From Handcrafted Features to LLMs: A Comparative Study in Native Language Identification

Aliyah C. Vanterpool
Montclair State University
New Jersey, USA
aliyahvan@gmail.com

Katsiaryna Aharodnik
CUNY Graduate Center
New York, USA
kaharodnik@gradcenter.cuny.edu

Abstract

This study compares a traditional machine learning feature-engineering approach to a large language models (LLMs) fine-tuning method for Native Language Identification (NLI). We explored the COREFL corpus, which consists of L2 English narratives produced by Spanish and German L1 speakers with lower-advanced English proficiency (C1) (Lozano et al., 2020). For the feature-engineering approach, we extracted language productivity, linguistic diversity, and n-gram features for Support Vector Machine (SVM) classification. We also looked at sentence embeddings with SVM and logistic regression. For the LLM approach, we evaluated BERT-like models and GPT-4. The feature-engineering approach, particularly n-grams, outperformed the LLMs. Sentence-BERT embeddings with SVM achieved the second-highest accuracy (93%), while GPT-4 reached an average accuracy of 90.4% across three runs when prompted with labels. These findings suggest that feature engineering remains a robust method for NLI, especially for smaller datasets with subtle linguistic differences between classes. This study contributes to the comparative analysis of traditional machine learning and transformer-based LLMs, highlighting current LLM limitations in handling domain-specific data and their need for larger training resources.

1 Introduction

The role of a learner's native language (L1) in second language (L2) acquisition has been widely addressed in second language acquisition (SLA) literature (Lado, 1957; Corder, 1975). SLA research has shown that the spelling, grammar, and lexicon used in L2 writing are often influenced by patterns and rules from a learner's L1. However, the extent of L1 impact on L2 performance remains difficult to determine precisely.

With the emergence of learner corpora, it has become possible to empirically test SLA hypotheses and explore how different L1s manifest in L2 writing. One application of this is the Native Language Identification (NLI) task, which uses automated methods to predict a learner's L1 based on their L2 writing. Prior studies have demonstrated high performance for feature-engineered machine learning (ML) approaches to NLI. However, research examining the applicability of large language models (LLMs) to NLI remains limited. Moreover, there is a lack of studies directly comparing LLMs with traditional feature-engineered pipelines within the same experimental paradigm.

The current study addresses this gap by comparing a traditional feature-engineering ML approach to transformer-based LLMs for the NLI task. As a secondary goal, we explore both methods using a relatively small but unique learner corpus composed of video-based written narratives. This corpus offers more structured and homogeneous data than the topic-based essays commonly used in prior NLI studies. We report the results of both NLI approaches and discuss their implications for SLA research.

This paper is structured as follows: Section 2 introduces previous research. Section 3 outlines the methodology, including a description of the COREFL corpus and training/testing techniques. Section 4 presents the results of both approaches. Section 5 discusses the findings and implications for SLA and NLI. Section 6 provides the conclusion and suggests future research directions.

2 Related Work

In NLI research, findings are often interpreted through the lens of Second Language Acquisition (SLA) and linguistic transfer. Several theoretical approaches from SLA have served as a founda-

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 144–153, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-015

tion for this task. One of the most influential is the Contrastive Analysis Hypothesis (CAH; Lado, 1957), which posits that difficulties in second language learning arise from differences between the learner's first language (L1) and the target language (L2). Language typology plays a key role in this process: the more similar two languages are, the more likely learners are to experience positive transfer that facilitates acquisition; conversely, typologically distant languages tend to result in more negative transfer and errors.

Linguistic transfer refers to the application of phonological, morphological, syntactic, or lexical rules from one language to another (Odlin, 1989). For instance, a native speaker of a pro-drop language, such as Spanish, may incorrectly omit subjects when constructing sentences in a non-pro-drop language like German. The likelihood and nature of transfer errors depend not only on structural differences between the languages but also on the learner's level of proficiency (Montrul, 2014). As learners become more proficient in the L2, they tend to make fewer transfer-based errors.

