# Influence of Language Proficiency on the Readability of Review Text and Transformer-based Models for Determining Language Proficiency

Salim Sazzed
ssazz001@odu.edu
Old Dominion University
Norfolk, VA, USA

## ABSTRACT

In this study, we analyze the influence of the English language proficiency of non-native speakers on the readability of the text written by them. In addition, we present multiple approaches for automatically determining the language proficiency levels of non-native English speakers from the review data. To accomplish the above-mentioned tasks, we first introduce an annotated social media corpus of around 1000 reviews written by non-native English speakers of the following five English language proficiency (ELP) groups: very high proficiency (VHP), high proficiency (HP), moderate proficiency (MP), low proficiency (LP), and very low proficiency (VLP). We employ the Flesch Reading Ease (FLE) and Flesch-Kincaid Grade (FKG) tests to compute the readability scores of the reviews written by various ELP groups. We leverage both the classical machine learning (ML) classifiers and transformer-based language models for deciding the language proficiency groups of the reviewers. We observe that distinct ELP groups do not exhibit any noticeable differences in the mean FRE scores, although slight differences are observed in the FKG test. The results imply that the readability measures do not possess high discriminating capabilities to distinguish various ELP groups. In the language proficiency determination task, we notice fine-tuned transformer-based approaches yield slightly better efficacy than the traditional ML classifiers.

## CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; • **Information systems** → *World Wide Web*.

## KEYWORDS

readability, language proficiency, user demographics, native language classification, BERT

**ACM Reference Format:**
Salim Sazzed. 2022. Influence of Language Proficiency on the Readability of Review Text and Transformer-based Models for Determining Language Proficiency. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487553.3524666


*WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.*

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524666

## 1 INTRODUCTION

As a universal language, the prevalence of English text written by non-native speakers located across the world is prominent on the web and social media [9]. Indeed, of the approximately 1.5 billion English-speaking people, fewer than 400 million people practice English as a first language, which indicates the existence of over 1 billion secondary English speakers[1]. A vast amount of web content is continuously being generated by these non-native speakers. Analyzing the linguistic characteristics of textual content written by non-native English speakers has significance for decision making in areas such as forensic linguistics, author profiling, and authorship identification [20].

The English language proficiency of non-native English speakers varies across the demography. The Education First (EF), an international education company that specializes in language training, publishes the English Proficiency Index (EPI), which describes proficiency of English of the non-native English speakers based on a set of criteria (described below)[2].

(1) Very High Proficiency (VHP): People belonging to this group are capable of using nuanced and appropriate language in social situations and can read advanced texts with ease.
(2) High Proficiency (HP): People of this group can make presentations at work, understand TV shows and read newspapers.
(3) Moderate Proficiency (MP): People of this group can participate in meetings in one's area of expertise, can understand song lyrics, capable of writing professional e-mails on familiar subjects.
(4) Low Proficiency (LP): A person representing this group is capable of navigating in an English-speaking country as a tourist, can engage in small talk with colleagues, and understand simple e-mails from colleagues.
(5) Very Low Proficiency (VLP): People of this group can introduce themselves with name, age, country of origin information, understand simple signs, and can give basic directions to a foreign visitor.

The 2021 EF EPI report categorized the overall English language proficiency (ELP) of 112 countries into the above-mentioned five groups.

Readability is the easiness of understanding a written text [6]. The readability of text largely depends on the content (e.g., vocabulary, syntax) and presentation. A number of approaches exist for determining the readability of a piece of text [3, 7]; However, most

[1]https://www.weforum.org/agenda/2019/11/countries-that-speak-english-as-a-second-language/
[2]https://www.ef.com/wwen/epi/

of them require text over 100 words to calculate the readability scores. The Flesch Reading Ease Formula [7] is a simple method to determine the readability of text for the reader of various grade-level. It is one of the few measures that can be applied to varied types of text. The Flesch–Kincaid Grade (FKG) level readability test [14] is a formula similar to FRE; however, it uses different weights for the various terms.

