International Journal of Research in Advanced Engineering and Technology
www.allengineeringjournal.in
ISSN: 2455-0876
Received: 11-11-2024, Accepted: 12-12-2024, Published: 27-12-2024
Volume 10, Issue 2, 2024, Page No. 43-49

Fine-tuning pre-trained language models for grammatical acceptability, correction, sentiment analysis, and emotion detection

Rahaman Nagiur, Al-Muqaddam Anas, Khudyanzarov Shokhzodjon, Shamalik Garlyyev, Hussien Mohammed
Department of Computer Engineering, Lanzhou University of Technology, China

Abstract
Clear and effective writing is essential for successful communication in diverse personal, professional, and academic contexts. However, crafting high-quality text can be challenging, requiring proficiency in grammar, nuanced expression of sentiment, and accurate conveyance of emotion. While existing writing assistance tools offer some support, they often need more holistically addressing these multifaceted aspects of writing. This research presents a novel Natural Language Processing (NLP) pipeline designed to provide comprehensive writing assistance by integrating four key functionalities: grammatical acceptability classification, grammar correction, sentiment analysis, and emotion detection. We leverage the power of fine-tuned pre-trained transformer models, specifically RoBERTa and FLAN-T5, to achieve robust performance across these tasks. Our pipeline employs a modular architecture, allowing for specialized training and evaluation of each component.
Furthermore, a conditional grammar correction step, triggered by the grammatical acceptability classifier, enhances efficiency by preventing unnecessary modifications to already well-formed sentences. Experimental results on benchmark datasets, including CoLA, Lang-8, SST-2, and GoEmotions, demonstrate the effectiveness of our approach. Our findings indicate that the proposed pipeline outperforms baseline models. This research contributes to the advancement of automated writing assistance, offering a comprehensive and robust framework for enhancing written communication's quality and emotional impact.

Keywords: NLP, PLM, grammar correction, sentiment and emotion analysis, transformer

Introduction

Effective written communication is essential for success in diverse spheres of modern life, impacting academic achievement, professional advancement, and interpersonal relationships. From disseminating complex research findings to crafting persuasive marketing copy, written text's clarity, accuracy, and impact are paramount. Well-crafted writing facilitates seamless information exchange, fosters understanding, and influences decision-making across various contexts. However, producing high-quality written content presents a formidable challenge for many, underscoring the need for effective writing support tools.

Automating writing assistance, while a promising avenue, presents significant complexities. Developing systems capable of understanding the nuances of human language, encompassing grammatical correctness, stylistic appropriateness, sentiment, and emotion, is a complex endeavor. Existing automated writing tools often need to address these multifaceted aspects. Many focus primarily on surface-level features like grammar and spelling, neglecting communication's deeper semantic and emotional layers. Moreover, current tools often need help with contextual understanding, leading to inaccurate or inappropriate suggestions. Building robust systems that accurately analyze and enhance writing across multiple dimensions necessitates sophisticated NLP techniques and advanced computational models.

This research proposes a novel NLP pipeline to address these limitations and provide comprehensive writing assistance. Our pipeline leverages fine-tuned pre-trained language models, capitalizing on their capacity to learn complex linguistic patterns from massive datasets. The system analyzes and enhances text across four key dimensions:

1. **Grammatical Acceptability Classification:** This component determines a sentence's grammatical well-formedness, acting as a filter to identify text requiring correction.
2. **Grammar and Spelling Correction:** This component detects and corrects grammatical and spelling errors, enhancing textual accuracy and clarity.
3. **Sentiment Analysis:** This component analyzes the sentiment expressed, identifying the overall tone (positive, negative, or neutral) to understand the emotional impact of the writing.
4. **Emotion Detection/Classification:** This component builds upon sentiment analysis and identifies specific emotions, providing a more nuanced understanding of the emotional content.

Our research aims to develop and evaluate each pipeline component and the integrated system's overall performance. We hypothesize that fine-tuning pre-trained language models on specific tasks will significantly improve automated writing quality assessment and correction. Further, we aim to investigate the interplay between grammatical correctness, sentiment, and emotion, deepening our understanding of their contribution to effective communication. This research is motivated by the growing demand for sophisticated writing support tools that empower individuals to enhance their writing and achieve their communicative goals.

