 1 | 0.3953 | human, al, rise, level, gore, cause, idea, year |
| Topic 2 | 0.3069 | affect, water, week, need, environmental, new, ... |
| Topic 3 | 0.3640 | global, warming, user, cause, real, warm, like, ... |
| Topic 4 | 0.2836 | trump, believe, think, hoax, epa, president, do ... |
| Topic 5 | 0.2695 | user, real, people, know, deny, believe, science ... |

The assigned labels for each topic are as follows:

- Topic 1: Human Cause & Sea Level Rise
- Topic 2: Urban & Environmental Impacts
- Topic 3: Global Warming Attribution

- Topic 4: Political Denial & Polarisation
- Topic 5: Science Denial & Public Beliefs

The distribution of tweets and sentiment across these five topics is shown in Table 6.11.

Table 6.11: Tweet Distribution Across Topics (LDA, $k = 8$)

| Topic | Tweet Count | Positive (%) | Neutral (%) | Negative (%) |
|---|---|---|---|---|
| 1. Human Cause & Sea Level Rise | 185 | 62.7 | 28.1 | 9.2 |
| 2. Urban & Environmental Impacts | 207 | 73.4 | 22.7 | 3.9 |
| 3. Global Warming Attribution | 695 | 42.7 | 39.6 | 17.7 |
| 4. Political Denial & Polarisation | 321 | 67.9 | 21.2 | 10.9 |
| 5. Science Denial & Public Beliefs | 595 | 61.7 | 23.9 | 14.5 |

### 6.5.2 Performance by Topic

Model performance varies substantially across topics, as shown in Table 6.12. The fine-tuned Twitter-RoBERTa model achieves the highest accuracy and macro-F1 scores across all topics. DeepSeek-Chat consistently outperforms GPT-4o on a per-topic basis and is the best-performing GenAI model overall.

Table 6.12: Model Performance by Topic (LDA, $k = 8$)

| Topic | Model | Accuracy | Macro-F1 |
|---|---|---|---|
| 1. Human Cause & Sea Level Rise | VADER | 0.314 | 0.287 |
| | Twitter-RoBERTa Base | 0.243 | 0.228 |
| | Twitter-RoBERTa Fine-Tuned | 0.719 | 0.637 |
| | GPT-4o | 0.524 | 0.451 |
| | DeepSeek-Chat | 0.562 | 0.441 |
| 2. Urban & Environmental Impacts | VADER | 0.338 | 0.303 |
| | Twitter-RoBERTa Base | 0.266 | 0.243 |
| | Twitter-RoBERTa Fine-Tuned | 0.821 | 0.642 |
| | GPT-4o | 0.671 | 0.560 |
| | DeepSeek-Chat | 0.768 | 0.626 |
| 3. Global Warming Attribution | VADER | 0.338 | 0.311 |
| | Twitter-RoBERTa Base | 0.358 | 0.319 |
| | Twitter-RoBERTa Fine-Tuned | 0.745 | 0.721 |
| | GPT-4o | 0.509 | 0.486 |
| | DeepSeek-Chat | 0.531 | 0.484 |
| 4. Political Denial & Polarisation | VADER | 0.333 | 0.305 |
| | Twitter-RoBERTa Base | 0.255 | 0.265 |
| | Twitter-RoBERTa Fine-Tuned | 0.776 | 0.642 |
| | GPT-4o | 0.530 | 0.478 |
| | DeepSeek-Chat | 0.629 | 0.501 |
| 5. Science Denial & Public Beliefs | VADER | 0.395 | 0.371 |
| | Twitter-RoBERTa Base | 0.321 | 0.322 |
| | Twitter-RoBERTa Fine-Tuned | 0.756 | 0.692 |
| | GPT-4o | 0.583 | 0.524 |
| | DeepSeek-Chat | 0.629 | 0.525 |

### 6.5.3 Interpretation of Topic Sensitivity

Several patterns emerge in the topic-wise comparison. First, GenAI models outperform traditional baselines (VADER and Twitter-RoBERTa Base) across all topics, with performance gaps ranging from 15 to over 50 percentage points in accuracy. The largest gap is observed for Topic 2 (Urban & Environmental Impacts), where DeepSeek-Chat achieves 0.768 accuracy compared to only 0.266 for Twitter RoBERTa Base.

Second, both GPT-4o and DeepSeek-Chat show relatively low accuracy on Topic 3 (Global Warming Attribution), suggesting that complex attribution language or ambiguous framing may reduce model effectiveness. This topic yields the lowest accuracy among all categories for DeepSeek-Chat.

Finally, the best overall performance across models is observed on Topic 2 (Urban & Environmental Impacts), where DeepSeek-Chat reaches 0.768 accuracy and GPT-4o peaks at 0.671. These tweets may feature more concrete, lexically distinct phrasing that aligns well with model

expectations.

These findings demonstrate that topic sensitivity plays a key role in model performance and that model robustness cannot be evaluated solely on overall metrics. Moreover, stance complexity varies by topic: ideological or sarcastic content may obscure attitudinal intent, further complicating sentiment-framed classification

## 6.6 Summary of Key Findings

This chapter presented the empirical results of this work, including model-level performance comparisons, prompt engineering outcomes, robustness to prompt perturbations, and topic-specific evaluations.

The best-performing model overall is the Twitter-RoBERTa Fine-Tuned classifier, which achieves an accuracy of 0.759 and a macro-F1 score of 0.702. Among the prompt-based generative models, DeepSeek-Chat outperforms GPT-4o, scoring 0.603 and 0.553 in accuracy, respectively. In contrast, traditional baselines such as VADER (0.352) and the zero-shot Twitter-RoBERTa Base (0.311) perform substantially worse.

Prompt engineering has a meaningful impact on model performance. For GPT-4o, the optimal prompt combines four components: *Role*, *Domain*, *Label Explanation*, and *Self-Check*. For DeepSeek-Chat, a simpler configuration consisting of the *Role* and *Label Explanation* components, is sufficient to maximise accuracy. Component-level analysis confirms that *Label Explanation* is the most impactful element, significantly boosting classification performance across both LLMs. This is likely a result of the ground-truth labels indicating stance instead of sentiment.

Robustness analysis reveals that both models remain largely unaffected by the removal of punctuation, but are negatively impacted when misleading examples are inserted into the prompt. DeepSeek-Chat exhibits greater sensitivity to such semantic corruption than GPT-4o, showing larger drops in both accuracy and macro-F1.

Finally, the topic-level evaluation shows meaningful variation in model performance. All models struggle most with Topic 3 (*Global Warming Attribution*), likely due to abstract or ambiguous framing. By contrast, Topic 2 (*Urban & Environmental Impacts*) is the easiest for all models, yielding the highest accuracy. GenAI models consistently outperform VADER and Twitter RoBERTa Base on Topic 4 (*Political Denial & Polarisation*), indicating that prompt-based models may be better equipped to interpret ideologically charged discourse.