In this study, we address the following two tasks:
i) Exploring the relationship between the readability of the text and English language proficiency of the writers.
ii) Determining the language proficiency groups of the writers from the social media review text.

To carry out the above-stated tasks, we introduce an annotated non-native English review corpus comprising around 1000 reviews where each review is labeled with one of the ELP groups (e.g., VHP, HP, MP, LP, VLP). The label of a review is determined based on the ELP group of the country from where the review is collected. The reviews are collected from the following countries: Finland (Very High Proficiency), Kenya (High Proficiency), China (Moderate Proficiency), Bangladesh (Low Proficiency), Myanmar (Very Low Proficiency). We then employ the FRE and FKG tests to see whether reviews written by different ELP groups differ in terms of readability. We observe that FRE readability scores are similar across diverse ELP groups. The FKG test shows some disparities between the top 3 and the bottom 2 English fluency groups. The results suggest that the readability measure is not an effective metric to discern various ELP groups. In addition, we aim to distinguish reviews of different ELP groups automatically by utilizing machine learning (ML) algorithms and annotated data. We leverage four popular classical ML (CML) classifiers and transformer-based language models. The fine-tuned transformer-based models yield slightly better results than best performing CML classifiers, obtaining macro F1 scores of 0.77.

## 1.1 Contributions

The main contributions of this study can be summarized as follows-

- We introduce a novel social media corpus of around 1000 reviews written by non-native English speakers and make it publicly available [3].
- We manually categorize the reviews into five English language proficiency (ELP) groups.
- We compare the readability scores of the reviews of various ELP groups by calculating the FRE and FKG scores.
- Finally, we employ traditional ML classifiers and transformer-based fine-tuned language models for categorizing reviews into various ELP groups.

## 2 RELATED WORK

A number of related studies tried to determine the demographic information of the textual content. For example, some studies tried to determine the native language of non-native English writers solely on the writing samples. [2, 15, 24]. However, although related, unlike this study which considers language proficiency, they investigated the native language (l1) of non-native English writers. Besides, the perspective of their study (most) was the second language acquisition (SLA) research, such as contrastive analysis, syntactic or grammatical errors made by non-native speakers [15, 26] based on corpus compiled from the sample essay of ESL (English as a Second Language) learners such as TOEFL (Test of English as a foreign language) [1], the international corpus of learner English [10]. Tetreault et al. [24] tried to identify the native language from small corpora generated by ESL students. The authors considered various character-level lexical features, words, POS tags, and document structures to build the classifier.

**Table 1: Statistics of various ELP groups**

| ELP group | Representative Country | #Samples |
|---|---|---|
| VHP | Finland | 185 |
| HP | Kenya | 194 |
| MP | China | 195 |
| LP | Bangladesh | 220 |
| VLP | Myanmar | 202 |

Jarvis et al. [12] obtained 83.6% accuracy in NLI-2013 shared task on the dataset of the 12,100 English essays from the TOEFL test. The authors employed features such as word n-grams, parts-of-speech, and lemmas. To normalize the frequencies, the authors applied the log-entropy weighting schema. As a classifier, the L2-regularized SVM classifier was used.

Gebre et al. [8] proposed a Native Language Identification (NLI) system based on TF-IDF weighting and several linear classifiers such as support vector machine (SVM), logistic regressions (LR), and perceptrons. Their proposed model obtained a high accuracy of 0.814 in NLI SharedTask-2013 for categorizing a set of 11 native languages.

Besides, a number of paper investigated the impact of demography for linguistic analysis [22]. Sazzed [21] studied the linguistic characteristics of the reviews of two demographically different groups contrasting two datasets.

Regarding readability, some studies investigated the impact of readability on various domains such as business, scientific research, health [4, 11, 13, 17, 18]. Pancer et al. [18] showed that text readability plays a critical role in propelling consumer engagement on social media. By analyzing 4,000 Facebook posts from a photography blog, the authors found that easy-to-read posts are more liked, commented on, and shared on social media.

Guerini et al. [11] examined a corpus of scientific abstracts and three feedback metrics: article downloads, citations, and bookmarks. The authors found that certain stylistic and readability features of abstracts have an impact on the success and virality of a scientific article.