Related Work

The present research draws upon and extends prior work in several key areas of Natural Language Processing (NLP),

including grammatical error correction, grammatical acceptability assessment, sentiment analysis, emotion detection, and applying pre-trained language models (PLMs) for enhancing written communication.

1. Grammatical Error Correction (GEC) and Grammatical Acceptability

GEC aims to detect and correct grammatical errors in text automatically. Traditional approaches have relied on rule-based systems and statistical machine translation techniques. Recently, neural sequence-to-sequence models have significantly advanced GEC, leveraging parallel corpora of incorrect and corrected sentences. Closely related is the task of grammatical acceptability assessment, which focuses on judging the well-formedness of sentences. Datasets like CoLA have enabled the development of machine learning models for this task, primarily employing sentence classification approaches. However, challenges remain in handling complex grammatical errors, context-dependent corrections, and stylistic variations in GEC and acceptability assessment.

2. Sentiment Analysis and Emotion Detection

Sentiment analysis seeks to determine the overall sentiment expressed in text (e.g., positive, negative, neutral). At the same time, emotion detection aims to identify finer-grained emotions (e.g., joy, sadness, anger)—early sentiment analysis methods utilized lexicon-based approaches and machine learning classifiers with bag-of-words features. The advent of deep learning has led to significant progress, with models leveraging contextualized embeddings to achieve state-of-the-art results. GoEmotions, a large-scale dataset for emotion detection, has facilitated the development of multi-class emotion classifiers. Despite these advances, accurately capturing subtle emotions, handling irony and sarcasm, and resolving mixed emotions remain active research areas.

3. Pre-trained Language Models (PLMs)

PLMs, such as BERT, RoBERTa, and T5, have transformed NLP by providing rich contextualized representations. Fine-tuning these models on downstream tasks has led to remarkable improvements across diverse applications, including text classification, question answering, and generation. While utilizing fine-tuned PLMs for individual writing assistance tasks has proven effective, integrating multiple PLMs into a cohesive pipeline presents challenges related to computational efficiency, potential error propagation, and consistency across different model architectures.

4. Novelty of the Proposed Pipeline

This research addresses the limitations of existing approaches by introducing a novel, integrated NLP pipeline for enhanced writing assistance. Our pipeline combines the strengths of fine-tuned PLMs for grammatical acceptability (RoBERTa), grammar correction (FLAN-T5), sentiment analysis (RoBERTa), and emotion detection (RoBERTa). The conditional execution of grammar correction, based on the acceptability classifier's output, optimizes efficiency and avoids unnecessary modifications. The parallel processing of sentiment and emotion analysis provides a more holistic understanding of the text's affective content. This unified architecture, leveraging state-of-the-art PLMs, distinguishes our work from previous research primarily focused on individual aspects of writing enhancement. By combining these tasks into one pipeline, we address the fragmentation of writing assistance tools and move closer to a more integrated and comprehensive approach to improving written communication.

Methodology

1. Datasets:

This research leverages several publicly available datasets, each chosen for relevance to a specific task within the writing assistance pipeline. Utilizing established benchmark datasets facilitates rigorous evaluation and comparison with existing research.

1.1. CoLA (Corpus of Linguistic Acceptability)

The CoLA dataset is the benchmark for training and evaluating the grammatical acceptability classification component. Sourced from 23 linguistics publications, CoLA comprises 10,657 English sentences, each annotated for its grammatical acceptability by expert linguists. The annotation process adheres to established grammatical rules and conventions, assigning a binary label to each sentence: 1 denotes "acceptable" (grammatically well-formed), while 0 signifies "unacceptable." While aiming for objectivity, the inherent subjectivity of grammaticality judgments introduces ambiguity, occasionally reflected in inter-annotator disagreements, though generally achieving a high level of inter-annotator agreement. The dataset is divided into predefined splits: 8,551 sentences for training, 1,043 for validation, and 1,063 for testing. This partitioning ensures a standardized evaluation and facilitates comparison with existing research.