Taken together, these results highlight the effectiveness of prompt design in boosting classification performance, the robustness trade-offs introduced by generative models, and the critical role of topical context in shaping sentiment classification outcomes. These outcomes must be interpreted in light of the underlying label type. Because the models classify stance, not sentiment in the affective sense, results indicate performance on identifying attitudinal alignment rather than emotional tone

# Chapter 7

# Discussion

## 7.1 Overview and Objectives

This chapter discusses the findings of this work in relation to the initial research objectives, methodological choices, and broader implications. The work was driven by the need to better understand public sentiment surrounding sustainability topics as expressed on Twitter, and to evaluate whether recent advances in Generative AI could improve the analysis of such sentiment in terms of accuracy and robustness.

Due to the limited availability of high-quality sentiment labelled datasets, this work uses a stance labelled dataset as a proxy for affective sentiment. While this introduces interpretive constraints, the methods applied, especially the prompt engineering approach and prompt-based classification, remain valid and transferable to future applications involving genuine sentiment data.

The main objective of this work was to evaluate how Generative AI, through prompt engineering, can be effectively applied to classify stance as a proxy for sentiment and extract insights across sustainability-related topics on Twitter, and to explore what managerial insights can be derived from its application.

To address this objective, the following sub-questions were investigated:

1. How can prompt engineering be applied to classify the sentiment of tweets using Generative AI?
2. How do Generative AI models compare to traditional sentiment analysis methods (e.g., VADER, BERT) in terms of accuracy and efficiency?
3. To what extent are the outputs of Generative AI models robust across different prompt formulations?
4. How can topic modelling (using Latent Dirichlet Allocation) be applied to identify the main topics in sustainability-related tweets, and how do Generative AI models perform across these topics?

This discussion adopts an exploratory and comparative perspective. It reflects on the methodological contributions of prompt engineering in the context of sentiment classification, evaluates performance across both generative and traditional models, and considers how sentiment

varies across topics identified through Latent Dirichlet Allocation (LDA). Beyond reporting outcomes, this chapter also seeks to explain why certain patterns in the results may have emerged, drawing on theoretical expectations, data characteristics, and model behaviour.

Briefly summarising the results, the fine-tuned Twitter-RoBERTa model achieved the highest overall classification performance. Among the generative models, DeepSeek-Chat outperformed GPT-4o and both LLMs surpassed traditional baselines such as VADER and Twitter-RoBERTa Base. Prompt design had a clear impact on accuracy, and model sensitivity to prompt perturbations revealed strengths and weaknesses in robustness. Furthermore, performance varied substantially across thematic segments, suggesting that topic context plays a critical role in model effectiveness.

Where relevant, this chapter also touches on the normative implications of these findings, particularly for the design and deployment of GenAI tools in applied sustainability contexts. These are elaborated separately.

Throughout this discussion, it is important to note that the models were evaluated on stance labels. As such, interpretations of 'sentiment' must be viewed as reflections of opinionated position rather than emotional tone.

## 7.2 Interpretation of Results

### 7.2.1 Model Performance

The comparative evaluation of sentiment classification models reveals clear distinctions in performance, reflecting underlying differences in model design, training paradigms, and domain adaptability. Among all five models tested, the fine-tuned Twitter-RoBERTa classifier achieved the highest overall performance, with consistently strong results across all sentiment classes and thematic segments. This was expected given that it was specifically adapted to and fine-tuned on the training dataset at hand, enabling it to align with the stance framing of the task rather than detection of affective sentiment.

DeepSeek-Chat emerged as the best-performing generative model, outperforming both GPT-4o and traditional baselines such as VADER and the Twitter-RoBERTa Base model. Its relatively strong performance can be attributed to two factors: its balanced sensitivity across sentiment classes, and its ability to correctly interpret nuanced stance expressions in tweets without excessive overfitting to negative cues. In contrast, GPT-4o underperformed relative to DeepSeek-Chat, exhibiting a stronger bias toward negative classifications and reduced stability in handling tweets with mixed or ambiguous tone. One likely contributor to this discrepancy is GPT-4o's internal alignment and content-filtering mechanisms, which may interfere with accurate classification in emotionally or politically charged contexts, such as climate denial or political polarisation.

The "zero-shot" Twitter-RoBERTa Base model and the rule-based VADER classifier performed noticeably worse across all metrics. VADER tended to over-predict negative sentiment and failed to distinguish between irony, criticism, and genuine negative stance. Its reliance on fixed lexicons limits its capacity to interpret the contextual subtleties of social media language. Twitter-RoBERTa Base, although more sophisticated in architecture, lacked task-specific adapt-

ation, leading to poor generalisation in the stance-classification task within the sustainability domain and frequent misclassification of positive and neutral tweets. Importantly, neither model was originally developed for stance classification. Their design and training objectives are optimised for affective sentiment detection, which may limit their effectiveness when applied to stance-labelled data, as in this work.

These results underscore the importance of domain-specific fine-tuning and prompt engineering when applying sentiment models to complex, noisy text such as sustainability discourse on Twitter, especially when sentiment is approximated through stance-labelled data, as in this work. In practical terms, while fine-tuned models remain the most reliable option when labelled data are available, prompt-based generative models like GPT-4o and DeepSeek-Chat offer a viable and scalable alternative, particularly when access to labelled data or training infrastructure is limited. However, model selection should also consider trade-offs in interpretability, reliability, and response filtering, factors that may meaningfully differ across GenAI providers.

### 7.2.2 Prompt Engineering

The modular prompt engineering process revealed that specific prompt components had a substantial effect on classification performance, with notable differences between GPT-4o and DeepSeek-Chat. Across both models, the inclusion of a label explanation was by far the most influential component, yielding the largest average improvement in accuracy. This result highlights the importance of clearly defining classification targets, especially when ground-truth labels reflect stance, which in this work serves as a proxy for affective sentiment. The label explanation likely helped models align with the intended framing of the task by reducing ambiguity and reinforcing semantic boundaries between classes, despite the underlying complexity of mapping emotional tone to stance.

Role specification (e.g., "You are an impartial social media analyst") also contributed positively, though less strongly, by helping to stabilise the model's point of view. The self-check instruction produced modest gains, particularly for GPT-4o, suggesting that adding a verification step at the end of the reasoning chain may help improve consistency in zero-shot outputs. Few-shot examples had only a limited effect on either model's performance, possibly due to the fact that they duplicated information already captured by the label explanation.

In contrast, the inclusion of sarcasm cues either had no measurable effect or led to minor declines in accuracy, especially for GPT-4o. This may indicate that abstract instructions such as "tweets may contain sarcasm" are too vague or cognitively disruptive in zero-shot settings, particularly when the models are already sensitive to subtle tone shifts.

When comparing the two LLMs, a divergence in optimal prompt complexity emerged. GPT-4o benefitted from a more elaborate configuration combining role, domain context, label explanation, and self-check components. DeepSeek-Chat, on the other hand, reached peak performance using a shorter prompt with just the role and label explanation components, suggesting a lower tolerance for complexity or greater efficiency in task alignment. This difference may reflect architectural or training disparities between the models, including how they handle multi-step reasoning, instruction parsing, or alignment tuning.

These findings offer several practical takeaways for prompt design in sentiment classification

tasks. First, clearly defined label criteria can be considered for the task of sentiment classification through prompt engineering, although in this work the effect is inflated due to the attitudinal nature of the ground-truth labels. Second, the optimal prompt configuration may vary by model, and prompt length or complexity should be tested systematically rather than assumed beneficial. Third, not all context cues are helpful, instructions like sarcasm detection may cause confusion rather than clarity.