In the context of NLI, the underlying assumption is that classification algorithms can detect subtle linguistic patterns in learners' L2 that reflect L1 influence, such as deviations in syntactic structure, part-of-speech usage, or inconsistencies in lexicon and use these cues to identify the writer's native language. These linguistic traces provide support for theoretical approaches in SLA and help explore the phenomenon of cross-linguistic influence or transfer (Jarvis and Crossley, 2012; Tsur and Rappoport, 2007).

Prior studies have consistently demonstrated the effectiveness of n-gram-based features for NLI. For instance, several studies have found character n-grams to be among the most discriminative features (Koppel et al., 2005; Markov et al., 2022), while others have reported high classification performance using lexical and part-of-speech (POS) n-grams. Jarvis et al. (2013), for example, achieved an accuracy of 83.6% using word n-grams, and Markov et al. (2022) reported accuracies ranging from 80% to 90% using character n-grams with high values of n (up to n=9). Furthermore, combinations of POS n-grams and error features have yielded precision and recall scores exceeding 80% (Aharodnik et al., 2013; Kochmar, 2011). For example, Kochmar (2011) reported 84% accuracy using a combined feature set of character n-grams, POS n-grams, and corpus-derived error rates for classifying Romance and Germanic languages. In contrast, fewer NLI studies have examined features that reflected language productivity and lexical diversity, such as function word and content word ratios, mean length of utterance in words, and type-token ratio. However, these features may also be informative, as learners may exhibit L1-influenced lexical and syntactic patterns in their writing. For example, some studies emphasized that function words have contributed to high-performing models when combined with n-grams and error features (Koppel et al., 2005; Wong and Dras, 2009).

Studies exploring NLI with LLMs have yielded mixed results. For example, Lotfi et al. (2020) reported an accuracy of 89% on the test set for TOEFL11 and 94.2% on 5-fold cross validation for ICLE Corpus using GPT-2. These results indicated that the open source GPT model (GPT-2) was higher than the traditional machine learning approaches, with the best performing model achieving 88.2% accuracy with the SVM (Malmasi et al., 2017). However, studies have shown lower performance for BERT-like LLMs compared to a GPT-2 model. For example, 80.8% accuracy was attained using BERT-base-uncased when tested on the TOEFL11 corpus test set (Lotfi et al., 2020). Importantly, few studies have directly compared traditional machine learning and LLM-based approaches within the same experimental framework.

Moreover, LLM performance appears to be sensitive to dataset size. For instance, Steinbakken and Gambäck (2020) found that BERT-based models reached 85.3% accuracy on the TOEFL11 dataset, but accuracy improved to 90.2% when using the larger Reddit-L2 dataset. These findings suggest that LLMs require larger and more diverse data to perform optimally, highlighting the need for further research that examines LLM effectiveness on datasets of varying sizes and content types.

The nature of the data itself also plays a critical role in classification performance. Most NLI studies have relied on the TOEFL11 corpus, which contains argumentative essays on various topics (Malmasi and Dras, 2015). While high performance has consistently been reported for this dataset (Koppel et al., 2005; Malmasi and Cahill, 2015), its topic-based structure introduces the risk of content bias, particularly when using content-sensitive features such as word and character n-grams. Studies on cross-corpora evaluation have found that

Features	Description
Narrative Microstructure	
MLU(w)	Mean Length of Utterance in Words: ratio of total word tokens to total number of sentences per text
FCR	Function-to-Content Word Ratio. *Function words:* auxilliaries, pronouns, determiners, prepositions, conjunctions, particles. *Content words:* nouns, verbs, adjectives, adverbs.
TTR	Type-Token Ratio: ratio of unique words to total words.
POS_fc	Part-of-speech frequency counts (e.g., the number of NOUNs, VERBs, ADJs, etc. per text).
N-gram Features	
POS n-grams	Sequences of POS tags (e.g., "DET NOUN", "NOUN VERB ADV").
Word n-grams	Sequences of words (e.g., "he walked", "the baby ate")
Character n-grams	Sequences of characters (e.g., "ing", "ys", "ies")

Table 1: Overview of linguistic productivity, language diversity, and n-gram features included in the study.

genre-diverse corpora produce a higher accuracy when tested on a genre-specific corpus than the reverse. However, overall accuracy remains relatively low, as many features useful for NLI are genre-dependent (Malmasi and Dras, 2015). More structured datasets, such as those based on picture- or video-based narratives, can be used as an alternative for a more consistent feature extraction across participants.