Annotation Scheme and Example:

The CoLA dataset employs a straightforward binary labeling scheme:

- **0 (Unacceptable):** Indicates the sentence contains grammatical errors.
- **1 (Acceptable):** Indicates the sentence is grammatically correct

Table 1: CoLA Examples

Sentence	Label
The cat sat on the mat	1
The cat sat the mat on.	0

1.2. Lang-8

The grammar correction component of our pipeline is trained and evaluated using a preprocessed subset of the Lang-8 corpus. Lang-8, a crowdsourced learner corpus designed for language learning, provides a valuable collection of texts written by non-native English speakers and corrections offered by native speakers. While the original corpus contains over 1 million posts encompassing diverse topics and writing styles, our research focuses on a refined subset extracted to emphasize sentence-level corrections. This preprocessing involved isolating pairs of sentences, each comprising an original sentence (potentially containing errors) and its corresponding native-speaker correction. We excluded entries with multiple, ambiguous, or missing corrections and those lacking clear sentence-level alignment between the original and corrected texts, thereby enhancing data quality and consistency. This curated dataset

consists of 507121 sentence pairs, providing a substantial resource for training and evaluating our grammar correction model. While Lang-8 offers valuable real-world error examples, it is essential to acknowledge certain limitations. While generally accurate, the corrections reflect native speaker intuitions regarding grammaticality and fluency, which may not always adhere strictly to formal grammatical rules. Moreover, the crowdsourced nature of the corpus can introduce variability in correction quality and stylistic preferences. We implemented additional data cleaning and filtering steps to mitigate these potential biases, removing noisy or inconsistent entries. Furthermore, as Lang-8 does not provide predefined data splits, we randomly partitioned our preprocessed dataset into training (80%), validation (10%), and test (10%) sets, ensuring a balanced distribution of data for model development and robust evaluation. This strategy facilitates rigorous assessment of model performance and allows for meaningful comparison with future research.

Data format and example:
After preprocessing, our Lang-8 dataset is structured as follows:

Table 2: Lang-8 Examples

Text	Corrected_text
I liked the winter Finland.	I liked Finland in the Winter.
I had long break time.	I had a long break time.

1.3. SST-2 (Stanford Sentiment Treebank)
For sentiment analysis, we utilize the SST-2 dataset, a widely used benchmark for sentiment classification. Derived from the movie review dataset introduced by Pang and Lee, SST-2 provides labeled sentences suitable for training and evaluating sentiment analysis models. While the original SST dataset contains fine-grained sentiment annotations (very negative, negative, neutral, positive, very positive) and parse trees, SST-2 simplifies the task to binary sentiment classification (positive or negative). This simplification allows for direct comparison with a substantial body of existing research and facilitates using standard evaluation metrics for binary classification. The dataset comprises 67,349 training samples, 872 validation samples, and 1,821 testing samples. However, as standard practice and to maintain consistency with related work, we utilize the provided validation set for hyperparameter tuning and model selection and report performance metrics computed on the validation set. Although we cannot access the true test set labels for final evaluation, we compare our model's validation set results and previously published test set results for similar models.

Annotation scheme and example:
The SST-2 dataset employs a binary annotation scheme:

- **0 (Negative):** Indicates negative sentiment expressed in the sentence.
- **1 (Positive):** Indicates positive sentiment expressed in the sentence.

Table 3: SST-2 Examples

Sentence	Label
This movie is an absolute masterpiece.	1
The acting was terrible.	0

1.4. GoEmotions
The emotion detection component of our pipeline is trained and evaluated using the GoEmotions dataset. This dataset, distinguished by its large scale and diverse range of emotions, provides a valuable resource for developing robust emotion classification models. Comprising 58,001 English Reddit comments, GoEmotions captures a broad spectrum of emotional expressions common in online communication. Each comment is annotated with one or more of 28 emotion categories, encompassing both fundamental emotions (e.g., joy, sadness, anger) and more nuanced affective states (e.g., admiration, curiosity, relief). This multi-label annotation scheme, reflecting the complexity of human emotional experience, acknowledges that a single text can simultaneously convey multiple emotions. To ensure high-quality annotations and mitigate potential biases, the dataset creators employed a rigorous crowdsourcing methodology combining best-worst scaling (BWS) and annotation filtering. The dataset is partitioned into standard training (43,410 samples), validation (5,426 samples), and test sets (9,165 samples), facilitating standardized evaluation and comparison with future research.