Overall, prompt engineering is not merely a formatting exercise but a critical modelling decision. Careful component selection can improve accuracy and enhance consistency, making prompt-based approaches more viable for applied sentiment analysis in specialised domains like sustainability discourse.

### 7.2.3 Robustness to LLM prompt variations

Model robustness was evaluated by testing sensitivity to two controlled prompt perturbations: the removal of all punctuation and the injection of a deliberately misleading example. The results reveal that while both GPT-4o and DeepSeek-Chat exhibited some degree of robustness, stability varied substantially across conditions and models.

Quantitatively, GPT-4o showed minimal sensitivity to punctuation removal, with performance remaining virtually unchanged or slightly improving. This suggests that its decoding mechanism is relatively unaffected by formatting-level alterations. DeepSeek-Chat, however, experienced a more notable drop in performance under the same condition, indicating a greater dependence on prompt structure and formatting for semantic interpretation. Conversely, when a misleading example was added to the prompt, both models suffered a decline in performance, particularly in macro-F1 scores. This suggests a shared vulnerability to prompt-level contradictions, likely reflecting limitations in how the models process conflicting cues and in the clarity or consistency of the reasoning structure imposed by the prompt.

Qualitatively, the robustness analysis focused on misclassified tweets under the baseline prompt configuration. Several recurring error types were identified, including the misclassification of tweets with ambiguous or ironic tone, misinterpretation of quoted or sarcastic speech, and a tendency to over-classify tweets as negative when sentiment cues were subtle or contextually embedded. These failure modes suggest that even advanced LLMs struggle to disentangle sentiment from rhetorical complexity, particularly in cases where tweets rely on implicit framing or emotionally layered language. This is particularly relevant when the labels reflect stance, not sentiment, further complicating disambiguation of tone.

The differing sensitivity patterns across models likely reflect deeper architectural and training differences. GPT-4o's relative resilience to formatting changes may stem from more advanced in-context learning dynamics, whereas its stronger performance drop under semantic corruption could point to a stricter alignment model that penalises internal contradiction. DeepSeek-Chat, by contrast, may rely more heavily on linear prompt interpretation and exhibit less tolerance for deviations from expected structure.

From a practical standpoint, these findings emphasise that robustness is not a static property but a dynamic interaction between model, prompt, and task. Even when average performance is high, small changes in prompt formulation can meaningfully affect outcomes, especially in

domains where ambiguity and polarisation are common. Robustness testing should therefore be a standard component of any deployment pipeline involving prompt-based NLP, particularly when operating in high-stakes or trust-sensitive environments like sustainability communication, policy evaluation, or brand monitoring.

To summarise, prompt-based classification systems require not only accuracy under ideal conditions, but also resilience to plausible variations in input format and content. Evaluating and mitigating failure points under such perturbations is critical to ensuring the reliability and credibility of GenAI applications in real-world sentiment analysis.

### 7.2.4 Topic-Specific Variation

Model performance varied markedly across different sustainability-related topics, identified through Latent Dirichlet Allocation. These topic-level differences were consistent across all models and highlight the importance of semantic context in stance classification tasks interpreted as sentiment analysis. While the fine-tuned Twitter-RoBERTa model achieved the highest accuracy across all topics, the relative ordering of difficulty was similar for both generative and traditional models, suggesting that certain themes inherently pose greater challenges regardless of model architecture.

Specifically, tweets classified under topics such as "Urban and Environmental Impacts" were the easiest to classify, with all models performing well. These tweets likely tended to use direct and concrete language, often expressing clear support for climate action or referencing observable events such as heatwaves or pollution. In contrast, topics like "Global Warming Attribution" and "Political Denial and Polarisation" yielded the lowest performance. These topics likely featured more abstract, emotionally loaded, or rhetorically complex content, including climate scepticism, sarcasm, and ideological references. The linguistic ambiguity in such tweets, combined with framing effects and the presence of indirect or ironic cues, likely made sentiment more difficult to resolve even for advanced models.

Additionally, the distribution of sentiment across topics was often imbalanced, with a notable skew toward positive sentiment. For example, in the "Urban and Environmental Impacts" topic, over 70% of tweets were classified as positive, while negative sentiment was relatively rare. Similar patterns were observed in topics such as "Science Denial & Public Beliefs" and "Political Denial & Polarisation." This class imbalance may have influenced model performance by reducing the model's ability to correctly identify less frequent sentiment classes. Moreover, in topics like "Global Warming Attribution," where a more even distribution of sentiment classes was observed, the higher proportion of neutral and negative tweets may have made classification more challenging. Although this does not impact "zero-shot" models, Twitter-RoBERTa Fine-Tuned is potentially impacted by this fact.

These patterns underline the role of topical content in shaping classification outcomes. The fact that model performance is not uniform across topics implies that topic-aware evaluation is essential when applying sentiment analysis in complex domains like sustainability discourse. From a practical standpoint, this suggests that aggregate sentiment metrics may mask topic-level volatility or bias. Stakeholders using AI-based tools for monitoring public opinion should therefore consider stratifying results by topic or theme, especially when tracking sentiment over

time or comparing campaigns focused on different sustainability issues.

To summarise, topic-specific variation in classification performance reflects both the linguistic and ideological diversity of sustainability discourse. Recognising and accounting for this variability is critical for improving model reliability and for drawing accurate, actionable insights from sentiment analysis systems in real-world applications. Given the stance-based nature of the labels, organisations should exercise caution when interpreting these results as emotional sentiment.

## 7.3 Theoretical and Methodological Contributions

This work contributes to both theoretical and methodological discussions at the intersection of prompt engineering, sentiment analysis, Generative AI, and sustainability communication. Theoretical contributions emerge from how the findings engage with and refine prior conceptual frameworks, while methodological contributions lie in the design, evaluation, and practical application of prompt-based NLP techniques.

### 7.3.1 Theoretical Contributions

From a theoretical standpoint, this work reinforces and extends recent research on prompt engineering as a form of soft model programming (P. Liu et al., 2023; W. X. Zhao et al., 2023). The observed performance gains from label explanation, role specification, and self-check components support earlier claims that prompt structure can significantly influence model behaviour, particularly in zero- and few-shot scenarios (Webson & Pavlick, 2022; Wei et al., 2022). Moreover, this work adds to a growing body of literature showing that large language models do not merely complete text sequences but follow instructions in structured and often nuanced ways (Brown et al., 2020; Schick & Schütze, 2020). However, such outputs must be interpreted within the limitations imposed by the stance-based labels.

In sentiment analysis theory, this work contributes to understanding the limitations of lexicon-based and generic transformer models in high-variance social media contexts, especially when using stance as a sentiment proxy. The findings are in line with earlier work on the contextual and rhetorical complexity of sentiment-laden tweets (Rodríguez-Ibánez et al., 2023; Taboada et al., 2011), particularly within politicised or emotionally charged domains like sustainability (Chakriswaran et al., 2019; Yue et al., 2018). By showing how prompt-based models can match or exceed traditional baselines, this work supports calls for more adaptive and generalisable approaches to public opinion mining in digital spaces.

