A STUDY ON SENTENCE-LEVEL ARGUMENT IDENTIFICATION IN IMBALANCED STUDENT ESSAY CORPORA

by

Rishabh Lingam

A Thesis Submitted to the Faculty of

The College of Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Florida Atlantic University

Boca Raton, FL

July 2025

A STUDY ON SENTENCE-LEVEL ARGUMENT IDENTIFICATION IN IMBALANCED STUDENT ESSAY CORPORA

by

Rishabh Lingam

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Xingquan Zhu, Department of Electrical Engineering and Computer Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements for the degree of Master of Science.

SUPERVISORY COMMITTEE:

Xingquan Zhu, Ph.D.
Thesis Advisor

Abhijit Pandya, Ph.D.

Borko Furht, Ph.D.

Sipai Klein, Ph.D.

Wendy Hinshaw, Ph.D.

Hari Kalva, Ph.D.
Chair, Department of Electrical Engineering and Computer Science

Stella Batalama, Ph.D.
Dean, The College of Engineering and Computer Science

Robert W. Stackman, Jr., Ph.D.
Dean, Graduate College

Date

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everyone who has supported me throughout my Master's journey. The support and encouragement I have received has truly made me push towards my goal.

I am extremely grateful to my advisor, Dr. Xingquan Zhu, for accepting me as his student. This wouldn't have been possible if it wasn't for his guidance, and patience. I thank him for making me reach my potential and for believing in me when I did not myself. Under his mentorship, I have learnt much more than knowledge, I learnt hard work and discipline. He has inspired me in many ways.

I am equally grateful to Dr. Sipai Klein for his unwavering support, both during this project and beyond. His guidance has shaped my academic and professional growth in many ways, and I deeply appreciate his encouragement and confidence in me.

I would also like to thank Dr. Abhijit Pandya,Dr. Borko Furht, and Dr. Wendy Hinshaw for graciously serving as members of my thesis committee.

My heartfelt thanks go to my parents, Ravindra Babu Lingam and Swarnalatha Lingam, and to my brother, Saurabh Lingam, whose love and unwavering support have been the source of my strength. I am also thankful to Vivek Sonar, Sai Sathvika Chalasani, Aditya Waghmode, and Vineet Kate for their encouragement, suggestions and friendship during my time at FAU.

A special thank you to my best friend Sakshi Dugar and to Bhanu Mohindra. Their encouragement echoed all the way from India and Sweden and reminded me that I was not alone in this journey.

This research is partially sponsored by the U.S. National Science Foundation under grant Nos. IIS-2236579, IIS-2302786, the Florida Atlantic University Dorothy F. Schmidt College of Arts and Letters seed grant, and Florida Atlantic University's Writing Across the Curriculum Program. Any opinions, findings, conclusions, or recommendations expressed in this research are those of the author and do not necessarily reflect the views of the afore mentioned sponsoring agencies.

ABSTRACT

Author:	Rishabh Lingam
Title:	A Study on Sentence-Level Argument Identification in Imbalanced Student Essay Corpora
Institution:	Florida Atlantic University
Thesis Advisor:	Dr. Xingquan Zhu
Degree:	Master of Science
Year:	2025

Writing is an essential skill that affects success in nearly every academic subject and professional field. For undergraduate students, strong writing helps them organize ideas, communicate clearly, and perform better in both written assignments and overall coursework. Good writing also supports critical thinking, which is key to problem-solving and academic growth. Beyond school, writing continues to be important in the workplace, where it is used for emails, reports, presentations, and formal documents. Yet, despite its importance, many students and graduates do not have strong writing skills, and this gap is noticed by employers. A recent survey by Ashley Finley [1] found that while 90% of employers value written communication, only 44% believe graduates are prepared.

At the same time, recent progress in artificial intelligence has made tools like neural language models useful for supporting writing instruction and grading. These models offer faster and more objective ways to assess student writing. In this study, we explore how automated writing assessment can work at the sentence level, focusing on Writing Across the Curriculum (WAC) categories used to assess student writing in College Writing 1 and 2 at Florida Atlantic University. We collected final argumentative essays written by students and analyzed them using a neural language model to assess writing quality, at both sentence by sentence level and across the whole essay. Our findings show that the model can recognize patterns in writing and provide useful evaluations, but there are still challenges with scoring consistency. This research shows possible improvements to address these issues and highlight key takeaways from the case study that support using sentence-level assessment in writing instruction.

A STUDY ON SENTENCE-LEVEL ARGUMENT IDENTIFICATION IN IMBALANCED STUDENT ESSAY CORPORA

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 MOTIVATION AND CONTRIBUTION

Writing is a universal skill that affects every discipline, from arts to engineering. Strong skills are especially important for undergraduate students, as they impact their academic performance, career prospects, and overall communication ability.

In academia, writing enables students to organize their thoughts, evaluate information, and develop coherent and well-supported arguments. Courses that traditionally do not focus on writing still include essays and reports that greatly impact grades. Even to pursue higher education, writing skills are important for dissertations, thesis, grant proposals, and capstone projects, where clarity and quality of writing are essential for sharing ideas and making contributions in different fields of study. Through there are support tools like ChatGPT, QuillBot, Claude, Gemini, etc., that help in the process of writing, they undermine a crucial outcome of writing, which is critical thinking. Unassisted manual writing is a cognitive exercise that stimulates mental growth and problem solving. Developing these skills not only affects grades but, by extension, also increases chances for scholarships, internships, and jobs.

Writing is as important for professional success as it is for academics. In a typical work environment, good writing skills are essential for regular tasks like creating reports, writing emails, and making presentations. It also helps in formulating opinions, which empowers collaboration. Professional fields like business, law, and medicine also require a high level of writing competency to draft contracts, write patient reports, or create business proposals. For professionals in non-profit roles, a significant portion of the day is dedicated to community outreach and showing the public why they should care about a particular issue. The common thread running through all of these scenarios is the ability to write comprehensively and succinctly. Despite such a huge requirement for quality writing, there is a widening skill gap in the current work force. According to a recent study [1], 90% of employers ranked written communication as either "very" or "somewhat" important. Yet only 44% believed graduates were adequately prepared in this area.

Recent advancements in neural language models have enabled a range of applications in writing instruction and assessment, including automated evaluation in secondary and post-secondary education. These models offer the promise of increased accuracy, efficiency, and objectivity in assessing student writing, benefiting both learners and educators. Based on our case study involving sentence-level classification for Writing Across the Curriculum (WAC) assessment in College Writing 1 and 2 courses at Florida Atlantic University (FAU), we argue that automated writing assessment should incorporate fine-grained, sentence-level analysis of argumentative structure. We collected end-of-semester argumentative essays and evaluated them at both the sentence and document levels. Our findings demonstrate that while transformer-based models can capture surface-level argumentative patterns, their performance suffers from class imbalance and lack of contextual or syntactic awareness. To address these challenges, we explored context-pairing, syntactic augmentation, and distribution-aware hierarchical modelling. These strategies led to measurable improvements in identifying under-represented argumentative elements. Our study contributes to practical modelling strategies and architectural configurations that enhance the robustness and fairness of automated argumentation assessment in student writing.

1.2 BACKGROUND KNOWLEDGE

1.2.1 WAC Assessment Overview

In order to evaluate student writing from across the university and allow for different disciplinary conventions, WAC assessment employs a hybrid rubric to assess student writing across 11 categories. This approach combines analytical and holistic rubrics in place of a single holistic score per essay, which provides more detailed feedback on students' writing skills. As seen in Table 1.1, the assessment evaluates core writing skills including opening strategy, argumentative features, organizational structure, concluding strategy, disciplinary concerns, grammar, and syntax. These core writing skills are divided into the following categories: thesis, organizational framework, reasoning, evidence, rhetorical structure, implication and consequences, academic tone, disciplinary conventions, clarity, style and, mechanics.

Each category is assigned an integer score on a 4-point continuum between 1 and 4 (4 – Extremely Effective, 3 – Effective, 2 – Adequate and, 1 – Inadequate). Each essay is scored by three raters using an Agreement-Disagreement method. This is a simple percentage of agreement between the set of raters. The advantages of using this statistic is that it is physically easy to calculate and understand. The disadvantage is that raters could agree by chance or by factors that have little or nothing to do with the criterion measured. On a 4-point scale, such as the one used in this study, raters could agree 25% of the time by chance. Thus, this procedure tends to over-estimate Inter-Rater Reliability.

Final scores for each of the 11 categories are determined through a process of central tendency. What this means is that modal scores are used as the default. Median scores are used secondarily. For example, if all three raters award the same score (e.g., "3"), it is considered the final score for that paper and for that category. If only two raters award the same score, that score becomes the final score for that

Table 1.1: Eleven categories of WAC Assessment

WAC Categories	Definitions
Thesis/purposes/argument	Persuasive purpose of the paper
Organizational Statement	A statement describing the building of the argument
Reasoning	Analysis of evidentiary materials and demonstrated comprehension of ideas
Evidence	Integration of data, quotations, visuals, and counterarguments
Rhetorical Structure	Sustained focus on argument's development and its progression
Implications and Consequences	Development of argument's conclusion
Academic Tone	Formality of specialized terms and concepts
Disciplinary Conventions	Discipline-specific formatting and citation
Clarity	Sentence-level comprehension and consistent usage of discipline-specific terminology
Style	Linguistic variation between sentences
Mechanics	Mechanical sentence level error patterns

paper and for that category, regardless of the level of disagreement by the third rater. If all three raters award different scores, the middle (median) rater's score is used as the final score for that paper and for that category.

1.2.2 The BERT Model

BERT first introduced by Devlin et al [5], stands for Bidirectional Encoder Representation from Transformer. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditional on both left and right context in all layers. BERT's Model architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described by Vaswani et al [6]. It uses WordPiece embedding with a 30,000 token vocabulary. To handle a variety of down-stream tasks, the input representation unambiguously represent both a single sentence and a pair of sentences in one token sequence.

The first token of every sequence is always a special classification token ([CLS]). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. Sentence pairs are packed together into a single sequence. The sentences are differentiated in two ways- first, the sentences are separated with a special token ([SEP]). Second, a learned embedding is added to every token indicating whether it belongs to sentence A or sentence B. BERT was pre-trained using two unsupervised tasks.

1. Masked Language Modeling (MLM): Masked Language Modeling (MLM) is a training task where certain words in a sentence are hidden (masked), and the model learns to predict them based on the surrounding context. This helps the model understand grammar, context, and word relationships.

2. Next Sentence Prediction (NSP): Next Sentence Prediction (NSP) is a task where a model learns to determine whether one sentence logically follows another in a given text. It helps language models understand sentence relation-

ships and improve coherence in tasks like summarization and question answering.

The model is pre-trained on the BookCorpus dataset [26] (which is a collection of 11,038 unpublished books) and Wikipedia-English corpus.

BERT for Question-Answering

In the question answering task, the input question and passage are represented as a single packed sequence, with the question using the A embedding and the passage using the B embedding. A start vector $S \in \mathbb{R}^H$ and an end vector $E \in \mathbb{R}^H$ are introduced during fine-tuning. The probability of word i being the start of the answer span is computed as a dot product between T_i and S followed by a softmax over all of the words in the paragraph:

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}.$$

A similar formula is used for the end of the answer span. The score of a candidate span from position i to position j is defined as $S \cdot T_i + E \cdot T_j$, and the maximum scoring span where $j \geq i$ is used as a prediction. The training objective is the sum of the log-likelihoods of the correct start and end positions.

There also exits a possibility that no answer exists in the provided paragraph. We treat questions that do not have an answer as having an answer span with start and end at the `[CLS]` token. The probability space for the start and end answer span positions is extended to include the position of the `[CLS]` token. For prediction, the scores of the no-answer span:

$$s_{\text{null}} = S \cdot C + E \cdot C$$

is compared to the score of the best non-null span

$$\hat{s}_{i,j} = \max_{j \geq i} S \cdot T_i + E \cdot T_j.$$

A non-null answer is predicted when $\hat{s}_{i,j} > s_{\text{null}} + \tau$, where the threshold τ is selected on the dev set to maximize F1.

1.2.3 Data Balancing

Imbalanced data is a common problem in the real world. It happens when a few of the categories make up a large portion of the dataset leaving the rest underrepresented. The problem becomes more severe if the underrepresented categories are equally or more important. The model trained on imbalanced data learns that it can achieve high accuracy by constantly predicting the major classes. Balancing data is a crucial step as it prevents the model from becoming biased towards dominant classes. The following table shows the generally accepted ranges of degrees of imbalance,

Table 1.2: Degree of Imbalance

% of data belonging to Minority Class	**Degree of Imbalance**
20-40% of the dataset	Mild Imbalance
1-20% of the dataset	Moderate Imbalance
$\leq$ 1% of the dataset	Extreme Imbalance

One obvious way to deal with imbalanced data is to collect more data, but this is usually expensive and time consuming. There are programmatic ways to overcome imbalanced data, like,

1. Down-sampling: We only use a portion of the majority class for training the model.

2. Over-sampling: It involves increasing the number of samples within the minority class by taking duplicates.

3. Perfect Balancing: Perfect balancing involves using operations to make all the classes equal in count.

While these techniques are effective, they come with potential problems. Down sampling causes loss of information and generalization, over-sampling may cause the model to overfit and develop unintended bias towards the minority classes and perfect balancing is not representative of the real-world distribution of the data.