Temnikova et al. [23] investigated the readability of governments, non-governmental organizations, and mainstream media tweets related to crisis communications during the years between 2012 and 2013. The authors illustrated factors that adversely impact comprehension. Besides, they provided guidelines about how understanding can be improved.

[3]https://github.com/sazzadcsedu/EnglishLanguageProficiency.git

| | |
|---|---|
| VHP | Incredibly hard to get a table, for a reason! Hidden away in the ouskirts, in an amazing old factory building, head chef Nicholas and his team delivers amazing food with a passion that is hard to find in this rugged country. Share four starters on two people, abd go with whatever middle and maincourse you like from the menu, I still had room for dessert. The wines are amazing, trust the recommendations they give you! Book a table well in advance, this is one of the hardest ones to bpok, and being a local that is very frustrating! |
| | The folks I went to restaurant with were just amazed, one said it was one of the TOP food experiences of her life. Being mostly vegan and sometime vegetarian, as it was a must in this case; Bas Bas was not for me. The cuisine is all butter and cream, so having not being accustomed to that heavy cow-taste anymore, it was too rich for me. Nice atmosphere in the modern and elegant environment. The staff is fantastic and the wine ofcourse is top notch. |
| HP | I loved the ambiance, the customer service was great and so was the food. My children, sisters and I really enjoyed ourselves. My children really wanted to try out sea food so I took them there. We ordered the sea food platter and boy oh boy, we loved the assortment. We are definitely going back.We would like to personally appreciate @sophiehostess and @mercywaitress for your readiness to serve and guide when called upon. @mercywaitress helped us pick our food which we loved. @sophiehostess ensured we got a seating space . |
| | Had a brunch meeting here. The ambience is pleasant and the food is good! A bit expensive but it's value for money. Godfrey especially attended to us perfectly. I'd recommend the place anytime. The specialty is sea food but other cuisines are available as well. |
| MP | A vegetarian but tasty restaurant. It is located in an old elegant Chinese style house and next to subway. I am a meat lover. But I find the dishes here very tasty and not boring at all. For those who are no familiar with the food served here, it is best to order set meal. |
| | The service is superb, one of the best in Beijing. Carter really helped us during our dinner, he tried to explain everything to us to ensure the quality of our experience. Such a wonderful restaurant staff, who's attentive, gentle with great communication skill. Thank you Carter for making our experience wonderful! |
| LP | Food quality below average, behaviour of staffs is worst, all are careless. I think they have no control. Staffs are not well trained about good manners. |
| | Good place for family and friends...last day Sunday I went with my family member to attend a family get-together party.I really enjoyed there.. |
| VLP | Excellent fried tofu and Service. Excellent Service come from the heart. Genuine Smile. Quick Service. Strongly Recommended. |
| | This place it totally overrated, I mean the food and tea was good but:- expensive- staff could be nicer- they messed up my order and then tried to charge me for it too - they brought me a salad without chicken but I initially ordered it with, had to pay the price with chicken- I was sitting next to a small room with the door open and there was a strange smell coming out of it. Would recommend any local place with reasonable prices! |

**Figure 1: A list of sample reviews from different groups**

However, none of the existing works investigated the relationship between English language proficiency and the readability of the text. To the best of our knowledge, this is the first work that aims to map two important aspects of social media text.

# 3 DATASETS

## 3.1 Data Collection and Annotation

All the review data used in this study are manually collected from the TripAdvisor website [4]. TripAdvisor is the world's largest travel platform that contains millions of traveler reviews and opinions regarding places, hotels, restaurants, flights.

We leverage restaurant reviews written by non-native English speakers of diverse ELP groups. To collect the ELP group-specific sample data, we utilize the country-specific categorization of ELP groups provided by EF. Based on the EF categorization, we select reviews written towards restaurants located in five different countries where each country represents a particular ELP group. Table 1 shows the ELP groups, the representing countries, and the number of samples for each ELP group.

