2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Stylometric and Semantic Analysis of Demographically Diverse Non-native English Review Data

Salim Sazzed
Department of Computer Science
Old Dominion Univesrsity
Norfolk, USA
ssazz001@odu.edu

***Abstract*—The demographic knowledge facilitates a fine-grained interpretation of the user-generated review text and enables better decision-making. In this study, we aim to comprehend how various attributes of non-native English text vary across demographically distinct groups. We introduce a non-native English corpus of around 1150 reviews representing four demographically diverse country-specific groups: Finland, Kenya, Bangladesh, and China. The reviews differ in various contexts, including geography, native language family, race and culture, and English proficiency levels of the reviewers. We then perform stylometric and semantic analysis on these distinct sets of reviews to unveil how the linguistic characteristics differ across the demography. The investigation reveals that stylometric features are mostly similar across the reviews of various groups; nevertheless, dissimilarities are observed in attributes, such as review length, presence of articles, or prepositions. We employ classical machine learning (ML) algorithms and transformer-based fine-tuned language models for categorizing the reviews into distinct demographic groups. We observe that semantic features yield slightly better efficacy than syntactic features for distinguishing the demography-specific reviews.**

***Index Terms*—demographic analysis, linguistic attributes, semantic features, syntactic features, demography prediction.**

I. Introduction

The stylistic and semantic aspects of the review texts may vary across the demography, such as geography and socio-cultures. As a universal language, the presence of English text written by non-native speakers of diverse demographics is prevalent on the web and social media [1]. A vast amount of web content is continuously being generated by these non-native speakers. Analyzing the linguistic characteristics of the demographically diverse textual content has significance for decision-making in areas such as forensic linguistics, author profiling, and authorship identification [2], [3].

In this study, we focus on accomplishing the following two tasks:

1) Exploring linguistic attributes of reviews written by people of different demographic groups.
2) Automatically determining the demography of the reviewer from the review text.

To carry out the above-mentioned tasks, we first introduce an annotated English review corpus comprised of around 1150 restaurant reviews written by non-native English speakers. The restaurants are located in the following four countries: Finland, Kenya, China, and Bangladesh. The reviews are manually retrieved from a popular travel website, TripAdvisor. Each review is annotated with geographical information (i.e., country label) based on the location of the corresponding restaurant. For each review, we check the reviewer's TripAdvisor profile to ensure that it represents content written by a native person.

The four countries are selected considering the diverse characteristics of non-native English speakers (L2 speakers) from numerous perspectives. For example, the English language fluency level of people in these countries differ. Finland, Kenya, China, and Bangladesh represent very high proficient, high proficient, moderate proficient, and low proficient groups, respectively [1]. Besides, the native languages (L1) of these four countries were originated from different language families. The Finnish language originated from the Finno-Ugric language family, the Swahili language came from the Niger-Congo language, the Chinese language derived from the Sino-Tibetan language family, and the Bengali language originated from the Indo-Aryan language family. Moreover, the native people of these four countries differ in ethnicity levels.

We analyze the linguistic and semantic characteristics of various country-specific review groups. Various text statistics and stylistic features such as review length, percentages of different types of part-of-speech (POS), and sentence structure are considered in the reviews representing diverse groups. Furthermore, we investigate the presence of sentiment and emotion across the reviews of different groups. In addition, we seek to automatically differentiate reviews of various groups leveraging various machine learning (ML) classifiers with both syntactic and semantic features. We employ four popular classical ML (CML) classifiers and two pre-trained transformer-based language models. We observe that among the CML classifiers, the SVM performs best, using both the syntactic and semantic features. The fine-tuned transformer-

[1] https://www.ef.com/wwen/epi/

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

based models such as BERT [4] and RoBERTa [5] yield slightly better performance than the best-performing CML classifiers (i.e., SVM), obtaining a macro F1 score of 0.81.

A. Contributions

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

- We introduce a social media corpus of around 1150 user reviews and annotate them with demography (location) information. We make the corpus publicly available.[2].
- We provide details of the various stylistic and semantic features of the reviews belonging to demographically diverse groups.
- Finally, we employ CML classifiers and transformer-based language models utilizing various extracted stylistic and semantic features for the automatic demography prediction task.