The current study addresses these gaps by comparing a traditional feature-engineering approach with supervised machine learning classifiers and the fine-tuning of LLMs within a single experimental setup. We examine both previously validated feature sets, such as n-grams, and a complementary set of language productivity and diversity measures. This approach aims to assess whether these additional linguistic features enhance classification performance and provide deeper insights into L1-specific patterns in learner writing. To minimize topic-related bias in L1 identification, we apply both methods to a more homogeneous dataset.

3 Methods

3.1 Dataset

We used the COREFL corpus (Lozano et al., 2020). The corpus contained English L2 learner data of Spanish and German L1 backgrounds. Only learners with a lower advanced level of English proficiency (C1) were included in the study. The writers' age ranged from 18 to 60 years old. The data consisted of 84 German and 79 Spanish files with one file per participant. The participants watched a 4-minute video clip about Charlie Chaplin and summarized the story in a written essay.

Language	Total Files	VS	BDS
German	84	13	7
Spanish	79	17	7
Total	163	30	14

Table 2: Total number of files and the number of files used for validation and blind test sets for both language groups. VS = Validation Set. BDS = Blind Dataset.

3.2 Feature-Engineering Approach

The feature-engineering step focused on selecting and automatically extracting specific features that best characterized the data. The features for this study included two sets described in detail in Table 1.

The pre-processing step involved data cleaning and feature extraction. Data cleaning consisted of basic steps: removing special characters, removing punctuation, lowercasing, tokenization, and POS tagging. All features were extracted from narratives using bash scripts. The POS tagging was implemented using *en_core_web_trf* with Spacy Python package. All bash scripts and Python code is available on GitHub[1]:

The extracted features were used as input for supervised machine learning binary classification. We implemented the Support Vector Machine clas-

[1]https://github.com/AliyahVanterpool/ml_features_vs_llm.git

Testing Set	Feature	Accuracy	F1
VS	All	0.70	0.70
BDS	All	**0.79**	**0.78**
VS	MTF	0.67	0.64
BDS	MTF	0.64	0.64
VS	POS_fc	0.63	0.63
BDS	POS_fc	0.71	0.71
VS	MLU(w)	0.67	0.62
BDS	MLU(w)	0.50	0.48
VS	TTR	0.60	0.57
BDS	TTR	0.50	0.33
VS	FCR	0.60	0.55
BDS	FCR	**0.79**	**0.78**

Table 3: Highest accuracy and F1 for language productivity and diversity features. VS = Validation Set. BDS = Blind Dataset. All = MLU(w), TTR, FCR, POS_fc. MTF = MLU(w), TTR, FCR.

sifier (SVM; Cortes and Vapnik, 1995) with linear and rbf kernels, Logistic Regression (Cox, 1958; Hosmer Jr et al., 2013), and K-Nearest Neighbors (KNNs) classifier (Cover and Hart, 1967). We compared the performance across feature sets and classifiers. Table 2 shows the total number of files and those allocated to the validation and blind test sets for both language groups. The following models were included in the analysis: 1) all features (ALL = MLU(w) + TTR + FCR + POS_fc); 2) each feature from ALL models individually; 3) MTF feature set (MLU(w) + TTR + FCR); 4) word n-grams (bigrams, trigrams); 5) POS n-grams (bigrams, trigrams); and 6) character n-grams (four-grams to nine-grams).

The training dataset was created using 90% of the entire dataset, while 10% was held out for the blind test set. 80% of the training data was used for training and 20% for validation. The classifier training and parameter tuning was implemented using scikit-learn package in Python (Pedregosa et al., 2011). The kernels and c-parameter were explored to evaluate which models performed the best.

We also looked at Sentence-BERT embeddings (Reimers and Gurevych, 2019). We implemented *all-MiniLM-L6-v2*, a distilled BERT-based model from the Sentence Transformers. These embeddings were used as feature vectors for downstream binary ML classification with SVM and Logistic Regression. We evaluated the performance of both classifiers and reported the accuracy for the blind test set.

