Mitigating Learnerese Effects for CEFR classification

Rricha Jalota*, Peter Bourgonje+, Jan van Sas+, and Huiyan Huang+

*Universität des Saarlandes
+Morningsun Technology GmbH

Abstract

The role of an author's L1 in SLA can be challenging for automated CEFR classification, in that texts from different L1 groups may be too heterogeneous to combine them as training data. We experiment with recent debiasing approaches by attempting to devoid textual representations of L1 features. This results in a more homogeneous group when aggregating CEFR-annotated texts from different L1 groups, leading to better classification performance. Using iterative null-space projection, we marginally improve classification performance for a linear classifier by 1 point. An MLP (e.g. non-linear) classifier remains unaffected by this procedure. We discuss possible directions of future work to attempt to increase this performance gain.

1 Introduction

The need for automated methods in establishing both the readability of a piece of text and the level of linguistic proficiency of its author has been recognised decades before most students started writing essays, compositions and other homework assignments on computers. Motivations for creating such automated methods are diverse. Seminal work by Page (1966) focused on alleviating work load of language teachers and fast turn-around of writing feedback to language students. Since then, much progress has been made, and a comprehensive overview of original and still standing challenges in this field is presented by Beigman Klebanov and Madnani (2020). Related to this is the line of research on grammatical error correction (Leacock et al., 2010; Bryant and Ng, 2015), accompanied by a number of shared tasks (Ng et al., 2013, 2014; Bryant et al., 2019).

Much of the work in this sub-field of NLP is usually aggregated under the label *Automated Essay Scoring*[1]. Scoring an essay, however, depends on a number of factors related to the background of the author and moreover is not just about grading the quality of language usage, but usually also about the quality of content. The same essay about basic concepts of quantum physics may receive a high grade when written by a child in elementary school, but a considerably lower grade when written by a post-graduate physics student. A framework focusing solely on second language (L2) level skills, attempting to propose an objective (i.e., independent of native language) six-point scale is represented by the CEFR[2] levels. Since our use case is establishing the proficiency level of L2 language learners and providing them with feedback on how to improve, we experiment with CEFR classification.

While the nature of the influence of one's native language (L1) on Second Language Acquisition (SLA) is a topic of ongoing debate (Richards and Rodgers, 2014) and the terms being used are dependent on the assumed framework (*interference* (Weinreich, 2010), *transfer* (Lado, 1957; Selinker, 1969), *influence* (Smith and Kellerman, 1986)) the fact that there is interaction is uncontroversial. This L1 interaction is problematic in the sense that a classifier trained on texts written by native speakers of Chinese may perform poorly on texts written by native speakers of Portuguese, for example.

Inspired by recent successes in debiasing embeddings-based representations for particular traits (Manzini et al., 2019; Sun et al., 2019; Ravfogel et al., 2020; Karimi Mahabadi et al., 2020; Chowdhury et al., 2022), we set out to dispose the representations that feed into the classifier of traits that can be taken as signs of L1 influence, to train a single CEFR classifier -devoid of L1 features (i.e. *learnerese*)- that improves its performance when trained on aggregated data from different native

[1] Or variations thereof: Automated Essay Grading, Automated Writing Evaluation, etc.

[2] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 14 - 21

speaker groups.

The rest of this paper is structured as follows: Section 2 discusses earlier work on both CEFR classification and debiasing strategies. Section 3 explains the data we used in our experiments. Section 4 explains the classification setup. Section 5 discusses our results and provides pointers to future work. Finally, Section 6 sums up our main findings.

2 Related Work

The task of Automated Essay Scoring itself has received a fair amount of attention over the last decades, see Beigman Klebanov and Madnani (2020) for a comprehensive overview of the current state of the art. Individual sub-tasks that can be taken as indicative for proficiency in a given language, such as Grammatical Error Correction (GEC), have been accompanied by a number of popular shared tasks (Ng et al., 2013, 2014; Bryant et al., 2019). The task of CEFR classification itself however, seems to have received fewer attention. Among the studies that address this problem for various languages are Santucci et al. (2020) (Italian), Hancke and Meurers (2013) (German), Vajjala and Lõo (2014) (Estonian) and Volodina et al. (2016) (Swedish). Earlier work on English (our language of interest) is represented by Tack et al. (2017), who create their own annotated corpus and experiment with automated classification using several classification algorithms.