II. Related Work

Determination of demographic attributes such as age, gender, and language based on the content available in blogs and social media posts have been explored by various authors [6]–[9]. A number of studies tried to determine the native language (L1) of non-native English writers solely based on their writing samples [10]–[12]. However, the perspective of these studies was mainly the second language acquisition (SLA) research, such as contrastive analysis, syntactic or grammatical errors made by non-native speakers [10], [13]. Tetreault et al. [11] tried to identify the native language of ESL students considering characters-level lexical features, words, POS tags, and document structures on the corpus compiled from TOEFL [14] and the international corpus of learner English [15] essay samples.

Sazzed [16] studied the relationship between the English language proficiency levels of non-native speakers and the readability of the review texts. The author found that readability tests do not have the much-distinguishing capability to differentiate reviews of different proficiency levels. Rosenthal and McKeown [17] integrated various blog-specific features, such as user behavioral patterns and interest, with logistic regression (LR) to predict the user age from blog content. Another work related to age prediction was performed by Nguyen et al. [18] that also employed linear regression. The authors utilized data from three sources: blogs, telephone conversations, and forum posts. The authors found both stylistic and content features influenced the prediction. Integrating shallow text features into the linear regression (LR) model, the authors obtained mean absolute errors (MRE) within a range of 4.1-6.8 years.

Besides, a number of works tried to predict genders from textual content collected from various sources. Schler et al. [19] analyzed a micro-blog corpus of around 300 million words and found significant differences in content and style levels with respect to gender. Besides, the authors noticed ages of the bloggers affected their writing style. Peersman et al. [20] applied the SVM classifier to infer the genders of Twitter users by extracting features based on word and character n-grams. Phuong et al. [21] employed the SVM classifier to predict the genders of the users based on their news website browsing data. The authors utilized various hand-crafted features such as news categories, topic features, access time, and sequential features based on website browsing sessions. Rozen et al. [7] analyzed multiple real-world datasets to understand the contribution of both browsing data and user-generated content for determining three types of user attributes: gender, location, and mobile device. The authors proposed a BERT-based model, ProfBERT, that encodes user-generated texts in the context of their associated content for creating user profiles. In addition to social text, various authors also tried to predict demographics from the scientific text [3].

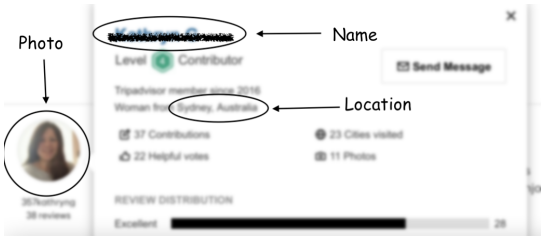

Fig. 1. Example of user profile in TripAdvisor Website

III. Data Collection and Annotation

The review data are manually retrieved from the TripAdvisor website[3], a popular travel platform. As this study aims to interpret the traits of reviews written by demographically diverse groups, data are collected from four different countries: Finland, Kenya, Bangladesh, and China. These four countries are selected considering their diversity in geographical locations (i.e., Asia, Africa, Europe), native language family, race and culture, and English language fluency levels [4].

For each country-specific group, we collect reviews of eight highly-rated restaurants of the corresponding country (TripAdvisor mainly contains reviews of highly-rated restaurants). The selected restaurants serve varieties of menus, including local cuisines and seafood. As the goal of this study is to investigate the impact of socio-cultures on the reviews, we try to ensure that each review is written by native people. We consider multiple user-specific attributes to exclude reviews composed by non-native people, such as tourists.

We consider the following TripAdvisor user attributes: i) city and country, ii) name, and iii) profile picture during the annotation (Figure 1). However, one or multiple of these attributes may not be available in the profile if a user prefers to conceal them. For example, the location information may be hidden, or the username can be arbitrary (i.e., a placeholder name that does not represent country/race/culture-specific naming conventions). Besides, the user profile picture may not be available or may not be meaningful (e.g., a picture of an

[2]https://github.com/sazzadcsedu/MultipleDemography.git

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

[4]https://www.ef.com/wwen/epi/

object). Since we are only interested in reviews written by the native people of a particular country, unless we are convinced about the native country of a user, we do not include the user and corresponding review(s) in the dataset. The names of the users representing each of the country-specific groups are checked by a graduate student of the same country studying in the USA (e.g., the Finland group is checked by a Finish graduate student).

