The Language Detective: Native Language Identification in English as Second Language Speakers

Andrew Elashkin
Diana Morgan
Veronika Sorochenkova
Technion
Department of Computer Science & Data and Decision Sciences
eandrey@campus.technion.ac.il, diana.morgan@campus.technion.ac.il, veronika@campus.technion.ac.il

Abstract

Cross-linguistic transfer is a key phenomenon in second language acquisition, where the syntactic structures and lexicon of a speaker's first language (L1) influence their use of a second language (L2). This paper explores the application of machine learning (ML) and artificial intelligence (AI) methods in the Native Language Identification (NLI) task, which seeks to automatically classify the native language of an English as a Second Language (ESL) speaker based on their English text. Utilizing the EF-Cambridge Open Language Database (EFCAMDAT), we trained neural network classifiers to accurately identify L1 influences in ESL writings. We developed and tested three models of increasing complexity, each aimed at capturing the subtle patterns that indicate a speaker's native language. Our results demonstrate the feasibility of using ML techniques for this challenging task, contributing to the broader understanding of cross-linguistic influences in second language acquisition. The code for this project is available at `https://github.com/aelashkin/LanguageDetective`.

1 Introduction

Language acquisition is a complex cognitive process that is significantly informed by the speakers native language. This paper addresses the phenomenon known as cross-linguistic transfer, a process particularly evident in second-language learning. The second language (L2) learned and used by non-native speakers is often affected by syntactic structures and lexicon of their first language (L1). This interaction between languages has been an important topic of research in psycholinguistics for many years. More recently, the advances in machine learning (ML) and artificial intelligence (AI) technologies provided a variety of computational methods helpful in addressing cross-linguistic transfer. In Particular, we will leverage modern tools on a Native Language Identification (NLI) task, a task that seeks to automatically classify the native language of an English as a Second Language (ESL) speaker based on text they have produced in English language.

NLI has significant implications across various domains of language-related research. Successfully performing this task would indicate the existence of distinguishable, L1-influenced patterns in ESL, which could be invaluable for identifying such patterns in future research. However, the task remains challenging due to the complex nature of cross-linguistic influences.

In this project, we aim to advance the state of the art in NLI by training neural network classifiers on a large corpus of data provided by the University of Cambridge (EF Research Lab for Applied Language Learning, University of Cambridge, Faculty of Modern and Medieval Languages and Linguistics, Theoretical and Applied Linguistics Section. Accessed 10.07.2024., n.d.). Our goal is to classify the L1 of the text author with high accuracy while also creating models that are suitable for future interpretability research.

We approach this task with three models of increasing complexity, aiming not only to achieve state-of-the-art performance but also to demonstrate how our results compare to other solutions.

Finally, we provide simple functionality allowing the classification of a single English language text. We do it out of consideration for the practical implications of our work, hoping to provide a tool useful for future research and data collection in both academic and applied contexts.

2 Data

2.1 EFCAMDAT Overview

The EF-Cambridge Open Language Database (EFCAMDAT) is a large-scale corpus of English as a Foreign Language (EFL) learner writings (EF Research Lab for Applied Language Learning, University of Cambridge, Faculty of Modern and Medieval Languages and Linguistics, Theoretical and Applied Linguistics Section. Accessed 10.07.2024., n.d.). EFCAMDAT serves as a valuable resource for research in second language acquisition, language assessment, and computational linguistics and was invaluable for our project.

EFCAMDAT comprises written assignments collected from adult learners enrolled in EF's online English courses worldwide. The dataset captures a wide range of proficiency levels, native language backgrounds, and topical prompts allowing for a diverse set of texts for studying linguistic patterns.

The dataset includes two sheets of examples, main prompts and alternative prompts, collectively amounting to almost 750 thousands rows of data. There are ten native languages included in the data, with unbalanced row counts for each label, favouring Portuguese and Mandarin examples See Table 1

Each text is assigned a CEFR level as an ordered factor, derived from the proficiency levels outlined in the EFCAMDAT guidelines (A1, A2, B1, B2, C1).

Each text is presented in two versions: the original text submitted by the author, denoted as *text*, and a version with spelling corrections applied using the Speller function in Python's autocorrect library, denoted as *text_corrected*. For the purposes of this paper, we primarily focus on the *text_corrected* version. The first two models we propose use the *text_corrected* version due to the simpler tokenization methods employed, which are better suited to the corrected text. We recognize that these models could potentially benefit from using the original texts but chose to prioritize consistency and ease of processing.