3.3 LLM Approach

For the LLM approach, we explored BERT-like models (Devlin et al., 2019). These models were ALBERT (Lan et al., 2019), BERT-base-multilingual-cased, BERT-base-uncased, Distil-RoBERTa-base, DistilBERT-base-uncased, and XLM-RoBERTa-base. We fine-tuned these pre-trained models for sequence classification using the learner corpus. The fine-tuning process involved training each model on 80% of the entire dataset, with 20% validation for a maximum of 3 epochs with a learning rate of 1e-5 and a batch size of 8. We experimented with frozen layers, however the models with all layers demonstrated better results and thus were reported in our study.

Additionally, we evaluated GPT-4 performance across three runs in two ways - 1) when tested on the blind dataset with class labels provided; and 2) no labels given. When GPT-4 was provided with labeled data, the prompt was: *The following English text is written by either a native German speaker or native Spanish speaker. What is the native language of the writer of this text: German or Spanish? Explain your choice in 1-2 sentences.* The prompt for unlabeled data was: *The following English text is written by a non-native speaker. What is the native language of the writer of this text? Explain your choice in 1-2 sentences.*

3.4 Testing and Evaluation Metrics

For the feature-engineering approach, we used three testing techniques: validation set, blind dataset, and k-fold cross validation (CV). The validation split was 20% of the training dataset. The blind dataset consisted of 10% of the entire dataset held out for testing and not included in the training. The blind dataset included 7 random files for each label (14 files in total). For K-fold CV, k ranged from 5-10 and the best k (k = 7) was reported. We reported the results for the SVM classifier since it demonstrated the best performance. We evaluated the best accuracy for linear and rbf kernels, and for C-parameter value. We also calculated feature importance scores with Random Forest Classifier for word bigrams and trigrams from the blind test set to identify those n-grams that impacted the classifier's decisions.

For the LLMs approach, we looked at both the validation and blind dataset results and reported the blind test results. Cross-validation techniques was computationally expensive for the BERT-like mod-

els, hence those were not reported for this study.

4 Results

4.1 Feature Engineering Approach

For the feature engineering approach, the best performing model was the model with all productivity and diversity features combined (ALL; 79% accuracy and 78% F1-score). K-fold CV for all feature models produced the highest mean accuracy of 72.5%. The productivity and diversity measures are described in Table 3. Models with individual features showed the highest accuracy (79%) and F1-score (78%) for function-to-content ratios with k-fold CV at 57.7%.

Among n-gram features, word bigrams and trigrams as well as character four- and five-grams attained the highest accuracy and F1 for both the validation and blind datasets. These results are shown in Table 4. The highest accuracy of 100% (95% CI [0.78, 1.00], Wilson interval) was achieved by word bigrams when tested on the blind dataset. The k-fold CV accuracy with k = 7 was 90.6% for word bigrams. The best models for the validation set were word bigrams and trigrams, as they acquired an accuracy of 93% (95% CI [0.79, 0.98], Wilson interval). The k-fold CV accuracy for trigrams was 91.3%. POS bigrams had the highest accuracy when tested on the validation set (87%; 95% CI [0.70, 0.95], Wilson interval) and POS trigrams acquired the highest accuracy of 79% (95% CI [0.52, 0.92], Wilson interval) when tested on the blind dataset. The K-fold CV accuracy was 81.2% for POS bigrams and 82.5% for POS trigrams. Overall, the results for n-gram features demonstrated the highest accuracy and stable results across different testing techniques (validation, blind test, and K-fold CV).

The highest accuracy for sentence embeddings with SVM was 93% and 78% with logistic regression when tested on the blind dataset. Additionally, the SVM embedding results performed better than the language productivity and diversity measures. However, sentence embeddings results were lower than the word bigram results. The best performing models for the feature-engineering approach, including sentence embeddings, are displayed in Figure 1.