Even though the selected restaurants of each group are located in a particular country, reviews could be written by non-native people such as tourists. Since our study heavily relies on the implicit characteristics of the ELP groups, it is crucial to make sure that each ELP group contains only the representative reviews.

For annotation, we consider the following three user attributes obtained from the TripAdvisor user profiles: i) city and country, ii) name, and iii) profile picture. However, in the social media profile, it is not unusual to have one or multiple of the above-stated attributes missing. For example, many users prefer to hide the location information or use arbitrary names (e.g., placeholder names that do not resemble country/race/culture). Besides, the profile picture may not be available or may not be meaningful (e.g., pictures of various objects). Since we are only interested in reviews written by the native people of a country, unless we are sure about the native country of the user, we do not include the user and corresponding review(s) in the dataset.

Figure 1 shows sample reviews from different groups.

# 4 READABILITY ASSESSMENT OF REVIEWS OF DIFFERENT ELP GROUPS

The Flesch Reading Ease (FRE) test is applied to see if the mean readability scores of five ELP groups differ. This readability test indicates how difficult a passage in English is to understand. In the FRE test, a higher score indicates that the content is easier to read. The following formula is used to calculate the FRE score of in a textual content-

$$FRE = 206.835 - 1.015 \times \frac{total\ word}{total\ sentence} - 84.6 \times \frac{total\ syllables}{total\ words}$$

The two main criteria of the FRE calculation are sentence length (i.e., the average number of words per sentence) and the presence of syllables in words.

Figure 2 provides the interpretation of FRE scores of different ranges.

[4]https://www.tripadvisor.com

| Score | School level (US) | Notes |
|---|---|---|
| 100.00–90.00 | 5th grade | Very easy to read. Easily understood by an average 11-year-old student. |
| 90.0–80.0 | 6th grade | Easy to read. Conversational English for consumers. |
| 80.0–70.0 | 7th grade | Fairly easy to read. |
| 70.0–60.0 | 8th & 9th grade | Plain English. Easily understood by 13- to 15-year-old students. |
| 60.0–50.0 | 10th to 12th grade | Fairly difficult to read. |
| 50.0–30.0 | College | Difficult to read. |
| 30.0–10.0 | College graduate | Very difficult to read. Best understood by university graduates. |
| 10.0–0.0 | Professional | Extremely difficult to read. Best understood by university graduates. |

**Figure 2: The interpretations of FRE scores [7]**

Flesch-Kincaid Grade Level (FKG) is another formula used to test the readability of text. Although the FRE and FKG tests employ the same core criteria (i.e., word and sentence length), different weighting factors are employed. The results of the two tests correlate approximately inversely: a text with a comparatively high score on the FRE test should have a lower score on the FKG test. The distinct weighting factors for words per sentence and syllables per word in each scoring system imply that the two schemes are not directly comparable and cannot be converted. The following equation is used to calculate the FKG score-

$$FKG = 0.39 \times \frac{total\ word}{total\ sentence} + 11.8 \times \frac{total\ syllables}{total\ words}$$

- 15.59

The Flesch-Kincaid Grade (FKG) level is equivalent to the US grade level of education. It indicates the education level required to understand a piece of text. For example, an FKG score of 6 for a text means that the reader needs to have at least grade-6 level reading knowledge to understand it.

# 5 ELP GROUP PREDICTION TASK

## 5.1 Classical ML Classifiers

We employ four classical ML (CML) classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (k-NN) for predicting the ELP groups of the reviewers.

We extract the word-based unigrams and bigrams from the review texts. An *n*-gram is a contiguous sequence of *n* items from a sample piece of text. The tf-idf (term frequency-inverse document frequency) scores of the extracted *n*-gram features are computed and then used as input for the CML classifiers. For all the CML classifiers, the default parameter settings of the scikit-learn library [19] are used. For all the classifiers, the class *balanced* weight is used. For the k-NN, the value of *k* is set to 5.