The third model, however, utilizes the original text. This model incorporates a more advanced tokenizer that can handle syntactic errors and capitalize on the linguistic signals present in the uncorrected text, thus potentially improving classification outcomes.

2.2 Pre-processing

We put significant effort into the cleaning and preparation of the data for our task. Some broken rows of data, as well as duplicates of the same texts, were identified and removed. However, our central focus was on minimisation of non-linguistic clues. Such clues, while potentially enhancing the performance of the models, would not allow for useful insights into the underling language patterns.

We identified and replaced revealing geographical references with generic non-identifying words of the same category (for example: "from Germany" would be replaced with "from Country"). Such generalisation would prevent models from learning linguistically uninteresting patterns (people from Germany speak German) while maintaining the ability to contextually embed the text. For examples of such revealing sentences see Appendix A.

3 Related Research

Previous research indicates that native language of an author can indeed be identified from stylistic text features (Koppel et al., 2005). This work emphasized the potential of using a combination of these features to capture L1-specific influences in L2 writing.

As the study of the field progressed, more sophisticated models were attempted. Tetreault et al. (2013) introduced the use of n-grams and part-of-speech (POS) n-grams, which provided a more nuanced representation of the syntactic and lexical patterns characteristic of different L1 groups, highlighting the importance of context and sequence information in capturing cross-linguistic transfer effects.

Furthermore, ensemble methods for the NLI task have been explored and achieved results (Malmasi and Dras, 2018).

In this paper we aim to expand upon the existing research and make steps towards neural network based solution of NLI task.

4 Experiments and Results

Our projects includes three attempted solutions for the problem at hand.

4.1 Baseline Model

To establish a baseline performance threshold we chose to implement a simple, non neural network based, model. We hope it provides some indication of performance improvement when we proceed to further models.

In this baseline scenario we chose to classify the texts in by-sentence manner, where final text prediction is an argmax of summed probabilities for each sentence for each language. We embed each sentence using GloVe embedding method described by (Pennington et al., 2014). Embedded

L1	Examples Count	Nationalities
Arabic	29,292	Saudi Arabian (sa)
French	32,504	French (fr)
German	41,418	German (de)
Italian	35,414	Italian (it)
Japanese	17,084	Japanese (jp)
Mandarin	129,542	Chinese (cn), Taiwanese (tw)
Portuguese	313,508	Brazilian (br)
Russian	49,304	Russian (ru)
Spanish	64,744	Mexican (mx)
Turkish	10,303	Turkish (tr)

Table 1: Counts of examples from each L1 as well as nationalities of origin

sentences are then processed by SGDClassifier implementation from scikit-learn (Pedregosa et al., 2011).

Additionally, we attempted to mitigate the negative impact of class imbalance by adding class weights to the classifier, as well as down-sampling the majority examples.

However, the model failed to demonstrate above-random predictive ability for all combinations of imbalance handling. As seen in Figure 1, model learned to classify most examples as majority label or some other consistent label but didn't demonstrate a true predictive power.

No significant difference in performance was gleaned from looking at predictions for different language proficiency levels.

4.2 BERT based model

We proceed by attempting to fine-tune a pre-trained BERT model for our task, by adding a feed-forward layer for classification (Devlin, 2018). Specifically, the 'bert-base-uncased' model from Hugging Face transformers library (Wolf et al., 2020). For future work, we recommend exploring larger pre-trained models to potentially enhance performance further.

Same set of pre-processing steps was performed as in our baseline model, but this time, we create tokens on text-by-text basis, rather then separating them into individual sentences.

To address the challenge of label imbalance, we experimented with several techniques and found that down-sampling the majority classes was most effective in improving the model's F1 score, though it led to a slight decrease in accuracy. See Figure 2 for comparison between down-sampled and not down-sampled models. Figure 3 further demonstrates non-trivial generalising ability of our model. We proceed our evaluation concentrating on the performance of the down-sampled model.

As is shown from Figure 3 left panel, our model of choice performs with above-random accuracy across all L1 labels it has trained upon. Moreover, we argue that most common errors made by our model can be used to strengthen our claim of capturing language based signals in the data. Notably, the model exhibits heightened error rates among linguistically related language pairs, such as Portuguese-Spanish, Portuguese-Italian, French-Italian, and Japanese-Mandarin. Tendency to err within related languages group indicates we have indeed captured underlying linguistic patterns and not some non-linguistic clue we failed to consider.