In this paper, we interpret the influence of L1 as an issue of bias in the embeddings-based representation of the English texts. Particular word order, article- or gender-based preferences or errors that can be traced back to the native language of the author, are likely to be more ubiquitous within the same group of native speakers. To the best of our knowledge, the CEFR classification problem has not been combined before with methods attempting to debias embeddings for L1 features.

Bias in NLP has attracted a lot of interest recently (Bender et al., 2021; Costa-jussa et al., 2021; Bokstaller et al., 2021; Garrido-Muñoz et al., 2021), and the specific mitigation approach that we follow in our work is that of Ravfogel et al. (2020), who propose **INLP** - an iterative nullspace projection algorithm to debias gender stereotypes in text. Unlike previous approaches (Bolukbasi et al., 2016; Dev and Phillips, 2019) that solely rely on a contrastive wordlist to identify a linear direction for debiasing, INLP debiases all linearly present gender directions in a data-driven manner. Considering that a classification task relies on a certain feature that we want to remove, INLP iteratively trains a series of probing (linear) classifiers to predict that feature until the probing classifier is confounded. For more details, we encourage the readers to read the original paper. In Chowdhury et al. (2022), the potential of this approach was explored for debiasing translation artifacts (which carry similar stylistic differences as learnerese) in human/machine-translated documents. Building upon this work, we employ the algorithm for our use-case.

3 Data

Our first experiments were done on the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2019), a data set comprising essays from 2.800 authors from over 10 different native speaker groups, annotated for different metrics indicating skill levels (TOEIC, TOEFL, IELTS, etc.), including CEFR labels. When aggregating all data from non-native English speakers, using a vanilla BERT (Devlin et al., 2019) classifier, we obtained a classification accuracy of 0.51, for 2.600 essays[3]. For individual native language groups, however, we achieved comparable performance while using considerably fewer training instances (for example, 0.50 on just 200 essays whose authors are from Indonesia). At the same time, some native speaker groups in ICNALE are heavily imbalanced, resulting in simple majority vote classifiers outperforming the trained classifier for those native speaker groups. While these preliminary findings initially inspired us to apply debiasing strategies, we decided to use a larger, less imbalanced corpus for the majority of the experiments reported on in this paper.

We extracted a subset of the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2014), consisting of 191,969 texts from authors from China, Japan and Korea. Since all texts in EFCAMDAT are from language learners, we combined this with 200 texts from native English speakers from ICNALE to get debiasing directions. Furthermore, EFCAMDAT only provides information on the author's country of origin. Information on native language would be more accurate, but unfortunately is not specified in

[3] In a 10-fold cross-validation setup.

the corpus. For the purpose of this paper, we will assume the country of origin and native language to align. Table 1 summarises the key figures of the subsets of EFCAMDAT and ICNALE we used in our experiments.

In addition to the aforementioned English-L2 datasets, we conducted experiments on a subset of the MERLIN (Boyd et al., 2014) corpus, specifically the subset with German-L2 learners (henceforth called MERLIN_DE). This subset consists of 652 learner texts from 13 known nationalities and 275 Target Hypotheses (i.e. texts expected from the native speakers and written by annotators.) Due to the skewness of this dataset, we only consider data from the top three represented nationalities. Table 2 summarises the subset of the MERLIN_data we used for our experiments.

4 Method & Results

As mentioned in Section 3, we were initially inspired by the fact that adding more training data did not seem to improve classification performance. In addition, earlier work indicated that classifying the country of origin of an author based on their English text provides good results, with Tang et al. (2021) reporting an accuracy of 87% on all of ICNALE for this task. We argue that this points at signals of L1 in the English learner texts that a classifier can pick up on, and that consequently, finding a way to make input text more homogeneous to a classifier through debiasing (Section 4.1) can lead to CEFR classification performance gains (Section 4.2).

