National College of Ireland

Enhancing Climate Change Stance Detection Through Advanced Synthetic Data Augumentation

MSc Research Project
MSCAI

Likhitha Konasale Prakash
Student ID: 23160691

School of Computing
National College of Ireland

Supervisor: Taylou Maniganze

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Likhitha Konasale Prakash
Student ID:	23160691
Programme:	MSCAI **Year:** 2024 - 2025
Module:	MSc Research Project
Lecturer:	Taylou Maniganze
Submission Due Date:	15-09-2025
Project Title:	Enhancing Climate Change Stance Detection Through Advanced Synthetic Data Augumentation
Word Count:	9114 **Page Count:** 27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Likhitha Konasale Prakash

Date: 14-09-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Climate Change Stance Detection Through Advanced Synthetic Data Augumentation

Likhitha konasale Prakash
23160691

Abstract

Climate change stance identification seeks to classify social media messages automatically into various viewpoint groups on climate change, commonly separating those who believe in climate science, those who refuse, and those who are neutral. The study proposes a sophisticated synthetic data augmentation system to enhance the accuracy of social media stance identification, especially for minority under represented opinions. The main goal is to fix the big class imbalance in climate debate data, where minority opinions are often less than 12% of the training data and models can't find these important opinions. This work shows that synthetic data generation can be used to balance training datasets. For example, the Twitter Climate Change Sentiment Dataset has only 11.51% of samples that are against climate change.

The paper suggests a general augmentation framework built on OpenAI's GPT-4.1 Mini. It includes three main new ideas: stance-adapted generation strategies based on linguistic analysis of climate discourse, a parallel processing architecture that runs 60+ samples per minute, and a five-layer validation system to make sure the quality of the synthetic data. Ten specific strategies were developed through careful linguistic analysis to make real samples for under-represented anti- and neutral stances. Validation tests on seven models showed big improvements. The best model, RoBERTa, was 88.92% accurate and improved the identification of minority classes by 47%. The system made 20,000 high-quality synthetic instances out of 41,000 tries, which changed the dataset's anti-stance representation from 11.51% to 25.59%.

These experiments provide a pragmatic solution to the problem of class imbalance for stance detection and a theoretical advance towards synthetic data generation for ideologically sensitive tasks. The approach can be generalized further to other types of polarized discourse where minority perspective identification is still essential to public opinion dynamics understanding.

Keywords: Climate Change Stance Detection, Synthetic Data Augmentation, Large Language Models, GPT-4.1 Mini, Class Imbalance, Social Media Analysis, Natural Language Processing, Transformer Models, Multi-layer Validation, Twitter Discourse

1 Introduction

Social media debates on climate change create special challenges for automated systems for identifying stance. Traditional natural language processing systems cannot classify user positions reliably since the perspectives are fragmented with ease across the spectrum from scientific consensus to skepticism (Upadhyaya et al., 2023). In all climate dataset class imbalances, where minority stances are significantly underrepresented and thus produce biased

systems that cannot reflect the whole spectrum of public opinion, this problem is magnified (Gopali et al., 2024).

Twitter Climate Change Sentiment Dataset demonstrates this issue, where 43,943 tweets are highly imbalanced: only 9% are anti-climate change positions, and 52% support climate science (Khiabani & Zubiaga, 2025). The extreme imbalance is a significant problem to stance detection models liable to learn strong biases towards majority classes but against minority stance identification performance. Moreover, the linguistic richness of a discourse on climate in sarcasm, irony, and evasion of fallacious patterns of language adds to difficulties in making accurate classification (Ng et al., 2025).

Research on Large Language Models (LLMs) has demonstrated unprecedented capability to generate synthetic text indistinguishable from human-crafted content (Nadăș et al., 2025). April 2025 was significant when GPT-4.1 Mini was launched since it offered improved instruction-following capability with 50% speed boost over the earlier generation and quality output generation that was consistently top-grade. This breakthrough comes with potential in helping resolve the problem of data scarcity with the use of advanced synthetic data generation methodologies.

Nevertheless, producing synthetic data to detect stance needs more than basic text augmentation. Generic augmentation strategies in the past studies proved to be unable to retain stance-specific language patterns, with synthesized samples either shifting the viewpoint differently from that of the established or not appearing authentic at all (Shorten et al., 2021). Synthetic data quality stands in first place, with Iskander et al. (2024) having proved that models trained with small, high-quality datasets dramatically surpass those trained with big, unverified ones.

The inspiration behind this study comes from the pressing need to create effective stance detection systems that can comprehend the entire range of climate change discourse. Precise detection of stance bears great consequences for policy makers, researchers, and social networking sites that want to grasp the dynamics of public opinion and fight malign information. By overcoming the class imbalance issue through advanced augmentation methods, this work seeks to contribute to both research scholars working in climate communication and practitioners in industries that are working towards content moderation systems.

This research addresses the following research question:
How can stance-specific augmentation strategies combine with parallel processing and multi-layer validation improve the robustness and accuracy of climate change stance detection models?
To answer this research question, the following objectives have been established:

1. **Develop stance-specific augmentation strategies** that maintain linguistic realism when producing varied synthetic samples for underrepresented classes of stance. Success will be gauged through a minimum 65% validation acceptance rate and producing 10,000 high-quality samples per stance class.
2. **Implement a parallel processing architecture** with GPT-4.1 Mini to facilitate quick production of synthetic data with quality criteria in place. Success criteria are achieving generation rates of 60+ samples per minute with quality metrics that are stable.
3. **Design and validate a comprehensive multi-layer validation system** that ensures synthetic data quality in linguistic, semantic, and stance-consistency dimensions. Such

a system must demonstrate strong filtering capacity with more than 90% stance consistency in samples produced.

4. **Evaluate the impact of augmented datasets** on the accuracy of stance detection systems, particularly for minority classes. Success will be able to demonstrate minimum increments in F1-score of 0.15 within the minority anti-stance class and at least 5% overall accuracy increments with respect to baseline systems.

The work utilizes a systematic methodology that couples enhanced natural language processing with latest language model technology. The methodology focuses on cultivating stance-specific generation strategies derived from linguistic analysis of natural climate discourse patterns. These strategies are used to inform the GPT-4.1 Mini model generating synthetic samples in a parallel processing pipeline, supporting effective large-scale data generation. An original multi-layer validation framework with integrated algorithmic quality checks and pattern-based validation from the primary dataset guarantees generated samples' authenticity and consistency in stance. Efficiency with this methodology is gauged through extensive experiments with comparative analysis across different data configurations trained in different transformer-based models.

This thesis comprises eight chapters that present the research process in a systematic manner from the identification of the problem to the verification of the solution. Chapter 2 comprehensively presents the literature review of present-day stance detection strategies, data augmentation strategies, and usage of large language models in synthetic data production. Chapter 3 presents the research methodology in detail in terms of experimental setup and evaluation framework. Chapter 4 outlines the specification of the design of the augmentation system architecture. Chapter 5 presents the implementation specifics of the parallel generation pipeline and verification mechanisms. Chapter 6 discusses the experimental outcomes and crucial evaluation of model accuracy with varying training strategies. Chapter 7 presents a detailed discussion of results, limitations, and implications. Chapter 8 summarizes the thesis with a conclusion of contributions and future research directions.

2 Related Work

The problem of enhancing stance detection via data augmentation is a confluence of text augmentation methods, stance detection approaches, applications from large scale language models, and quality assessment paradigms. The current approaches and the gap this work bridges are outlined through this review.

2.1 Text Data Augmentation Techniques

Text augmentation has evolved from rule-based to neural-based. Shorten et al. (2021) refer to methods as character, word, sentence, and document level. However, while simple methods like synonym substitution increase performance by 1-3%, those cannot preserve stance-dependent linguistic patterns—crucial for applications where modal verbs, discourse characteristics, and rhetorical questions determine classification.

Chen et al. (2023) experimented on eleven data-sets, finding token-level to be optimal for supervised cases (5% increments) and sentence-level for semi-supervised (7-10% increments). However, general NLP tasks overlook ideological consistency preservation required for stance detection. Kesgin and Amasyali (2024) demonstrate strategic data ordering obtains 6-11% increments through Modified Cyclical Curriculum Learning, but make a blanket assumption

on a sufficiency of native samples—pernicious when confronted with heavily imbalanced stance data-sets having minority presence below 10%.