TABLE I
DISTRIBUTIONS OF REVIEWS ACROSS VARIOUS DEMOGRAPHIC GROUPS

Representative Country	#Samples
Finland	288
Kenya	270
China	301
Bangladesh	292

Table I shows group-specific statistics of reviews in the corpus. As we can see, the dataset is mostly class-balanced and contains around 300 reviews for each class.

IV. LINGUISTIC ANALYSIS

A. Stylometric Analysis

We investigate the presence of a list of stylistic attributes in the reviews of various demographic groups [22].

1) Text Length: We consider three text-length related features: i) the number of words per review, ii) the number of sentences per review, and iii) the number of words per sentence.

2) Grammatical and Negation Features: The following grammatical and negation features are analyzed-

i. Percentage of negation words: The percentage of negative words in the reviews of the different demographic groups is computed. The VADER [23] negative words list, shown in Table II), is used as a reference.
ii. Percentage of articles: The percentage of articles (i.e., *a, an, the*) present in the reviews of each group is computed.
iii. Percentage of adjectives and verbs: The percentages of adjectives and verbs in the reviews of each group are computed. The spaCy [24] library is employed to identify adjectives and verbs in the text.
iv. Subordinating conjunctions: Besides, the presence of subordinating conjunctions that represents complex sentence is considered. A subordinating conjunction is a word or phrase that connects a dependent clause to an independent clause. A list of commonly occurred 50 subordinating conjunctions are considered (Table III).
v. Percentage of prepositions: This metric provides the percentage of prepositions present in the reviews of different groups. A list of commonly used prepositions is considered (see Table IV).

3) Sentiment Features: The presence of sentiment words is investigated for all the groups based on two popular English sentiment lexicons: Opinion lexicon [25] and VADER [23]. The Opinion lexicon consists of roughly 6800 positive and negative words with polarity scores of +1 and -1, respectively.

TABLE II
THE LIST OF NEGATIVE WORDS CONSIDERED IN THIS STUDY

aint	can't	mightnt	neednt	shant
arent	couldn't	mustnt	needn't	shouldnt
cannot	daren't	neither	never	uhuh
cant	didn't	don't	none	wasnt
couldnt	doesn't	hadn't	nope	werent
darent	dont,	hasn't	nor	oughtn't
didnt	hadnt	haven't	not	shan't
doesnt	hasnt	isn't	nothing	shouldn't
ain't	havent	mightn't	nowhere	uh-uh
aren't	isnt	mustn't	oughtnt	wasn't
weren't	without	wont	wouldnt	won't
wouldn't	rarely	seldom	despite	weren't

TABLE III
THE LIST OF SUBORDINATE CONJUNCTIONS

even though	who	whoever	whom
after	as soon as	as long as	as much as
although	assuming that	as though	as if
before	by the time	because	than
how	whether	whereas	that
whatever	which	whichever	now that
once	since	till	until
when	whenever	while	though
whomever	whose	where	wherever
if	only if	unless	provided that
even if	in case (that)	rather	so that
in order that	provided	least	even though
in order	as		

TABLE IV
THE LIST OF PREPOSITION

above	away	from	outside	time
across	before	in	over	down
after	behind	in	through	on
against	below	inside	till	until
along	beneath	into	to	out
among	beside	near	toward(s)	up
around	between	next	under	onto
at	by	off	underneath	

The VADER lexicon comprises around 7500 opinion words and emoticons with a polarity strength between -4 and +4. The sentiment feature reveals how sentiments are expressed across the demography.

B. Top Adjectives and Verbs

We report the most frequently occurring adjectives and verbs in reviews of various demography-specific groups. The goal is to see if people across the demographics have similar patterns in choosing adjectives or verbs while writing reviews.

V. DEMOGRAPHY PREDICTION TASK

A. Classical ML Classifiers

We leverage four classical ML (CML) algorithms: i)Logistic Regression (LR), ii) Support Vector Machine (SVM), iii)Random Forest (RF), and iv) K-Nearest Neighbor (k-NN) for the country-specific group prediction tasks. For all