4.1 Country of Origin Classification and Debiasing

To classify the country of origin of the author of a learner text, we use multiple binary classifiers (for example, China vs. EN, Japan vs. EN, Korea vs. EN). In particular, we first derive BERT document-level representations (by mean-pooling the token-level embeddings) of the text and then feed them to a Logistic Regression classifier for the binary classification task. Recall that country of origin classification is just an intermediate step in order to find directions to debias our embeddings. For this task, we randomly sample 200 texts from China, Japan and Korea to compare against the 200 from native English speakers (to keep the data balanced) and we used a static train/dev/test split of 70/15/15, respectively. Following Ravfogel et al. (2020), we proceed to get rid of any signals (in the embeddings) that the classifier exploits to base its decision on and found that this works surprisingly well. After 300 iterations for null-space projection, the perfect performance of 100 for country of origin classification for all three language pairs (to be compared to 87% for all of ICNALE as reported by Tang et al. (2021)) drops to approximately random performance after debiasing (Table 3).

We follow similar steps for the country of origin classification for the MERLIN_DE dataset. Recall that in this setup, the direction for native-German comes from Target Hypotheses (TH) and the number of Target Hypotheses (275) exceeds the number of text samples coming from the three nationalities. In order to achieve a balanced dataset for the binary classification, we randomly sample TH texts equal to the number of Russian-DE, Polish-DE and Spanish-DE texts, respectively. Thereafter, we apply INLP for 7 iterations on all three language pairs and achieve classification accuracies as shown in Table 4.

4.2 CEFR classification

As illustrated in Table 1, our data is fairly unbalanced, with most texts belonging to the A1 category. A majority vote classifier would result in an accuracy of 55%. To improve over this, as a baseline, we apply a multinomial Logistic Regression classifier and an MLP classifier having a hidden layer of 256-dimensions.

We then attempt to improve over this baseline by applying debiasing conditional on the country of origin of the author. BERT-encoded document-level representations of native-EN and L2-EN[4] (200 each) are fed to the INLP algorithm for bias removal. As stated earlier in section 4.1, to carry out this procedure, the data are first combined and shuffled, and then split into train, test and dev (70/15/15), followed by 12 iterations of INLP. By applying the INLP procedure on the training split, as one of the three outputs, we get the nullspace projection, which is devoid of any learnerese-signal. So, we simply project this nullspace onto the whole of respective L2-EN BERT embeddings to get debiased L2-EN embeddings.

We combine all data (i.e. BERT embeddings) from China, Japan and Korea for EFCAMDAT_NATIVE_EN, and for Russian, Polish and

[4]Where L2 corresponds to Japan/Korea/China.

	A1	*A2*	*B1*	*B2*	*C1*	*C2*	total
China	94,494	48,564	17,613	3,946	504	51	**165,162**
Japan	8,567	6,396	4,390	1,601	395	25	**21,374**
Korea	1,966	1,697	1,277	379	103	11	**5,433**
EN	-	-	-	-	-	-	**200**
total	*105,027*	*56,657*	*23,280*	*5,916*	*1,002*	*87*	

Table 1: Number of texts in native speaker groups and skill levels in EFCAMDAT_NATIVE_EN dataset.

	A1	*A2*	*B1*	*B2*	*C1*	*C2*	total
Russia	7	35	45	48	8	0	143
Polish	1	22	27	41	5	0	96
Spanish	3	23	31	27	1	0	85
total	11	80	103	116	14	0	**324**

Table 2: Number of texts in native speaker groups and skill levels in MERLIN_DE dataset.

	before debiasing		**after debiasing**	
	LR	**MLP**	LR	MLP
China	100	100	48.33	95.00
Japan	100	100	46.47	93.34
Korea	100	100	40.00	96.67

Table 3: EFCAMDAT_NATIVE_EN dataset: Accuracy for country of origin classification.

	before debiasing		**after debiasing**	
	LR	**MLP**	LR	MLP
Russia	83.72	93.02	51.16	53.48
Polish	89.66	93.10	65.52	82.76
Spanish	84.61	92.31	57.69	73.07

Table 4: MERLIN_DE dataset: Accuracy for country of origin classification.

Spanish for MERLIN_DE, and proceed to classify the CEFR levels. The results are illustrated in Table 5, where in the *after debiasing* column, the debiased embeddings, conditional on the author's country of origin, are used in classification. The numbers are the result of 5-fold cross-validation.