Another way to deal with imbalanced data without losing information or overfitting is to use Inverse Weighted Loss. It assigns higher weight to loss incurred by misclassifying minority samples and relatively lower weight to the loss from misclassifying majority samples, essentially making the model more sensitive towards minority classes and forcing it to improve its performance on them.

1.2.4 Data Augmentation

Data augmentation is a technique of increasing the training data by creating modified copies of existing data. It helps the machine learning models to perform better by preventing overfitting, increasing accuracy, generalization and injecting diversity into the data, especially when the data is smaller and imbalanced. Data augmentation also helps to reduce operational cost of cleaning and labeling raw data. Another technique that is distinct yet complementary to data augmentation is Synthetic Data Generation. Data augmentation involves creating modified versions of existing data, for example, for images, we can apply transformations like rotations, flips, cropping, inversions or hue adjustments to create new data points. In case of synthetic data generation, we use artificially created data, for example, images created by GANs or Diffusion models. Data augmentation is a careful process as it can lead to enhancing biases present in the original data.

When augmenting textual data, the transformation and operation should be carefully selected to maintain the semantic and syntactic meaning of the text. In our series of experiments, we used two types of textual data – single sentence and continuous sequence of sentences aka contextual text. Singles sentences are augmented using synthetic data generation technique. We use a text paraphrasing BERT model to create new sentences that carry the same semantic meaning as the original sentence. For augmenting contextual text, we create pairs of contexts and swap randomly single sentences from each context based on the maximum or minimum similarity of the embedding of the sentence.

1.3 STRUCTURE OF THE THESIS

This thesis is organized into five chapters:

- **Chapter 1 – Introduction**

  This chapter outlines the motivation behind the study, the main research questions, and the contributions of this work. It introduces the challenges in automated sentence-level classification of argumentative components in student essays and provides an overview of the dataset used.

- **Chapter 2 – Related Work**

  This chapter presents a survey of previous research on automated writing assessment, argumentation mining, and the use of neural models for educational text analysis. It highlights prior modelling strategies, limitations in existing datasets, and the relevance of sentence-level approaches.

- **Chapter 3 – Proposed Model**

  This chapter describes the experimental design and proposed modelling strategies, including baseline, context-aided, syntactically enhanced, and distribution-aware hierarchical models. It explains the rationale and configuration for each approach.

- **Chapter 4 – Experiments and Results**

  4.1 Data Preprocessing and Preparation: This section covers the dataset collection, annotation consistency checks, preprocessing techniques, and label consolidation process.

  4.2 Results and Analysis: This section presents and interprets the results of each experiment, comparing performance across model variants, and evaluating improvements in low-resource argumentative categories.

- **Chapter 5 – Conclusion and Future Work**

  This chapter summarizes the main findings, outlines the limitations of the current study, and proposes future directions such as incorporating document-level features, soft hierarchical routing, and instruction-tuned large language models for better generalization.

CHAPTER 2
RELATED WORK

2.1 AES AND AWE

As artificial intelligence and literacy researcher McNamara [13] argues, AI research lacks large data sets by which to investigate key questions about the role of AI in education. This challenge complicates prior research on automated writing evaluations (AWE) and automated essay scoring (AES) systems which has demonstrated limitations in these systems' ability to provide detailed feedback and attend to higher-order writing concerns [14]. However, emerging research on AI-enabled writing tools suggests such tools can aid writing teachers and learners in metalinguistic awareness and development of writing skills ([14], [8], [18]). In addition, research into AI-driven assessment and intelligent tutoring has shown potential in AWE systems adopting NLP tools to improve students' persuasive writing skills [16]. Neural language modeling, such as BERT and ChatGPT, has the capacity to extend research into Automated Essay Scoring (AES) because it has the potential to identify linguistic patterns, be trained on human rating practices to generalize probabilistic assessments, and generate responses to prompt writing revision [19].

Prior AWE and AES research does not use argumentative, thesis-driven, reading-centered, long-form student essays as the corpus, leading to a substantial limitation in the training data; for example, AES scoring based on data sets of tenth grade ([15]) or seventh grade essays with an average length of approximately 250 words ([19]). This study proposes to use data from FAU WAC assessment of writing samples from a first-year composition course, ENC 1101: College Writing 1, as the basis for developing a neural language model for AWE that can accurately score student writing. Our objective is to develop a model that can assess and also provide feedback to students on their writing according to the categories of the learning model.

The above observations motivate the proposed research, which seeks to answer the following question: Can the elementary units of argumentation in thesis-driven, reading-centered, long-form essays taught in College Writing 1 and 2 be determined automatically? By using a real-world dataset as the case study material, our study essentially advances AI-integrated AWE systems for improving core writing skills in writing-intensive university courses.

2.2 ARGUMENTATION MINING

Wang, Hao, et al. [28] try to show that different types of argumentation components exist at different granularity levels within an essay, and traditional approaches like single sequence tagging problem at the word level, while suitable for integrating local word-level information, fail in inference on long-distance text. Older methods handled the task at the sentence level in a crude way, failing to effectively capture components that serve as core opinions. They propose a model that is designed to identify different types of argumentation components at different levels. This multi-scale approach addresses the aforementioned limitations by mining different component types at their corresponding levels. Experimental results indicate that mining at these different levels improves performance on identifying all types of argumentation components. They identified three levels shown below, with the argumentation components associated with them,

1. Essay Level: This level is used for identifying Major Claims (MC). Major claims serve as the core opinions on essays and are at the essay level. MCs are mined using an essay-level argumentation extraction submodule based on a multi-span extraction strategy. This submodule takes the entire essay as input to a BERT

encoder, uses a pointer network to score candidate spans, and then ranks and filters spans based on scores and rules.

2. Paragraph Level: This level is used for identifying Claims (C). Claims serve as core opinions on paragraphs and are considered to be at the paragraph level. Claims are mined using a paragraph-level argumentation extraction submodule based on a randomized extraction strategy. This submodule processes each paragraph separately with a BERT encoder to mine result spans, which are then gathered for the essay. Filtering rules are applied to remove apparently wrong and overlapped candidate spans.

3. Word Level: This level is used for identifying Premises (P). Premises consist of elements like logical statements, survey results, typical examples, public thoughts, or expert suggestions. Premises are at word level. Premises serve as evidence for major claims and claims, are mined using a word-level argumentation tagging submodule. This uses a BERT-CRF sequence tagging model with the whole essay as input to the BERT encoder and CRF as the decoder.

A Coarse-to-Fine Argumentation Fusion Mechanism is used to handle potential overlaps between identified spans of different argumentation types. It assigns priorities: Major Claim > Claim > Premise. If spans from different sets overlap, the span from the set with higher priority is kept, and the overlapping span from the lower priority set is removed. They the Persuasive Essay dataset (PE 2.0), which is based on PE 1.0 and annotates major claim, claim, and premise. Evaluation metrics included span-based precision, recall, and F1 score for specific types, as well as macro-F score and micro-F score for overall performance. Baselines used for comparison were TARGER (a BiLSTM-CNN-CRF sequence tagging model) and a BERT-CRF sequence tagging model.

The key findings and challenges are listed below,

- It is effective to handle different argumentation types at their corresponding levels. Essay-level extraction was better for Major Claim, paragraph-level for Claims, and word-level tagging for Premise. But the task is not scale-independent; components like Major Claims and Claims require processing at larger scales than Premises.

- Sequence tagging methods rely on local context words and struggle to effectively mine components at higher levels. Word-level sequence tagging models show extremely limited performance on mining major claims and claims compared to the multi-scale approach.

- Identifying exact boundaries of premises can be difficult. Distinguishing component types and non-argumentative text poses challenges. A disambiguation system, like the coarse-to-fine argumentation fusion mechanism, is required to further improved performance.

Isaac Persing and Vincent Ng [33] addressed the end-to-end task of argument mining in persuasive student essays by dividing the task into two subtasks - Argument Components Identification (ACI) and Relations Identification (RI) between them. Like Wang, Hao, et al., they used three argument categories – Major Claims, Claims, Premises with an additional category called Non-Argumentative. For relation types – Attack, Support and No-Relation. They present a pipeline approach as a baseline and then proposed an improvement method using joint inference over the outputs of ACI and RI classifiers within an Integer Linear Programming (ILP) framework to mitigate error propagation and enforce consistency.

The independent ACI classifier predicts the probability of each Argument Component Candidate (ACC) belonging to a specific type (premise, claim, major claim, or non-argumentative). It is a two-step process: extract argument component candidates (ACCs) heuristically using low precision, high recall rules based on syntactic parse trees, and then classify each ACC as premise, claim, major claim, or non-argumentative using a maximum entropy classifier with various features (structural, lexical, syntactic, indicator, contextual). Then, the Relationship Identification classifier predicts the probability of each Relation Candidate (RC) having a specific relation type (support, attack, or no-relation) between pairs of ACCs that were identified by the ACI step. These probability scores are inputs to the ILP model.

The ILP framework uses binary indicator variables to represent the final decisions made by the system. For example, X_{mi} might be a variable that is 1 if ACC_i is predicted to be a major claim and 0 otherwise. Similarly, $Y_{i,j}$ might be 1 if a support relation is predicted from ACC_i to ACC_j and 0 otherwise.

The aim of the ILP system is to find an assignment of values to these binary variables that maximises an objective function. Unlike typical NLP objectives, which might maximise the sum of confidence scores, this work proposes a novel objective function designed to directly maximise a measure related to the F-score , which is the primary evaluation metric. This involves calculating expected values for True Positives, False Positives, and False Negatives based on the classifier probabilities and the variable assignments.

A crucial part of the ILP framework is the inclusion of global consistency constraints. These are rules that the final predicted structure for the entire essay must satisfy. These constraints can be:

- Within-task constraints: Rules applying to one task across the entire essay, such as ensuring there is exactly one major claim per essay.

- Cross-task consistency constraints: Rules linking decisions across tasks, such as requiring that if a relation is predicted between two ACCs, both must be predicted as argumentative components. Specific examples of these constraints include:

→ Each ACC is assigned exactly one type.

→ Each RC is assigned exactly one type.

→ If a relation is predicted, both participating ACCs must be predicted as argumentative.

→ There is exactly one major claim per essay.

→ Major claims occur only in the first or last paragraph.

→ Major claims have no parents (are not supported or attacked).

→ A premise must have at least one parent (support or attack another component).

→ A premise is related only to components in the same paragraph.

→ A claim has no more than one parent.

→ If a claim has a parent, it must be a major claim.

→ Argumentatively labelled ACCs cannot overlap in the text.

→ Each paragraph contains at least one claim or major claim.

→ A sentence must not have more than two argumentatively labelled ACCs.

An ILP program is constructed for each test essay based on the ACCs and RCs generated for that essay, their classifier probabilities, the objective function, and the consistency constraints. This program is then solved using an ILP solver to find the optimal assignment of binary variables that satisfies all constraints and maximises the objective function.

The ILP framework is used primarily to overcome key limitations of the standard pipeline approach:

1. Addressing Error Propagation: In a pipeline, errors from the initial ACI step are strictly passed to the RI step. If the ACI classifier incorrectly labels a component as non-argumentative, the RI system cannot possibly identify any relations

involving that component. The joint inference enabled by ILP mitigates this. By considering the confidence scores of both ACI and RI classifiers simultaneously within the ILP, a strong prediction from the RI classifier (e.g., high probability of a support relation) can potentially override a weaker prediction from the ACI classifier, allowing components to be labelled as argumentative if they participate in highly probable relations.

2. Enforcing Global Consistency: Independent classifiers typically make decisions locally on individual instances (ACCs or RCs) without considering the overall argumentative structure of the entire essay. However, student essays have specific structural properties. The ILP framework allows for the explicit enforcement of these global consistency constraints across the whole essay, ensuring that the final predicted structure is more coherent and conforms to expected essay argumentation patterns.

The use of the ILP framework for joint inference has a positive impact on performance compared to the standard pipeline approach. On a corpus of 90 student essays, the joint-inference approach using ILP yielded an 18.5% relative error reduction in F-score over the pipeline system. Experiments on a larger corpus (402 essays) also show that the ILP approach outperforms the pipeline baseline. While still being outperformed by a state-of-the-art neural sequence tagging model in the evaluation setup, the ILP system s F-scores on ACI and RI were notably better than the pipeline baseline. In approximate matching, the ILP system achieved an average F-score of 52.7%, compared to the pipeline s 45.8%.

The following summarizes their work in terms of challenges and key findings,

Challenges:

- Error propagation is inherent in the pipeline approach, where errors in ACI affect the performance of RI.

- Enforcing global consistency constraints across different components and relations within an essay is difficult for classifiers that treat instances independently.
- The typical objective function used in ILP for NLP tasks is not ideal for tasks whose primary evaluation metric is F-score.
- Identifying the exact boundaries of argument components is a challenging task.
- Developing high-quality features for relation identification is difficult, and relationships are often not triggered by discourse markers, especially between non-adjacent sentences.