## 5.2 Pre-trained Language Models

*5.2.1 BERT-based language models.* We employ two variants of transformer-based language models, BERT [5] and RoBERTa [16]. The Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model created from a huge amount of unlabeled data. BERT utilizes Transformer to learn contextual relationships between words in a piece of text. The BERT-base-uncased [25] model utilized in this study consist of twelve layers of transformer blocks, where each block contains twelve-head self-attention layers and 768 hidden layers.

The other transformer-based language model, RoBERTa (Robustly optimized BERT), was introduced to overcome some of the limitations of the BERT. The RoBERTa-base model consists of twelve transformer layers with 768-hidden layers, twelve attention heads, and 125 million parameters. Unlike BERT, which uses static masking, RoBERTa uses dynamic masking. RoBERTa generates new masking patterns every time a sequence is fed to the model.

*5.2.2 Fine Tuning.* We fine-tune the pre-trained models for categorizing reviews into five classes (i.e., number of ELP groups). Since this is a classification task, we utilize the classification module of pre-trained models. The hugging face library [25] is used for fine-tuning all the pre-trained models.

Since the initial layers of pre-trained models only learn very general features, we keep them intact. Only the last layers of the pre-trained models are fine-tuned for the classification task. We add one more layer on top of the pre-trained model for classification. For fine-tuning, we tokenize and feed the input training data to the language model and train the model for some steps; The trained model is subsequently used for classifying the testing data.

A mini-batch size of 16 and a learning rate of 0.00004 are used. During the training, 20% samples are utilized as a validation set. The Adam optimizer is used for optimization, and the loss function is set to sparse-categorical-cross-entropy. The training process runs for 3 epochs, and an early stopping criterion is employed.

# 6 EVALUATION, RESULTS AND DISCUSSION

We perform 10-fold cross-validation to assess the performances of various approaches. The 10-fold cross-validation splits data into 10-mutually independent subsets. The training process runs for 10 iterations; in each iteration, a new subset is selected as a testing set, and the other 9 subsets are used as the training set. We report the overall precision, recall, macro F1, and accuracy of the various methods.

**Table 2: Mean and standard deviation (Std.) of the FRE scores of various English language proficiency groups**

| ELP Group | Mean | Std. |
|---|---|---|
| VHP | 68.69 | 15.96 |
| HP | 67.67 | 12.39 |
| MP | 66.78 | 15.94 |
| LP | 68.16 | 18.15 |
| VLP | 72.78 | 13.98 |

Table 2 and 3 show the readability scores of various ELP groups. We report the mean and standard deviation (Std.) values of the readability scores for different ELP groups.

We find that the readability tests can not distinguish the ELP groups sufficiently since they can not capture morphological differences, the diversity in the vocabulary usage, or the complexity of sentences ( it only counts the sentence length and the number of syllables). Besides, for short text, where less information is available, determining the readability is highly challenging. The various

**Table 3: Mean and standard deviation (Std.) of the FKG scores of various English language proficiency groups**

| ELP Group | Mean | Std. |
|---|---|---|
| VHP | 6.92 | 3.37 |
| HP | 7.08 | 2.88 |
| MP | 7.61 | 4.52 |
| LP | 5.94 | 3.41 |
| VLP | 5.68 | 2.59 |

**Table 4: Performances of various approaches for ELP group determination task**

| Classifier | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| LR | 0.78 | 0.73 | 0.75 | 75.91% |
| SVM | 0.79 | 0.71 | 0.75 | 74.63% |
| RF | 0.78 | 0.60 | 0.68 | 65.81% |
| K-NN | 0.69 | 0.64 | 0.67 | 66.01% |
| BERT | 0.78 | 0.76 | 0.77 | 77.32% |
| RoBERTa | 0.77 | 0.76 | 0.77 | 76.92% |

readability tests were not designed for short text; they are more appropriate for determining the readability of large and formal textual content. The short informal comments found in social media, usually less than 100 words, are not the best samples to compare or check the readability.