As shown in Table 5, the debiasing strategy improves performance by 1 point for the Linear Regression classifier, whereas the Multi-Layer Perceptron classifier remains unaffected for the EFCAMDAT_NATIVE_EN dataset. The performance of both classifiers drops for the MERLIN_DE dataset. We refer to Appendix A.1 for hyper-parameter settings. In the next section, we discuss these results, discuss promising directions for future work and summarise our main findings.

		before debiasing	after debiasing
EN-CEFR	LR	82	83
	MLP	96	96
DE-CEFR	LR	58	43
	MLP	73	63

Table 5: Weighted F1-scores for CEFR classification.

5 Discussion

For the EFCAMDAT_NATIVE_EN dataset, we observe a marginal performance gain when using a linear classifier (LR), but not when using a non-linear classifier (MLP). This can be explained from the results in Table 3, wherein the accuracy for MLP drops only marginally after debiasing. This means the non-linear classifer is still able to tell whether a sample comes from native or non-native speaker. The effects of debiasing on linearly separable vs. non-linearly separable problems is also discussed in Ravfogel et al. (2020), who state that their method is designed for *"removal of linear information regarding a protected attribute"*. This may explain why our setup with an MLP classifier shows no difference. Furthermore, the MLP classifier having better performance in the baseline setup already may suggest that the specific surface realisations of learnerese may be less prone to linear separation. Alternatively, Ravfogel et al. (2020) focus on guarding the classifier against gender and race. These dimensions might be expected to correlate to individual words or short phrases. The effects of learnerese may surface more on syntactic (phrase- or sentence-) level, which may just need more training data than we have available to us. As for hyperparameter settings; we have experimented with various different numbers of iterations (ranging from 8 to 300) for finding the directions for

debiasing, but beyond a certain point (12 for EFCAMDAT_NATIVE_EN and 7 for MERLIN_DE) the INLP classifier started to overfit and the quality of embeddings start to decrease.

Furthermore, in the EFCAMDAT data, there appears to be a strong correlation between sentence length and CEFR level, with the average text length in words for levels A1 to C2 being, respectively, 45, 74, 97, 128, 161 and 164. This may be a strong indicator to the classifier, and one we have not compensated for. We decided against simply sampling individual sentences from the different CEFR levels, as we argue that (the ability to implement) overall text coherence is an important part of mastering a language. Any such text structure or coherence features would in most cases be lost when considering individual sentences. We consider experimenting with more sophisticated techniques to compensate for the differences in text length an important part of future work.

As illustrated in Table 1, we only have 200 native English texts to find directions for debiasing. This works surprisingly well (Table 3), but we get a comparatively small performance gain of 1 point for CEFR classification. Perhaps the ICNALE essays are easily distinguishable from the EFCAMDAT ones on other grounds (lay-out, topic, length) than just native vs. non-native. The EFCAMDAT corpus contains data from English-speaking countries, but since these originate from language learners, it is a heterogeneous L1 group. Using this would thus result in finding, for example, Chinese-specific vs. many-different-L1-specific traits, as opposed to finding Chinese-specific vs. native English-specific traits. In order to find out if the additional data (42,442 texts from authors from the USA and Great Britain from EFCAMDAT, compared to 200 from ICNALE) would compensate for the heterogeneity in L1 background however, we experimented with this setup too and got comparable results to the ones reported on in Table 5.

Compared to earlier work, the overall performance of our system scores well. Tack et al. (2017) also work on English and report an accuracy of 53% on their data set[5]. In other related work however, performance seems to depend highly on the specific data set (and language), with reported accuracy figures between 64.5% (Hancke and Meurers, 2013) and 79% (Vajjala and Lõo, 2014).

From Table 5, both the CEFR classifiers perform poorly on MERLIN_DE corpus. This comes as no surprise since we had only a few hundred samples for training and the data-class ratio was too skewed to begin with. Even though in Table 4, the accuracies of country-classifiers drop significantly after debiasing, it does not translate to a performance gain during CEFR-classifation and instead has the opposite effect. This means that the directions that are being removed by INLP are rather significant and perhaps to achieve gains on the downstream CEFR-classification task, INLP requires lot more training samples to find more reliable learnerese directions.