Key Findings:

- The novel objective function enables F-score to be maximized directly by an ILP solver and is applicable to other ILP-based joint inference tasks.
- The ILP framework successfully enforces essay-level within-task and cross-task consistency constraints.
- The RI task is inherently more difficult than the ACI.

In a 2016 paper, Huy Nguyen and Diane Litman [29] focused on another important aspect of argument mining i.e. cross-topic evaluation. Cross-topic evaluation is an evaluation strategy (in tasks like text classification and argumentation mining) where a model is trained on data from one or more topics and tested on data from entirely different topics not seen during training. Why is it important ? Because argument mining systems need to reliably identify argument components irrespective of the specific topic of an essay. Student essays are often written in response to various assignments on diverse subjects. Standard approaches relying heavily on lexical features can become overly tied to the specific vocabulary of the training topics, leading to degraded performance when applied to essays on new topics. Their study addresses this by proposing and evaluating new features that go beyond standard lexical approaches to better model argumentation cues while abstracting away from topic-specific language. These features include,

- Common Word Counts: Two features measuring the number of words a given sentence shares with the immediately preceding sentence and with the essay's title.
- Plural First Person Pronouns: Five binary features indicating the presence of each of five plural first person pronouns.
- Discourse Relations: Three binary features indicating the presence of specific discourse relations (Comparison, Contingency, Expansion).
- RBR Part-of-Speech: This includes common topic-independent comparative adverbs like 'more' which were identified as a new feature that captures comparison more broadly.

This study mainly focuses on the feature design and evaluation for identifying argument components (ACI), particularly concerning topic independence. It does not describe the Integer Linear Programming (ILP) framework used for joint inference to mitigate error propagation between ACI and Relation Identification (RI) as discussed in another source, nor does it detail a multi-scale argumentation mining model processing components at essay, paragraph, and word levels as described in a different source. The methodology here is primarily about building better features and proving their effectiveness across topics.

Experimental results show that models enhanced with the proposed features significantly improve argument mining performance in both 10-fold cross-validation and cross-topic validation settings. The new models demonstrate topic-robustness, with cross-topic performance levels even higher than 10-fold cross-validation in some cases.

While baseline features (derived from the extracted words) are effective, the new features are a necessary supplement to achieve the best performance, especially in cross-topic validation. This suggests that explicitly designing features to abstract over topic-specific language is beneficial. However, the conclusion that the new features are necessary in addition to the learned argument and domain words suggests that the unsupervised extraction algorithm might yield "noisy" lists of words that are not fully sufficient on their own to capture topic independence. We can safely assume that relying solely on such learned lists without carefully designed, more abstract features might be a limitation.

In 2020, Isaac Persing and Vincent Ng presents a novel unsupervised approach [32] to end-to-end argumentation mining (identify both ACI and RI) in persuasive student essays. Their study was to check how well argument mining can be performed without requiring argument-annotated data. The key idea is to bootstrap from a small set of argument components that are automatically identified and labelled using simple heuristics and reliable contextual cues. The process starts by identifying ACCs (Argument Component Candidates) by applying a set of low precision, high recall heuristic rules to the parsed sentences of the essay. These rules define potential left and right boundaries for ACCs within a sentence, often focusing on clause-level structures. These heuristics are capable of finding ACCs that exactly match the boundaries of 92% of all argument components in the corpus. Once the ACCs are identified, a subset of them are heuristically labelled with specific argument component labels (Major Claim, Claim, or Premise). These heuristics are designed to be high precision but low recall. They rely on three main factors,

- Paragraph Location: Whether the ACC is in the first, last, or a middle paragraph. Major Claims, for instance, are only searched for in the first and last paragraphs. Claims and Premises are primarily labelled in middle paragraphs (body).

- Sentence Location: The location of the sentence containing the ACC within its paragraph. For example, Claim labelling heuristics exclude ACCs outside a body paragraph's first or last sentences. Premise labelling heuristics exclude ACCs in the last sentence of a body paragraph.

- Context N-grams: Specific preceding or succeeding n-grams that frequently indicate the presence of an argument component. These n-grams were identified based on observations of unannotated essays. Examples include phrases like "contend that", "first of all", "therefore". The heuristics look for these specific phrases and label nearby ACCs based on the expected component type associated with the phrase. Additional rules might apply if specific markers are absent, or for short final paragraphs.

This heuristically labelled data serves as the initial training data. ACCs that were heuristically labelled get their assigned argument component type (Major Claim, Claim, or Premise), while all other ACCs are initially labelled as "non-argumentative". A four-class maximum entropy classifier is then trained using this labelled data for the ACI task. The classifier uses features developed in prior supervised work on the same corpus, including structural, lexical, syntactic, indicator, and contextual features. The purpose of this is to train a system that is more general than the initial, low-recall heuristics. The classifier is further improved iteratively via self-training. The classifier is applied to the ACCs initially labelled "non-argumentative". Instances where the classifier is highly confident ($\geq$ 80% probability for instance) about an argumentative label are added to the labelled dataset. This process repeats until no more instances meet the confidence threshold for relabelling. The final ACI classifier is applied to all paragraphs. Based on the ACI predictions, a Candidate Argument Tree (CAT) is constructed for each paragraph,

- For first/last paragraphs: The most likely ACC (if probability > 0.5) is labelled as a major claim, all others are non-argumentative.

- For body paragraphs: The highest probability ACC is labelled as the claim for that paragraph. Remaining ACCs are ranked by premise probability and labelled as premises if they meet certain conditions (no overlap, premise probability higher than claim/MC probabilities). For each ACC labelled as a premise, a support relationship is added between it and the paragraph's claim. Attack relations are not covered in this heuristic tree building.

Performance is evaluated using span-based F-score. This requires defining true positives, false positives, and false negatives by comparing predicted spans to gold-standard annotations. Both exact matches, when boundaries are identical, and approximate matches, when spans share over half their tokens are used, with extra focus primarily on approximate match due to the difficulty of exact boundary identification. F-scores are calculated for ACI and RI, and an overall average F-score across ACI and RI is used. Evaluation using the approximate match metric, Unsupervised system outperformed supervised baseline (ILP) in overall average F-score with 50.7% to 47.8%. It also achieved better F-scores on both the ACI and RI subtasks (64.6% and 36.8%) compared to the ILP baseline (63.8% and 31.8%). In exact match metric, Unsupervised system's performance was only marginally better than the baseline systems.

The following are the key findings of the study,

- A non-trivial level of argumentation mining performance can be achieved in student essays without argument-annotated data.

- The initial heuristic labelling and training of a classifier on this heuristically labelled data is largely responsible for the ACI results and provide the majority of the learned knowledge.

- Unsupervised approach reduces reliance on language-specific annotated data and can be fairly easily adapted to identify argumentative structures in essays written in languages where such resources are not readily available.

Henning Wachsmuth et al. [31] investigated whether, to what extent, and how the output of argument mining can be leveraged to assess the argumentation quality of persuasive student essays. It focuses on assessing specific quality dimensions based on the mined argumentative structure. This work fits into a three-step writing support system,

1. Mining of Argumentative Structure

2. Assessment of quality dimensions based on mined structures

3. Synthesis of suggestions for quality improvements

This is important because while many approaches to the mining step have been developed, the benefit of the mined structure for downstream applications has rarely been evaluated. They built upon the essay-oriented argumentation model of Stab and Gurevych [34], which defines four types of ADUs (Argumentative Discourse Units) - Thesis, Conclusion, Premise, and None. To capture structure at an abstract level suitable for assessing quality and to ease pattern recognition, and to focus on mining units effectively by omitting structures such as attack-support relations (as majority of available data is insufficient for reliable training), the authors made the following two assumptions,

1. Each sentence corresponds to exactly one ADU. This avoids the need for ADU segmentation.

2. Each paragraph corresponds to exactly one argument. This avoids the need to identify relations between ADUs.

The identification of ADUs is treated as four-class classification task using supervised machine learning. They used six types of features to capture the content, style, and position of a sentence,

1. Prompt Similarity: cosine similarity to the essay prompt.
2. Token n-Grams: frequency of 1-gram to 3-gram.
3. POS n-Grams: frequency of part-of-speech 1-gram to 3-gram.
4. General Inquirer Classes: frequency of word classes.
5. 1st Token n-Grams: indicators for initial n-grams.
6. Sentence Position: position within paragraph and essay.

The analysis revealed common patterns, such as paragraphs often starting with a Conclusion followed by Premises. They also found patterns which were different for the first and last paragraphs than body paragraphs. This showed that the essays can be differentiated based on the combination, ordering, and number of ADU types. Based on these patterns found in the analysis, the authors proposed the following three novel shallow feature types designed to capture structural variations,

1. ADU Flows: Frequencies of different ADU sequences in an essay.
2. ADU n-Grams: Frequencies of ADU type n-grams (n=1, 2, 3) and indicators for the first and last ADU n-grams.
3. ADU Compositions: Percentages of paragraphs with a specific number (0, 1, 2, > 2) of each ADU type, plus summary statistics (min, max, mean, median) for each type per paragraph, and percentages of each type in the first and last paragraph.

These features can be used for a variety of downstream tasks like Argumentation, Analysis of Argumentative Structure and Argumentation Quality Assessment (Essay Scoring).

The following points summarize the key points of this study,

- The ADU compositions features consistently performed best among all structure-oriented features across all tasks.

- Matching structure-oriented features outperforms standard content and POS features in scoring Organization. This shows organization is about argumentative structure, not just discourse functions.

- Structure-oriented features were less effective for Thesis Clarity which relies more on content.

- Combining ADU compositions with standard features improved results over the SOTA. This indicates that argument strength benefits from both structure and content.

In essence, this study showed that mining and analysing the abstract argumentative structure of student essays provides significant benefits for automatically assessing the argumentation quality, particularly for dimensions related to structure like organization and, to some extent, argument strength, even without explicitly modelling relations or fine-grained units.

In this thesis, we focus exclusively on the task of Argumentation Component Identification. This is because the WAC dataset, which was collected and annotated in-house by the University Centre of Excellence in Writing at Florida Atlantic University, contains a broader range of annotation categories compared to the three-category schemes (Major Claim, Claim, and Premise) commonly used in prior research. Consequently, it is difficult to define the nature of relationships between components, as simple labels such as support or attack that do not adequately represent the complex interactions among the WAC categories. To further simplify the analysis, we adopt the assumption proposed by Persing and Ng [32] that treats each sentence as an independent argumentation unit, avoiding complications caused by overlapping annotation boundaries of multiple categories. Unlike Wang et al. [28], we do not treat the WAC categories as ordinal, since they are not comparable in nature and purpose.

CHAPTER 3

PROPOSED MODEL

To ensure cross-topic evaluation, we do not split the dataset directly based on the size. We do it based on Essay IDs so that while training the model does not have access to the topic-information of the essays used in the validation and testing set. This data split is constant throughout all the experiments conducted as part of this research to ensure fair performance comparison.

3.1 CONTEXT AIDED CLASSIFICATION MODEL

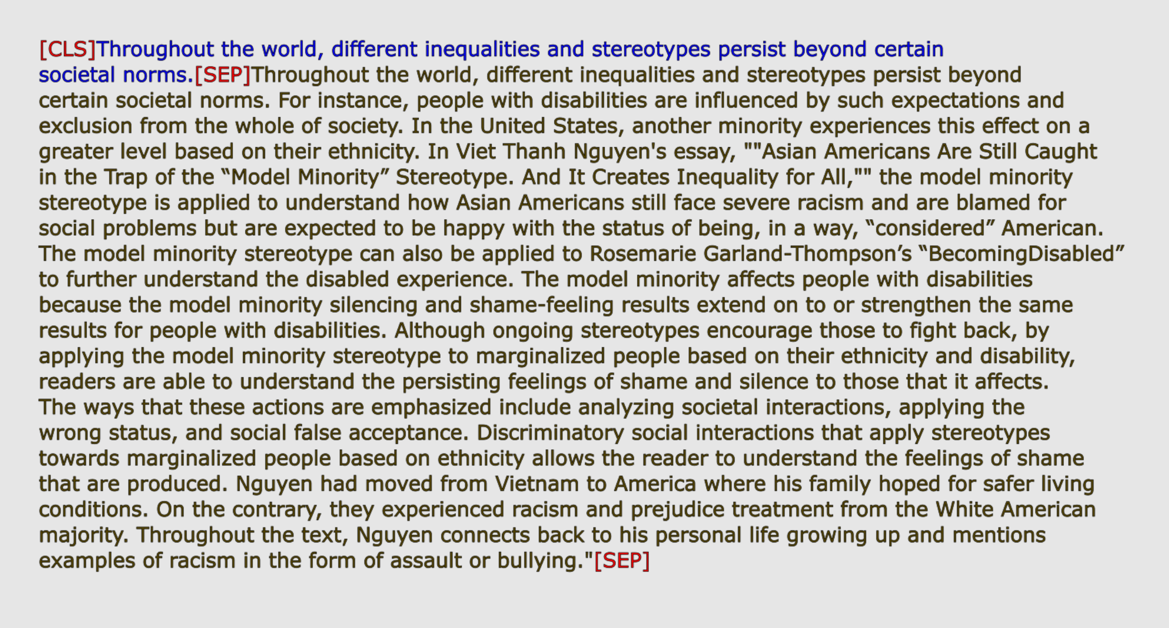
In this experiment, we investigated whether incorporating contextual information and class rebalancing techniques can improve performance on minority categories in sentence-level argument classification.