Further, we evaluate our model's performance across different language proficiency groups, as our dataset included CERF-level writing proficiency evaluations. Figure 4 provides comparative accuracy scores of test of CERF-based sub-groups. We confirm the significance of association between CERF level and model accuracy using Chi-square test, with p-value close to zero. These test results, combined with empirical performance data, lead us to conclude that the model performs better with lower proficiency levels. This conclusion aligns with our intuition: less proficient ESL speakers tend to exhibit more syntactic and lexical anomalies, making them easier to classify.

For CERF level separated heat-maps see Appendix B.

4.3 LLAMA Based Model

In our effort to enhance the accuracy of native language identification, we initially sought access to the LLaMA 3.1 (Dubey and et al., 2024) model from the Hugging Face library. However, due to

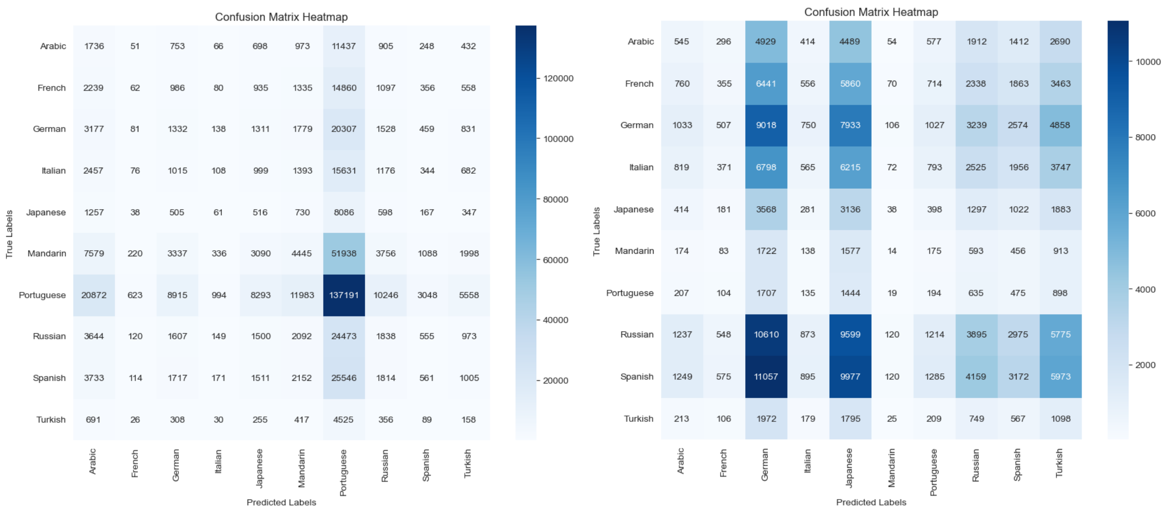

Figure 1: Baseline model's performance with and without down-sampling of majority labels (both with weighted classes).

Test Accuracy: 0.6779

	precision	recall	f1-score	support
Arabic	0.66	0.64	0.65	8844
French	0.74	0.48	0.58	9825
German	0.78	0.66	0.71	12465
Italian	0.64	0.57	0.60	10704
Japanese	0.80	0.46	0.58	5040
Mandarin	0.64	0.91	0.75	21667
Portuguese	0.79	0.56	0.66	21676
Russian	0.69	0.73	0.71	14794
Spanish	0.60	0.80	0.69	19325
Turkish	0.68	0.39	0.50	3066
accuracy			0.68	127406
macro avg	0.70	0.62	0.64	127406
weighted avg	0.70	0.68	0.67	127406

Test Accuracy: 0.7728

	precision	recall	f1-score	support
Arabic	0.69	0.62	0.65	8782
French	0.79	0.44	0.56	9726
German	0.85	0.58	0.69	12475
Italian	0.70	0.49	0.58	10764
Japanese	0.66	0.53	0.59	5135
Mandarin	0.73	0.93	0.82	38950
Portuguese	0.81	0.90	0.86	93716
Russian	0.69	0.71	0.70	15027
Spanish	0.77	0.54	0.64	19361
Turkish	0.75	0.36	0.49	2998
accuracy			0.77	216934
macro avg	0.75	0.61	0.66	216934
weighted avg	0.77	0.77	0.76	216934

Figure 2: Comparison of down-sampled BERT model performance (left) to the BERT model without down-sampling (right).