Unfortunately, we suspect that the majority of freely available datasets for CEFR-classification are too small (in the order of 10^2 or 10^3) to see any improvements from debiasing with INLP.

In future work, besides experimenting with other debiasing approaches, we plan to address this bottleneck by curating data for language-families (instead of considering languages in isolation for debiasing) and investigating if a combined debiasing approach on aggregated data from the same language family works better.

6 Conclusion

In this paper, we experiment with compensating for L1 influence in CEFR classification by applying a debiasing approach, the idea being to debias the embeddings for learnerese features in any specific L1-related direction. By doing so, we obtain a small performance improvement with a linear classifier. CEFR classification performance seems to be highly dependent on the particular corpora/data used, with earlier work reporting accuracy figures between 53% and 79%. On the EFCAMDAT dataset, results look promising - best weighted F1-score of 83 via Logistic Regression and even higher (96) with MLP classifier without any debiasing. Our code is available on GitHub[6].

References

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7796–7810, Online. Association for Computational Linguistics.

[5]Moreover, they aggregate the C1 and C2 levels, resulting in 5-way classification, compared to 6-way in our setup.

[6]https://github.com/mst-sb/AES

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Jonas Bokstaller, Georgios Patoulidis, and Aygul Zagidullina. 2021. Model bias in NLP - application to hate speech classification using transfer learning techniques. CoRR, abs/2109.09725.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in Neural Information Processing Systems, 29:4349–4357.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 697–707, Beijing, China. Association for Computational Linguistics.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef van Genabith. 2022. Towards debiasing translation artifacts.

Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster, editors. 2021. Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, Online.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 879–887. PMLR.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ismael Garrido-Muñoz , Arturo Montejo-Ráez , Fernando Martínez-Santiago , and L. Alfonso Ureña-López . 2021. A Survey on Bias in Deep NLP. Applied Sciences, 11(7).

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines, pages 240–254.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. pages 54–56.

Shin'ichiro Ishikawa. 2019. The ICNALE Spoken Dialogue: A New Dataset for the Study of Asian Learners' Performance in L2 English Interviews. English teaching, 74:153–177.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8706–8716, Online. Association for Computational Linguistics.

Robert Lado. 1957. Linguistics across cultures: Applied linguistics for language teachers. University of Michigan press.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated Grammatical Error Detection for Language Learners. Morgan and Claypool Publishers.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Ellis B. Page. 1966. The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5):238–243.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7237–7256, Online. Association for Computational Linguistics.

Jack C Richards and Theodore S Rodgers. 2014. Approaches and methods in language teaching. Cambridge university press.

Valentino Santucci, Filippo Santarelli, Luciana Forti, and Stefania Spina. 2020. Automatic classification of text complexity. Applied Sciences, 10(20).

Larry Selinker. 1969. Language transfer. General linguistics, 9(2):67.

M Sharwood Smith and Eric Kellerman. 1986. Crosslinguistic influence in second language acquisition: An introduction. Crosslinguistic Influence in Second Language Acquisition, New York: Pergamon.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Zixin Tang, Prasenjit Mitra, and David Reitter. 2021. Are BERTs sensitive to native interference in L2 production? In Proceedings of the Second Workshop on Insights from Negative Results in NLP, pages 36–41, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In Proceedings of the third workshop on NLP for computer-assisted language learning, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Elena Volodina, I. Pilán, and David Alfter. 2016. Classification of Swedish learner essays by CEFR levels.

Uriel Weinreich. 2010. Languages in Contact: Findings and Problems. De Gruyter Mouton.

A Appendix

A.1 Hyper-parameter settings

LinearSVC (iterative debiasing):

- penalty='l2'
- C=0.01
- fit_intercept=True
- class_weight=None
- dual=False

Logistic Regression (country of origin):

- penalty = 'l2'
- warm_start = True
- solver="saga"
- random_state=23
- max_iter=7

Logistic Regression (CEFR):

- penalty = 'l2'
- warm_start = True
- solver="saga"
- random_state=23
- max_iter=7
- multi_class='multinomial'
- fit_intercept=True

MLP (CEFR):

- hidden_layer_sizes = 256
- activation = relu