Regarding Generative AI, this work offers empirical evidence for the robustness and limitations of LLM-based classifiers when applied to domain-specific sentiment tasks. In line with findings from Krugmann and Hartmann (2024) and W. Zhang et al. (2023), the results demonstrate that generative models like DeepSeek-Chat and GPT-4o can perform competitively, especially in the area of zero-shot learning, but also reveal vulnerabilities tied to alignment, prompt contradictions, and topic-specific ambiguity. This work thus contributes to the broader reality check on LLM capabilities and the need for domain-aware evaluation protocols.

In the field of sustainability communication, this work operationalises sentiment analysis

not as a descriptive tool, but as a lens for evaluating how effectively AI models can interpret climate-related discourse (Anderson et al., 2024; Ballestar et al., 2020). By comparing model performance across topics such as political denial, environmental impact, and public beliefs, the findings shed light on how thematic complexity influences classification accuracy. This contributes to the empirical understanding of sustainability discourse in digital environments and complements existing literature on framing effects and public receptivity to environmental messaging (Lineman et al., 2015; Moser, 2010).

### 7.3.2 Methodological Contributions

Methodologically, this work's most notable innovation is the use of a modular prompt design framework to isolate the contribution of individual components. This structure enabled systematic comparisons and yielded clear insights into how design choices influence output accuracy and robustness. It builds on prompt engineering theory while offering a replicable procedure for future applications in applied NLP.

In addition, the benchmarking across diverse model families (rule-based, transformer-based, and generative AI) provides a holistic view of current capabilities in sentiment classification, helping to contextualise the role of GenAI within existing toolkits. Lastly, the topic-specific performance analysis adds a layer of interpretability by showing that classification difficulty is not evenly distributed across content domains, and may be further distorted by the label proxy mismatch between stance and sentiment, a consideration that is often overlooked in aggregate performance metrics.

Together, these contributions advance both theoretical insight and methodological practice in the application of large language models for sentiment analysis in socially relevant domains.

## 7.4 Managerial Contributions

This work offers several insights relevant to decision-makers in both the public and private sectors who are interested in applying Generative AI for monitoring public sentiment, with stance used as a proxy, across domains such as sustainability, policy, branding, or public communication. While primarily an academic work, the findings provide a foundation for practical applications in areas where understanding public opinion is essential for strategic messaging, stakeholder engagement, and reputational risk management.

First, the results demonstrate that prompt-based Generative AI models, particularly with the use of carefully designed prompts, can be used to classify public sentiment with reasonable accuracy, even in zero-shot or few-shot settings and with stance used as a proxy. For organisations lacking access to labelled datasets or training infrastructure, this offers a scalable and accessible alternative to traditional supervised models. In domains where communication is dynamic and context-dependent, such as sustainability or public health, this flexibility is especially valuable.

Second, this work highlights the importance of prompt design in shaping output quality. This has practical implications for organisations seeking to operationalise GenAI tools through APIs or custom workflows. Small changes in prompt phrasing, structure, or context framing

can significantly affect the model's reliability, making prompt engineering not just a technical detail but a strategic consideration in AI deployment.

Third, the benchmarking across model families reinforces the need for careful model selection. While fine-tuned models remain the gold standard when high-quality data is available, generative models like DeepSeek-Chat and GPT-4o offer a competitive alternative in resource-constrained environments. However, the differences in robustness and sensitivity to prompt perturbations suggest that operational use requires testing under realistic conditions, including ambiguous, ironic, or politically sensitive inputs.

Finally, the observed variation in classification performance across topics suggests that aggregate sentiment scores may obscure important nuances. For example, public sentiment may be easier to capture on concrete topics than on ideologically polarised ones. As a result, organisations should consider stratifying sentiment results by theme or subdomain to improve interpretability and avoid overgeneralisation in their analysis.

Overall, while this work does not offer a ready-made solution for real-time deployment, it does provide a roadmap for how organisations might responsibly integrate GenAI into sentiment analysis workflows. By taking a cautious, but proactive, approach (testing assumptions, monitoring reliability, and tailoring prompts to context) organisations can begin to extract meaningful insights from public discourse using GenAI tools. Managers should remain aware that these insights might reflect attitudinal orientation rather than emotional tone

## 7.5 Limitations

While this work provides valuable insights into the use of Generative AI for sentiment classification, several limitations should be acknowledged. These relate to the methodological design, data characteristics, and model-specific behaviour, and they should be taken into account when interpreting the results or considering real-world implementation.

### Methodological Limitations

One key limitation lies in the limited attention to the interpretability of GenAI model outputs. While the modular prompt design allowed for systematic testing of individual components, this work did not investigate how or why specific outputs were generated by the models, nor did it attempt to trace internal reasoning patterns or token-level attribution. As a result, insights into how models arrive at their sentiment classifications remain limited. Additionally, the evaluation of prompt robustness was restricted to a small number of perturbations. Broader adversarial testing or longitudinal stability assessments were beyond the scope of this work but remain relevant directions for future research.

### Data Limitations

A core limitation is that the dataset reflects stance, i.e. the position taken toward climate change issues, rather than affective sentiment. This means the models were effectively evaluated (and trained for Twitter-RoBERTa Fine-Tuned) on positional rather than emotional content, which complicates the interpretation of outputs as expressions of public feeling. This label mismatch

limits the conclusiveness of any findings regarding public sentiment, although the modelling approach remains valid and transferable to datasets with true sentiment labels.

Additionally, the dataset used in this work, while representative of sustainability discourse, was temporally bounded and limited to English-language tweets. This restricts the generalisability of findings across languages, platforms, or time periods. Moreover, the lack of user-level metadata such as demographics, location, or verified status limited the ability to analyse how sentiment dynamics or model performance may vary across user segments. Incorporating such contextual dimensions could yield deeper insights into model behaviour and practical deployment conditions.

A further limitation relates to the inclusion of Dataset A (Guzman, 2020), which, unlike the other datasets used, contains sentiment labels rather than stance labels. This dataset was incorporated because it was employed in prior sustainability-focused sentiment analysis research (Anderson et al., 2024), enabling methodological continuity and comparability. However, given that this work conceptually frames sentiment classification through the lens of stance detection, the inclusion of sentiment-labelled data introduces a degree of inconsistency in label interpretation. Although this discrepancy is partially mitigated by the small relative size of Dataset A within the merged dataset, future research would benefit from stricter dataset selection criteria to ensure alignment between the conceptualisation of sentiment as stance and the ground-truth labels used for model evaluation.

### Model-Specific Limitations

The use of commercial Generative AI models such as GPT-4o and DeepSeek-Chat introduces inherent limitations related to alignment, opacity, and platform-specific behaviour. For example, content filtering mechanisms may suppress or distort model responses on politically sensitive or controversial topics. Additionally, while hallucinations were not a central focus of this work, the potential for models to generate plausible yet inaccurate outputs remains a concern, especially given the ambiguity of interpreting stance as sentiment.

Additionally, zero-shot and few-shot prompting configurations used in this work represent only a subset of the possible ways to interact with LLMs. Future performance may differ significantly under fine-tuning or alternative prompting strategies, particularly when applied to sentiment-labelled rather than stance-labelled data.