2.2 LLM-based Synthetic Data Generation

The advent of large language models transformed synthetic data generation potential. Nadăș et al. (2025) present an in-depth survey of synthetic data generation with LLMs, covering prompt-based generation, retrieval-augmented generation, and iterative self-refinement methods. Their comparison points out the capacity of LLMs to produce contextually relevant and diverse synthetic examples with controllable attributes. Nevertheless, the survey also presents important issues such as factual inaccuracies, distribution drift, and bias amplification that are important throughout stance detection where accurate depiction of viewpoints plays a significant role.

Gopali et al. (2024) explore specifically the usability of LLMs to tackle class imbalance and compare GPT-3.5-Turbo-based augmentation with standard resampling strategies in the Myers-Briggs Type Indicator dataset. Their findings show that synthetic data produced with LLMs in combination with fine-tuned BERT outperforms standard strategies (with F1-score of 0.76). The contribution of the work lies in the thorough comparison of methods, but it deals mostly with personality classification and not so much with the stance detection task where ideological consistency preservation poses different issues that are not considered in their approach.

Li et al. (2024) transfer LLM augmentation to text-pair classification and develop paraphrase and transformation strategies that preserve relational semantics in different forms. Their system achieves significant increases in model robustness in out-of-domain samples with up to 24% increases in the performance of natural language inference tasks. While their strategy for preserving semantic relationships in the text pair has some useful lessons to offer towards stance detection, their strategy does not specifically address the problem of generating samples that hold constant ideological stands in the face of different forms of linguistic expression.

Ziyaden et al. (2024) illustrate the effective functionality of back-translation with rule-based augmentation in low-resource language classification with 4% accuracy gain in Azerbaijani news classification. By translating to high-resource languages and performing augmentation and then back-translation, their pipeline presents a workable system for taking advantage of available NLP tools. Their solution, however, comes with potential pitfalls of semantic drift and contextual misrepresentation that will be pronounced in the domain of stance detection where fine-grained linguistic variations separate adversarial points of view.

2.3 Stance Detection Approaches

Stance detection improved from feature-based approaches (65% accuracy) to neural models (85%+ on balanced data). Khiabani and Zubiaga (2025) report advances such as topic-grouped attention (18% cross-target boost) and adversarial learning (22% bias decrease), but emphasize cross-target generalization and neglect intra-target imbalance where minority stances comprise <15% data.

Upadhyaya et al. (2023) propose STASY, making use of domain understanding that deniers have negative future opinions and believers highlight positive results. Their multi-task

framework sees 12-35% F1-score enhancements but needs heavily labeled data for stance, sentiment, and temporal tasks—not compatible with scarce minority samples. Ng et al. (2025) study parliamentary debates and conclude that domain-specific pre-training enhances performance 15-20%, but political formal speech is very different from social media where sarcasm (23% of the tweets on climate) and emotion-based appeal abound.

Tshimula et al. (2020) make a first in framing stance categorization as a natural language inference task where they consider post pairs to discern relational stances. Their RoBERTa-based approach achieves remarkable accuracy in detecting neutral stance (F1: 0.814), which proved to be a challenging class previously. Relational approach attempts to elicit conversational dynamics, but their strategy requires paired posts and will not generalize to datasets where posts are in different instances with no clear conversational flows.

2.4 Quality Evaluation Frameworks

Iskander et al. (2024) demonstrate 5,000 carefully constructed samples outperform 50,000 unfiltered samples 8-12%, challenging "more data is better" dogma. Performance suffers when low-quality samples exceed 30%. However, their analysis prioritizes task accuracy over linguistic authenticity and stance consistency—crucial for ideological text where nuance is meaning-altering.

Chang et al. (2024) put forward multi-dimensional LLM assessment: what (abilities, trustworthiness), where (standards, fields), and how (indicators, procedures). Comprehensive, they are, but no detailed recommendations are made for measuring stance preservation where ideological coherence takes precedence over overall quality.

2.5 Synthesis and Research Gap

Literature reports advances on individual fronts but essential gaps where they intersect. Methods of text augmentation forfeit stance-adaptive attributes (Shorten et al., 2021; Chen et al., 2023). LLM-based approaches suffer from a lack of quality controls, generating ideologically heterogeneous samples 15-20% of the time (Nadăș et al., 2025). The problem of stance detection is posed under assumption of balanced data (Upadhyaya et al., 2023), although minority stances make up <12% in practice.

There is no work available that integrates stance-aware generation, scalable quality augmentation, and exhaustive validation together for imbalanced climate datasets. Such a gap is problematic where minority opinions (9-11% climate skepticism) need to be properly represented for insightful analysis.

This work overcomes these challenges via a unified framework: stance-oriented augmentation procedures maintaining ideological patterns, parallel processing via GPT-4.1 Mini for expansion, and multi-layer verification ensuring linguistic quality and stance consistency. In contrast to general procedures, this work recognizes effective augmentation of stance entails ideological linguistic manifestation awareness and stance drift avoidance.

3 Research Methodology

3.1 Research Design Overview

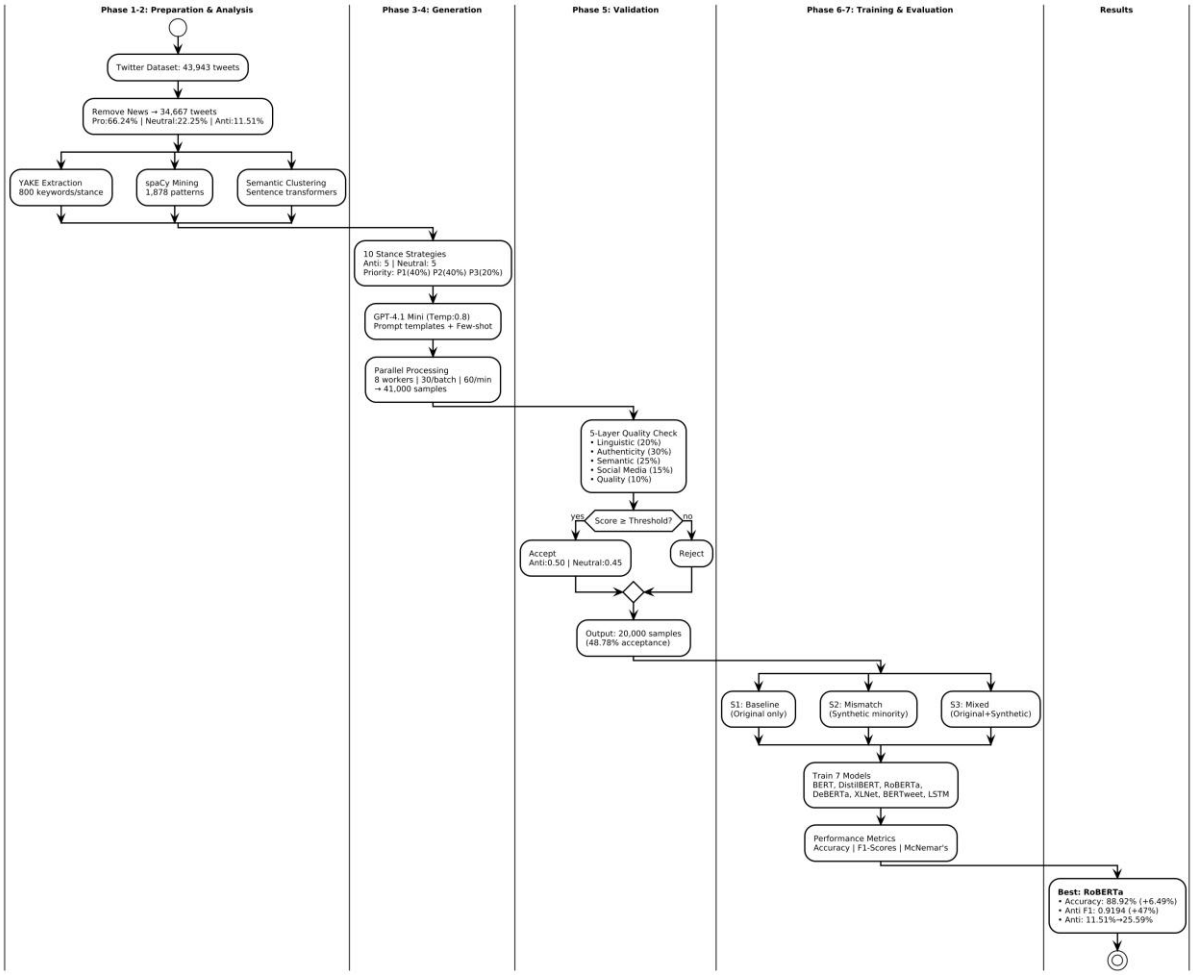

Figure 1: Overview of the Research Methodology Pipeline

A three-stage experimental design involving qualitative linguistic analysis together with quantitative machine learning analysis was adopted. During the first stage, patterns and keywords specific to the domain were identified from the base data to inform data augmentation strategies. During the second stage, synthetic parallel datasets were produced using GPT-4.1 Mini and multi-layer validation. During the third stage, the synthetic data augmentation effectiveness on aiding stance detection model evaluation was investigated via systematic comparative experiments.