TABLE V
MEAN VALUES OF VARIOUS ATTRIBUTES IN DEMOGRAPHICALLY DIVERSE GROUPS

Feature Type	Attribute	Finland	Kenya	China	Bangladesh
Text length	Words/review	46.34	59.06	44.02	17.95
	Sentence/review	3.62	5.23	3.17	2.17
	Words/sentence	12.34	11.12	13.97	8.02
Grammatical	Negation/review (%)	0.93	0.94	0.95	1.07
	Articles/review (%)	6.83	7.14	7.40	3.17
	Adjectives/review (%)	12.21	11.65	10.86	12.21
	Verbs/review (%)	10.86	10.65	11.32	10.67
	Preposition/review (%)	6.9	7.18	7.42	5.50
	Subordinate conjunction/review (%)	2.67	2.52	2.33	2.21
Sentiment	Opinion lexicon (%)	7.21	7.82	5.93	7.16
	VADER (%)	7.85	8.17	6.67	7.44

CML classifiers, the default parameter settings of the scikit-learn library [26] are used with class-*balanced* weight. Both stylometric and semantic features are used as input for the ML classifiers, separately.

1) Stylometric Features: In this scenario, various extracted stylometric features of the reviews are used as inputs for the classifier ML classifiers. As described in subsection IV (A), for each review, we extract features such as the review length, the ratio of negations, articles, adjectives, verbs, prepositions, subordinate conjunctions, and the percentage of opinion conveying words and use them as features.

2) Semantic Features: We extract the word n-grams (i.e., unigrams and bigrams) from the review text. An n-gram represents 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. As named entities, such as restaurant name or location, may have a positive influence on the classification, we report the results in the following two settings: i) using the original review text and ii) by excluding the named entities from the review text. The spaCy [24] library is utilized to identify named entities in the reviews.

TABLE VI
PERFORMANCES OF VARIOUS CLASSICAL ML CLASSIFIERS FOR DEMOGRAPHY-SPECIFIC GROUP PREDICTION TASK USING STYLOMETRIC FEATURES

Classifier	Precision	Recall	Macro F1	Accuracy
LR	0.427	0.452	0.4391	0.435
SVM	0.746	0.730	0.738	0.724
RF	0.649	0.655	0.652	0.642
K-NN	0.662	0.635	0.649	0.634

B. Fine-Tuned Language Models

We fine-tune two pre-trained language models, BERT and RoBERTa. BERT is a pre-trained language model that can capture contextual relationships between words in the unlabeled text. The basic BERT model, BERT-base-uncased, includes 12 layers of transformer blocks, 768 hidden and embedding layers, and 110M parameters. RoBERTa is another transformer-based model that follows a similar architectural design to BERT. The main addition of the RoBERTa is dynamic masking; it generates new masking patterns in each epoch.

TABLE VII
TOP 5 ADJECTIVES AND VERBS ACROSS FOUR DEMOGRAPHIC GROUPS

Group	Top 5 adjectives and verbs
Finland	(adj.) good, great, best, nice, perfect (verb) had, get, go, have, see
Kenya	(adj.) good, great, best,amazing, excellent (verb) had,recommend, go,have, enjoyed
China	(adj.) great, good, nice, best, amazing (verb) had, have, go, try, want
Bangladesh	(adj.) good, tasty,great,best,delicious (verb) have, went, love, had, like

We fine-tune these two language models for categorizing reviews into four classes (i.e., the number of country-specific groups). We leverage the classification module of the pre-trained models of the Hugging Face library [27]. A mini-batch size of 16, a learning rate of 0.00002, and a training/validation split of 80%:20% are used. The optimization process employs the Adam optimizer, and the loss function is set to sparse-categorical-cross-entropy. We train the model for 3 epochs with an early stopping criterion set. Similar to classical ML classifiers, we report results in both settings, using the original text and excluding named entities.

VI. RESULTS AND DISCUSSION

We perform 5-fold cross-validation to assess the performances of various approaches (both classical ML classifiers and transformer-based models). We report precision, recall, macro F1, and accuracy of various classifiers.

As we can see from Table V, most of the stylistic attributes are very similar across the country-specific groups. The major discrepancies are observed in the length of the review, both in terms of words and sentences. For example, the review length in the Bangladesh group is much shorter than the three other groups, which could be related to the low English proficiency levels of Bangladeshi people. Regarding the presence of sentiment and opinion words in the reviews, we find reviews of Finland, Kenya, and Bangladesh have identical percentages; around 7% based on the opinion lexicon and around 8% based on the VADER lexicon. The reviews from China constitute fewer sentiment words compared to the other groups.