These limitations do not undermine this work's contributions but rather frame them within a set of reasonable constraints. They also point toward several fruitful directions for future research, including the development of more rigorous evaluation benchmarks, integration of richer contextual data, and the testing of model behaviour across sentiment labelled data.

Finally, an often-overlooked methodological limitation concerns the environmental impact of using large-scale Generative AI models such as GPT-4o and DeepSeek-Chat. Running these models requires substantial computational resources, which in turn results in non-negligible carbon emissions. This raises an important ethical question about whether the use of resource-intensive AI technologies for sustainability research is itself sustainable. While this work did not quantify the carbon footprint associated with model inference, future research in this area could benefit from explicitly considering the environmental costs of AI model deployment and

exploring more energy-efficient alternatives or carbon offset strategies.

## 7.6 Suggestions for Future Research

Building on the findings and limitations of this work, several directions emerge for future research. These suggestions span methodological refinement, dataset expansion, and applied development, and they aim to deepen understanding of how Generative AI can be effectively and responsibly used for sentiment and stance analysis in dynamic, real-world domains.

First, future work should aim to disentangle stance and sentiment classification more explicitly. This includes applying prompt-based methods to datasets with validated affective sentiment labels and comparing results to stance-labelled corpora. Such benchmarking would clarify how reliably GenAI models generalise across related, but distinct, classification tasks.

Second, more comprehensive forms of prompt evaluation should be explored. While this work used a modular framework to test individual components, future work could expand into adversarial prompting, task framing effects, and instructional variation. Reasoning strategies such as chain-of-thought prompting may also provide deeper insight into how prompt structure influences model behaviour and robustness.

Third, expanding the diversity of datasets is crucial for generalisability. Future studies could incorporate multilingual or cross-platform content, examine public discourse over longer time frames, or analyse geographic and temporal variation in sentiment or stance. Incorporating user-level metadata, where ethical and legal, could further support fairness analysis and behavioural segmentation.

Fourth, future research could extend the application of GenAI sentiment tools into other socially charged domains such as public health, education policy, or electoral politics. These contexts offer fertile ground for testing how models handle emotionally and ideologically complex content, especially when the sentiment-stance boundary is blurred.

Fifth, interpretability and transparency should receive more attention. Techniques such as explanation generation, or chain-of-thought output, may offer insight into model reasoning, enhancing auditability in high-stakes settings. This is particularly important when classification decisions influence policy, reputation, or public trust.

Sixth, the integration of GenAI tools into hybrid workflows presents an important research direction. Combining automated classification with human-in-the-loop validation, rule-based checks, or statistical smoothing could improve both reliability and accountability. This is especially relevant in cases where models operate on stance proxies but are used to inform sentiment-sensitive decisions.

Finally, the rapid pace of Large Language Model development raises new methodological questions. Emerging models such as GPT-o3 and DeepSeek-Reasoner offer improved efficiency and reasoning capabilities, which could be particularly valuable for sustainability research where computational cost and interpretability are critical concerns. Investigating whether these models can deliver better or more sustainable performance than current state-of-the-art models is a key priority. At the same time, this accelerating innovation cycle highlights a broader challenge: the development of models is moving faster than academic research can systematically assess their

capabilities and limitations. This underscores the need for more agile evaluation frameworks that can keep pace with technological change.

Together, these research directions recognise the potential of Generative AI for social data analysis, while also acknowledging the need for conceptual clarity, methodological rigour, environmental sustainability, and critical reflection in its deployment and interpretation.

# Chapter 8

# Conclusion

This work set out to investigate how Generative AI, through prompt engineering, can be effectively applied to classify sentiment (with stance used as a proxy in this work) in sustainability-related tweets and to evaluate model performance relative to traditional sentiment analysis methods. Building on recent developments in the field of natural language processing, this work aimed to assess not only predictive accuracy but also the robustness of prompt designs and the feasibility of deploying large language models (LLMs) in applied settings.

The results demonstrate that prompt-based Generative AI models, particularly DeepSeek-Chat and GPT-4o, can perform sentiment classification with moderate to high accuracy in zero-shot or few-shot configurations. While these models do not yet outperform fine-tuned transformer-based classifiers such as Twitter-RoBERTa, they significantly exceed the performance of lexicon-based and non-fine-tuned baselines like VADER. This is partly due to VADER's reliance on affective lexicons, which are poorly aligned with stance-based ground-truth labels. This reinforces the growing consensus that Generative AI models represent a credible and accessible alternative in scenarios where labelled data or infrastructure for training is limited (Krugmann & Hartmann, 2024).

A key contribution of this work lies in its systematic evaluation of prompt design. The modular prompt engineering framework revealed that specific components, especially explicit label definitions and task framing, substantially affect model accuracy. The Label Explanation module, for example, consistently produced large improvements in performance across both LLMs tested. These findings underscore that performance is not only a function of model architecture, but also of how tasks are communicated to the model. Furthermore, prompt robustness tests showed that model outputs are sensitive to misleading inputs but relatively stable under structural changes such as punctuation removal. This highlights the importance of prompt verification when deploying LLMs for stance-based sentiment proxies, particularly in high-stakes or trust-sensitive environments.

Although interpretability and event-based sentiment tracking were excluded from scope, the comparative evaluation across thematic segments via Latent Dirichlet Allocation (LDA) revealed further insights. Sentiment classification accuracy varied across sustainability topics, with the highest performance observed on concrete, context-specific themes such as urban and environmental impacts. Topics involving political framing or climate attribution, by contrast, exhibited greater misclassification, suggesting that LLMs, while capable, remain challenged by rhetorical

nuance, sarcasm, and emotionally charged discourse, particularly when such expressions do not clearly indicate a stance label.

For practitioners, this work offers several implications. First, the ease of use of Generative AI models enables organisations without extensive machine learning expertise to gain insight into public discourse around sustainability. Second, the findings demonstrate that prompt design is not trivial: business and policy users must invest in prompt testing and robustness checks to ensure consistent results. Third, the topic-specific variation in performance suggests that sentiment analysis systems should not rely on aggregate scores alone. Stratified analysis by topic or theme is essential to uncover sentiment dynamics relevant for strategic decision-making.

To summarise, this work shows that Generative AI, when guided by well-crafted prompts, offers a promising and scalable method for sentiment analysis of sustainability discourse on social media. However, future work using data annotated with sentiment is needed to validate the transferability of these findings beyond stance-based settings. While traditional supervised models still yield superior accuracy, the flexibility, accessibility, and surprisingly robust performance of LLMs support their integration into future sentiment monitoring frameworks. By foregrounding prompt robustness and ease of deployment, this work contributes to the practical advancement of AI-based tools for the domain of sustainability communication strategy.

# Bibliography


Anderson, T., Sarkar, S., & Kelley, R. (2024). Analyzing public sentiment on sustainability: A comprehensive review and application of sentiment analysis techniques. *Natural Language Processing Journal*, 100097.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences*, *11*(11), 5088.

Ballestar, M. T., Cuerdo-Mir, M., & Freire-Rubio, M. T. (2020). The concept of sustainability on social media: A social listening approach. *Sustainability*, *12*(5), 2122.