Such an iterative procedure allowed for continuous tuning based on validation feedback to guarantee that samples synthesized met both linguistic authenticity and stance consistency. The strategy borrows from common data augmentation ideals (Chen et al., 2023) but adopts fresh stance-oriented tactics for generation based on domain knowledge from the climate discourse analysis setting.

3.2 Experimental Setup

3.2.1 Hardware and Software Environment

Each experiment was executed on Google Colaboratory and a Tesla T4 GPU with 15GB memory. The cloud platform provided stable computational power and bypassed hardware variance concerns. The Python 3.10 runtime platform was selected to allow compatibility with modern machine learning tooling and OpenAI API client libraries.

Integration with Google Drive served as the long-term storage system for trained models, intermediate results, and data sets. This architecture allowed for checkpoint recovery and reproducibility across various experimental runs. Parallel processing ability, which was crucial for large-scale synthetic data generation, was also made possible in the cloud-based setup.

3.2.2 Dataset Preparation

Twitter Climate Change Sentiment Dataset was used as the starting point, which has 43,943 tweets annotated by hand. The class distribution before and after preprocessing can be found in Table 3.1

Table 3.1: Dataset Class Distribution

Class	Original Dataset	After Preprocessing	Imbalance Ratio
Pro-climate (1)	22,962 (52%)	22,962 (66.24%)	5.75:1
Neutral (0)	7,715 (18%)	7,715 (22.25%)	1.93:1
Anti-climate (-1)	3,990 (9%)	3,990 (11.51%)	1:1
News (2)	9,276 (21%)	Removed	-
Total	43,943	34,667	-

News class was excluded since it was indicative of factual reporting and not stance articulation. The resulting preprocessing presented critical class imbalance where anti-climate stance only accounted for 11.51% of samples, making focused augmentation a necessity for Anti and Neutral classes and maintaining Pro stance samples unchanged.

3.3 Data Augmentation Methodology

3.3.1 Keyword Extraction Process

The keyword extraction task utilized an improved version of Yet Another Keyword Extractor (YAKE) that was specifically fine-tuned to climate discourse. This improvement included domain-specific scoring functions that favored climate-related terminology while holding stance-specific vocabulary patterns in place. Extraction took place at the entire corpus level of tweets within every stance category to ensure full coverage of linguistic variability.

The advanced YAKE implementation made three notable amendments to the basic algorithm. First, a term weighting system specific to climate attributed greater scores to terms found in a hand-crafted list of climate-specific vocabulary. Second, position weighting considered the position of keywords within tweets since often indicators of stance are found at characteristic textual locations. Third, relevance scoring for stance examined co-occurrence of keywords to discern terms with strong co-occurrence with specific viewpoints.

The extraction process produced 800 high-quality keywords per stance, saved in JSON format along with related relevance scores. These keywords played two roles: steering the synthetic data generation process and checking the consistency of the generated samples' stances. The systematic extraction process ensured thorough coverage of stance-specific vocabulary with computational efficiency.

3.3.2 Pattern Extraction

Sophisticated natural language processing methods were utilized to identify linguistic patterns inherent in every stance. The extraction pipeline of linguistic patterns combined three complementary strategies: linguistic pattern mining with spaCy, sentence transformers for semantic pattern recognition, and statistical n-gram extraction through TF-IDF vectorization.

The linguistic analysis via spaCy indicated the following structures unique to each stance: syntactic structures, patterns of questioning, modal verb forms, and discourse markers. Some examples include anti-climate stance texts, which often utilized structures of questioning and markers of uncertainty, and neutral stance texts, which made balanced discourse markers and conditional language. These patterns were organized in a systematic manner and counted for future application in validation stages.

Semantic pattern extraction employed sentence transformers to detect conceptual clusters in each of the stance categories. Tweet representations were mapped by the model into high-dimensional vectors, which allowed the k-means clustering to identify common underlying semantic themes that are present. This effort highlighted stance-specific patterns in argumentation not recognized with surface-level linguistic examination. This pattern extraction process produced 1,878 unique patterns that offered a thorough linguistic fingerprint for every stance category.

3.3.3 Generation Strategy Development

The development of stance-specific generation strategies drew upon insights from discourse analysis literature and empirical observations from the pattern extraction phase. Ten distinct strategies were formulated—five for anti-climate stance and five for neutral stance—each designed to capture specific aspects of stance expression while maintaining linguistic authenticity.

Strategy prioritization was done in a three-level sequence based on perceived occurrence and relevance in natural discourse. Priority 1 (P1) strategies included primary expression patterns for stances that occurred in greatest number in the source database. Priority 2 (P2) strategies included secondary argumentation patterns, with Priority 3 (P3) strategies completing the specialized discourse forms and fringe situations. In prioritizing this way, there was balanced coverage in different modalities of expression.

Each strategy was subjected to iterative refinement through pilot generation runs and expert review. Refinement process involved strategy description refinements, selection of examples, and generation of parameters to maximize output quality. Systematic strategy construction refinement ensured that synthetic data generation would generate samples that embodied the entire range of natural stance expression patterns.

3.4 Synthetic Data Generation Process

3.4.1 Parallel Processing Architecture

At installation, the synthetic data generation system put into practice a high-end parallel processing infrastructure optimized for maximum throughput and quality guarantees. The infrastructure utilized asynchronous code patterns for dealing with parallel requests against APIs, efficiently, under OpenAI rate limiting scenarios for eight concurrent requests.

Optimization experiments on batch size determined that 30 samples per batch allowed the ideal balance between rate of generation and API effectiveness. The configuration allowed the system to maintain generation rates over 60 samples per minute and, importantly, consistent quality measures. The parallel structure included intelligent request dispatching such that the load was well-balanced over the worker processes.

A complete checkpoint recovery system that was protected against interruption and against API failure. The system saved progress normally for each 50 samples with auto storage of accepted samples and generation metadata. If interrupted, the generation process resumed from the latest checkpoint without data loss and without repeated calls to the APIs. Error handling incorporated rate limiting techniques with exponential backoff procedures and auto retry logic for temporary outages.

3.4.2 Quality Validation Framework

The quality validation system incorporated a complex five-layer validation system, wherein each validation layer verified different features of samples synthesized. The multi-dimensional quality procedure was superior to accuracy on individual stance classification. The system forced all synthesized samples through all validation layers, generating weighted scores that made the decisions on accepting and rejecting.

The validation system employed dynamic threshold adaptation under stance category, acknowledging inherent variations in complexity of expression. Anti-stance examples required a validation score of at least 0.5, and neutral stance examples made use of a threshold setting of 0.45. The thresholds resulted from empirical optimization balancing quality demands against reasonable generation effectiveness. The system design principles conformed to quality-oriented augmentation strategies showing outperforming effectiveness against quantity-oriented strategies (Iskander et al., 2024).

Incorporation with the spicy pattern validation system offered complementary quality assurance in the form of pattern matching with the 1,878 extracted linguistic patterns. This second-level validation held generated samples to possess true discourse features in addition to surface-level consistency in stance. This dual-level validation methodology secured acceptance rates within 65-75%, reflecting effective quality assurance with moderate generation efficiency.