First, we applied inverse class-weighted cross-entropy loss during training. This re-weighting penalizes misclassification of minority samples more heavily than majority ones, thereby encouraging the model to focus on under-represented categories. We avoided other balancing techniques like perfect balancing, under-sampling, and over-sampling due to the extreme disparity between class sizes. Perfect balancing would result in severe data loss from dominant categories, while oversampling could introduce excessive synthetic noise.

Second, we introduced contextual pairing by concatenating each sentence with a portion of the essay that includes the sentence. This reformulated the task as a text-pair classification problem, where the first input sequence contains the focal sentence and the second sequence provides contextual information. These pairs were passed into BERT as

[CLS] sentence [SEP] context [SEP]

Figure 3.1 shows a sample sentence and context pair used in the experiment. The text in blue is the sentence that will be classified into argumentation categories. The text enclosed between the two [SEP] tokens in red is the context. The sentence is part of the context, in this case it is the very first line of the context.



[CLS]Throughout the world, different inequalities and stereotypes persist beyond certain societal norms.[SEP]Throughout the world, different inequalities and stereotypes persist beyond certain societal norms. For instance, people with disabilities are influenced by such expectations and exclusion from the whole of society. In the United States, another minority experiences this effect on a greater level based on their ethnicity. In Viet Thanh Nguyen's essay, ""Asian Americans Are Still Caught in the Trap of the “Model Minority” Stereotype. And It Creates Inequality for All,"" the model minority stereotype is applied to understand how Asian Americans still face severe racism and are blamed for social problems but are expected to be happy with the status of being, in a way, “considered” American. The model minority stereotype can also be applied to Rosemarie Garland-Thompson’s “BecomingDisabled” to further understand the disabled experience. The model minority affects people with disabilities because the model minority silencing and shame-feeling results extend on to or strengthen the same results for people with disabilities. Although ongoing stereotypes encourage those to fight back, by applying the model minority stereotype to marginalized people based on their ethnicity and disability, readers are able to understand the persisting feelings of shame and silence to those that it affects. The ways that these actions are emphasized include analyzing societal interactions, applying the wrong status, and social false acceptance. Discriminatory social interactions that apply stereotypes towards marginalized people based on ethnicity allows the reader to understand the feelings of shame that are produced. Nguyen had moved from Vietnam to America where his family hoped for safer living conditions. On the contrary, they experienced racism and prejudice treatment from the White American majority. Throughout the text, Nguyen connects back to his personal life growing up and mentions examples of racism in the form of assault or bullying."[SEP]

Figure 3.1: Example - Sentence and Context Pairing

Figure 3.2 illustrates the flow of this experiment. Allowing the model to use bidirectional attention across both sequences. This setup reasons with the fact that argumentative role is often determined by discourse placement, sentence transitions, or rhetorical cues within local surrounding text. An important point to mention in our context-pairing setup is that the contexts were constructed with variable lengths, rather than using a fixed number. This design choice is backed by two key factors:

1. **Token limit constraints in BERT:** BERT models are limited to a maximum input length of 512 tokens. Since student sentences vary significantly in length,

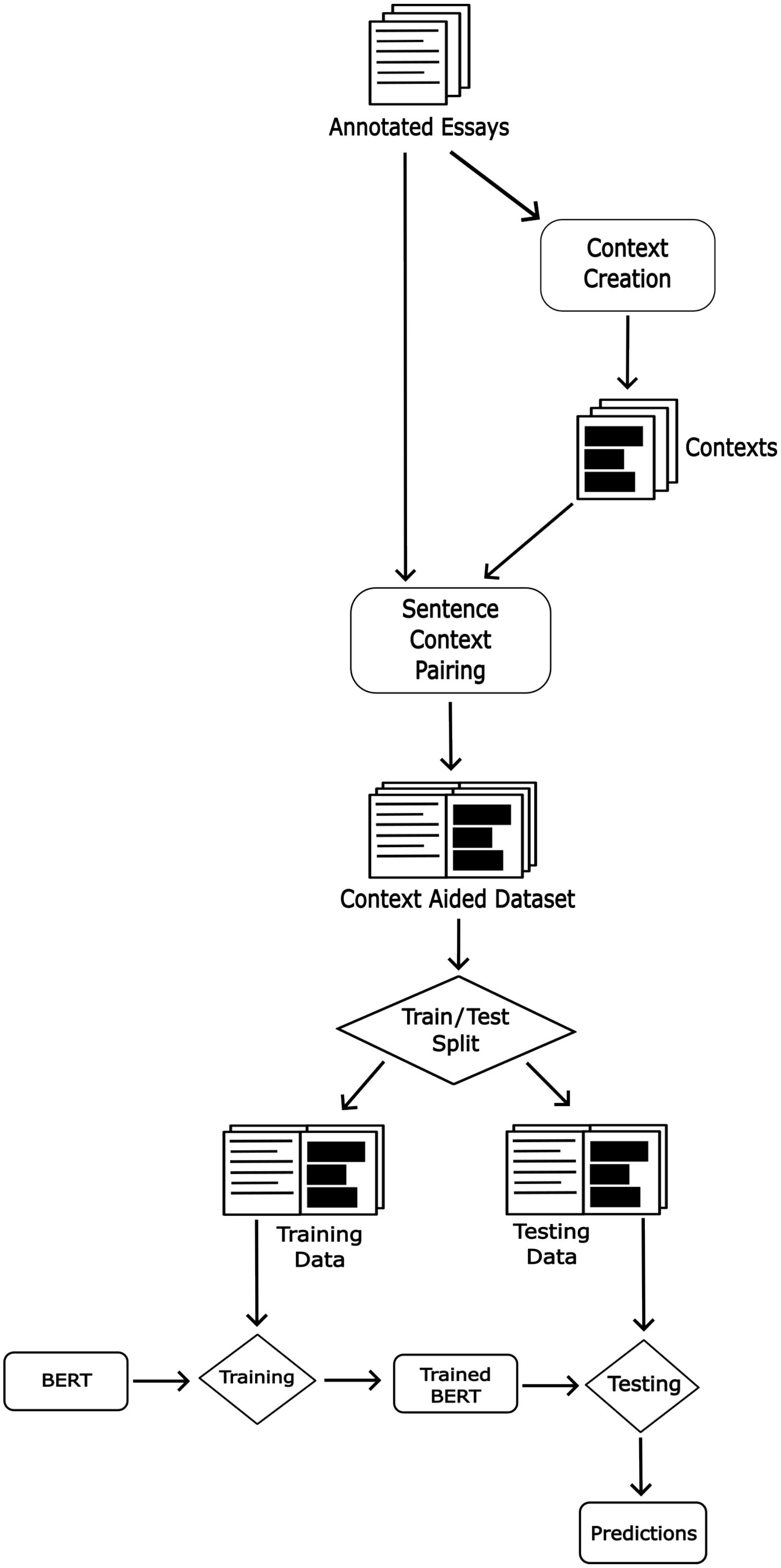


Figure 3.2: Context Aided Classification Model

using a fixed-size context window risks exceeding this limit, which would result in truncation and the loss of important information from the surrounding context.

2. **Generalizability and robustness:** Allowing variable-length inputs encourages the model to process sentences in a more context-aware and semantically driven manner, rather than learning artificial cues from fixed input structures. Fixed-length contexts may inadvertently teach the model to rely on positional heuristics or token count as a predictive signal. Studies have shown that transformer models are susceptible to length bias, where models inadvertently associate specific output classes with input length or position rather than meaning [53, 54]. Adding variability in input length forces the model to attend to content and meaning rather than form, enhancing its ability to generalize to structurally diverse inputs.

We also did a third experiment which combines both these approaches of adding weighted loss and pairing each sentence with a context. (Table 4.5 summarizes the results of these experiments across all categories.)

3.2 SYNTACTICAL INFORMATION AIDED CLASSIFICATION MODEL

In this experiment we explore if there are ways to improve the performance based on individual sentences only. Mainly by making the model focus on syntactic information, such as parts-of-speech in addition to semantic information. The general motivation is that syntactic features can offer additional structural cues that complement BERT's contextual embeddings. Syntactic information, such as part-of-speech (POS), tags play a significant role in improving argumentation mining, particularly in persuasive essays. Early models used syntactic features as inputs to classifiers or sequential models for identifying argumentative components and their relationships [35,46]. These structural cues were helpful for recognizing argument boundaries and distinguishing roles like claims and evidence. With the rise of transformer-based models like BERT, research began incorporating syntactic information into the architecture itself. Approaches include feature augmentation, where syntactic indicators are embedded with token representations [48], and graph-based methods, such as Tree-Constrained GNNs, that encode syntactic hierarchies through constituency parses [49]. These methods have shown empirical gains, particularly in structured academic writing.

However, graph-based and rule-driven approaches are often limited in generalizability. They tend to require corpus-specific designs that depend on the annotation schema and linguistic regularities of the dataset at hand. For instance, syntactic trees effective on one dataset may not be transferred well to others. This brittleness underscores the need for more adaptable architectures capable of learning structural and contextual features directly from data [47]. Generalizable models minimize reliance on handcrafted or tightly coupled features and better support domain transfer and scalability. Keeping generalizability in mind, we experimented with three ways of injecting syntactical information,

1. **Appending POS tags to each word:** To the input text we added its POS tag in the following format,

   WORD / POS-TAG

   Figure 3.3 shows an example of appending POS Tags to words.

2. **Replacing words with POS tags:** Instead of appending the words with POS tags, we replaced the words with their respective POS tags. Figure 3.4 shows an example of replacing words with POS Tags.

3. **Adding POS Embedding:** Originally the final BERT embedding is calculated in the following way:

   Initial Embedding = Token Embedding + Positional Embedding + Segment Embedding

   We modified this by adding a custom embedding that represents syntactical information:

   Final Embedding = Token Embedding + Positional Embedding + Segment Embedding + POS Embedding

To calculate the initial POS Embedding, I mapped all the tags to a unique integer and then added them to the word embeddings. The embedding values of these are learned in the training process. Figure 3.5 shows how the POS embedding is added BERT embedding.

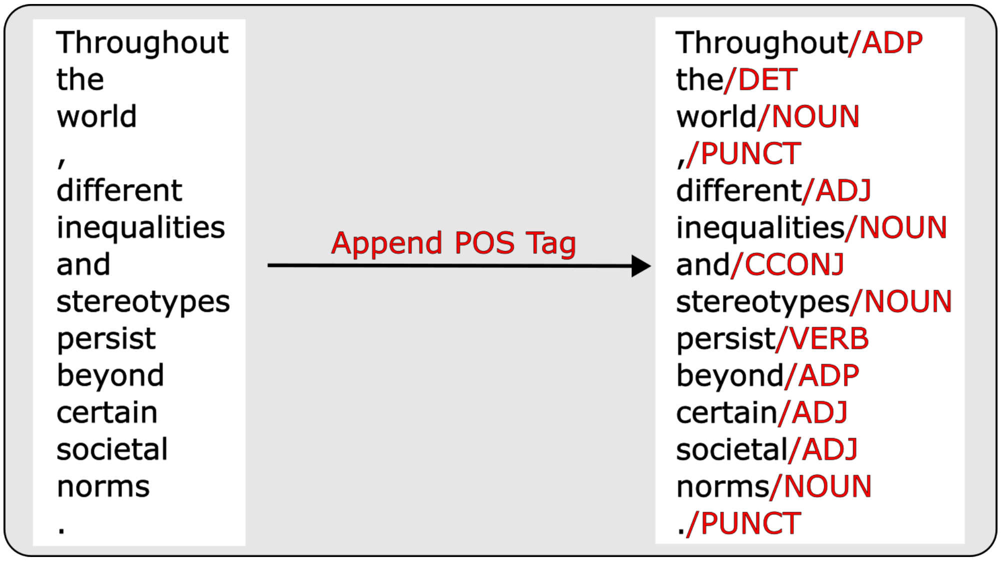


Figure 3.3: Example - Appending words with POS Tags

The first two approaches can be considered as Horizontal Augmentation of POS Tags, as we are adding the tags horizontally to the input text increasing the input