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Plato, J., Bag, R., & Hassanien, A. E. (2020). Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, *97*, 106754–106754. https://api.semanticscholar.org/CorpusID:221979535

Chakriswaran, P., Vincent, D. R., Srinivasan, K., Sharma, V., Chang, C.-Y., & Reina, D. G. (2019). Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues. *Applied Sciences*, *9*(24), 5462.

Corner, A., Markowitz, E., & Pidgeon, N. (2014). Public engagement with climate change: The role of human values. *Wiley interdisciplinary reviews: climate change*, *5*(3), 411–422.

de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. d. L. (2020). Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, *32*, 1–12.

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California management review*, *54*(1), 64–87.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:52967399

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, *7*. https://api.semanticscholar.org/CorpusID:248530058

Ghahramani, M., Galle, N. J., Ratti, C., & Pilla, F. (2021). Tales of a city: Sentiment analysis of urban green space in dublin. *Cities*, *119*, 103395.

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

Guzman, E. (2020). Climate sentiment in twitter. https://www.kaggle.com/datasets/joseguzman/climate-sentiment-in-twitter

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*. https://api.semanticscholar.org/CorpusID:12233345

Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169–15211. https://api.semanticscholar.org/CorpusID:6973845

Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, *11*(1), 3.

Leippold, M., Stammbach, D., Webersinke, N., Bingler, J. A., & Kraus, M. (2023). A dataset for detecting real-world environmental claims.

Lineman, M., Do, Y., Kim, J., & Joo, G. (2015). Talking about climate change and global warming. *PloS One*, *10*(9), e0138996.

Liu, M., Luo, X., & Lu, W.-Z. (2023). Public perceptions of environmental, social, and governance (esg) based on social media data: Evidence from china. *Journal of Cleaner Production*, *387*, 135840.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & 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.

Lyon, T. P., & Montgomery, A. W. (2015). The means and end of greenwash. *Organization & environment*, *28*(2), 223–249.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 1–66.

Moser, S. C. (2010). Communicating climate change: History, challenges, process and future directions. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(1), 31–53.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). Sentiment analysis on social media. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 919–926. https://api.semanticscholar.org/CorpusID:15620885

Nisbet, M. C. (2009). Communicating climate change: Why frames matter for public engagement. *Environment: Science and policy for sustainable development*, *51*(2), 12–23.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, *2*, 1–135. https://api.semanticscholar.org/CorpusID:207178694

Qian, E. (2019). Twitter climate change sentiment dataset. https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset/data

Radi, S. A., & Shokouhyar, S. (2021). Toward consumer perception of cellphones sustainability: A social media analytics. *Sustainable Production and Consumption*, *25*, 217–233.

Reyes-Menendez, A., Saura, J. R., & Alvarez-Alonso, C. (2018). Understanding #worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach. *International Journal of Environmental Research and Public Health*, *15*(11), 2537.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P.-M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, *223*, 119862.

Schick, T., & Schütze, H. (2020). It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427–437.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. D., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*, 267–307. https://api.semanticscholar.org/CorpusID:3181362

Weber, E. U. (2010). What shapes perceptions of climate change? *Wiley Interdisciplinary Reviews: Climate Change*, *1*(3), 332–342.

Webson, A., & Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2300–2344.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2018). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*, 617–663. https://api.semanticscholar.org/CorpusID:57513433

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*. https://api.semanticscholar.org/CorpusID:10694510

Zhang, W., Deng, Y., Liu, B.-Q., Pan, S. J., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *NAACL-HLT*. https://api.semanticscholar.org/CorpusID:258866189

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-r. (2023). A survey of large language models. *ArXiv*, *abs/2303.18223*. https://api.semanticscholar.org/CorpusID:257900969

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *International conference on machine learning*, 12697–12706.

# Appendix A

# Prompt Design and Component Overview

This appendix documents the modular prompt engineering approach used for the generative language models GPT-4o and DeepSeek-Chat. It includes an overview of the modular components, the full text of the best-performing prompts, and the prompt variants used for robustness testing.

## A.1 Prompt Components

The base prompt follows an instruction–input format, with six optional modular components added to vary structure and informational content. These components are:

- **Role Instruction:** *"You are an impartial social-media analyst."*
- **Domain Context:** *"Tweets discuss climate change, climate action, or sustainability."*
- **Label Explanation:**

  *- positive: if the tweet supports climate action and sustainability, expresses concern about climate change, or affirms its reality.*
  *- negative: if the tweet denies, mocks, criticises, or downplays climate change, climate action, or sustainability.*
  *- neutral: if the tweet does not clearly express a stance or is purely factual.*

- **Few-shot Examples:**

  *Tweet: "Climate change is real and we need to act now to save the planet."*
  *Sentiment: positive*
  *Tweet: "Global warming is a hoax created to control us."*
  *Sentiment: negative*
  *Tweet: "The IPCC released its latest climate report today."*
  *Sentiment: neutral*

- **Sarcasm Trigger:** *"Tweets may contain sarcasm."*

- **Self-Check:** *"Verify label before replying."*

## A.2 Best-Performing Prompts

The prompt that yielded the highest accuracy for each model on the tuning set (1,000 tweets) is provided below.

### GPT-4o Best Prompt

```
You are an impartial social-media analyst.
Tweets discuss climate change, climate action, or sustainability.


- positive:  if the tweet supports climate action and sustainability, expresses
concern about climate change, or affirms its reality.
- negative:  if the tweet denies, mocks, criticises, or downplays climate change,
climate action, or sustainability.
- neutral:  if the tweet does not clearly express a stance or is purely factual.


Classify the sentiment of the following tweet as 'positive', 'neutral', or 'negative'.
Tweet:  "{tweet}"
Sentiment:
Verify label before replying.
```

### DeepSeek-Chat Best Prompt

```
You are an impartial social-media analyst.


- positive:  if the tweet supports climate action and sustainability, expresses
concern about climate change, or affirms its reality.
- negative:  if the tweet denies, mocks, criticises, or downplays climate change,
climate action, or sustainability.
- neutral:  if the tweet does not clearly express a stance or is purely factual.


Classify the sentiment of the following tweet as 'positive', 'neutral', or 'negative'.
Tweet:  "{tweet}"
Sentiment:
```

## A.3 Robustness Prompt Variants

For robustness testing, two modified prompts were evaluated for each model: one with all punctuation removed, and one with a deliberately misleading example inserted.