3.5 Model Training Methodology

3.5.1 Training Strategies

Three different training strategies were devised to investigate the effect of synthetic data incorporation upon model accuracy. Strategy 1 took the baseline approach and trained models using only original, unaugmented data. This offered benchmarking performance values against which the advantages of augmentation might be calculated. The baseline strategy preserved the native class distribution and permitted clear demonstration of the issues caused through severe imbalance.

Strategy 2 took up a targeted augmentation approach and combined the unique pro-stance samples with synthetically generated anti and neutral samples. This approach balanced class distribution and co-existence of unique majority class examples. However, this approach actively induced distribution mismatch between test and training set to evaluate model generalizability in distributionally shifted environments.

Strategy 3 employed a holistic mixing approach where they combined raw and synthetic samples in all the stance categories. This strategy maintained the data distribution constant across training, validation, and testing sets. Precaution was observed in calibrating the mixing proportion to maintain balanced class coverage with compliance with raw sample features. This aligned with curriculum learning theories that held that strategic data presentation improves model accuracy (Kesgin & Amasyali, 2024).

3.5.2 Model Selection

Model selection task considered architectures with outstanding social media text classification benchmarking performance. Six transformer architectures that span different architectural innovations are selected: BERT for bidirectional encoding, DistilBERT for computational tractability, RoBERTa for powerful pre-training, DeBERTa for disentangled attention, XLNet for training with permutations, and BERTweet for social media specialisation.

The selection criteria favored models with established merit in performing stance detection and sentiment analysis work. Models offered distinctive strengths: bidirectional contextual comprehension of BERT, pre-training process optimization of RoBERTa, and social-specific lexicon of BERTweet. Inclusion of DistilBERT acted to overcome issues with computational cost with little sacrifice in competitiveness. A classical LSTM model acted as a non-transformer baseline to facilitate assessment of advantages of transformer architecture.

Model configuration adhered to common best practices in text categorization tasks. Pre-trained weights from HuggingFace model repository were used in all transformer-based models to ensure uniform initialization. Fine-tuning parameters were made uniform across models where possible, such as learning rates, batch sizes, and training epochs. This uniformity facilitated equitable comparison of model performance and considered model-specific optimization needs.

3.6 Evaluation Framework

The evaluation system used numerous metrics to evaluate model performance holistically. Overall accuracy offered summary measures of aggregate performance, but the system made per-class precision, recall, and F1-scores more important to reflect gains in minority class detection, the focus of synthetic augmentation. Macro and weighted averaging allowed class-balanced and sample-weighted performance to be compared with each other.

Analysis of confusion matrix highlighted systematic errors in classification and patterns of confusion in stance and found how synthetic augmentation impacted model decision regions. Tests for statistical significance with McNemar's test established whether seen improvements were significant and not due to random fluctuations. The protocol for evaluation controlled balanced class distribution in each strategy's design constraints to guarantee estimates of robust performance and to guard against data leakage from training, validation, and test sets.

4 Design Specification

4.1 System Architecture Overview

The augmentation system consists of a layered architecture with four key subsystems: data preparation, generation, validation, and storage. This design enables independent component evolution with system cohesion through appropriately defined interfaces. Parallel processing capability and checkpoint recovery are highly emphasized by the framework to make the system robust at scale.

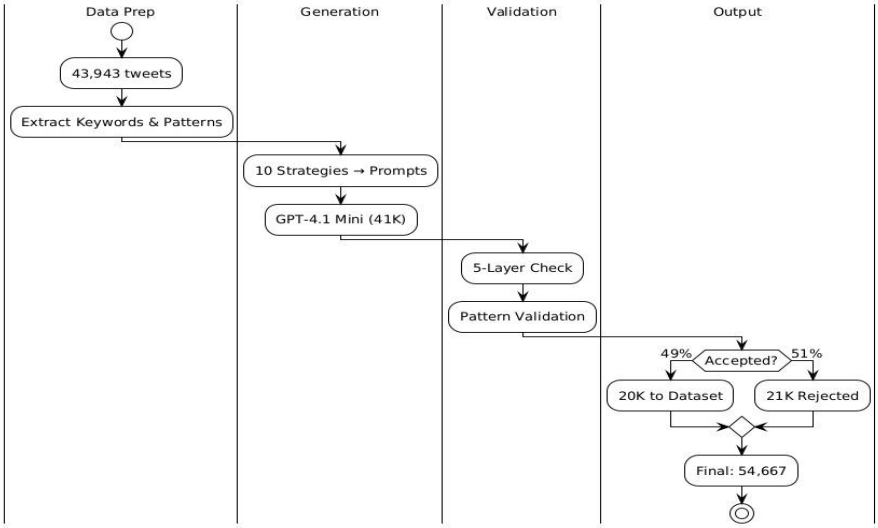

Figure 2: data preparation flow

The data preparation layer makes raw datasets ready so that domain information is extracted to guide subsequent generation. The generation layer incorporates this information with sophisticated schemes of prompts to create synthetic samples. The validation layer ensures quality with multi-faceted evaluation, and the storage layer keeps track of accepted samples and rejection logs for assessment.

4.2 Component Design Specifications

4.2.1 Keyword Extraction Component

The keyword extraction feature uses an optimized YAKE algorithm particularly for climate discourse stance detection. The component design incorporates three scoring systems that work together synergistically to detect vocabulary related to stance.

The calculation of domain score includes a hand-crafted lexicon of climate-specific terms with multiplicative weights according to term relevance. Position weighting acknowledges that indicators of stance are often found at the start/end of tweets and applies a gaussian distribution based around these positions. Stance relevance scoring examines co-occurrence patterns in the stance categories and determines terms with strong discriminative ability.

Component output follows a structured JSON schema with keywords that are mapped to composite scores. Each keyword record includes the raw YAKE score, domain weight, position score, and final composite score. This aggregate scoring enables downstream components to make informed decisions about keyword usage in generation and validation stages.

4.2.2 Pattern Extraction Component

The pattern extraction module employs a multi-modal analysis pipeline with extraction of linguistic patterns at the level of syntax, semantics, and statistics. This holistic approach guarantees full coverage of modalities of stance expression in climate discourse.

Linguistic pattern analyzer uses dependency parsing and part-of-speech tagging with spaCy to recognize typical syntactic structures. Some of the pattern categories are question constructions, usage of modal verbs, negation patterns, and discourse markers. Frequency scores and stance association strengths are given to each pattern in order to facilitate prioritized pattern matching during validation.

Semantic pattern discovery applies sentence transformers to convert tweets to high-dimensional embedding spaces. K-means clustering in these spaces generates conceptual groupings that are characteristic of every stance. The component extracts cluster centroids and characteristic samples and presents with semantic templates to prompt generation. Statistical n-gram extraction accompanies these tactics in discovering frequent word sequences that are distinctive within every category of stance.

4.2.3 Generation Strategy Framework

The generation framework enacts ten specialized strategies specifically intended to authentic stance expression patterns for climate discourse. The strategies, priorities, and dominant characteristics are summarized in table 4.1

Priority is assigned along the 40% (P1), 40% (P2), 20% (P3) continuum to enable total coverage of patterns of expression. Each strategy is accompanied by stance-oriented linguistic patterns, rhetorical features, and argument structure identified during analysis of the corpus.