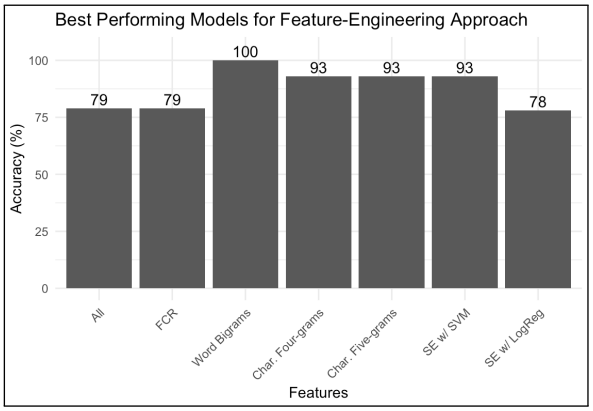

Figure 1: Best performing models for feature-engineering approach. ALL, FCR, word bigrams, SE w/ SVM, and SE w/ LogReg when tested on the blind dataset. Character four-grams and five-grams when tested on the validation set. SE w/ SVM = Sentence-embeddings with SVM. SE w/ LogReg = Sentence-embeddings with Logistic Regression.

4.2 LLM Approach

The LLM approach was separated into two parts: (1) BERT-like models with a classification layer, and (2) GPT-4 results. For the first part, we reported the performance of the blind test set. For the second part, we provided the average GPT-4 results across three runs for prompting with and without labels.

For BERT-like models, the highest accuracy of all six models is displayed in Figure 2. This included only models with all layers, as models with frozen layers demonstrated lower accuracy. The results show that two small BERT-like models and one large model performed with the highest accuracy: ALBERT (83%), DistilBERT-base-uncased (81%), and BERT-base-uncased (73%). As ALBERT and DistilBERT-base-uncased are lighter models, these results demonstrate that lighter models perform better than larger models for this studies data. Additionally, compared to previous BERT results, the BERT results in this study outperformed previously reported results 83% vs 80.8% (Lotfi et al., 2020), but lower than cross-corpora comparison accuracy of 85.3% when using SVM and FFNN base classifiers (Steinbakken and Gambäck, 2020).

For GPT-4, we performed 3 runs for with-label and no-label options with temperature set to 0.2. The accuracy when labels were provided was 92.9% for the first two runs – with only one file being mislabeled, and 85.7% for the third run. The average accuracy of the three runs was 90.48%. When no labels were given, GPT-4 attained an accuracy of 50% for all three runs. German was misclassified as Turkish and Russian, and Spanish was mislabeled as Italian, French, and Turkish.

Testing Set	N-gram Type	N-gram	Accuracy	F1-score
VS	Word	Bi	**0.93**	**0.93**
BDS	Word	Bi	**1.00**	**1.00**
VS	POS	Bi	0.87	0.86
BDS	POS	Bi	0.71	0.71
VS	Word	Tri	**0.93**	**0.93**
BDS	Word	Tri	0.86	0.85
VS	POS	Tri	0.80	0.80
BDS	POS	Tri	0.79	0.78
VS	Character	Four	**0.93**	**0.93**
BDS	Character	Four	0.79	0.78
VS	Character	Five	**0.93**	**0.93**
BDS	Character	Five	0.86	0.86

Table 4: Accuracy and F1 for n-gram models. The best models are in bold. VS = Validation Set. BDS = Blind Dataset.

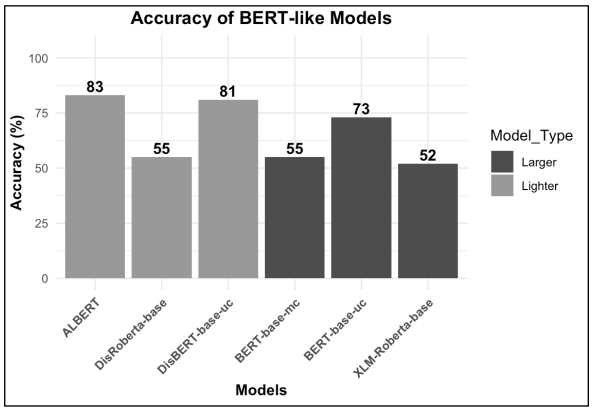

Figure 2: The results for BERT-like models. Dis = Distil. UC = Uncased. MC = Multilingual-cased.