TABLE VIII
PERFORMANCES OF VARIOUS APPROACHES FOR COUNTRY-SPECIFIC REVIEW GROUP PREDICTION TASK USING SEMANTIC FEATURES (TF-IDF BASED)

Classifier	w/o Named Entity				w/ Named Entity			
	Precision	Recall	Macro F1	Accuracy	Precision	Recall	Macro F1	Accuracy
LR	0.728	0.726	0.727	0.717	0.780	0.780	0.780	0.773
SVM	0.746	0.73	0.738	0.724	0.802	0.789	0.795	0.785
RF	0.649	0.655	0.652	0.642	0.688	0.692	0.690	0.680
K-NN	0.662	0.635	0.6485	0.633	0.689	0.671	0.680	0.670
BERT	0.734	0.731	0.733	0.734	0.81	0.81	0.81	0.807
RoBERTa	0.738	0.732	0.735	0.738	0.81	0.81	0.81	0.813

TABLE IX
TOP DISTINGUISHING ADJECTIVES AND ASPECT WORDS USED BY THE LR CLASSIFIER FOR THE DEMOGRAPHY CLASSIFICATION

Group	Top group-specific adjectives and aspects
Finland	**Adjective:** relaxed, worth it, perfect, friendly, fantastic, heavenly, decent, interesting **Aspects:** atmosphere, staff, ingredients
Kenya	**Adjective:** amazing,loved,top, excellent,pleasant,awesome **Aspects:** ambiance,ambience, service, food
China	**Adjective**: great, marvellous, relaxing, marvellous foods **Aspects:** location, staff,manager,services,dishes
Bangladesh	**Adjective:** poor, good,bad, worst,disappointing,disgusting, cheap, testy, delicious, yummy **Aspects:** food quality, taste, price, environment, behaviour, presentation

Table VI shows the performance of CML classifiers using various stylistic features. The best performance is obtained by the SVM classifier, an F1 score of 0.738, while other classifiers yield comparatively subpar performance.

Table VII shows that high similarity exists in different demographically distinct groups regarding the most frequently occurring adjectives and verbs. For example. the words *good*, *great* and *best* are present among the top 5 words in all the groups. Note that the top 5 adjectives in all the distinct groups are positive words; As we collect reviews of highly-rated restaurants from the TripAdvisor website, this outcome is expected. Similar to adjectives, frequently occurring verbs are common among all the four groups (e.g., *have, had, go*). However, the high similarity of the most frequently occurring adjectives and verbs in various groups should be interpreted cautiously, as it does not necessarily mean that all or most words are similar in different groups.

From Table VIII, we can see when semantic features (i.e., unigram and bigrams) are considered, LR and SVM both yield F1 scores of around 0.80, which is much higher than stylistic features. The RF and k-NN exhibit comparatively poor performances, although, still, it is higher than using stylistic input features. The transformer-based BERT and RoBERTa yield slightly better performances by attaining F1 scores around 0.81 using the semantic features. The minor performance improvement by the BERT-based models may be related to the small dataset size. The better efficacy of the semantic features compared to the stylistic features can be partially explained by the presence of restaurant or location-related information present in the reviews. When we exclude named entities like restaurant name or location, we observe that many classifiers, such as SVM or Tree-based classifiers, perform similarly using both stylistic and semantic features.

Finally, we perform an ablation study to find which semantic features (word n-grams) have a high influence (i.e., a high weight) on the classification. As a classifier, we select the LR classifier, which has high interpretability, at the same time, yields a high F1 score (Table VIII). We observe that a number of distinguishing adjectives that refer to distinct groups help LR classifiers achieve good performance. In addition, we notice these distinguishing adjectives refer to diverse aspects of the restaurants. For example, we find reviewers of Bangladeshi groups mention the 'cheap' price, which is missing for other groups. Similarly, the 'location' aspect is mentioned in the Chinese group with the distinguishing adjectives, which is not the case for the other three groups. Again, any mention regarding 'staff' or 'service' is missing for the Bangladeshi group. Furthermore, we notice people of various groups use different words to indicate the same aspect. For example, the Finland group uses the word- 'atmosphere', while people of group Kenya write 'ambiance' or 'ambience', and people of group Bangladesh employ the term 'environment'.