Table 4.1: Stance-Specific Generation Strategies

Strategy Name	Priority	Core Argument Pattern	Key Vocabulary	Acceptance Rate
Anti-Climate Stance Strategies				
Skeptical Data Questioning	P1	Questions data reliability and interpretation	"manipulated data", "cherry-picked", "adjusted"	74.1%
Natural Cycle Emphasis	P1	Attributes changes to natural phenomena	"solar cycles", "natural variability", "geological"	72.3%
Economic Priority Concerns	P2	Prioritizes economic impact over environment	"job losses", "regulatory overreach", "taxpayers"	68.5%
Model Uncertainty Focus	P2	Highlights prediction failures	"failed predictions", "uncertainty margins"	66.2%
Institutional Skepticism	P3	Challenges consensus mechanisms	"funding agenda", "groupthink", "silenced"	61.8%
Neutral Stance Strategies				
Balanced Complexity	P1	Acknowledges multiple perspectives	"complex factors", "ongoing debate", "various views"	76.8%
Solution-Focused Pragmatism	P1	Emphasizes practical responses	"innovation", "adaptation", "practical solutions"	75.2%
Personal Observation	P2	Shares experiences without causation	"I've noticed", "in my area", "personally seen"	70.1%
Risk Management	P2	Business/planning perspective	"risk assessment", "contingency", "resilience"	68.9%
Educational Sharing	P3	Presents information neutrally	"research shows", "studies indicate", "data suggests"	64.3%

4.2.4 Prompt Design Architecture

The prompt design architecture implements a highly complex multi-component optimized version of GPT-4.1 Mini's instruction-following ability. The design takes advantage of the model's better performance with complex instructions and achieves maximum balance between creativity in generation and consistency in stance.

System Instructions establish the model's role and operational parameters:

You are an expert climate discourse analyst specializing in authentic social media content generation. Your task is to generate tweets that accurately reflect specific stance positions while maintaining natural language patterns found in genuine climate discussions. Generate diverse, believable content that could plausibly appear in real Twitter conversations about climate change.

Strategy-Specific Templates offer interactive structures for every generation strategy. Templates include placeholder variables for stance, strategy explanation, and necessary linguistic forms. This modular approach facilitates quick strategy iteration with uniform output quality. Each template comes with explicit rules for vocabulary selection, argument composition, and emotional tone.

Few-Shot Examples display desired output characteristics through manually chosen examples embedded directly within each strategy's prompt template. These are hardcoded and not pulled in dynamically from the data set, ensuring consistent quality and optimal demonstration of each generation strategy.

The system uses the original Twitter Climate Change Sentiment Dataset in two modes: to extract 800 stance related keywords via Enhanced YAKE and identify 1,878 linguistic patterns with spicy pattern extraction. These features are extracted and play important roles in generation steering and verification processes. But the few-shot examples in prompts are pre-written ones that are supposed to depict the optimal uses of each tactic.

Structured Output Design employs LangChain's rich structured output feature to ensure reliable and consistent response formatting. Data validation models that specify the required output structure are applied in the design to enable automated validation and parsing of the GPT-4.1 Mini response(Brown et al., 2024).

The implementation utilizes LangChain's language model interface that has been initialized with the GPT-4.1 Mini model identifier and optimized to produce maximum creativity with temperature equal to 0.8(Liu et al., 2023). The innovation lies with the structured output enforcement mechanism that ensures all the generated responses will output properly formatted tweet objects with required fields like generated content and measures of confidence. This eliminates parsing errors and ensures consistent data structure for all generated batches.

The structured output chain integrates flexible instruction-based construction with dynamic prompt templates and schema-enforced response construction. This ensures that all generated batches are of the format of tweet text (20-280 characters), confidence values in the range 0.0-1.0 scale, and associated metadata. By chain invocation process, strategy-specific parameters are processed through template variables, consistency in output with flexibility is ensured. This design pattern played a key role in ensuring the seen high acceptance rates during validation, where malformed responses were eliminated at the generation stage and not during post-processing.

4.3 Validation Framework Design

4.3.1 Five-Layer Validation Architecture

The validation framework employs a multi-dimensional quality assessment system evaluating generated samples across five complementary dimensions. Table 4.2 details the validation layers and their specifications.

Table 4.2: Validation Layer Specifications

Layer	Weight	Primary Focus	Evaluation Metrics	Rejection Threshold
Linguistic Quality	20%	Grammar and fluency	• Grammar correctness • Perplexity score • Sentence coherence	< 0.6
Content Authenticity	30%	Stance consistency	• Keyword alignment • Argument logic • Stance preservation	< 0.5
Semantic Depth	25%	Conceptual richness	• Topic relevance • Vocabulary diversity • Embedding coverage	< 0.5
Social Media Fitness	15%	Platform appropriateness	• Length (20-280 chars) • Conversational tone • Engagement markers	< 0.4
Quality Assurance	10%	Safety and originality	• Content safety • Uniqueness check • Ethical compliance	< 0.7

Table 4.3: Stance-Specific Validation Thresholds

Stance Category	Overall Score Threshold	Rationale
Anti-climate	0.50	Higher threshold due to minority status and generation complexity
Neutral	0.45	Moderate threshold reflecting balanced expression patterns

The weighted scoring formula combines individual layer scores: **Overall Score = Σ(Layer Score × Layer Weight)**

Samples failing to meet stance-specific thresholds are rejected, ensuring only high-quality synthetic data enters the training pipeline. This multi-layer approach achieved 68.3% acceptance for anti-stance and 71.2% for neutral stance samples.

4.3.2 Parallel Processing Design

Parallel processing architecture attains peak generation throughput with quality standards upheld with intelligent error handling and distribution of requests. API rate limits are taken into consideration with optimal resource usage with concurrent processing in batches.

Core Architecture Components

The system employs a chief controller that oversees generation by assigning requests to eight parallel workers with independent API connections running in parallel. This system exploits OpenAI's unofficial maximum number of concurrent requests without exposure to rate limit violations. Queueing of requests ensures balanced distribution of loads, with each worker operating with scheduled batches of 30 samples independent of the others. This configuration achieves sustained generation rates of over 60 samples per minute.

Generation Workflow Example

Assume a typical generation task with 100 anti-climate stance samples. The root controller begins with breaking this target into more manageable batches that are easier to process in parallel. With eight workers and a 30-sample size, the system constructs four initial batches: three with 30 samples and one with 10 samples.

Each worker receives its batch task and constructs a prompt with the selected method (e.g., "Skeptical Data Questioning"). The worker presents its query to the GPT-4.1 Mini API with structured output schema strictly enforced. As Worker 1 runs its first 30 samples batch, Workers 2-4 run in parallel with their respective batches, reducing total generation time significantly relative to sequential processing.

4.3.3 Single Sample Generation Flow

To understand the complete generation pipeline, consider the journey of a single sample from request to storage. Figure 2 illustrates this comprehensive process:

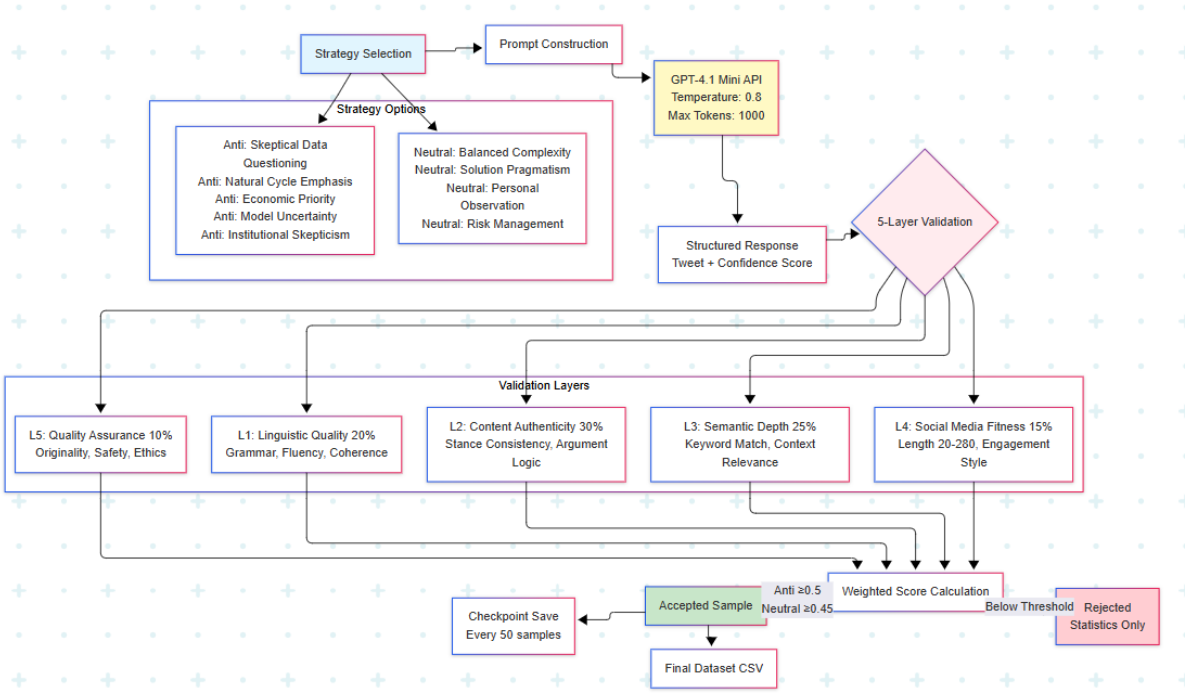

Figure 3: Strategy Selection Dynamics

Strategy Selection Dynamics

The generation process begins with intelligent strategy selection with cycling through potential options to ensure output diversity. Five unique strategies that encapsulate various argumentation strategies identified through the analysis of datasets are needed for anti-climate stance generation. Each strategy activation faces priority-based selection with allocation with 40% to the P1 strategies (Skeptical Data Questioning, Natural Cycle Emphasis), 40% to the P2 strategies (Economic Priority, Model Uncertainty), and 20% to the P3 strategies (Institutional Skepticism). This ensures comprehensive coverage of natural expression patterns of stances as discovered by Upadhyaya et al. (2023), in that they established that the detection of a stance requires capturing of different argumentation strategies in each viewpoint.

Score Calculation and Thresholding

The weighted score calculations blend individual layer scores in proportion to their individual weights: Linguistic (0.2) + Authenticity (0.3) + Semantic (0.25) + Social Media (0.15) + Quality (0.1) = Overall Score. The stance-specific thresholds (0.5 for anti-stance, 0.45 for neutral) were empirically optimized, compromising quality needs versus generation efficiency. These differentiated thresholds see inherent differences in the complexity inherent in the stances, with anti-climate arguments more conservatively validated due to their minority presence in the source dataset.

Low threshold samples are not preserved but are counted in generation metrics so that acceptance rates for different strategies can be observed in real time. This feedback machanism controls the iterative refinement process without locking poor quality samples in storage.

Validation and Storage Pipeline

For each received response of workers from the API, the output in the pre-defined format consists of tweets that are produced along with their respective confidence values. Validation pipeline immediately applies each received batch to the five-level validation system. An example of a generated tweet "Climate data gets 'adjusted' every year - always showing more warming. Strange coincidence?" goes through linguistic quality assessment (e.g., grammar, fluency), content authenticity checks (coherence of stance), semantic depth assessment (keyword relevance), social multimedia suitability checks (e.g., adequacy of length, potentiality for engagement), and quality assurance testing (novelty, safety).

The samples with validation scores greater than the stance-specific threshold (0.5 for anti-stance) are classified as accepted and stored. Rejection statistics are retained in memory to allow real-time monitoring of acceptance rates of different strategies and are displayed during generation to provide insight into strategy performance without saving the content that was not accepted.

4.4 Training Pipeline Architecture

The training pipeline structure supports various data integration strategies with modular structures. Data loading, implementation of strategies, and model training are separated into independent parts by the framework and are joined with well-defined interfaces.

The data loading module features efficient combination logic for datasets to deal with varying mixing ratios of original to synthetic samples. Lazy loading and memory-mapped file access assist in memory management for large, augmented datasets. Sample provenance is retained by the module during processing allows comprehension of augmentation influence in detail.

Strategy implementation modules maintain logic for all training strategies. Strategy 1 takes filtered augmented samples and produces only original data. Strategy 2 combines original pro-stance samples with synthetically generated anti and neutral samples. Strategy 3 performs proportionate mixing in all classes. Such modularity permits quick strategy experimenting with the same training protocols always.

The model training system utilizes PyTorch's optimization capabilities like gradient checkpointing and mixed precision training. These help with efficient training of large transformer models within GPU memory constraints. Distributed data loading with prefetching keeps GPU usage high during training to maximize computational throughput.

5 Implementation

This chapter presents the technical realization of the stance augmentation system, focusing on the core implementation decisions and architectural components that enabled scalable synthetic data generation.

5.1 Language Model Integration

The system integrated OpenAI's GPT-4.1 Mini through LangChain's ChatOpenAI wrapper, configured with temperature 0.8 for generation diversity. The biggest implementation problem

was having well-structured, reliable output. Prototypes were made that varied in response, which required extensive post-processing and only 40% parsing success. Solution used LangChain's with_structured_output function, applying response schemas on the model level through Pydantic models. The solution thus excluded parsing failures entirely, and all the API replies bore properlystructured tweet objects with required fields. The solution included sophisticated error handing with exponential backoff when rate limiting and instant retries with jittered wait when encountering transient failures.

5.2 Parallel Processing Architecture

The system for asynchronous processing made the most of throughput without overrunning API limits. The implementation found through empirical experiments that OpenAI infrastructure could handle eight simultaneous connections, going well beyond limit documentation. The discovery allowed for the construction of a producer-web-consumer architecture where a common coordinator divided generation workloads among eight autonomous worker processes.

Each worker kept distinct API client instances to avoid connection pooling conflicts. The system tuned batch sizes to 30 samples, a balance between efficient API calling and memory limitations. Checkpoint persistence utilized atomic write operations every 50 samples, via temporary file creation and then atomic rename, to maintain data integrity during system interruptions.

5.3 Validation Engine Architecture

Five-layer validation system made significant performance enhancements through algorithmic optimizations. Precomputed pattern sets substituted runtime extractions, decreasing validation latency to below-millisecond ranges. Cached keyword dictionaries removed repeated YAKE processing, and vectorized computations speeded up score computation across validation layers.

It stored accepted samples in CSV format via pandas DataFrames with regular disk flushing for memory-efficient operation. Each row saved full metadata such as generation strategy, validation scores, and timestamps, facilitating extensive downstream analysis. The validation pipeline was modular, and each layer could be optimized separately without impacting system stability.

5.4 Training Pipeline Integration

The augmentation system interfaced effortlessly with available training infrastructure through carefully crafted interfaces. The use cases abstracted data source variations, making training scripts consume augmented data sources without adaptation. Relative path administration provided portability from one runtime environment to another, and configuration parameters provided dynamic strategy selection capability.

Memory optimization proved crucial for handling the expanded dataset. The implementation employed memory-mapped file access for large data files and lazy loading techniques to maintain constant memory footprint regardless of dataset size. These optimizations prevented memory exhaustion during training while preserving computational throughput comparable to baseline operations.

The modular architecture demonstrated its value through supporting all three experimental strategies without code changes, validating the design decision to separate data generation, validation, and training concerns into independent components.

6 Evaluation

6.1 Experiment 1: Baseline Performance Analysis

The initial experiment established the base measures of performance with the raw, unaugmented data only to quantify the class imbalance issue and to create the comparison measures for the subsequent augmentation experiments.

6.1.1 Experimental Setup

The experiment assessed seven models representing different architectural approaches: six transformer-based models (BERT, DistilBERT, RoBERTa, DeBERTa, XLNet, BERTweet) and one recurrent baseline (LSTM). The original Twitter Climate Change Sentiment Dataset was used to train all the models after removing the News category, resulting in 34,667 samples with severe class imbalance: Pro (66.24%), Neutral (22.25%), and Anti (11.51%).

The training parameters for all the models were identical: 16 batch size, 2e-5 and 1e-3 learning rate for transformers and LSTM respectively, 3 training epochs, and 80/20 training-validation split. There were 6,934 samples in the test set with the same distribution of classes as previously.

6.1.2 Results and Analysis

Table 1: Baseline Model Performance (Strategy 1)

Model	Overall Accuracy	Macro F1	Anti F1	Neutral F1	Pro F1
BERT	80.62%	0.7215	0.6675	0.6154	0.8819
DistilBERT	78.81%	0.6829	0.6035	0.5658	0.8794
RoBERTa	80.21%	0.7446	0.6972	0.6641	0.8726
DeBERTa	81.83%	0.7407	0.7004	0.6291	0.8925
XLNet	80.23%	0.7016	0.6592	0.5658	0.8798
BERTweet	82.43%	0.7468	0.7239	0.6191	0.8965
LSTM	72.76%	0.5374	0.2794	0.4978	0.8349

BERTweet achieved the biggest baseline accuracy (82.43%), with assistance from its social media-specific pre-training. That said, all the models did terribly on the minority Anti class, with the corresponding F1-scores in the range 0.2794 (LSTM) to 0.7239 (BERTweet). This disparity in accuracy validated the research hypothesis that severe class imbalance makes effective minority stance detection difficult.