5 Discussion

The contribution of the current study is two-fold. First, we compared two approaches - feature engineering and fine-tuning BERT-like LLMs - within the same study. The results showed that the feature-engineering approach outperformed the LLM-based approach, highlighting the effectiveness of feature-engineering pipelines for the NLI task, particularly in scenarios with relatively small datasets. Second, we explored a type of data that differs from that used in most previous studies. Specifically, our dataset consisted of narratives written by participants in response to the same video-based stimulus, providing more consistency across texts than the corpora of topic-based argumentative essays commonly used in NLI research.

Word bigrams were the most effective features extracted from the data. This finding suggested that word bigrams can effectively distinguish between learners with Spanish L1 and German L1 backgrounds based on their English writing. These n-grams likely captured differences in vocabulary use, word choices reflecting possible morphosyntactic errors, and distinctive lexical-syntactic patterns (the combinations of word tokens) between the two groups, which could be evidence of language transfer from learners' native languages to their L2 English. For example, German L1 influences were seen in lexical choices such as 'small human being' instead of 'baby' (possibly influenced by 'kleines menschliches Wesen' in German) and 'perceives it' instead of 'notices it' (possibly from 'wahrnehmen' meaning both 'perceive' and 'notice' in German).

Spanish L1 transfer was also evident from morphosyntactic patterns, such as noun-pronoun gender disagreement (e.g., 'the baby… she'). The preposition use was another source of transfer for Spanish L1 writers. For instance, 'yells him' (from Spanish 'le grita') reflected the incorrect omission of a preposition possibly due to the Spanish verb allowing a direct object.

An analysis of function words (Figure 3) revealed no major quantitative differences in the frequency of POS categories between the two groups, except for prepositions: German L1 writers tended to use more prepositions in their narratives compared to the Spanish L1 group. Qualitative differences in preposition use were seen, for instance, in 'walking on the street' phrase, where Spanish L1 writers overused the preposition 'on' instead of 'in'. The above examples indicated instances of linguistic transfer which are in line with the previous

research on interlingual errors in Spanish-English bilinguals (Alonso Alonso, 1997). These patterns influenced the classifiers' decisions in disambiguating the two classes in the current study.

The Random Forest classifier also highlighted the bigrams that contributed to classification. For example, 'next to' was predominantly used by German L1 writers, while 'he is' and 'to leave' appeared more frequently in Spanish L1 texts. These features further illustrated the distinct lexico-syntactic choices between the two L1 groups. Overall, our results suggested that even when the differences between learner groups were subtle, traditional ML classifiers were capable of detecting them based on word n-grams and related surface-level patterns.

Importantly, our findings aligned with previous research that has identified word n-grams as effective features for NLI (e.g., Koppel et al., 2005; Jarvis et al., 2013) and demonstrated comparable or higher accuracy. For example, Jarvis et al. (2013) found that word and POS n-grams acquired an accuracy of 83.6% when using 10-fold cross validation and 90.1% when used on an ensemble classifier. However, our results cannot be directly compared because the number of classes and the nature of the data were different in the current study. In addition, word-based n-grams successfully captured class-specific differences from the dataset that consisted of written narratives based on the same video stimulus, thus reducing the risk of content bias from topic-related vocabulary.

Other n-gram types, including character n-grams and POS n-grams, also performed well. For character n-grams, we explored a range from four characters to nine characters. The best results were achieved with four- and five-grams. These likely captured class differences in short function words, such as prepositions, which are often markers of L1 influence (Jarvis and Odlin, 2000). The high performance of POS n-grams may be attributed to distinctive patterns in part-of-speech use and distribution across the two groups. For example, the qualitative analysis of the data suggested that German L1 writers relied more on subordinate clauses, a pattern consistent with transfer from German's preference for embedded structures (Swan and Smith, 2001).

Among lexical diversity and productivity features, the model combining all measures (function-to-content word ratio, MLU(w), TTR, and POS frequency counts) achieved the highest accuracy and

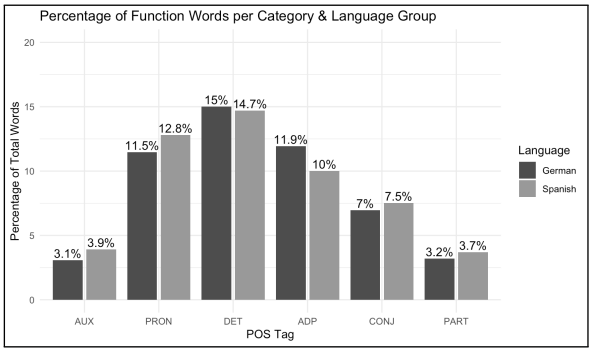

Figure 3: Percentage of function words per category and language group. AUX = Auxillaries. PRON = Pronouns. DET = Determiners. ADP = Adposition (Preposition). CONJ = Coordinating and Subordinating Conjunctions. PART = Particles.