TABLE X
CONFUSION MATRIX OF LR CLASSIFIER

Classifier	Finland	Kenya	China	Bangladesh
Finland	**211**	23	46	8
Kenya	18	**180**	70	2
China	38	44	**203**	16
Bangladesh	17	5	26	**244**

To comprehend the prediction patterns of various classifiers, we analyze the confusion matrix of the LR classifier (with name entities removed). From Table X, we notice that the overall language proficiency levels of the demographic groups have some influence on the predictions. For most groups, we observe that mispredictions primarily refer to the below or

above language proficiency groups. The results suggest that if the difference in the language proficiency level is high between the two groups, they are comparatively easier to distinguish.

VII. Summary and Future Work

In this study, we introduce an annotated review corpus comprised of around 1150 reviews labeled with demographic (e.g., country) information. Then, we perform a stylometric analysis of the reviews curated in different demographics to unveil similarities and differences. Finally, we employ classical ML classifiers and transformer-based pre-trained language models for the automatic demography prediction task. The linguistic analysis reveals that most stylistic features are consistent across reviews of different demographic groups, even though they represent people of various language proficiency levels; nevertheless, some differences are observed in the review length and usage of prepositions. The best-performing ML classifiers yield F1 scores of around 0.80 utilizing word n-gram features, which suggests some semantic differences exist in the reviews of different groups. However, the performance of the semantic feature-based classification degrades when named entities are removed. Our future work will expand the corpus size and include review data from multiple domains and more countries. In addition, we will thoroughly investigate how user preferences on aspects deviate across the demography.

References

[1] G. Goldin, E. Rabinovich, and S. Wintner, "Native language identification with user generated content," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 3591–3601.

[2] R. Sarwar, A. T. Rutherford, S.-U. Hassan, T. Rakthanmanon, and S. Nutanong, "Native language identification of fluent and advanced non-native writers," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 4, pp. 1–19, 2020.

[3] S. Sazzed, "Revealing the demographic attributes of the authors from the abstracts of scientific articles," in *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, 2022, pp. 209–213.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] W. Li and M. Dickinson, "Gender prediction for chinese social media data." in *RANLP*, 2017, pp. 438–445.

[7] O. Rozen, J. Oren, and A. Raviv, "Predicting user demography and device from news comments," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1995–1999.

[8] D. Nguyen, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. De Jong, "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment," in *25th International Conference on Computational Linguistics (COLING 2014)*. Dublin City University and Association for Computational Linguistics, 2014, pp. 1950–1961.

[9] M. Abdul-Mageed, C. Zhang, A. Rajendran, A. Elmadany, M. Przystupa, and L. Ungar, "Sentence-level bert and multi-task learning of age and gender in social media," *arXiv preprint arXiv:1911.00637*, 2019.

[10] M. Koppel, J. Schler, and K. Zigdon, "Automatically determining an anonymous author's native language," in *International Conference on Intelligence and Security Informatics*. Springer, 2005, pp. 209–217.

[11] J. Tetreault, D. Blanchard, and A. Cahill, "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*, 2013, pp. 48–57.

[12] J. Brooke and G. Hirst, "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, 2013, p. 37.

[13] S.-M. J. Wong and M. Dras, "Contrastive analysis and native language identification," in *Proceedings of the Australasian Language Technology Association Workshop 2009*, 2009, pp. 53–61.

[14] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow, "Toefl11: A corpus of non-native english," *ETS Research Report Series*, vol. 2013, no. 2, pp. i–15, 2013.

[15] S. Granger, "The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research," *Tesol Quarterly*, vol. 37, no. 3, pp. 538–546, 2003.

[16] S. Sazzed, "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*, 2022, pp. 881–886.

[17] S. Rosenthal and K. McKeown, "Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 763–772.

[18] D. Nguyen, N. A. Smith, and C. Rose, "Author age prediction from text using linear regression," in *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, 2011, pp. 115–123.

[19] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging." in *AAAI spring symposium: Computational approaches to analyzing weblogs*, vol. 6, 2006, pp. 199–205.

[20] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 2011, pp. 37–44.

[21] D. V. Phuong and T. M. Phuong, "Gender prediction using browsing history," in *Knowledge and Systems Engineering*. Springer, 2014, pp. 271–283.

[22] S. Sazzed, "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)*, 2021, pp. 1281–1288.

[23] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM*, 2014.

[24] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[25] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 168–177. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014073

[26] 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, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6