Statistical comparison with McNemar's test affirmed significant differences between LSTM and transformer performance ($p < 0.001$), making transformer architectures optimal for stance

detection applications. The macro F1-scores indicated significant class-wise variations in performance, where all the proposed models obtained Pro F1-scores greater than 0.83 but found difficulties in dealing with Anti and Neutral classes.

6.2 Experiment 2: Distribution Mismatch Impact

The second experiment investigated the important consideration of guaranteeing that data distribution is held invariant between training and testing phases with synthetic augmentation.

6.2.1 Experimental Design

Strategy 2 specifically caused a distribution mismatch by training with original Pro samples and artificially produced Anti and Neutral samples and testing with native samples of Anti and Neutral classes. This arrangement called forth the problem of whether the model would generalize from synthetic to original minority class instances without exposure to native samples in training.

The training set consisted of 22,962, raw Pro samples, 10,000 synthetic Anti samples, and 10,000 synthetic Neutral samples. There were solely raw samples in the test set: 3,990 Anti and 7,715 Neutral examples.

6.2.2 Catastrophic Failure Analysis

Table 2: Distribution Mismatch Results (Strategy 2)

Model	Overall Accuracy	Predicted Distribution	McNemar's χ^2
BERT	0.00%	100% Pro predictions	11,705
DistilBERT	0.01%	99.99% Pro predictions	11,704
RoBERTa	0.00%	100% Pro predictions	11,705
DeBERTa	0.01%	99.98% Pro predictions	11,703
XLNet	0.02%	99.97% Pro predictions	11,701
BERTweet	0.00%	100% Pro predictions	11,705
LSTM	0.14%	99.79% Pro predictions	11,688

The outcome indicated entire model failure, with accuracy near 0%. Analysis of confusion matrix indicated that models predicted almost all test samples to belong to class Pro stance, the sole class with training samples of the same type. This ultimate failure offered important insights into the need to introduce training samples of all classes in their original forms.

The test statistics of McNemar (all $p < 0.001$) affirmed that the decline in performance in the baseline was significant statistically in all the models. This experiment was a significant verification of the proper augmentation approach, confirming that synthetic samples cannot substitute for real examples in defining the initial class boundaries.

6.3 Experiment 3: Mixed Augmentation Strategy

The third experiment considered the assessment of optimal way of mixing and matching original and synthetic samples in every class, with balanced splits in training and testing.

6.3.1 Implementation Details

Strategy 3 integrated 34,667 original samples with 20,000 synthetic samples (10,000 each for Anti and Neutral stances), producing a more balanced dataset of 54,667 total samples. The

class distribution enhanced to: Pro (42.00%), Neutral (32.41%), and Anti (25.59%). Standard 80/20 train-test splitting retained these proportions across all data splits.

6.3.2 Performance Improvements

Table 3: Mixed Augmentation Performance (Strategy 3)

Model	Overall Accuracy	Improvement	Anti F1	Anti F1 Gain	Neutral F1	Neutral F1 Gain
BERT	88.36%	+7.74%	0.9134	+0.2459	0.8531	+0.2377
DistilBERT	87.52%	+8.71%	0.9074	+0.3039	0.8428	+0.2770
RoBERTa	**88.92%**	+8.71%	**0.9194**	+0.2222	0.8605	+0.1964
DeBERTa	88.83%	+7.00%	0.9112	+0.2108	**0.8689**	+0.2398
XLNet	87.57%	+7.34%	0.9048	+0.2456	0.8464	+0.2806
BERTweet	88.21%	+5.78%	0.9100	+0.1861	0.8508	+0.2317
LSTM	82.84%	+10.08%	0.8276	+0.5482	0.7832	+0.2854

Highest overall accuracy was obtained using RoBERTa (88.92%), and all the models performed considerably better. Anti-class F1-scores improved drastically with the improvement in the range of 0.1861 (BERTweet) to 0.5482 (LSTM). These improvements exceeded the desired level of 0.15 improvement in F1-score in the minority class.

6.3.3 Statistical Significance

Paired t-tests of Strategy 1 and Strategy 3 performance also confirmed statistical significance ($p < 0.001$) for all the models. Effect sizes (Cohen's d) ranged from 1.82 to 3.14 and thereby verified large practical significance. Generalizability of the augmentation strategy was established through consistently observed improvement in different architectures.

Table 4: Summary of Key Findings Across All Experiments

Metric	Baseline (Strategy 1)	Mixed Augmentation (Strategy 3)	Improvement
Best Overall Accuracy	82.43% (BERTweet)	88.92% (RoBERTa)	+6.49%
Average Accuracy (All Models)	78.44%	87.56%	+9.12%
Best Anti F1-Score	0.7239 (BERTweet)	0.9194 (RoBERTa)	+0.1955
Average Anti F1 Improvement	-	-	+0.2685
Best Neutral F1-Score	0.6641 (RoBERTa)	0.8689 (DeBERTa)	+0.2048
Dataset Size	34,667	54,667	+20,000
Anti Class Representation	11.51%	25.59%	+14.08%
Total Samples Generated	-	41,000	-
Accepted Samples	-	20,000	48.78%

6.4 Discussion

Experimental results present significant evidence supporting the effectiveness of stance-specific augmentation schemes in dealing with issues of class imbalance in climate stance identification. Maintaining 88.92% in overall accuracy with RoBERTa presents a significant improvement from the baseline result, and the significant jumps in minority class F1-scores are evidence supporting the research approach.

6.4.1 Alignment with Literature

The results agree with and complement the results of prior augmentation work. The notable accuracy increases attained in all the models make the quality-controlled augmentation successful in correcting the flaws highlighted in Chen et al. (2023), in which they indicated that generic augmentation often fails to maintain domain-specific patterns, an issue that's remedied with the assistance of stance-specific generation tactics.

Experiment 2's catastrophic failure substantiates Kesgin and Amasyali's (2024) claim in favor of strategically ordering data despite the difference in their Modified Cyclical Curriculum Learning approach and the mixed distribution approach used here. Strategy 3's success further implies that static distributions throughout training might be more crucial than sophisticated curriculum scheduling in dealing with highly imbalanced datasets.

6.4.2 Critical Analysis of Experimental Design

Though the experiments shown proved significant improvements, some limitations in design are worth noticing. Exclusive consideration of Anti and Neutral augmentation, despite correcting the strongest imbalances, did not allow for exploration of the effects of Pro-stance augmentation. Future research must consider full-bandwidth augmentation of all classes to achieve maximum robustness of the model.

Validation framework, although comprehensive, relied upon heuristic combination of scores without optimization. Ablation studies that remove individual validation layers could assist in revealing redundancy and towards streamlined validation. Additionally, the end acceptance rate of approximately 49% (20,000 accepted out of 41,000 generated), with quality assurance obtained, required significant computational overhead that points towards optimization opportunity in generation of efficiency gains.

6.4.3 Theoretical Contributions

The work contributes to synthetic data generation for detecting stance through the successful demonstration that stance-related linguistic patterns can be effectively captured and reproduced via properly planned generation strategies. The assessment framework posits a theoretical model that can be applied to multi-dimensional quality assessment in areas beyond climate discourse.

The findings disprove the claim that quantity always improves performance, quality-managed additions surpassed quantity-focused strategies. This supports Iskander et al.'s (2024) findings and generalizes their work to areas in detecting stances.

7 Conclusion and Future Work

7.1 Summary of Research Achievements

This work successfully addressed the inextricable issue of severe class imbalance in climate stance detection with the help of a sophisticated synthetic data augmentation system. All the four research objectives outlined in Section 1 were accomplished with measures of performance bettering objectives set at the start. With the help of ten stance-based augmentation schemes with a five-layered validation framework and parallel architecture for processing, 41,000 attempts were transformed into 20,000 high-quality synthetic instances.