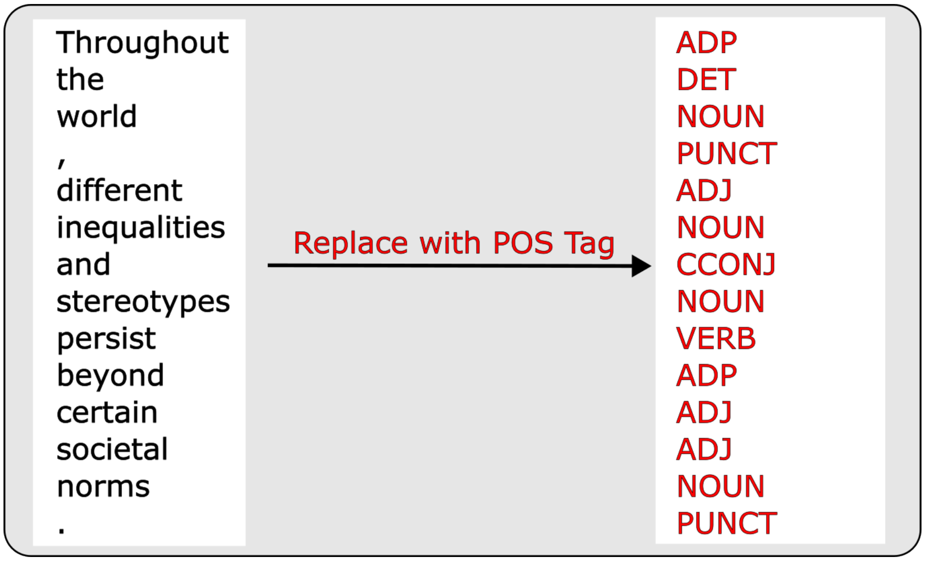


Figure 3.4: Example - Replacing words with POS Tags

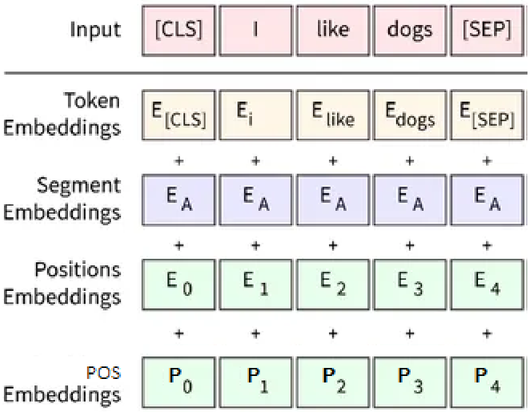


Figure 3.5: Adding Parts-of-Speech Embedding

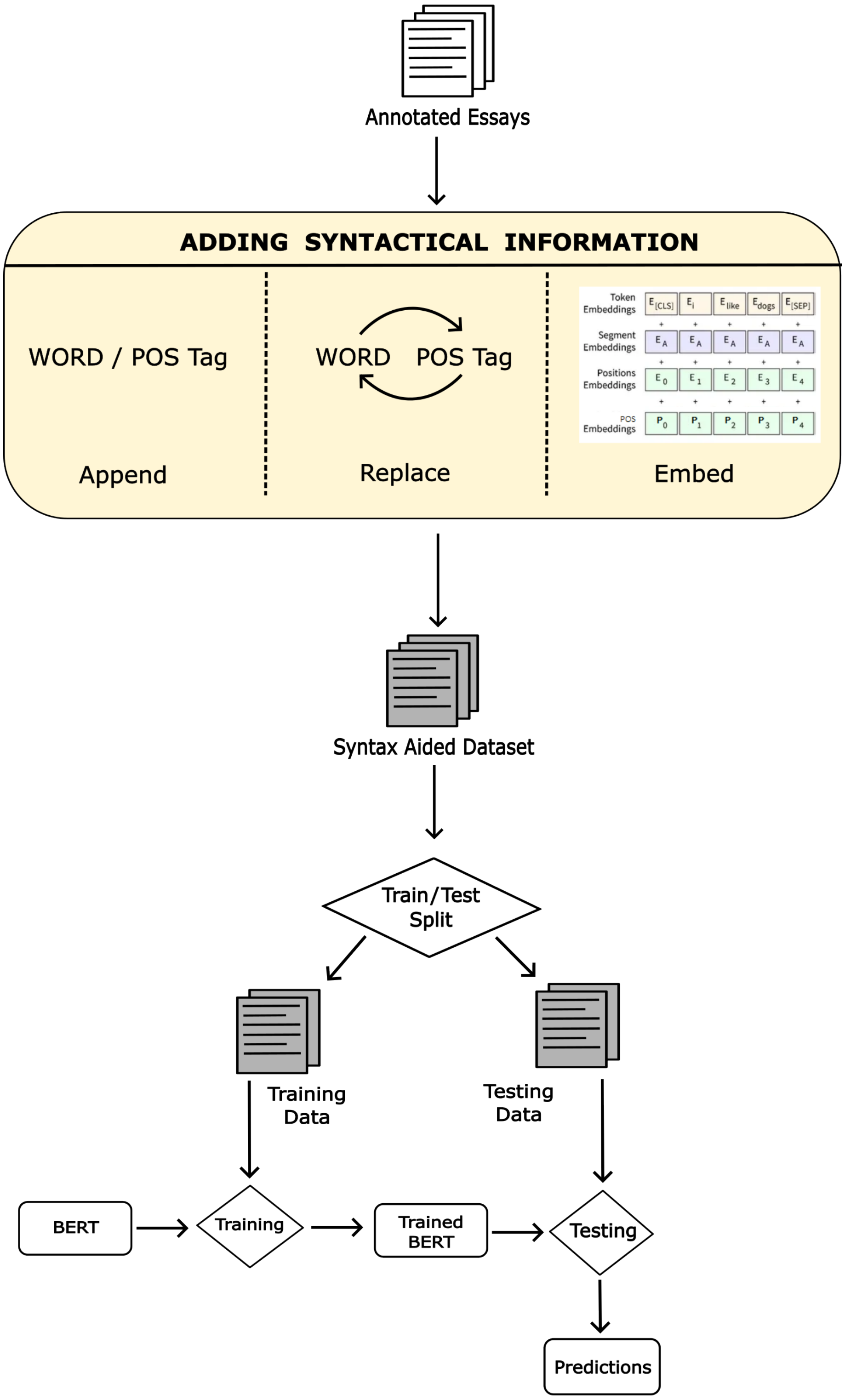


Figure 3.6: Syntax Aided Classification Model

text length. The third approach can be considered as Vertical Augmentation of the tags. In this, we numerically add the tag information to each word keeping the length of the input text the same. Figure 3.6 illustrates the flow of these experiments. These experiments use weighted loss to put extra emphasis on minority categories. (The results of these experiment are shown in Table 4.6)

3.3 DISTRIBUTION AWARE HIERARCHICAL CLASSIFICATION MODEL

Now we look at another approach which tries to tackle the imbalanced dataset problem, which is, distribution based or distribution aware hierarchical classification models. Distribution aware hierarchical models are ensemble classification systems that construct class hierarchies based on data distribution rather than predefined taxonomies. While general classifiers rely on semantic groupings or fixed label structures, these models reorganize classes to ensure more balanced sample distributions at each decision node. This approach is particularly effective in imbalanced classification tasks such as ours, as it prevents minority classes from being overshadowed by dominant ones and allows classifiers to focus on more binary distinctions. By recursively splitting labels to maintain distributional balance, distribution aware hierarchical models mitigate issues like data sparsity and class confusion common in flat or semantically clustered hierarchies. Compared to one-vs-rest or flat classifiers, distribution-aware ensembles provide improved robustness and modularity, especially when dealing with large and uneven label sets.

We implemented a three-level distribution based hierarchical classification model as shown in figure 3.7.

As the majority category is Normal Text, the top or the root classifier classifies a text into Normal Text and Other which is the group of all other categories. If this model predicts a sample as Normal Text, then it is taken as the final prediction, but if not, the sample is sent to the second classifier which classifies sample

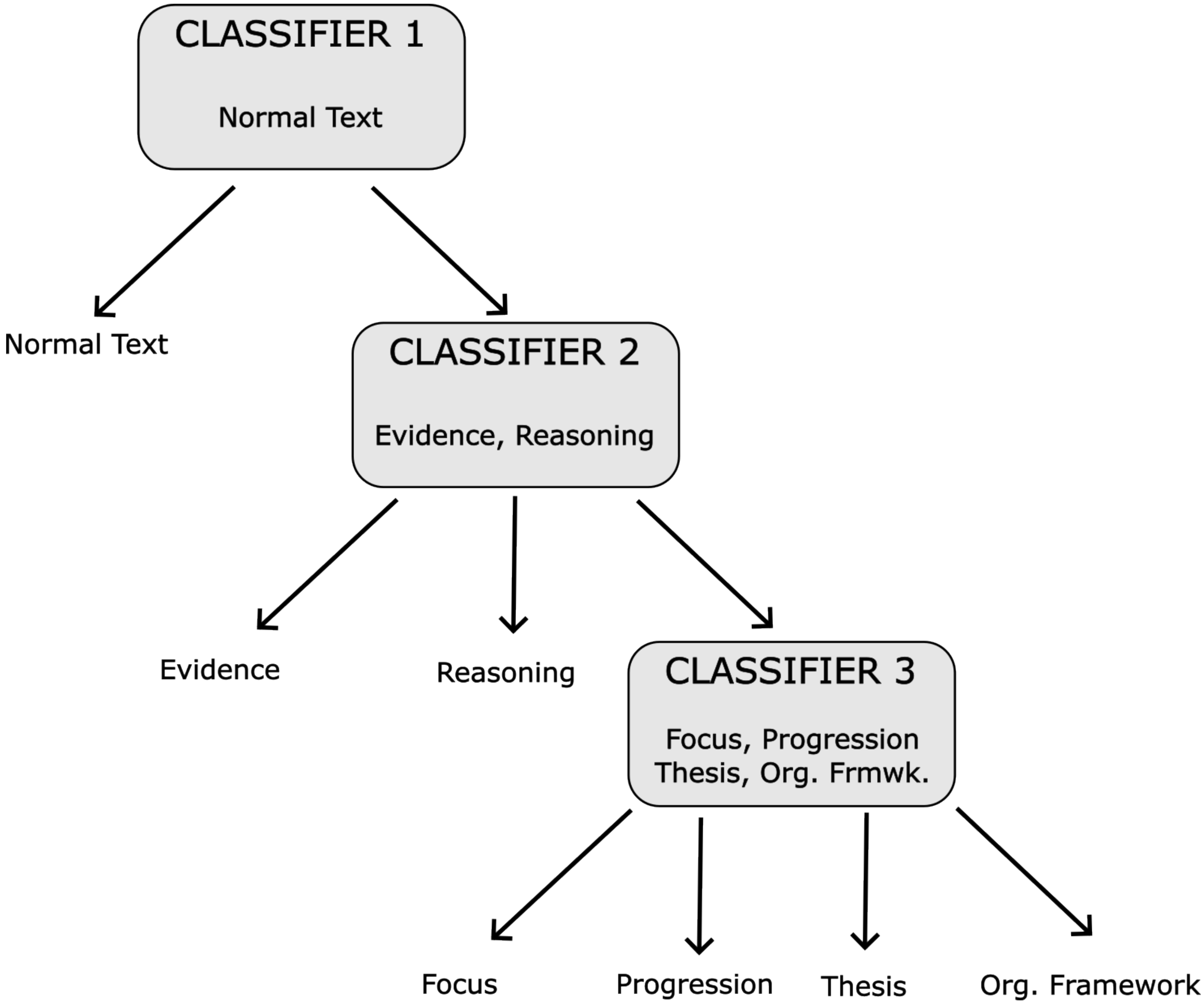


Figure 3.7: Distribution Aware Hierarchical Classification Model

into three categories, Reasoning, Evidence and Other. Similar to the root classifier, if the output prediction is Reasoning or Evidence it is taken as final and if not, the sample is sent to the third classifier which classifier the sample into the remaining four categories — Focus, Thesis, Progression and Organizational Framework. All the classifiers are designed such that the class labels they are predicting are uniformly distributed. The grouping of categories across the three levels was informed not only by label distribution but also by observing confusion patterns in the baseline model. Categories such as "Reasoning" and "Evidence" frequently overlap in linguistic function and position in student essays, making them ideal for joint classification. In contrast, highly under-represented categories like "Focus" or "Organizational Framework" benefit from being isolated at the final stage, where the classifier can attend to their finer distinctions without the noise of dominant classes. It is also important to note that the three classifiers are basic vanilla BERT based classifiers with no weighted-loss and are trained on the same training data. Figure 3.8 illustrates the flow of this experiment. (The results of this experiment are shown in Table 4.7)

In summary, we explored three distinct strategies in this study to address the challenges posed by sentence-level argument classification in a highly imbalanced educational corpus. The Context-Aided Classification Model emphasized the importance of discourse context by framing the task as a sentence-context pairing problem, showing how surrounding information influences argumentative role. The Syntactic Information Aided Model focused on aiding input representations with structural cues such as part-of-speech tags, leveraging both horizontal and vertical augmentation techniques to integrate syntactic awareness into BERT-based models. Finally, the Distribution Aware Hierarchical Model tackled the core imbalance problem through a tiered classification architecture driven by label frequency, enabling classifiers to make decisions within more uniform and manageable subsets. Each approach targeted a specific aspect of the task—contextual understanding, structural depth, or distribu-