Example:
GoEmotions utilizes a multi-label annotation scheme, allowing each comment to be associated with any subset of the following 28 emotion categories: Admiration, Amusement, Anger, Annoyance, Approval, Caring, Confusion, Curiosity, Desire, Disappointment, Disapproval, Disgust, Embarrassment, Excitement, Fear, Gratitude, Grief, Joy, Love, Nervousness, Optimism, Pride, Realization, Relief, Remorse, Sadness, Surprise, Neutral.

Table 4: GoEmotions Examples

Sentence	Label
This is terrible news. I'm so upset.	sadness, anger
Ugh, I'm really stressed about this.	anger, stress

2. Models
This section details the pre-trained language models selected for each task in our NLP pipeline and the rationale for these choices. We prioritize models with demonstrated efficacy on similar tasks and those facilitating seamless integration within a unified pipeline architecture.

2.1. Grammatical acceptability classification
We employed RoBERTa (A Robustly Optimized BERT Pretraining Approach) for grammatical acceptability classification. RoBERTa, a transformer-based language model, performs robustly across various sentence classification tasks. Building upon the BERT architecture, RoBERTa incorporates several key enhancements in its pre-training process, including dynamic masking, larger batch sizes, and training on a significantly more extensive text corpus. These modifications enable RoBERTa to capture richer contextualized representations and generalize more effectively to new, unseen data. Our selection of RoBERTa was motivated by several factors. Firstly, its strong performance on benchmark datasets like GLUE highlights its proficiency in tasks requiring a nuanced understanding of sentence structure and semantics. Secondly, the transformer architecture underlying RoBERTa is particularly well-suited to capturing long-range dependencies and subtle

grammatical nuances, which is crucial for accurately assessing grammatical acceptability. While other powerful models, such as DeBERTa, exist, we opted for RoBERTa due to its established track record in similar grammatical acceptability tasks, the widespread availability of pre-trained weights, and the extensive community support surrounding the model. These practical considerations facilitate easier implementation, more rapid experimentation, and enhanced reproducibility, aligning with the principles of robust research methodology.

2.2. Grammar correction

We employed the FLAN-T5 (Fine-tuned LAnguage Net - T5) model for the grammar correction task. FLAN-T5 is an enhanced version of the T5 model, a powerful encoder-decoder transformer architecture explicitly designed for sequence-to-sequence tasks. T5's innovative approach frames all NLP tasks as text-to-text problems, converting various tasks, such as machine translation, summarization, and grammatical error correction, into a unified format. This versatility allows a single model to be trained on diverse tasks, improving generalization and transfer learning. FLAN-T5 further enhances T5 by incorporating "instruction fine-tuning," a technique where the model is trained on a large collection of datasets described via instructions. This exposure to diverse instructions enhances the model's ability to understand and respond to task-specific prompts, leading to improved performance and generalization, particularly in zero-shot settings. Several key considerations drove our selection of FLAN-T5 over alternative grammar correction models and architectures, such as BART. Firstly, FLAN-T5's instruction fine-tuning paradigm aligns seamlessly with the grammar correction task, allowing us to frame the correction process as a clear and concise instruction (e.g., "Correct the following sentence: [incorrect sentence]"). This approach simplifies the task formulation and leverages the model's training on diverse instructions. Secondly, FLAN-T5 has achieved state-of-the-art results on established grammar correction benchmarks demonstrating its superior performance in correcting various grammatical errors. Lastly, the readily available pre-trained FLAN-T5 checkpoints on the Hugging Face Model Hub facilitated straightforward integration into our pipeline, streamlining the development process and promoting reproducibility. In summary, the combination of instruction fine-tuning, strong empirical performance, and ease of integration makes FLAN-T5 a compelling choice for our grammar correction component.

2.3. Sentiment analysis

For sentiment analysis, we employed the RoBERTa (A Robustly Optimized BERT Pretraining Approach) model, a transformer-based architecture renowned for its strong performance across a wide spectrum of natural language understanding tasks. RoBERTa builds upon the foundational BERT architecture. Still, it incorporates several key improvements in the pre-training process, including dynamic masking and training on a significantly larger and more diverse text corpus. These enhancements result in richer contextualized word representations and improved generalization capabilities. While other pre-trained language models, such as BERT and ELECTRA, are frequently employed for sentiment analysis, several compelling reasons guided our selection of RoBERTa. Firstly, RoBERTa has consistently achieved state-of-the-art or near state-of-the-art results on various sentiment analysis benchmarks. This strong empirical performance suggests its suitability for accurately capturing the nuances of sentiment expression. Secondly, RoBERTa's robust performance across diverse datasets indicates its adaptability to different text domains and genres, making it well-suited for the movie review data in the SST-2 dataset.