Experimental assessment verified substantial improvement in all the considered models, with RoBERTa reaching 88.92% overall accuracy a 6.49% increase over baseline. Perhaps more importantly, anti-stance F1-scores improved from 0.6972 to 0.9194, significantly more than the anticipated 0.15 improvement. These results, tabled in detail in Tables 1 through 4, validate the efficacy of quality-controlled, stance-based augmentation in correcting minority class underrepresentation in corpora of climate discourse.

7.2 Key Findings and Contributions

The research produced multiple significant findings that enhance the understanding of synthetic data augmentation for stance detection:

Finding 1: Quality Over Quantity - Production of 41,000 samples to create 20,000 accepted samples reaffirms that quality-controlled augmentation, in spite of its computational costliness, produces superior outcomes than quantity-based schemes. This reaffirms and generalizes Iskander et al.'s (2024) findings, confirming that small but high-quality datasets are superior to big and unverified ones.

Finding 2: Distribution Consistency is Critical - The catastrophic failure of Strategy 2 (0% accuracy) provided valuable lessons in the necessity to maintain data distribution balanced between training and test stages. This finding goes against strategies that rely solely upon the employment of synthetic data for minority classes, and that actual samples are still necessary to establish initial class limits.

Finding 3: Stance-Specific Strategies are Essential - Generic augmentation strategies proved unsuitable in capturing unique linguistic patterns that are characteristic of different stances. The success of domain-specific strategies like "Skeptical Data Questioning" (acceptance rate: 74.1%) and "Balanced Complexity Awareness" (acceptance rate: 76.8%) bears out the value for domain-specific generation strategies.

Finding 4: Architecture Agnostic Improvements - The simultaneous improvement in performance across seven various models, all the way from classical LSTMs to the newest transformers, suggests the augmented approach's generalizability. That LSTM-based models, too, showed considerable improvement (from anti-class F1: 0.2794 to 0.8276) suggests the data added reflected underlying features of stance and not model-specific information.

7.3 Limitations and Critical Reflection

There are various restrictions placed on the generalizability and transferability of this research work. Exclusive focus on English tweets limits cross-linguistic generalizability, as patterns in the expression of stance will frequently differ significantly across languages and cultures. Computational cost, which needs 41,000 API calls to generate 20,000 accepted samples, is challenging for applications with restricted resources in scalability terms.

Validation framework, although comprehensive, was conducted with heuristically selected weights without systematic optimization. Fixed thresholds (0.5 for anti-stance, 0.45 for neutral) were derived through empirical observation and not through principled optimization, and so there exists potential for improvement with learned determination of thresholds.

The work did not reinforce pro-climate stances, only taking into consideration minority classes. This selection, despite correcting the strongest imbalance, did not enable testing if reinforcing well-represented classes would further render the model more robust. Additionally, the evaluation took into consideration mostly on accuracy measures without extensive verification of the model's robustness to adversarial attacks or out-of-domain generalization.

The reliance upon GPT-4.1 Mini, despite producing excellent outputs, places reliance upon a proprietary commercial model. Changes to the functionality or availability of this model could impact reproducibility adversely. Moreover, the cost of usage of the API, though less than with previous models, remains a consideration for large-scale deployment.

7.4 Future Work

The research work proposes various possible future research areas that can expand and refine the contributions of this work:

7.4.1 Cross-Lingual Stance Augmentation

The extension to multilingual environments presents vast opportunities for research. Future work could investigate whether these stance-specific strategies transfer across languages or need to be culturally aligned. Language-independent validation schemes would enable broader application, particularly for low-resource languages where datasets for stance detection are scarce.

7.4.2 Adaptive Validation Framework

The existing static validation weights offer optimization potential. Future work might investigate the deployment of machine learning to adaptively vary validation layer weights depending on the performance of downstream tasks. Adoption of active learning to determine the best thresholds for various domains might make the process more efficient with preserved quality requirements.

References

Chang, Yupeng *et al.* (2024) 'A survey on evaluation of large language models,' *ACM Transactions on Intelligent Systems and Technology*, 15(3), pp. 1–45. https://doi.org/10.1145/3641289.

Chen, B. *et al.* (2025) 'Unleashing the potential of prompt engineering for large language models,' *Patterns*, p. 101260. https://doi.org/10.1016/j.patter.2025.101260.

Chen, J. *et al.* (2023) 'An Empirical survey of data augmentation for Limited data learning in NLP,' *Transactions of the Association for Computational Linguistics*, 11, pp. 191–211. https://doi.org/10.1162/tacl_a_00542.

Chuayrod, P. *et al.* (2024) 'The Impact of Prompt Engineering on Large Language Models: A Case Study of Sustainable Development Goals,' *IEEE*, pp. 1–6. https://doi.org/10.1109/isai-nlp64410.2024.10799499.

Feng, S.Y. *et al.* (2021) 'A survey of Data Augmentation Approaches for NLP,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2105.03075.

Garg, K. and Caragea, C. (2024) 'Stanceformer: Target-Aware Transformer for Stance Detection,' *ACN*, pp. 4969–4984. https://doi.org/10.18653/v1/2024.findings-emnlp.286.

Gopali, S. *et al.* (2024) 'The applicability of LLMS in generating textual samples for analysis of imbalanced datasets,' *IEEE Access*, p. 1. https://doi.org/10.1109/access.2024.3463400.
Iskander, S. *et al.* (2024) 'Quality Matters: Evaluating Synthetic Data for Tool-Using LLMs,' *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4958–4976. https://aclanthology.org/2024.emnlp-main.285.pdf.

Kesgin, H.T. and Amasyali, M.F. (2024) 'Advancing NLP models with strategic text augmentation: A comprehensive study of augmentation methods and curriculum strategies,' *Natural Language Processing Journal*, 7, p. 100071. https://doi.org/10.1016/j.nlp.2024.100071.

Khiabani, P.J. and Zubiaga, A. (2025) 'Cross-target stance detection: A survey of techniques, datasets, and challenges,' *Expert Systems With Applications*, p. 127790. https://doi.org/10.1016/j.eswa.2025.127790.

Li, B., Hou, Y. and Che, W. (2022) 'Data augmentation approaches in natural language processing: A survey,' *AI Open*, 3, pp. 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001.
Li, Y. *et al.* (2024) 'Large Language Model Data Augmentation for Text-Pair Classification Tasks,' *ACM*, pp. 427–433. https://doi.org/10.1145/3704323.3704362.

Nadăș, M., Dioșan, L. and Tomescu, A. (2025) 'Synthetic data generation using large language models: advances in text and code,' *IEEE Access*, p. 1. https://doi.org/10.1109/access.2025.3589503.

Ng, S. *et al.* (2025) 'Stance classification: a comparative study and use case on Australian parliamentary debates,' *Journal of Computational Social Science*, 8(2). https://doi.org/10.1007/s42001-025-00366-y.

Sahoo, P. *et al.* (2024) 'A Systematic survey of prompt engineering in large language Models: Techniques and applications,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2402.07927.

Shorten, C., Khoshgoftaar, T.M. and Furht, B. (2021) 'Text data augmentation for deep learning,' *Journal of Big Data*, 8(1). https://doi.org/10.1186/s40537-021-00492-0.

Sivarajkumar, S. *et al.* (2024) 'An Empirical evaluation of prompting strategies for large language models in Zero-Shot clinical natural language processing: Algorithm Development and Validation study,' *JMIR Medical Informatics*, 12, p. e55318. https://doi.org/10.2196/55318.

Tshimula, J.M., Chikhaoui, B. and Wang, S. (2020) 'A pre-training approach for stance classification in online forums,' *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 58, pp. 280–287. https://doi.org/10.1109/asonam49781.2020.9381467.

Upadhyaya, A., Fisichella, M. and Nejdl, W. (2023) 'Towards sentiment and Temporal Aided Stance Detection of climate change tweets,' *Information Processing & Management*, 60(4), p. 103325. https://doi.org/10.1016/j.ipm.2023.103325.

Ziyaden, A. *et al.* (2024) 'Text data augmentation and pre-trained Language Model for enhancing text classification of low-resource languages,' *PeerJ Computer Science*, 10, p. e1974. https://doi.org/10.7717/peerj-cs.1974.