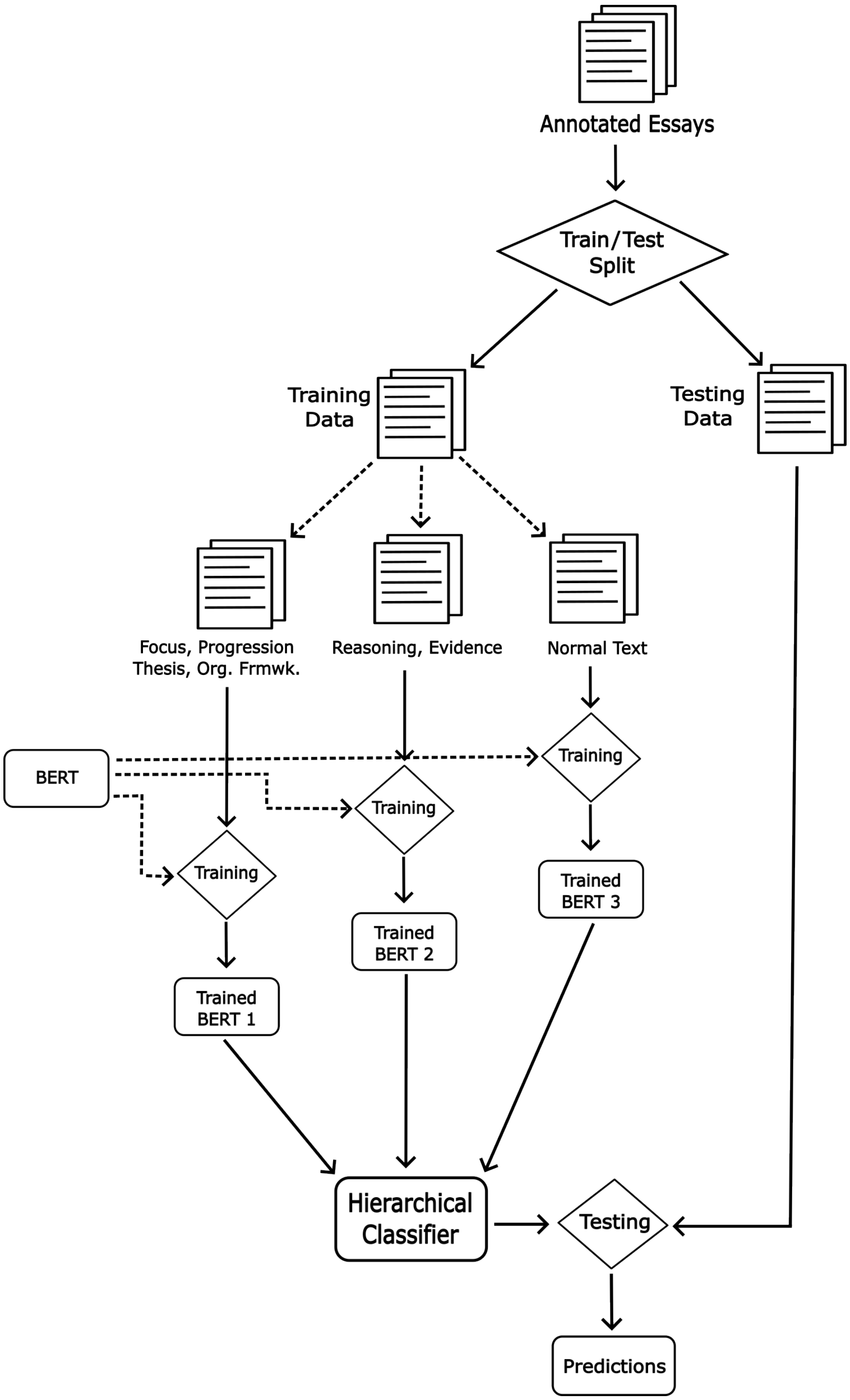


Figure 3.8: Distribution Aware Classification Model

tional fairness—and was evaluated on its ability to improve classification performance, particularly for under-represented argumentative categories.

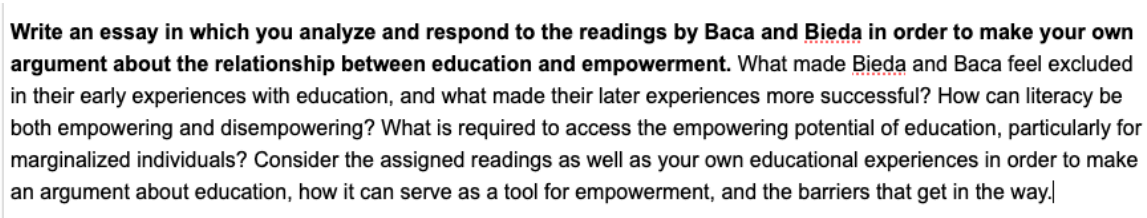
CHAPTER 4
EXPERIMENTS AND RESULTS

4.1 DATA COLLECTION

In this section, we will describe the data collected for this study, the instrument used to assess sample student papers, the WAC rating process, and the annotation process. In order to fulfil state-mandated General Education Curriculum outcomes for written communication, the WAC program at FAU provides a set of writing guidelines that can be adopted across curriculum of diverse disciplines, and assesses thesis-driven, research-based, near-end-of-term writing assignments. The WAC program defines "thesis-driven" as "papers with a thesis in which you build a case for a particular analysis, interpretation, or evaluation of data that leads to recommendations or specific conclusions," and defines "reading-based" as papers that "draw upon argument-driven articles or book chapters or in some cases works of literature. Typically, papers that are thesis-driven and reading-based are research projects of some kind" [24].

4.1.1 Data Set: College Writing 1 and 2

While the WAC program assesses student writing from a variety of disciplines and course levels, which we intend to examine in future work, in this study we specifically examined thesis-driven, reading-based student writing in ENC 1101: College Writing 1 and ENC 1102: College Writing 2. This approach sought to focus on core writing competencies that form the basis of writing in the general education curriculum. In College Writing 1 and 2, students write analytical argument-based essays of 1000+ words in order to demonstrate skills in analytical thinking and reasoning in response



Write an essay in which you analyze and respond to the readings by Baca and Bieda in order to make your own argument about the relationship between education and empowerment. What made Bieda and Baca feel excluded in their early experiences with education, and what made their later experiences more successful? How can literacy be both empowering and disempowering? What is required to access the empowering potential of education, particularly for marginalized individuals? Consider the assigned readings as well as your own educational experiences in order to make an argument about education, how it can serve as a tool for empowerment, and the barriers that get in the way.

Figure 4.1: Example of an ENC 1101 Essay Prompt

to readings on contemporary topics in genres including memoir, creative non-fiction, editorials, and long-form journalism.

Student essays are evaluated for demonstration of rhetorical skills tied to course outcomes including argument and reasoning, evidence and support, organization, language and style, and meeting audience needs. Success in these assignments requires students to interpret assigned readings, research additional sources appropriate for their analytical purpose, and use textual evidence from the sources by paraphrasing and quotating to support their analytical argument. In such essays, students typically state their central argument in their introductory paragraph in the form of a thesis statement, and also indicate the shape of their argument and the evidence they will use as part of their introduction and thesis. Arguments are analytical, not empirical, and so evidence will primarily desire from the texts, but students may also include personal experience as part of their interpretation of textual evidence. Students may shift between registers of formality throughout their essay, introduce slang or writing from different languages or dialects, and alternate between first, second and third persons in order to persuade the reader of their analytical argument.

4.1.2 WAC Data Annotation

In this study we focused on core writing skills specific to College Writing 1 and 2 courses. As with [17] where a large textual corpora is manually analyzed and annotated, we also annotated texts. Unlike [17], however, we did not adopt argumentative structures of premises and conclusions. Instead, we adopted annotation labels that aligned with elementary units identified in WAC and English Composition assessment rubrics. As noted in Table 4.1, these elementary units aligned with writing genre common to English Composition courses and WAC assessment objectives.

Because argumentation is the primary skill taught in College Writing 1 and 2 courses, the annotation procedure adopted in our study identified six core writing skills in College Writing 1 and 2 essays that align with five WAC rubric categories: thesis, organizational framework, use of evidence, analysis of evidence, reasoning, evidence, and rhetorical structure. Table 4.1 provides descriptions of the six core writing skills annotated by the raters.

This annotation procedure enabled us to break the WAC category of rhetorical structure into two components: raters annotated 1) rhetorical focus to identify where a student paper returned the reader's attention to the argument's central idea, and 2) rhetorical progression where the paper identified the argument's organization. This is an especially appropriate distinction in writing intensive courses where students write thesis-driven and reading-based extended essays.

In order to develop a fixed model of argumentation that could be applied to our corpus, we adopted a monological model (or Dialectic Argumentation), whereby we sought to identify the elementary units. In other words, each rater who scored a student essay also annotated it by labelling every sentence into the six categories mentioned, as defined in Table 4.1. The sentences which did not belong to any of the categories were labelled as "Normal Text".

Although, in few cases it so happened that the raters found that few sentences could not be labelled under a single category. A portion of the sentence, then, was annotated under one category and the remaining portion was labelled as another category. At times, some sentences contained three categories. We will discuss how

Table 4.1: Annotated Categories

Core Writing Skills	**WAC Categories**	**Descriptions**
Thesis	Thesis	Persuasive purpose of the paper
Organizational Framework	Organizational Framework	A statement identifying the argument’s organization
Inclusion of Textual Evidence	Evidence	Paraphrasing and quotations
Analysis of Evidence	Reasoning	Analysis of evidentiary materials, especially quotations
Rhetorical Focus	Rhetorical Structure	Statements identifying the argument’s central idea
Rhetorical Progression	Rhetorical Structure	Statements identifying the argument’s organization

we tackled this issue in the following subsection.

4.2 DATA PREPARATION

Fig. 4.2 lists the proposed framework for consistency evaluation and filtering.

4.2.1 Data Preprocessing

As each essay is analysed by more than one rater, it is natural to have competing opinions on sentence annotation. One sentence can be categorized under multiple categories by different raters. This problem became more severe when a sentence, in itself, contributed towards multiple categories (usually long sentences). We addressed this issue by breaking down the sentence into individual words and associated each word with its corresponding label. Then, we counted the number of words associated with each label and assigned that label to the sentence which had the maximum word count. In a similar fashion, after a label was assigned to a sentence, we compared the final label assigned to the sentence by each rater and selected the label that was the most frequent. A natural question one might ask is, "what if none of the raters agree upon a single category?" For cases like these, we proceeded with the annotation of the rater who showed the greatest consistency in their annotations.

4.2.2 Consistency Evaluation

Evaluating the consistency of the raters was an important step as erroneous annotations can cause improper training of the machine learning model and faulty predictions. It was crucial, then, to check the similarity between the annotations of raters and consider only the group of raters whose similarity reached a certain threshold. This process of measuring the consistency and agreement between two or more raters in their assessments, judgements or evaluation of any phenomenon or behavior is called Inter-Rater Reliability (IRR). And in the context of annotation, this is known

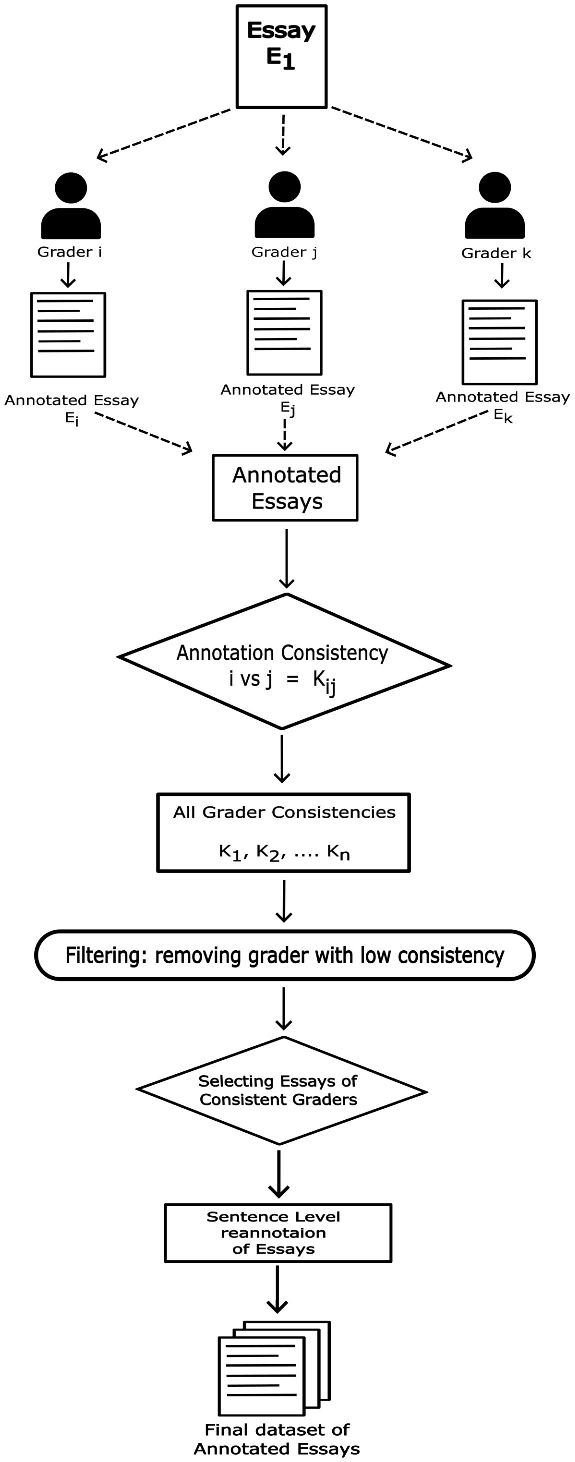

Essay E_1
Grader i
Grader j
Grader k
Annotated Essay E_i
Annotated Essay E_j
Annotated Essay E_k
Annotated Essays
Annotation Consistency i vs j = K_{ij}
All Grader Consistencies
K_1, K_2, ... K_n
Filtering: removing grader with low consistency
Selecting Essays of Consistent Graders
Sentence Level reannotaion of Essays
Final dataset of Annotated Essays

Figure 4.2: Annotation Consistency Evaluation

as Inter-Annotation Agreement (IAA).