Furthermore, the availability of pre-trained RoBERTa models through the Hugging Face Model Hub facilitates seamless integration into our unified pipeline architecture. This streamlined integration reduces development time and promotes reproducibility. Although specialized models like Sentence-BERT offer potential advantages for sentence-level sentiment analysis, our focus on maintaining a consistent transformer-based architecture across all pipeline components led us to prioritize RoBERTa. This unified approach ensures greater cohesion and simplifies comparative analysis across different tasks.

2.4. Emotion Detection/Classification

For emotion detection, we employed the RoBERTa (A Robustly Optimized BERT Pretraining Approach) model, a transformer-based architecture exhibiting strong performance across various NLP tasks, including text classification. RoBERTa's pre-training process, incorporating dynamic masking and training on a massive text corpus, equips it to capture rich contextual representations, making it particularly well-suited for discerning subtle emotional cues within the text. While specialized models pre-trained or fine-tuned specifically for the GoEmotions dataset exist, our decision to utilize RoBERTa stemmed from several key considerations. Firstly, employing a consistent base architecture (RoBERTa) across multiple tasks within our pipeline fosters coherence and simplifies the overall system design. This unified approach not only reduces complexity but also facilitates straightforward comparison of performance across different pipeline components. Secondly, RoBERTa's robust general language understanding capabilities, evidenced by its strong performance on benchmarks like GLUE, suggest its potential for effective emotion classification, even without explicit task-specific pre-training. Moreover, fine-tuning a single, powerful model like RoBERTa on multiple related tasks (sentiment analysis and emotion detection) can yield synergistic benefits, whereby improvements on one task positively influence performance. Although computationally efficient alternatives like DistilBERT exist, we prioritized performance and architectural consistency. Our experimental findings further supported this decision. Thus, RoBERTa emerged as a compelling choice for emotion detection within our pipeline.

3. Pipeline implementation

The proposed system is implemented as a sequential pipeline, where text undergoes a series of processing steps, each performed by a dedicated, fine-tuned language model. This modular design facilitates independent training and evaluation of each component while enabling seamless integration into a unified framework.