F1 score (see Table 3). However, these results were still lower compared to the n-gram-based models. Notably, the function-to-content word ratio (FCR) emerged as the strongest individual predictor in this group, showing the highest performance on the blind test set. These patterns suggest that both n-gram features and FCR effectively captured differences in language productivity and distributional tendencies across German and Spanish L1 groups. Lexical diversity features, such as TTR, did not show high accuracy (50%) for the blind dataset. Exploring other TTR metrics (e.g., Moving-Average Type-Token Ratio (MATTR)) might provide a different result given the length-sensitive nature of the feature.

The sentence embeddings approach also outperformed the fine-tuning of BERT-like classification models with 93% accuracy. By encoding contextual relationships and sentence-level semantics, these embeddings were able to capture subtle differences in linguistic patterns between the two L1 groups in their English L2. These findings are in line with the previous research that indicated the utility of the embeddings approach for the NLI task and demonstrated that word embeddings together with string kernels were effective for L1 classification (Franco-Salvador et al., 2017).

Taken together, the results of the feature-engineering approach highlighted the robustness of both sparse vector surface-level features, such as n-grams, and dense sentence embeddings approach. Both methods were effective for distinguishing advanced learners' L1 backgrounds in written narratives.

The classification with BERT-like LLMs did not

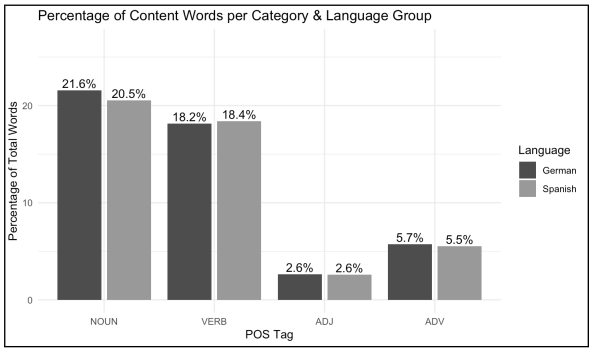

Figure 4: Percentage of content words per category and language group. ADJ = Adjective. ADV = Adverb.

perform on par with the feature-engineering approach. The highest accuracy within this group was achieved by the ALBERT model (83% accuracy, Figure 2), suggesting that lighter and more parameter-efficient architectures may be better suited for this task.

One possible explanation for the lower performance of BERT-like models is their sensitivity to dataset size and domain mismatch. Effective fine-tuning of these models typically requires large, diverse datasets to generalize better. In contrast, the relatively small and domain-specific nature of our dataset may have limited their ability to adapt. Additionally, while BERT models are designed for deep contextual understanding, this level of complexity may not be necessary for the current NLI task. Surface-level patterns, such as n-gram distributions and POS frequencies in our study, appear sufficient for distinguishing between L1 groups.

Furthermore, the results for the closed-source GPT-4 model revealed an average accuracy of 90.48%, which is similar to the sentence embeddings and word n-gram models. This performance was achieved using prompts that included labels, resembling a supervised approach. These findings align with previous studies investigating GPT models. For example, Zhang and Salle (2023) reported that GPT-4 achieved an accuracy of 91.7% on the TOEFL11 dataset. Similarly, Ng and Markov (2024) found that closed-source LLMs such as GPT-4 consistently outperformed open-source LLMs, regardless of fine-tuning. However, without labels, the closed-source GPT-4 performed poorly in our study.

Although both open- and closed-source models have demonstrated promising results for NLI, an important limitation of closed-source LLMs lies in the lack of transparency regarding their training data which raises concerns about reproducibility and potential biases in their outputs.