There are numerous metrics by which one can gauge the level of agreement between raters, such as Cohen's Kappa [9], Fleiss's Kappa [10] and Krippendorff's Alpha. Cohen's Kappa is used to compute the agreement between no more than two raters, while Fleiss's Kappa and Krippendorff's Alpha can be used to compute the agreement between multiple raters for a given set of annotation. We chose Cohen's Kappa as our consistency evaluation metric as we needed to measure the average degree of consistent annotation of each rater with every other rater as a tie-breaking criterion for ambiguous annotation rather than just the overall agreement of raters for every essay. It is given by,

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.1}$$

where $\mathrm{p_o}$ is the relative observed agreement among raters, i.e., the fraction of total number of sentences that same (or agreed) annotations by both raters.

$$p_o = \frac{\text{\# of agreement sentences}}{\text{Total number of sentences}} \tag{4.2}$$

and $\mathrm{p_e}$ is the hypothetical probability of chance agreement,

$$p_e = \sum_{i=1}^{K} \left(\frac{\mathrm{C}_1^i \times \mathrm{C}_2^i}{\text{Total number of sentences}^2} \right) \tag{4.3}$$

where K is the total number of categories (the labels of sentences in our case). Where $\mathrm{C_1}^{\mathrm{i}}$, $\mathrm{C_2}^{\mathrm{i}}$ are, for any given essay, the number of sentences annotated as label i.

To demonstrate the application of Cohen's Kappa for IAA, a brief example is shown below. Table 4.2 illustrates four sentences where two raters, G_1 and G_2, are asked to annotate the essay based on two categories — A and B.

Calculations for the level of agreement of the two raters, G_1 and G_2, using Cohen's Kappa metric employ the following steps: We initially calculate the relative observed

Table 4.2: An example of Cohen's Kappa score calculation

Sentences	**Rater** (G_1)	**Rater** (G_2)
Sentence 1	A	A
Sentence 2	B	B
Sentence 3	A	A
Sentence 4	A	B

agreement $\mathrm{p_o}$ between raters. Since both raters agreed on three sentences in total, sentence 1, 2 and, 3, the relative observed agreement is first set-up as follows:

$$p_o = \frac{3}{4}$$

The next step is to compute the probability of agreement by chance. According to equation 3 we proceed with the following:

$$p_e = \sum_{i \in [A,B]} \left[\frac{\mathrm{C}_1^i \times \mathrm{C}_2^i}{\text{Total number of sentences}^2} \right]$$

$$= \frac{1}{4^2} \times [(3 \times 2) + (1 \times 2)]$$

$$p_e = \frac{1}{2}$$

Now that we have $\mathrm{p_o}$ and $\mathrm{p_e}$, the values are substituted in equation 1 to arrive at the Kappa value:

$$\kappa = \frac{\frac{3}{4} - \frac{1}{2}}{1 - \frac{1}{2}}$$

$$\kappa = \frac{1}{2}$$

Hence, in this brief illustration, the raters G_1 and G_2 have an agreement of 50%. In our study, this process was carried out for every pair of essay raters for every single essay.

Table 4.3 shows the interpretation for kappa values introduced by Landis and Koch [12]. In our study, then, we adopted the approach discussed above to calculate the consistency of each rater in the following way:

1. For each essay, we calculated Cohen's Kappa Score for every pair of raters. Every score contributed to the consistency of both raters.

2. We then averaged the all the Cohen's Kappa values associated with each rater to compute their individual consistency scores.

Figure 4.3 shows the consistency scores of all the raters involved in the scoring and annotation process. We see that all the raters are in the range of Fair to Moderate Agreement, with the exception of one rater who was in the level of Slight Agreement. As explained above, these consistency scores were also used to filter out raters who showed poor annotation since low consistency score means their annotations were considerably and frequently different from other raters. By only considering annotations from raters who showed consistency, we increased the quality of data.

4.2.3 Sentence Level Assessment

Since the nature of data we used is nominal, we treated this as a classification problem. As we were using a dialectic model of argumentation, it was crucial to use a model

Table 4.3: Interpreting Kappa Values

Kappa Range	**Interpretation**
<0.00	No Agreement
0.00 - 0.20	Slight Agreement
0.21 - 0.40	Fair Agreement
0.41 - 0.60	Moderate Agreement
0.61 - 0.80	Substantial Agreement
0.81 - 1.00	Perfect Agreement

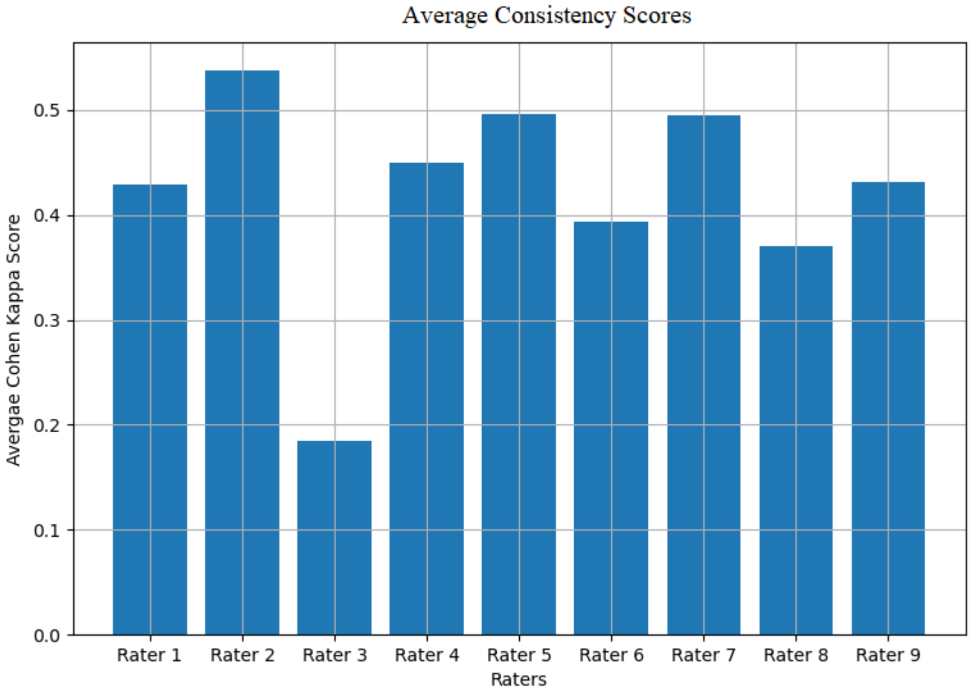


Figure 4.3: Consistency score of Raters. x-axis denotes individual rater, and the y-axis denotes each rater's average Cohen Kappa score.

that could capture the underlying patterns in a unit of text, i.e., an argument. Hence, we used a BERT model introduced by Devlin [5] which is built upon on a Transformer model introduced by Vaswani [6] in 2017. We fine-tune the model with our data of 5,064 labelled text sentences.

After calculating the consistencies of all the raters, we attempted to identify raters with a low consistency scores. As mentioned earlier, this helped us improve the quality of data. After identifying a list of consistent raters, we created a data pool of essays from these raters. This pool consisted of multiple versions of every single essay, which we then analyzed in order to finalize the category label for each sentence. We resolved labeling discrepancies by assigning each sentence to the category for which there was the highest rater agreement. In cases where none of the raters agreed on any category, sentences were labeled according to the label assigned by the rater with the highest consistency value. This process created a new data pool of essays consisting of a single version of each essay and final annotations that were used in our experiments.

Table 4.4 lists the dataset collected for the study, which consisted of sentences labeled in six categories (the table lists only sentences after low consistency raters were removed). The "Normal Text" category denotes sentences that were not labeled by raters. The dataset showed a clear imbalance, with "Normal Text" as the predominate category and "Organizational Framework" being the least represented category (which is 1/7 of the second dominant class "Evidence"). Due to this category distribution, we adopted precision, recall, and F-1 scores to assess the model performance. We use this refined data pool to conduct experiments in this study.

4.3 BASELINE MODEL

For the baseline model we used a simple BERT classifier that classifies the sentences into the WAC categories. The baseline model showed ghastly overall performance. The model was able to classify the majority categories like "Evidence" and "Normal

Table 4.4: A summary of number of sentences assessed by raters and the respective assessment categories

Sentence Assessment Label Categories	# of Sentences
Normal Text	2,590
Evidence	721
Reasoning	656
Rhetorical Structure (Focus)	563
Rhetorical Structure (Progression)	234
Thesis	192
Organizational Framework	108

Text" with a decent F1 score of 0.63 and 0.735 respectively, and a zero-performance on rest of the classes. Surprisingly, it did not identify a single sentence of "Reasoning which has more datapoints compared to "Evidence". The average precision, recall and F1 across all the classes are 0.173, 0.225 and, 0.195.

We deduced that this low performance can be attributed to two reasons. One obvious reason is the fact that the data is highly imbalanced. Another reason which inclines towards the linguistic aspect of the problem is that the model lacks contextual information when classifying the sentences. It does not have access to the words surrounding the sentence which play a crucial role in defining the nature and role of the sentence in the essay. In persuasive undergraduate essays, the surrounding context, especially discourse markers, plays a critical role in defining the argumentative function of individual sentences. Discourse markers such as "however", "but", "and" on the other hand typically signal counterarguments [35, 36], while causal or additive markers like "because", "in addition", and "furthermore" indicate supporting evidence [35, 36]. Concessive markers (e.g., although, admittedly) acknowledge limitations or opposing views [36]. Sequence markers such as “First of all” often introduce new argumentative points [36]. Corpus analyses of student essays confirm that such markers frequently indicate whether a sentence supports, opposes, qualifies, or transitions an argument [37]. Opitz and Frank [38] demonstrated that classifiers can often infer whether a sentence attacks or supports another based purely on local context cues (e.g., preceding conjunctions), without relying on deep sentence semantics. These findings underscore that context is vital for defining the argumentative nature of individual sentences in persuasive writing.

4.4 RESULTS AND ANALYSIS

Category	Baseline	Weighted Loss	Context Aided	Mixed
Evidence	0.630	0.588	0.658	0.489
Organizational Framework	0.000	0.054	0.000	0.029
Reasoning	0.000	0.251	0.182	0.294
Focus	0.000	0.091	0.000	0.045
Progression	0.000	0.000	0.000	0.007
Thesis	0.000	0.138	0.040	0.154
Normal Text	0.735	0.517	0.715	0.511

Table 4.5: F1 scores of Baseline, Weighted, Context-Paired and Mixed Model

The Mixed BERT model (using inverse-class weighting and contextual sentence pairs) improved F1 scores for categories with very few training instances. In the baseline model, low-resource classes such as *Thesis*, *Progression*, and *Org. Framework* had F1 scores of 0.000, indicating that the model never correctly identified these categories. This is typical in class-imbalanced learning scenarios where models are biased towards the majority classes and effectively ignore underrepresented ones [42].

After applying class weighting and adding context, the mixed model achieved non-zero F1 scores for all previously missed categories, with Reasoning rising to 0.294 and Thesis to 0.154. Even the smallest classes like *Org. Framework* showed tiny improvement (F1 = 0.029). These modest improvements are practically meaningful because they indicate the model is beginning to correctly recognize minority classes it previously ignored [42]. Contextual sentence input played a key role in these improvements. Prior research shows that including surrounding sentences helps disambiguate difficult cases. For example, [43] demonstrated improvements in medical sentence classification using neighboring context, and [44] achieved a significant F1 gain for rare discourse phenomena using context-aware BERT. These results support the notion that adding sentence context helped the model generalize better to low-resource categories in our setting.

While minority class performance improved, there was a drop in F1 scores for frequent categories such as *Normal Text* (from 0.735 to 0.511) and *Evidence* (from 0.630 to 0.489). This trade-off is a known consequence of cost-sensitive learning, emphasizing rare class detection can reduce the precision for well-represented classes [45]. [42] explains that models often overfit to frequent classes unless re-balanced, and that increasing attention to rare classes typically results in higher macro-F1 but lower micro-F1. Similarly, [45] observe that while macro-F1 increases through reweighting, the overall accuracy or performance on dominant classes can suffer.

This suggests that the mixed model achieves a more equitable performance distribution, even if it slightly compromises performance on majority categories. In tasks that value fairness or balanced prediction (such as argumentation component identification), such trade-offs are often desirable [44].