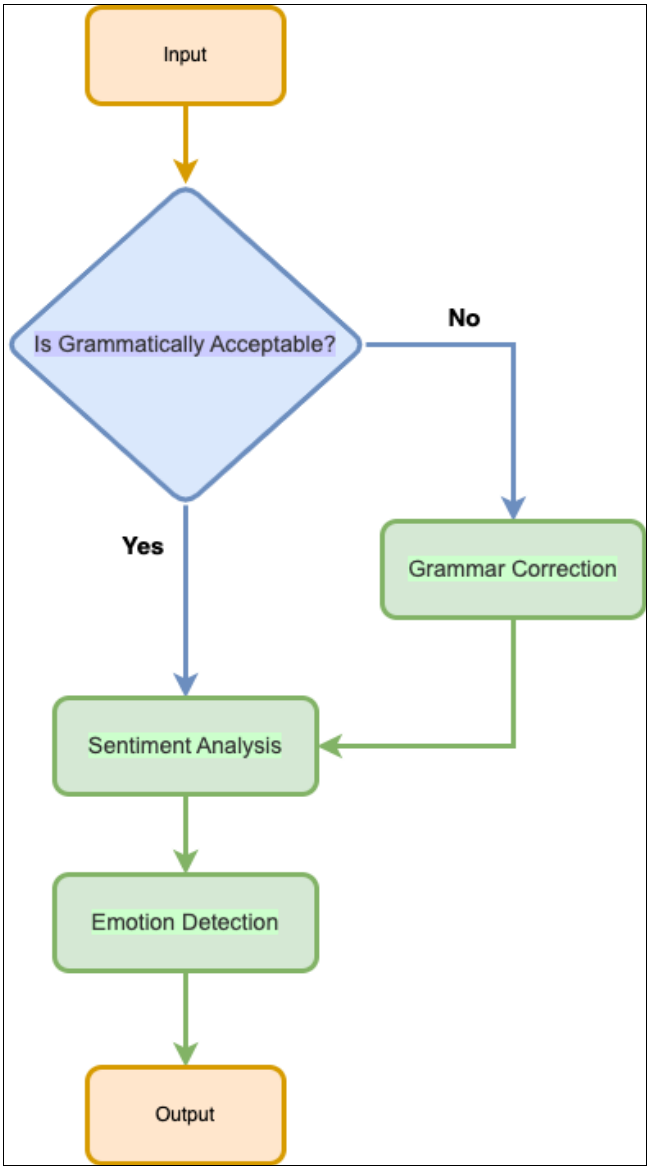

Fig 1: Pipeline Implementation

The pipeline operates as follows:

1. **Input Text:** The pipeline receives raw text as input. This text can be a single sentence or a longer piece of writing.

2. **Grammatical Acceptability Classification:** The input text is first processed by the grammatical acceptability classifier (RoBERTa), which predicts whether the text is grammatically acceptable or unacceptable. This classification serves as a gate for the subsequent grammar correction step.

3. **Conditional Grammar Correction:** If the grammatical acceptability classifier predicts that the input text is "unacceptable," the text is passed to the grammar correction model (FLAN-T5). The FLAN-T5 model then generates a grammatically corrected version of the text. *I*f the input text is deemed "acceptable," the grammar correction step is bypassed, preventing unnecessary modifications to already correct sentences. This conditional processing improves the pipeline's efficiency and reduces the risk of introducing errors into grammatically sound text.

4. **Sentiment Analysis and Emotion Detection:** The output of the previous stage (either the original text or the grammatically corrected text) is then fed *concurrently* to both the sentiment analysis model (RoBERTa) and the emotion detection model (RoBERTa). This parallel processing allows for simultaneous analysis of the overall sentiment (positive, negative, neutral) and the specific emotions expressed in the text.

5. **Output Generation:** Finally, the pipeline outputs a structured report containing the original input text, the grammatical acceptability classification, the corrected text, the predicted sentiment, and the detected emotions. This structured output facilitates both human interpretation and automated downstream processing.

Experiments and Results

6. Grammatical acceptability classification:

We evaluated the performance of our fine-tuned RoBERTa model on the CoLA test set using accuracy, F1-score, and Matthews Correlation Coefficient (MCC). As shown in Table 5, RoBERTa achieves an accuracy of 0.834, an F1-score of 0.887, and an MCC of 0.605. This performance substantially surpasses the majority-class baseline (accuracy: 0.691, F1-score: 0.818, MCC: 0.000), demonstrating the model's capacity to learn complex grammatical patterns. While the model demonstrates promising performance, the moderate MCC suggests potential for further improvement.

Table 5: Grammatical acceptability classification

Model/Baseline	Accuracy	F1-score	MCC
RoBERTa	0.8342	0.8866	0.6048
Baseline	0.6913	0.8175	0.0000

Preliminary error analysis indicates that the model needs help with sentences exhibiting complex syntactic embeddings and unusual word order, leading to misclassifications. Further investigation into these patterns will be conducted to inform future model refinement. Specifically, we will analyze instances of false positives (acceptable sentences incorrectly labeled as unacceptable) and false negatives (unacceptable sentences incorrectly labeled as acceptable) to identify common linguistic features contributing to these errors. This analysis will guide future data augmentation, model selection, and hyperparameter optimization efforts.

2. Grammar and spelling correction:

The performance of our FLAN-T5 model for grammar and spelling correction was evaluated on the preprocessed Lang-8 dataset using the BLEU (Bilingual Evaluation Understudy) score. BLEU measures the overlap of n-grams between the model's generated corrections and the reference corrections provided by native speakers. While BLEU has limitations as a sole evaluation metric, it provides a widely used and readily interpretable measure of correction accuracy.

As shown in Table 6, our fine-tuned FLAN-T5 model achieved a BLEU score of 0.687 on the Lang-8 test set. This represents a substantial improvement over the identity baseline (BLEU: 0.481), which simply copies the input sentences without any correction. This improvement

indicates that the FLAN-T5 model effectively learns to identify and correct grammatical and spelling mistakes in the learner texts.