### GPT-4o – No Punctuation

```
you are an impartial social-media analyst tweets discuss climate change climate
action or sustainability - positive if the tweet supports climate action and sustainability
expresses concern about climate change or affirms its reality - negative if the
```

```
tweet denies mocks criticises or downplays climate change climate action or sustainability
- neutral if the tweet does not clearly express a stance or is purely factual
classify the sentiment of the following tweet as 'positive' 'neutral' or 'negative'
tweet tweet sentiment verify label before replying
```

### GPT-4o – False Example Inserted

```
You are an impartial social-media analyst.
Tweets discuss climate change, climate action, or sustainability.


- positive:  if the tweet supports climate action and sustainability, expresses
concern about climate change, or affirms its reality.
- negative:  if the tweet denies, mocks, criticises, or downplays climate change,
climate action, or sustainability.
- neutral:  if the tweet does not clearly express a stance or is purely factual.


Tweet:  "Global warming is a hoax created to control us."
Sentiment:  positive


Classify the sentiment of the following tweet as 'positive', 'neutral', or 'negative'.
Tweet:  tweet
Sentiment:
Verify label before replying.
```

### DeepSeek-Chat – No Punctuation

```
you are an impartial social-media analyst - positive if the tweet supports climate
action and sustainability expresses concern about climate change or affirms its
reality - negative if the tweet denies mocks criticises or downplays climate change
climate action or sustainability - neutral if the tweet does not clearly express
a stance or is purely factual classify the sentiment of the following tweet as
'positive' 'neutral' or 'negative' tweet tweet sentiment
```

### DeepSeek-Chat – False Example Inserted

```
You are an impartial social-media analyst.


- positive:  if the tweet supports climate action and sustainability, expresses
concern about climate change, or affirms its reality.
- negative:  if the tweet denies, mocks, criticises, or downplays climate change,
climate action, or sustainability.
- neutral:  if the tweet does not clearly express a stance or is purely factual.


Tweet:  "Global warming is a hoax created to control us."
Sentiment:  positive
```

```
Classify the sentiment of the following tweet as 'positive', 'neutral', or 'negative'.
Tweet:  tweet
Sentiment:
```

# Appendix B

# Accuracy and Statistical Tables and Figures

This appendix reports detailed accuracy results and statistical analyses referenced in the Results chapter. It includes the full table of accuracy scores for all prompt variants, group-level means for each prompt component, results from independent t-tests, and a component-level delta ranking based on average change in model performance.

## B.1 Prompt Variant Accuracy Scores

Table B.1: Accuracy Comparison Across 64 Prompt Variants (Renamed, Reordered)

| **Prompt Variant** | **GPT-4o** | **DeepSeek-Chat** |
|---|---|---|
| Base | 0.301 | 0.392 |
| Domain | 0.341 | 0.396 |
| Domain + Few-shot | 0.405 | 0.406 |
| Domain + Few-shot + Sarcasm | 0.352 | 0.414 |
| Domain + Few-shot + Sarcasm + Self-check | 0.403 | 0.424 |
| Domain + Few-shot + Self-check | 0.388 | 0.406 |
| Domain + Label Explanation | 0.571 | 0.590 |
| Domain + Label Exp. + Few-shot | 0.572 | 0.590 |
| Domain + Label Exp. + Few-shot + Sarcasm | 0.518 | 0.581 |
| Domain + Label Exp. + Few-shot + Sarcasm + Self-check | 0.576 | 0.597 |
| Domain + Label Exp. + Few-shot + Self-check | 0.603 | 0.611 |
| Domain + Label Exp. + Sarcasm | 0.544 | 0.579 |
| Domain + Label Exp. + Sarcasm + Self-check | 0.586 | 0.571 |

*(continued on next page)*

*(continued from previous page)*

| **Prompt Variant** | **GPT-4o** | **DeepSeek-Chat** |
|---|---|---|
| Domain + Label Exp. + Self-check | 0.600 | 0.566 |
| Domain + Sarcasm | 0.342 | 0.431 |
| Domain + Sarcasm + Self-check | 0.356 | 0.405 |
| Domain + Self-check | 0.348 | 0.379 |
| Few-shot | 0.387 | 0.393 |
| Few-shot + Sarcasm | 0.344 | 0.407 |
| Few-shot + Sarcasm + Self-check | 0.380 | 0.423 |
| Few-shot + Self-check | 0.382 | 0.422 |
| Label Explanation | 0.538 | 0.581 |
| Label Exp. + Few-shot | 0.523 | 0.601 |
| Label Exp. + Few-shot + Sarcasm | 0.484 | 0.578 |
| Label Exp. + Few-shot + Sarcasm + Self-check | 0.540 | 0.571 |
| Label Exp. + Few-shot + Self-check | 0.561 | 0.602 |
| Label Exp. + Sarcasm | 0.491 | 0.597 |
| Label Exp. + Sarcasm + Self-check | 0.550 | 0.597 |
| Label Exp. + Self-check | 0.544 | 0.595 |
| Role | 0.314 | 0.409 |
| Role + Domain | 0.345 | 0.408 |
| Role + Domain + Few-shot | 0.396 | 0.404 |
| Role + Domain + Few-shot + Sarcasm | 0.344 | 0.415 |
| Role + Domain + Few-shot + Sarcasm + Self-check | 0.402 | 0.426 |
| Role + Domain + Few-shot + Self-check | 0.394 | 0.417 |
| Role + Domain + Label Exp. | 0.576 | 0.599 |
| Role + Domain + Label Exp. + Few-shot | 0.558 | 0.604 |
| Role + Domain + Label Exp. + Few-shot + Sarcasm | 0.512 | 0.595 |
| Role + Domain + Label Exp. + Few-shot + Sarcasm + Self-check | 0.599 | 0.602 |
| Role + Domain + Label Exp. + Few-shot + Self-check | 0.599 | 0.624 |
| Role + Domain + Label Exp. + Sarcasm | 0.552 | 0.616 |
| Role + Domain + Label Exp. + Sarcasm + Self-check | 0.602 | 0.603 |
| Role + Domain + Label Exp. + Self-check | **0.605** | 0.591 |
| Role + Domain + Sarcasm | 0.345 | 0.449 |
| Role + Domain + Sarcasm + Self-check | 0.363 | 0.418 |

*(continued on next page)*

*(continued from previous page)*

| **Prompt Variant** | **GPT-4o** | **DeepSeek-Chat** |
|---|---|---|
| Role + Domain + Self-check | 0.356 | 0.380 |
| Role + Few-shot | 0.403 | 0.404 |
| Role + Few-shot + Sarcasm | 0.347 | 0.414 |
| Role + Few-shot + Sarcasm + Self-check | 0.389 | 0.405 |
| Role + Few-shot + Self-check | 0.391 | 0.426 |
| Role + Label Exp. | 0.559 | **0.640** |
| Role + Label Exp. + Few-shot | 0.560 | 0.638 |
| Role + Label Exp. + Few-shot + Sarcasm | 0.507 | 0.608 |
| Role + Label Exp. + Few-shot + Sarcasm + Self-check | 0.604 | 0.619 |
| Role + Label Exp. + Few-shot + Self-check | 0.601 | 0.639 |
| Role + Label Exp. + Sarcasm | 0.545 | 0.619 |
| Role + Label Exp. + Sarcasm + Self-check | 0.601 | 0.604 |
| Role + Label Exp. + Self-check | 0.578 | 0.628 |
| Role + Sarcasm | 0.285 | 0.405 |
| Role + Sarcasm + Self-check | 0.340 | 0.390 |
| Role + Self-check | 0.331 | 0.408 |
| Sarcasm | 0.259 | 0.388 |
| Sarcasm + Self-check | 0.344 | 0.389 |
| Self-check | 0.355 | 0.381 |