We see that incorporating syntactic information into BERT substantially improved

Category	Baseline	Appended POS	Replaced POS	POS Embedding
Evidence	0.630	0.613	0.634	0.711
Organizational Framework	0.000	0.076	0.109	0.086
Reasoning	0.000	0.328	0.289	0.359
Focus	0.000	0.060	0.157	0.206
Progression	0.000	0.000	0.017	0.020
Thesis	0.000	0.421	0.132	0.176
Normal Text	0.735	0.499	0.483	0.440

Table 4.6: F1 scores of Baseline, Appended POS, Replaced with POS and POS Embedding model

F1 scores for low-resource categories like Thesis, Focus, Reasoning, and Organizational Framework. These argument components often follow distinct structural patterns, which are not easily captured through semantics alone. Explicitly including part-of-speech (POS) tags helps the model better identify such patterns. Prior research shows that POS features aid in distinguishing argumentative roles, particularly when training data is limited or imbalanced [35]. By providing clear structural cues, syntactic signals enhance the model's sensitivity to under-represented argument components.

However, performance on the dominant class like Normal Text declined across all syntactically enhanced models. This class is broad, semantically diverse, and lacks consistent structural signatures. Introducing syntactic emphasis shifts the model's attention from meaning to form, potentially misclassifying semantically rich but structurally neutral sentences as argument components. Replacing words entirely with POS tags further degrades semantic understanding of the model.

As discussed earlier, trade-offs such as these reflects a common pattern in imbalanced classification, improving minority class recognition often comes at the cost of majority class performance. Still, for tasks like argument mining where structural understanding is crucial, syntactic augmentation proves valuable. It acts as an inductive bias that guides the model toward form-based distinctions while balancing the effects of class imbalance.

Category	**Baseline**	**DHCM**
Evidence	0.630	0.733
Organizational Framework	0.000	0.089
Reasoning	0.000	0.366
Focus	0.000	0.166
Progression	0.000	0.000
Thesis	0.000	0.307
Normal Text	0.735	0.588

Table 4.7: F1 scores of Baseline and Distribution aware hierarchical classification model

Distribution-aware hierarchical classification breaks the task into smaller subtasks, helping the model attend more effectively to low-resource categories. In flat classification, the dominance of “Normal Text” leads to skewed decision boundaries that suppress minority classes [50]. By filtering out “Normal Text” at the root, subsequent classifiers operate on more balanced data, allowing them to learn fine distinctions among underrepresented labels like “Reasoning” and “Thesis”. This layered structure enables the model to specialize in subsets of labels without interference from majority class noise. Each classifier in the hierarchy handles fewer, more evenly distributed classes, thus improving recall and precision for rare categories [51]. This setup ensures that minority classes are no longer misclassified due to overwhelming presence of dominant ones. Research confirms that such hierarchical decompositions reduce class confusion and make minority classes more learnable [51].

The drop in "Normal Text" F1 is expected due to a trade-off between recall of minority and majority classes. At the root level, the classifier becomes more sensitive to potential minority instances to reduce false negatives, sometimes misclassifying true "Normal Text" as "Other". These errors propagate down the hierarchy and cannot be corrected later, a known issue in hierarchical models [51]. Additionally, prioritizing minority class detection inherently shifts the model's capacity and decision thresholds, leading to a drop in majority class performance [52]. This is similar to the effect seen with weighted loss or sampling methods: gains for rare categories often come at the cost of majority accuracy [52]. In this case, the hierarchical structure implicitly performs that rebalancing.

4.5 SUMMARY

This study explored multiple modelling strategies to improve the identification of argumentation components in undergraduate persuasive essays, addressing key challenges such as extreme class imbalance and the lack of contextual or syntactic cues in standard BERT-based classification. The baseline model demonstrated reasonable performance for dominant categories like Normal Text and Evidence but failed to recognize any instances from low-resource categories such as Reasoning or Thesis. To overcome this, the research introduced context-aided models using sentence-context pairs and inverse-weighted loss. These approaches notably improved the model's sensitivity to minority classes, with the mixed variant achieving measurable F1 gains in Reasoning and Thesis, though at the cost of reduced performance on majority categories, an expected behaviour in cost-sensitive learning.

The study also tested the effect of syntactic information through POS-tagging and embeddings. Among the syntactic models, the POS Embedding approach showed the most promising improvements for under-represented argumentative categories, further supporting the utility of structural cues. Finally, a distribution-aware hierarchical classification model was implemented to restructure the prediction task based on label distribution rather than semantics. This method yielded the most balanced improvements, significantly boosting F1 scores for Reasoning and Thesis while avoiding reliance on weighted loss. Although performance on Normal Text declined due to the cascading nature of hierarchical errors, this method offered a compelling structural solution to the class imbalance problem.

CHAPTER 5
CONCLUSION & FUTURE WORK

5.1 CONCLUSION

This thesis explored the task of argumentation component identification in undergraduate persuasive essays, a complex and under-addressed problem in computational linguistics. Using the WAC dataset, which includes multiple nuanced argument categories, the work examined several strategies to overcome the two key challenges: severe class imbalance and limited contextual awareness in sentence-level models.

The baseline BERT model served as a reference point, highlighting its inability to detect minority argument categories like Progression or Thesis, despite decent performance on dominant classes such as Evidence and Normal Text. To address this, multiple approaches were evaluated. Context-aided models paired sentences with surrounding text to provide discourse-level cues, and syntactic information was incorporated through POS tagging and custom embeddings to enrich structural understanding. Both strategies demonstrated that structural and contextual signals significantly aid in recognizing low-resource categories. Among these, the POS Embedding variant and the Mixed context-aware model showed notable improvements. However, the most balanced and scalable performance came from the proposed distribution-aware hierarchical classification framework. By restructuring the prediction task around data distribution rather than semantic similarity, this model improved recognition of rare argumentative elements without requiring class reweighting or architectural complexity.

The thesis demonstrates that incorporating linguistic indicators and distribution-aware architectural design can substantially enhance performance in challenging, imbalanced, and semantically nuanced classification tasks like argumentation mining in student essays.

5.2 FUTURE WORK

There are several ways to extend the work of this thesis. While the hierarchical classifier improved minority class detection, it still suffered from error propagation, misclassifications at higher levels could not be corrected downstream. Future work could explore soft-hierarchy or multi-path routing strategies, allowing model outputs to be reconsidered jointly across levels. Additionally, integrating uncertainty estimation or confidence-based thresholds might help reduce false positives in the hierarchical splits.

Another promising direction is the application of prompt-based learning and instruction-tuned large language models (LLMs), which can generalize well across low-resource settings without extensive retraining. Leveraging pre-trained language models (like T5 or GPT) with in-context learning may show promising results, especially for zero-shot or few-shot classification of underrepresented categories.

And as this thesis focused on sentence-level classification, future research could tackle the task at the paragraph or document level, using multi-granularity modelling. Incorporating discourse structure or graph-based representations (e.g., rhetorical structure theory trees or dependency graphs) may also enable deeper understanding of argumentative flow.

BIBLIOGRAPHY

[1] Finley, A. (2021). How College Contributes" to" Workforce Success: Employer Views on What Matters Most. Association of American Colleges and Universities.

[2] Bazerman, Charles and Little, Joseph (2005). Reference guide to writing across the curriculum, *Parlor Press LLC*, 2005.

[3] Michelle Cox, Jeffrey R. Galin, Dan Melzer. (2018). *Sustainable WAC: A Whole Systems Approach to Launching and Developing Writing Across the Curriculum Programs,* National Council of Teachers of English.

[4] Van Eemeren, F. H., Grootendorst, R., Johnson, R. H., Plantin, C., & Willard, C. A. (2013). Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments. Routledge.

[5] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.'

[7] Lindahl, A.,& Borin, L. (2024). Annotation for computational argumentation analysis: Issues and perspectives. Language and Linguistics Compass, 18(1), e12505.

[8] Ling, G., Elliot, N., Burstein, Jill., McCaffrey, D., MacArthur, C., & Holtzman, S. (2021) Writing Motivation: A Validation Study of Self-Judgment and Performance. Assessing Writing, 48, 100509.

[9] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

[10] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

[11] Conference on College Composition and Communication Executive Committee. (2023). Principles for the Postsecondary teaching of writing. National Council of Teachers of English.

[12] Landis JRKoch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159174.

[13] McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written communication, 27(1), 57-86.

[14] Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning.

[15] Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal Intelligent Tutoring System: Usability Testing and Development. *Computers and Composition*, *34*(Complete), 39–59.

[16] Butterfuss, R., Roscoe, R. D., Allen, L. K., McCarthy, K. S., & McNamara, D. S. (2022). Strategy uptake in writing pal: Adaptive feedback and instruction. Journal of Educational Computing Research, 60(3), 696-721.

[17] Palau, R. M., & Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 98–107. https://doi.org/10.1145/1568234.1568246

[18] Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in Argumentative Writing Classrooms. Assessing Writing, 57, 100752.

[19] Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. Heliyon, 10(14).

[20] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[21] Chen, Y. Y., Liu, C. L., Lee, C. H., & Chang, T. H. (2010). An unsupervised automated essay-scoring system. IEEE Intelligent systems, 25(5), 61-67.

[22] Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined?. International Journal of Artificial Intelligence in Education, 31, 538-584.

[23] Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. Psych, 3(4), 897-915.

[24] Florida Atlantic University WAC Assessment Student Information. https://www.fau.edu/wac/assessment/students/

[25] Zhu, W., & Sun, Y. (2020, October). Automated essay scoring system using multi-model machine learning. In CS & IT Conference Proceedings (Vol. 10, No. 12). CS & IT Conference Proceedings.

[26] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision (pp. 19-27).

[27] Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R., & Lan, A. (2022, July). Automated scoring for reading comprehension via in-context bert tuning. In International Conference on Artificial Intelligence in Education (pp. 691-697). Cham: Springer International Publishing.

[28] Wang, Hao, et al. ”Argumentation mining on essays at multi scales.” Proceedings of the 28th International conference on computational linguistics. 2020.

[29] Nguyen, Huy, and Diane J. Litman. ”Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics.” FLAIRS. 2016.

[30] Nguyen, Huy, and Diane Litman. ”Extracting argument and domain words for identifying argument components in texts.” Proceedings of the 2nd Workshop on Argumentation Mining. 2015.

[31] Wachsmuth, Henning, Khalid Al Khatib, and Benno Stein. ”Using argument mining to assess the argumentation quality of essays.” Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers. 2016.

[32] Persing, Isaac, and Vincent Ng. ”Unsupervised argumentation mining in student essays.” Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.

[33] Persing, Isaac, and Vincent Ng. ”End-to-end argumentation mining in student essays.” Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[34] Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In Proceedings of the 25th International Conference on Computational Linguistics, pages 1501–1510.

[35] Stab, C., / Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. Computational Linguistics , 43(3), 619–659.

[36] Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: Looking back and moving ahead. Discourse Studies , 8(3), 423–459.

[37] Becker, M., & Koller, A. (2014). Discourse structure and support vector machines for classifying argument components in persuasive essays. Proceedings of the First Workshop on Argumentation Mining , 33–42.

[38] Opitz, J., & Frank, A. (2019). Argument mining with structured SVMs and RNNs. Proceedings of NAACL-HLT 2019 , 909–921.

[39] Fang, B., & Koto, F. (2022). Context-Aware Sentence Classification in Evidence-Based Medicine. In *Proceedings of ALTA 2022 Shared Task*.

[40] Hou, Y. (2020). Fine-grained Information Status Classification Using Discourse Context-Aware BERT. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 6101–6112.

[41] De Angeli, K., Gao, S., Danciu, I., et al. (2022). Class Imbalance in Out-of-Distribution Datasets: Improving the Robustness of the TextCNN for the Classification of Rare Cancer Types. *Journal of Biomedical Informatics*, 125, 103957.

[42] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27.

[43] Fang, B., & Koto, F. (2022). Context-Aware Sentence Classification in Evidence-Based Medicine. In *Proceedings of ALTA 2022 Shared Task*.

[44] Hou, Y. (2020). Fine-grained Information Status Classification Using Discourse Context-Aware BERT. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 6101–6112.

[45] De Angeli, K., Gao, S., Danciu, I., et al. (2022). Class Imbalance in Out-of-Distribution Datasets: Improving the Robustness of the TextCNN for the Classification of Rare Cancer Types. *Journal of Biomedical Informatics*, 125, 103957.

[46] Eger, S., Daxenberger, J., & Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11–22).

[47] Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4129–4138).

[48] Mushtaq, U., & Cabessa, J. (2022). Argument classification with BERT plus contextual, structural and syntactic features as text. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)* (pp. 622–633). Springer.

[49] Ruggeri, F., Lippi, M., & Torroni, P. (2021). Tree-constrained graph neural networks for argument mining. *arXiv preprint arXiv:2110.00124*.

[50] Abdalaziz, H. S., & Saeed, F. A. (2017). New hierarchical model for multiclass imbalanced classification. *Journal of Theoretical and Applied Information Technology, 95*(16), 3861–3870.

[51] del Moral, P., Nowaczyk, S., Sant'Anna, A., & Pashami, S. (2022). Pitfalls of assessing extracted hierarchies for multi-class classification. *Pattern Recognition, 136*, 109225.

[52] Blanchard, A. E., Gao, S., Yoon, H.-J., et al. (2022). A keyword-enhanced approach to handle class imbalance in clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, *26*(6), 2796–2803.

[53] Murray, K., & Chiang, D. (2018). Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 212–223).

[54] Zhou, C., Ma, X., & Neubig, G. (2020). A closer look at the robustness of neural text classifiers over varying sentence lengths. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5769–5780).