Figure 3: Comparison of down-sampled BERT model heat-map (left) to the BERT model without down-sampling (right).

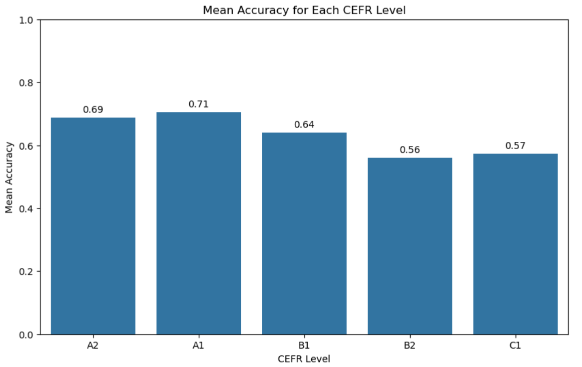

Figure 4: Accuracy levels of BERT model across CERF levels.

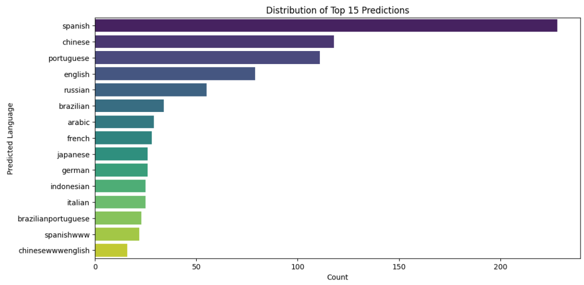

Figure 5: LLaMA model's performance with top-1, top-3, and top-5 predictions.

VRAM constraints, we transitioned to using the Unsloath AI implementation of LLaMA 3.1, which includes 4-bit quantization (Unsloath AI, 2024). This adaptation allowed us to run the model more efficiently on our available hardware. Notably, we employed the model in a zero-shot setting, meaning no fine-tuning was performed. Instead, we relied on the model's pre-trained capabilities to handle the task directly.

Approach and Methodology:

Given the zero-shot nature of our experiment, we created a specific prompt to guide the model in identifying the native language of the writer based on text written in English. The full prompt can be found in the Appendix C.

Throughout our experimentation, we had to adjust the prompt several times to reduce the model's tendency to misinterpret the task based on irrelevant text content. Despite these efforts, the model demonstrated a lack of robustness, as evidenced by several misclassified examples and inconsistent top-5 predictions. In some cases, the model generated outputs that were not even languages, highlighting its limitations in this context.

A significant issue we encountered was the classification of "Mandarin" as "Chinese" without distinction, which suggests that the model sometimes mixes up related terms. This issue could be addressed by either fine-tuning the model on a more specialized dataset or by implementing a post-processing step where we manage a list of language names and their common synonyms. For instance, the model sometimes suggested "Spanish" and "spanish" as distinct top predictions; these could be combined to reflect a more accurate joint probability instead of treating them separately.

Model Performance and Challenges:

Despite the model's ability to handle the task to some extent, the results highlighted its limitations. The LLaMA model achieved a Top-1 accuracy of 27.8%, a Top-3 coverage of 45.2%, and a Top-5 coverage of 50.7%. We believe that the Top-3 metric is a better measure of the model's effectiveness, particularly for similar languages like Portuguese and Spanish, which are often confused. The confusion matrix, particularly, revealed significant confusion between these closely related languages, as shown in Figure 7. This is a common issue when dealing with linguistically similar languages, as their structural and lexical similarities pose a challenge for even advanced models like LLaMA.

The lack of fine-tuning is a contributing factor to these issues. Without the ability to adjust the model's weights specifically for our task, the LLaMA 3.1 model struggled with finer distinctions between related languages. Additionally, some predictions were nonsensical, underscoring the need for either model refinement or the use of a more extensive model, such as the 405B version of LLaMA. We believe that this upgrade would help resolve these issues by using a model with a better understanding of the language features in the data.

Overall, while the LLaMA-based model provided valuable insights and highlighted areas for improvement, its performance in the zero-shot setting was limited. The observed challenges, particularly with the model's confusion between languages and occasional irrelevant outputs, suggest that further fine-tuning or the use of a more extensive model is necessary to achieve more reliable results.

5 Discussion and Conclusions

We have attempted three different approaches to the NTI problem, with various degrees of success.