Table 4 shows the performances of classifiers for determining various language proficiency groups. As we can see, the best performing CML classifiers are LR and SVM, both yield F1 scores of around 0.75. The RF and k-NN exhibit comparatively poor performances. The transformer-based BERT and ROBERTA models yield the best performances by attaining F1 scores around 0.77. Since this ELP group identification task is a 5-class classification problem, F1 scores close to 0.8 can be considered promising results. Besides, here, we use limited annotated samples of around 1000 reviews. If more annotated data are incorporated into the BERT-base models, it is expected that classification performance will improve. However, the high efficacy of the classifiers should be interpreted carefully; since each language proficiency group is represented by a single country, various country-specific features may positively influence the results.

## 7 SUMMARY AND CONCLUSIONS

In this study, we introduce a corpus consisting of around 1000 reviews annotated with the English language proficiency of the reviewers. We investigate the readability of text written by various English proficiency groups. Moreover, we leverage classical ML classifiers and transformer-based pre-trained language models for determining the language proficiency of the reviewers. Our results and finding suggest that the readability score is not a good predictor of the language fluency of the writers when the informal social media text is concerned. Utilizing limited labeled data, both the classical ML classifiers and transformer-based fine-tuned language models demonstrate efficacy for determining the English language proficiency groups of the writers. Although, the results could be positively influenced by various country-specific features of language proficiency groups. Our future will focus on increasing the corpus size and include reviews collected from multiple countries for each ELP group.

## REFERENCES


[1] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series* 2013, 2 (2013), i–15.

[2] Julian Brooke and Graeme Hirst. 2013. Native language detection with 'cheap'learner corpora. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, Vol. 1. Presses universitaires de Louvain, 37.

[3] Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60, 2 (1975), 283.

[4] Scott W Davis, Csilla Horváth, Anaïs Gretry, and Nina Belei. 2019. Say what? How the interplay of tweet readability and brand hedonism affects consumer engagement. *Journal of Business Research* 100 (2019), 150–164.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] William H DuBay. 2004. The principles of readability. *Online Submission* (2004).

[7] Rudolf Flesch. 1979. How to write plain English. *University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.[Retrieved 5 February 2016]* (1979).

[8] Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*. 216–223.

[9] Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 3591–3601.

[10] Sylviane Granger. 2003. The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly* 37, 3 (2003), 538–546.

[11] Marco Guerini, Alberto Pepe, and Bruno Lepri. 2012. Do linguistic style and readability of scientific abstracts affect their virality?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6. 475–478.

[12] 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*. 111–118.

[13] Wooksoo Kim, Isok Kim, Krisztina Baltimore, Ahmed Salman Imtiaz, Biplab Sudhin Bhattacharya, and Li Lin. 2020. Simple contents and good readability: Improving health literacy for LEP populations. *International Journal of Medical Informatics* 141 (2020), 104230.

[14] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. Naval Technical Training Command Millington TN Research Branch.

[15] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics*. Springer, 209–217.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] Jo Mackiewicz and Dave Yeats. 2014. Product Review Users' Perceptions of Review Quality: The Role of Credibility, Informativeness, and Readability. *IEEE Transactions on Professional Communication* 57, 4 (2014), 309–324. https://doi.org/10.1109/TPC.2014.2373891

[18] Ethan Pancer, Vincent Chandler, Maxwell Poole, and Theodore J Noseworthy. 2019. How readability shapes social media engagement. *Journal of Consumer Psychology* 29, 2 (2019), 262–270.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[20] Raheem Sarwar, Attapol T Rutherford, Saeed-Ul Hassan, Thanawin Rakthanmanon, and Sarana Nutanong. 2020. Native language identification of fluent and advanced non-native writers. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 4 (2020), 1–19.

[21] Salim Sazzed. 2021. A Hybrid Approach of Opinion Mining and Comparative Linguistic Analysis of Restaurant Reviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 1281–1288.

[22] Salim Sazzed. 2022. Stylometric Analysis and Classification of Demographically Diverse Review Data (in press).

[23] Irina Temnikova, Sarah Vieweg, and Carlos Castillo. 2015. The case for readability of crisis communications in social media. In *Proceedings of the 24th international conference on world wide web*. 1245–1250.

[24] 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*. 48–57.

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

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