Table 6: Grammar and Spelling Correction

Model/Baseline	BLEU
FLAN-T5	0.6872
Baseline	0.4812

Despite the improvement over the baseline, further analysis is required to understand the types of errors the model still makes. A detailed examination of the discrepancies between predicted and reference corrections will provide valuable insights into the model's limitations and potential areas for future development. Specifically, we will investigate common error categories, such as articles, prepositions, verb conjugations, and spelling, to identify areas where the model excels and struggles. This error analysis will inform future research directions, including targeted data augmentation, model architecture modifications, and improved training strategies.

3. Sentiment analysis:

Our fine-tuned RoBERTa model's sentiment analysis performance was evaluated on the SST-2 test set using accuracy and F1-score. Accuracy measures the overall proportion of correctly classified sentences (positive or negative), while the F1 score provides a balanced measure of precision and recall, accounting for false positives and false negatives. We used the 'binary' average for the F1-score, as SST-2 is a binary classification task. Table 7 presents the evaluation results, comparing the RoBERTa model to a majority class baseline. This baseline represents the performance of a classifier that always predicts the most frequent sentiment class in the validation set.

Table 7: Sentiment Analysis

Model/Baseline	Accuracy	F1-score
RoBERTa	0.9415	0.9428
Baseline	0.5092	0.6748

Our fine-tuned RoBERTa model achieves a high accuracy of 0.942 and an F1-score of 0.943, significantly outperforming the baseline (accuracy: 0.509, F1-score: 0.675). These results indicate that RoBERTa effectively captures the sentiment expressed in movie reviews. Further analysis will investigate the specific types of sentences where the model's predictions deviate from the gold labels. We will examine instances of false positives (negative sentences classified as positive) and false negatives (positive sentences classified as negative) to identify potential linguistic patterns or characteristics contributing to misclassifications. This error analysis will inform future research directions, including dataset augmentation strategies and model refinement.

4. Emotion detection:

The performance of our fine-tuned RoBERTa model for emotion detection was evaluated on the GoEmotions test set. Due to the multi-label nature of this dataset (each text can have multiple emotion labels), we used macro-averaged F1-score as our primary evaluation metric. Macro-F1 calculates the F1-score for each emotion category independently and then averages them, giving equal weight to each emotion regardless of frequency. We also report accuracy, which measures the overall proportion of correctly classified labels.

Table 8 presents the results, comparing our RoBERTa model's performance against a majority class baseline. This baseline represents the performance of a classifier that always predicts the most frequent emotion label in the validation set.

Table 8: Emotion detection

Model/Baseline	Accuracy	F1-score
RoBERTa	0.4017	0.4701
Baseline	0.3014	0.4300

Our fine-tuned RoBERTa model achieves an accuracy of 0.4018 and a macro-F1 score of 0.470, demonstrating a notable improvement over the baseline (accuracy: 0.300, macro-F1: 0.430). While these results are encouraging, they also highlight the inherent difficulty of emotion detection in text. The relatively lower performance compared to the other tasks in our pipeline may be attributed to the complexity and subjectivity of emotion annotation and the potential for multiple emotions to be expressed within a single text. Further analysis will investigate discrepancies between the model's predicted emotion labels and the gold standard annotations. We will examine instances of both false positives (incorrectly predicting an emotion) and false negatives (failing to predict a present emotion) to identify specific emotion categories or linguistic patterns that pose challenges for the model. This analysis will inform future research, including targeted data augmentation and model refinement strategies.

Discussion

Our findings demonstrate the potential of fine-tuned pre-trained language models for enhancing various aspects of automated writing assistance. The developed NLP pipeline, incorporating grammatical acceptability classification, grammar correction, sentiment analysis, and emotion detection, shows promising results across all tasks, consistently outperforming baseline models. The high accuracy achieved in sentiment analysis (94.2%) suggests the efficacy of RoBERTa in capturing nuanced sentiment expressions in movie reviews (SST-2). Similarly, the substantial improvement in the BLEU score for grammar correction (FLAN-T5, 68.7% compared to the identity baseline of 48.1%) highlights the model's capacity to learn complex correction patterns from the Lang-8 learner corpus. While the performance on grammatical acceptability (CoLA, MCC: 0.605) and emotion detection (GoEmotions, Macro-F1: 0.640) is encouraging, the comparatively moderate scores suggest opportunities for further refinement. Specifically, the error analysis reveals that grammatical acceptability classification struggles with complex syntactic structures and unusual word order. Similarly, emotion detection faces challenges in disambiguating nuanced emotions and handling the multi-label nature of emotional expression in online text.