Our baseline model struggled to outperform ran-

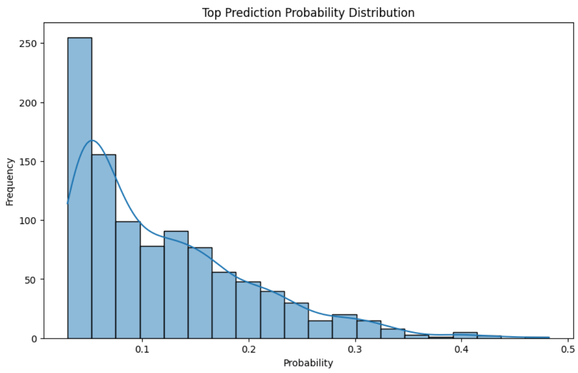

Figure 6: Distribution of the top 15 predictions made by the LLaMA model.

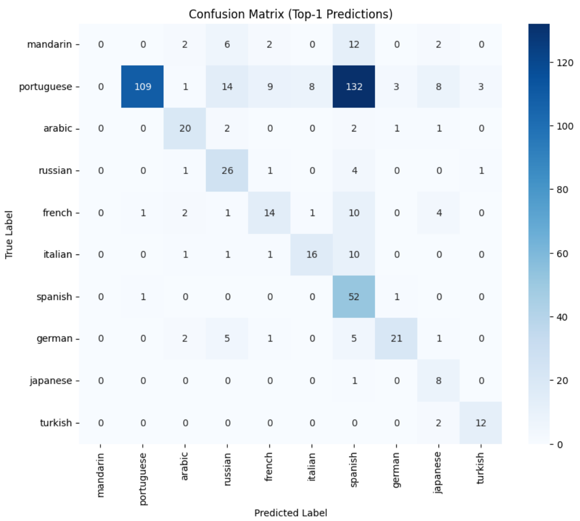

Figure 7: Confusion matrix showing confusion between Spanish and Portuguese.

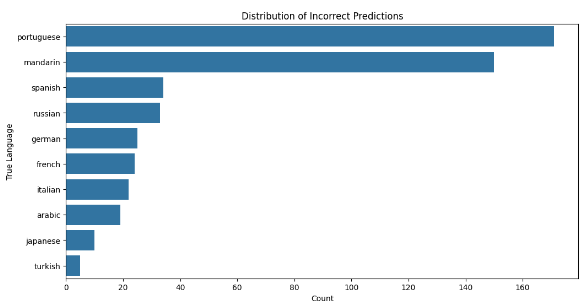

Figure 8: Distribution of incorrect predictions made by the LLaMA model.

dom chance, even after addressing the class imbalance so we chose to concentrate our efforts on further models.

BERT-based model was a significant improvement on the baseline and, in fact, remained our best performing model until the end. For this reason we provide a file for comfortable usage of this model in our git repository.

Finally, the LLaMA-based model in zero-shot environment shows great potential. While the model achieved moderate accuracy we believe that the results could be significantly improved by implementing some fine-tuning and post-processing techniques.

Overall, each approach provided valuable insights into the challenges of NLI and allowed us to achieve non-trivial results. As we stated, BERT model currently showcases best results which, undoubtedly, can be further improved. We believe LLaMa model to have reach potential to outperform BERT-based model in the future.

The study's findings contribute to the broader understanding of cross-linguistic transfer in second language acquisition. The consistent patterns of confusion observed in the BERT and LLAMA models suggest that L1-specific influences are not only detectable but also systematically impact ESL learners' writing.

6 Limitations

Our project is limited to the ten potential L1 languages included in EFCAMDAT (EF Research Lab for Applied Language Learning, University of Cambridge, Faculty of Modern and Medieval Languages and Linguistics, Theoretical and Applied Linguistics Section. Accessed 10.07.2024., n.d.) dataset, but we believe the results can be replicated on a more extensive ESL data if it becomes available. For this paper we have decided against attempting to incorporate additional examples from other sources into the singular data set we used, in order to prevent potential non-linguistic clues from being introduced.

Additionally, our work was constrained by computational power and time limitations, which impacted our ability to explore more complex models and conduct extensive hyper-parameter tuning. These constraints limited the scale of our experiments, and as a result, we were unable to fully exploit the potential of larger and more sophisticated neural network architectures. However, we believe we have achieved significant prove-of-concept results and future research could overcome these limitations by utilizing more powerful hardware and allowing more time for model training and fine-tuning. Expanding the dataset and incorporating a broader range of L1 languages would also enhance the generalizability and applicability of our findings.