Overall, our results highlighted that traditional supervised machine learning techniques (e.g., SVM classifier) remain highly robust for low-resource NLI tasks. These models not only outperformed BERT-like LLMs but also achieved performance on par with the GPT-4 model. The lower results for BERT-like LLMs underscore their limitations in settings with domain-specific and scarce training data, including issues of limited interpretability and a higher risk of overfitting during fine-tuning.

6 Conclusion & Future Directions

In this paper, we compared two approaches for the NLI binary classification task: the traditional ML feature-engineering method and fine-tuning of BERT-like LLMs with a classification head. Our findings suggested that studies working with smaller, domain-specific datasets may benefit more from feature-engineering pipelines than from fine-tuning BERT-like LLMs. Frequency-based surface-level features were more sensitive to subtle differences in written narratives of similar content. While BERT-like models were less robust, lighter variants performed noticeably better than their larger counterparts on the small NLI dataset. Nonetheless, including other fine-tuning methods (e.g., DAPT, LoRA) could produce different results. The GPT-4 model also showed promising results when provided with labels; however, since the sources of its training data are not transparent, it is difficult to assess the generalizability and reliability of its performance. By evaluating both feature-engineering and BERT-like LLM approaches within the same study, we offered a direct comparison of their effectiveness for NLI.

Future studies could focus on datasets with structurally and topically consistent content across classes, which may reveal more subtle linguistic cues relevant for classification. It would also be valuable for future work to explore robust cross-validation techniques for LLMs, particularly when sufficient computational resources are available. Future research should continue to explore both traditional feature-engineering and LLM approaches, including closed-source LLM models without given labels, within the same experimental framework to better understand their comparative advantages across diverse domain-specific datasets.

References

Katsiaryna Aharodnik, Marco Chang, Anna Feldman, and Jirka Hana. 2013. Automatic identification of learners' language background based on their writing in czech. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1428–1436.

María Rosa Alonso Alonso. 1997. Language transfer: Interlingual errors in spanish students of english as a foreign language. *Revista alicantina de estudios ingleses, No. 10 (Nov. 1997); pp. 7-14*.

Stephen Pit Corder. 1975. Error analysis, interlanguage and second language acquisition. *Language teaching*, 8(4):201–218.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Marc Franco-Salvador, Greg Kondrak, and Paolo Rosso. 2017. Bridging the native language and language variety identification tasks. *Procedia computer science*, 112:1554–1561.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118.

Scott Jarvis and Scott A Crossley. 2012. *Approaching language transfer through text classification: Explorations in the detectionbased approach*, volume 64. Multilingual Matters.

Scott Jarvis and Terence Odlin. 2000. Morphological type, spatial reference, and language transfer. *Studies in second language acquisition*, 22(4):535–556.

Ekaterina Kochmar. 2011. *Identification of a writer's native language by error analysis*. Ph.D. thesis, Master's thesis, University of Cambridge.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628.

Robert Lado. 1957. *Linguistics across Cultures: Applied Linguistics for Language Teachers*. The University of Michigan Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A deep generative approach to native language identification. In *Proceedings of the 28th international conference on computational linguistics*, pages 1778–1783.

Cristóbal Lozano, Ana Díaz-Negrillo, and Marcus Callies. 2020. Designing and compiling a learner corpus of written and spoken narratives: Corefl. *What's in a Narrative*, pages 21–46.

Shervin Malmasi and Aoife Cahill. 2015. Measuring feature diversity in native language identification. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 49–55.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2022. Exploiting native language interference for native language identification. *Natural Language Engineering*, 28(2):167–197.

Silvina Montrul. 2014. Interlanguage, transfer and fossilization: Beyond second language acquisition. In *Interlanguage*, pages 75–104. John Benjamins Publishing Company.

Yee Man Ng and Ilia Markov. 2024. Leveraging open-source large language models for native language identification. *arXiv preprint arXiv:2409.09659*.

Terence Odlin. 1989. *Language transfer*, volume 27. Cambridge University Press Cambridge.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stian Steinbakken and Björn Gambäck. 2020. Native-language identification with attention. In *Proceedings of the 17th international conference on natural language processing (icon)*, pages 261–271.

Michael Swan and Bernard Smith. 2001. *Learner English: A teacher's guide to interference and other problems*, volume 1. Cambridge University Press.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.