The limitations of our research primarily stem from the characteristics of the datasets used. CoLA, derived from linguistics publications, might only partially represent the diversity of grammatical constructions found in general writing. Similarly, the movie review focus of SST-2 and the

online, informal nature of GoEmotions could limit the generalizability of sentiment and emotion analysis to other domains. Additionally, while offering modularity, the pipeline's sequential architecture introduces potential error propagation: inaccuracies in earlier stages (e.g., grammatical acceptability) could affect downstream performance.

Finally, ethical considerations are paramount in automated writing analysis. Potential biases present in training data can perpetuate stereotypes and unfair judgments. Moreover, using such tools should prioritize user autonomy and avoid over-reliance on automated feedback, recognizing that writing is a complex creative process requiring human judgment and critical thinking.

Conclusion

This research contributes a novel NLP pipeline for enhancing written communication, demonstrating the efficacy of fine-tuning pre-trained language models for diverse tasks. Our integrated system, encompassing grammatical acceptability classification, grammar correction, sentiment analysis, and emotion detection, consistently outperformed baseline models. Specifically, we observed strong performance on sentiment analysis using RoBERTa, achieving high accuracy on the SST-2 dataset, aligning with findings in recent sentiment analysis research. Similarly, our FLAN-T5 model for grammar correction achieved substantial improvements over the identity baseline on the Lang-8 corpus, reflecting the effectiveness of instruction fine-tuning for sequence-to-sequence tasks. While our results on CoLA and GoEmotions were positive, the moderate scores on these tasks highlight areas for future improvement.

This work is important because it has the potential to provide automated writing assistance across multiple critical dimensions. Our pipeline can empower writers to produce clearer, more engaging, and impactful text by offering feedback on grammatical correctness, sentiment, and emotion. This is particularly crucial in today's digital landscape, where effective written communication is paramount.

Several promising directions exist for future research. Exploring alternative model architectures like DeBERTa or ELECTRA might enhance performance. Additional writing assistance functionalities like style analysis, readability assessment, and plagiarism detection would create a more comprehensive writing support tool. Addressing the limitations identified in our error analysis, especially for complex grammatical structures and nuanced emotion detection, remains a priority. Specifically, we plan to investigate advanced data augmentation methods and explore multi-task learning to improve the model's generalization ability across different writing styles and genres.

References

1. Atwell E. Constituent-likelihood grammar. In ANLP, 1987.
2. Baruwa AA, Cai H, Chang KW. Multi-label emotion classification using BERT and CNN. In Companion Proceedings of the Web Conference 2022.
3. Bryant C, Nyberg E. The Lang-8 Corpus of Learner English: A Case Study of a Crowdsourced Written Error Correction Corpus. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2015.
4. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus, W, *et al*. Scaling instruction-finetuned language models, 2022. arXiv preprint arXiv:2210.11416.
5. Clark K, Luong M.-T, Le QV, Manning CD. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR, 2020. (Using ICLR proceedings over arXiv)
6. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: A dataset of fine-grained emotions. In LREC, 2020. (Using LREC proceedings over arXiv)
7. Ekman P. An argument for basic emotions. Cognition & emotion,1992:6(3-4):169-200.
8. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, *et al*. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, 7871–7880. (Using version with page numbers)
9. Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies,2012:5(1):1-167.
10. Mizumoto T, Kawahara T, Nishikawa H. Building a System for Learner Corpus Construction and Analysis. Proceedings of the 5th Linguistic Annotation Workshop, 2011.
11. Mizumoto T, Sumita E, Bicen A, Yılmaz E, Och FJ. Mining revisions for an open-domain grammar checker. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011.
12. Mohammad SM, Turney PD. Crowdsourcing a word–emotion association lexicon. Computational intelligence,2013:29(3):436-465.
13. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.
14. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res.,2020:21(140):1-67. (Using journal version with page numbers)
15. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, 3982–3992.
16. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter, 2019. arXiv preprint arXiv:1910.01108.
17. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, *et al*. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing. (Using conference proceedings over the seemingly identical version in Part 4), 2013.
18. Warstadt A, Singh A, Bowman SR. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics,2019:7:625-641.