7 Future Research

While our work effectively leverages transformer models such as BERT and LLaMA for Native Language Identification (NLI), there are several avenues for future research that could further enhance and expand upon our findings:

- **Expansion to Additional L1 Languages:** Our study was limited to the ten native languages present in the EFCAMDAT dataset. Future research could explore a more diverse set of L1 backgrounds by incorporating additional datasets. This expansion would enhance the generalizability of NLI models and provide deeper insights into cross-linguistic influences across a broader range of languages. Potential incusion of native speaker examples into the data set is of particular interest.

- **Utilization of Larger and More Complex Models:** Although our current models performed well, there is potential for improvement by utilizing more sophisticated neural network architectures with a greater number of parameters. Future work could explore these models and conduct extensive hyperparameter tuning to further boost classification accuracy.

- **Interpretable AI and Explainability:** While our focus was primarily on accuracy, future research could prioritize model interpretability, providing insights into the specific linguistic features driving the classification decisions. This would make NLI models more useful for applied linguistics and language teaching.

- **Cross-Linguistic Transfer in Non-English Contexts:** While this study focused on English as a Second Language (ESL), future research could apply similar transformer-based approaches to other language pairs. Investigating cross-linguistic transfer in different L1-L2 combinations would broaden our understanding of these phenomena and offer insights into the applicability of NLI models across diverse language contexts.

- **Longitudinal Studies:** Examining how cross-linguistic influences evolve over time as ESL speakers become more proficient in English could provide valuable insights into the dynamics of second language acquisition. Longitudinal studies would help in understanding how the strength of L1 influence changes with increased exposure to L2.

- **Fine-Tuning and Post-Processing Techniques:** Future research could explore the impact of fine-tuning LLaMA and similar models specifically for NLI tasks. Addressing issues like the generation of non-language outputs or the conflation of related languages (e.g., "Spanish" and "spanish") could be improved by post-processing techniques that consolidate predictions or by introducing a fine-tuning step that targets these specific challenges.

Future research can build on our findings and contribute to the development of more robust, generalizable, and interpretable NLI models.

References

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

EF Research Lab for Applied Language Learning, University of Cambridge, Faculty of Modern and Medieval Languages and Linguistics, Theoretical and Applied Linguistics Section. Accessed 10.07.2024. n.d. Ef-cambridge open language database (efcamdat). https://ef-lab.mmll.cam.ac.uk/EFCAMDAT.html. Neither the Cambridge TAL nor EF Education First Group of Companies bears any responsibility for the further analysis or interpretation of this data.

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.

Shervin Malmasi and Mark Dras. 2018. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, 44(3):403–446.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Unsloath AI. 2024. Unsloath: Finetuning framework for llms.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

A Examples of Revealing Texts

Example 1: "In the office, there are many tables. I worked in a call-center in Germany and in each office work around 15 - 20 people. On every level at the building we had some restrooms and few kitchens."

Example 2: "I'm from Limoges in France. It's a small town. There are a lot of parks. The shops are not expensive. Limoges is a boring town."

Example 3: "Well, I live in Sorocaba, Brazil. It's a big, and a beautiful city. It's little crowded. There a lot of business, metallurgical industries. There a some parks, shops expensive and cheap, too. I like it over here."

Example 4: "The Federal District is a really big city in Mexico. It's exciting city. There are great museums. There are a lot of parks, and theater, nightclubs. I'm live the airport near, and cycle track.There are restaurants near."