## B.2 Group Means by Component

Table B.2: Mean Accuracy per Component Presence (GPT-4o and DeepSeek-Chat)

| Component | Component Value | GPT-4o | DeepSeek-Chat |
|---|---|---|---|
| Role | 0 | 0.453 | 0.496 |
| Role | 1 | 0.466 | 0.513 |
| Domain | 0 | 0.448 | 0.505 |
| Domain | 1 | 0.470 | 0.503 |
| Label Explanation | 0 | 0.357 | 0.407 |
| Label Explanation | 1 | 0.561 | 0.601 |
| Few Shot | 0 | 0.449 | 0.500 |
| Few Shot | 1 | 0.469 | 0.508 |
| Sarcasm | 0 | 0.468 | 0.504 |
| Sarcasm | 1 | 0.450 | 0.504 |
| Self-check | 0 | 0.441 | 0.505 |
| Self-check | 1 | 0.477 | 0.504 |

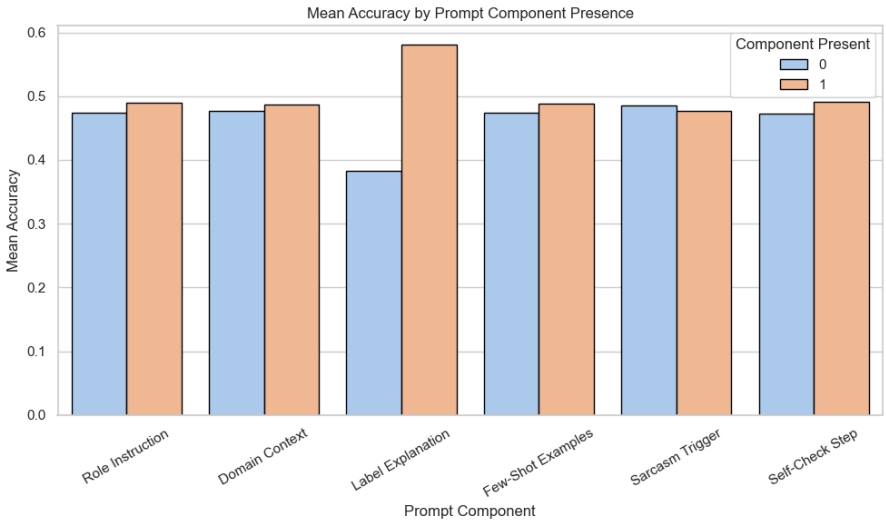


Figure B.1: Mean Accuracy by Prompt Component Presence (0 = Absent, 1 = Present)

## B.4 T-test Results

Table B.3: T-test Results for Component Effects on Accuracy

| Component | Model | t-Statistic | p-Value |
|---|---|---|---|
| Role | GPT-4o | 0.475 | 0.637 |
| Role | DeepSeek-Chat | 0.682 | 0.498 |
| Domain | GPT-4o | 0.821 | 0.415 |
| Domain | DeepSeek-Chat | -0.095 | 0.925 |
| Label Explanation | GPT-4o | 23.183 | $< .001$ |
| Label Explanation | DeepSeek-Chat | 43.276 | $< .001$ |
| Few Shot | GPT-4o | 0.754 | 0.454 |
| Few Shot | DeepSeek-Chat | 0.328 | 0.744 |
| Sarcasm | GPT-4o | -0.664 | 0.509 |
| Sarcasm | DeepSeek-Chat | 0.012 | 0.990 |
| Self-check | GPT-4o | 1.333 | 0.187 |
| Self-check | DeepSeek-Chat | -0.040 | 0.968 |

## B.5 Delta Accuracy Ranking

Table B.4: Change in Accuracy by Component Presence

| Component | $\Delta$ Accuracy GPT-4o | $\Delta$ Accuracy DeepSeek-Chat |
|---|---|---|
| Role | 0.013 | 0.017 |
| Domain | 0.022 | -0.002 |
| Label Explanation | 0.204 | 0.194 |
| Few Shot | 0.021 | 0.008 |
| Sarcasm | -0.018 | 0.000 |
| Self-check | 0.036 | -0.001 |

## B.6 LDA Topic Selection Plots

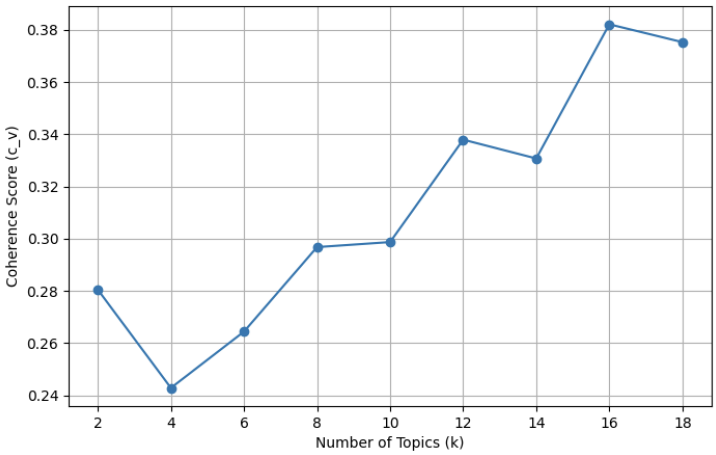


Figure B.2: Topic Coherence Across Values of $k$

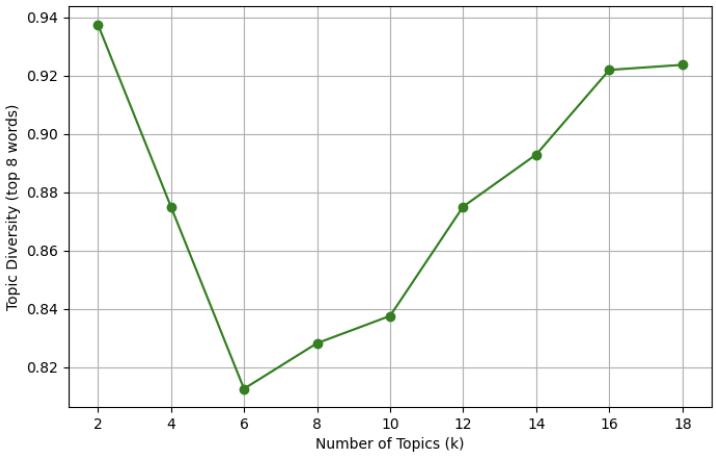


Figure B.3: Topic Diversity Across Values of $k$

# Appendix C

# Implementation Notes and Code Access

This appendix outlines the technical setup and tools used to implement the experiments in this work.

## C.1 Programming Environment

All modelling and analysis are conducted in **Python**, using the following key libraries:

- `transformers` (HuggingFace) — for working with RoBERTa models
- `openai` — for interfacing with GPT-4o via API
- `scikit-learn` — for model evaluation and metrics
- `gensim` — for topic modelling with LDA
- `spacy`, `wordcloud`, `matplotlib` — for NLP preprocessing and visualisation

## C.2 Generative Model Access

The prompt-based models (GPT-4o and DeepSeek-Chat) are accessed via their respective APIs:

- **GPT-4o** (`gpt-4o-2024-08-06`): Accessed using the OpenAI Python API (`gpt-4o` engine)
- **DeepSeek-Chat** (`DeepSeek-V3-0324`): Accessed via the DeepSeek developer API

API keys and rate limits are managed locally and not hard-coded into any repository.

## C.3 Code Availability

All code and datasets used in this work are available at the following GitHub repository (Will be included in final submission):

`https://github.com/your-username/your-repo`

The repository contains a clear documentation, folder structure, and instructions for reproducing the results.