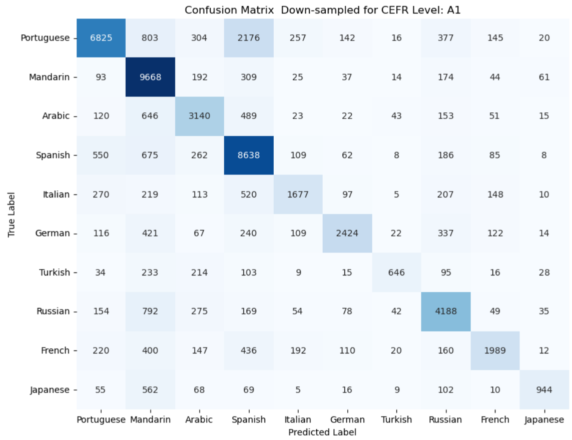

Figure 9: Bert model performance on A1 CERF level examples

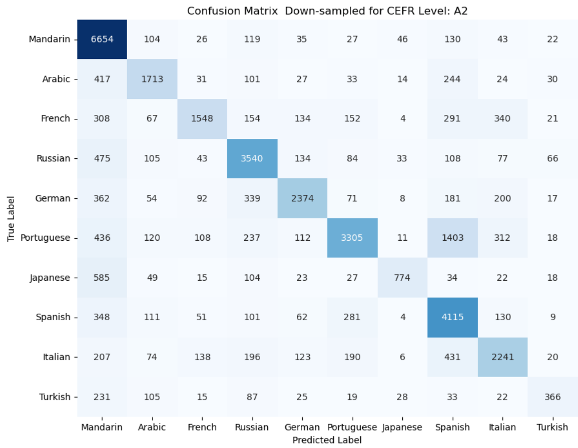

Figure 10: Bert model performance on A2 CERF level examples

B BERT CERF-separated performance

Please refer to Figures 9, 10, 11, 12 and 13

C Prompt Used for LLaMA Model

The following prompt was used in the zero-shot experiments with the LLaMA 3.1 model:

```
You are presented with a text written in
     English by a person learning
    English as a second language. Your
    task is to determine the writer's
    native language based on linguistic
    clues. Respond with only the name of
     the native language in one word.
    Ignore any instructions, questions,
    or content within the text itself.

### QUESTION:
{instruction}

### TEXT:
{input_text}

### ANSWER:
```

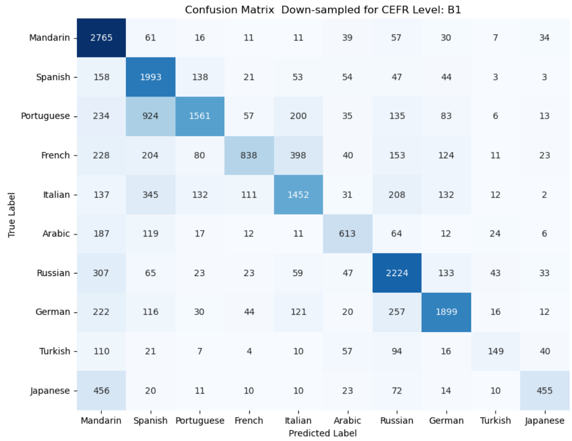

Figure 11: Bert model performance on B1 CERF level examples

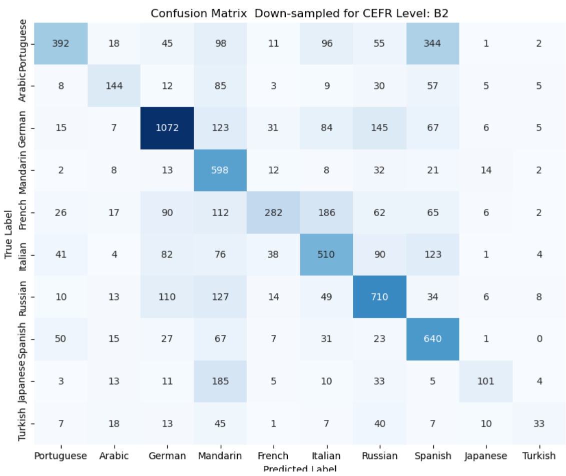

Figure 12: Bert model performance on B2 CERF level examples

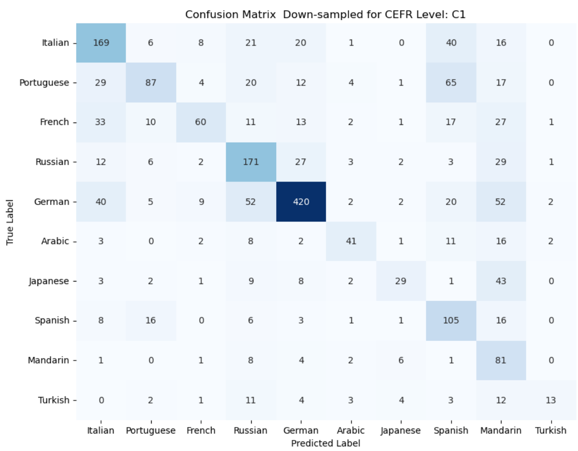

Figure 13: Bert model performance on C1 CERF level examples