Comprehending Lexical and Affective Ontologies in the Demographically Diverse Spatial Social Media Discourse

Salim Sazzed

Department of Computer Science, Old Dominion University, Norfolk, VA USA
Department of Computer Science, University of Memphis, Memphis, TN, USA
saim.sazzed@gmail.com

Abstract—This study aims to comprehend linguistic and socio-demographic features, encompassing English language styles, conveyed sentiments, and lexical diversity within spatial online social media review data. To this end, we undertake a case study that scrutinizes reviews composed by two distinct and demographically diverse groups. Our analysis entails the extraction and examination of various statistical, grammatical, and sentimental features from these two groups. Subsequently, we leverage these features with machine learning (ML) classifiers to discern their potential in effectively differentiating between the groups. Our investigation unveils substantial disparities in certain linguistic attributes between the two groups. When integrated into ML classifiers, these attributes exhibit a marked efficacy in distinguishing the groups, yielding a macro F1 score of approximately 0.85. Furthermore, we conduct a comparative evaluation of these linguistic features with word n-gram-based lexical features in discerning demographically diverse review data. As expected, the n-gram lexical features, coupled with fine-tuned transformer-based models, show superior performance, attaining accuracies surpassing 95% and macro F1 scores exceeding 0.96. Our meticulous analysis and comprehensive evaluations substantiate the efficacy of linguistic and sentimental features in effectively discerning demographically diverse review data. The findings of this study provide valuable guidelines for future research endeavors concerning the analysis of demographic patterns in textual content across various social media platforms.

I. INTRODUCTION

Demographic data concerning user traits enable an understanding of user behaviors, ultimately facilitating improved decision-making for various social and business challenges. For instance, the analysis of demographic data assists businesses in making informed decisions regarding marketing, product development, customer experiences, and competitive positioning. In a social context, diverse demographic review data empowers policymakers to gain insights into the experiences and perspectives of different social groups, which, in turn, facilitates decision-making that promotes equity, inclusive representation, targeted interventions, evidence-based decision-making, and accountability. Furthermore, demographically tagged social media plays a significant role in computational social science by highlighting differences in beliefs and behaviors among demographic groups [1].

Recent studies have shown that demographic traits can be inferred from the linguistic characteristics of written content [2], [3], [4], [5]. The determination of various demographic attributes of users, such as age or gender, from written comments, has been investigated by [2], [6]. For example, Rosenthal and McKeown [7] attempted to predict the users' ages from blog content by incorporating various features specific to the blog and the behavior and interest of users. Schler et al. [8] observed significant differences in content and style levels between male and female bloggers by analyzing a large corpus of blogs of around 300 million words.

In addition, some studies tried to identify the English language nativeness of the writers from demographically diverse data. Although the perspective of their study was the second language acquisition (SLA) research, such as contrastive analysis, syntactic or grammatical errors made by non-native speakers [9], [10] 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) [11], the international corpus of learner English [12]. Research pertaining to demographically diverse informal reviews, such as those found in social media and prominently affected by English language nativeness and fluency level, remains mostly unexplored [13] (except a few [14], [15], [16]).

Therefore, in this study, we aim to understand how linguistic and semantic attributes of text vary across demographically different groups in the context of social media, taking into account factors such as English language nativeness, geography, and socio-culture. In particular, we aim to provide insight into the following research questions-

- RQ1: Do the variations of the linguistic features (e.g., synthetic, lexical) render sufficient signals to distinguish diverse demographic groups when incorporated in classical ML classifiers?
- RQ2: Whether the linguistic or n-gram lexical features perform better for the demography prediction task when incorporated into ML classifiers.

The two distinct demographic groups considered here represent individuals of two different socio-economic cultures.

More importantly, the English reviews written by these two groups differ significantly in terms of English language nativeness and fluency levels. The first review group represents restaurant review data collected from Bangladesh, a country with mostly low proficient non-native English speakers [1]. In Bangladesh, almost 98% of people are Bengali native speakers [2]. The other group contains restaurant reviews written by users located in the USA (mostly English native speakers). We extract various statistical and syntactic features such as review length, frequency of opinion words, and usage of POS (part-of-speech) from the reviews of both groups. We find that linguistic features exhibit sufficient distinguishing signals to differentiate between the two groups of reviews when used as input for classical machine learning (CML) classifiers. Additionally, we explore the performance of lexical word n-grams-based features for classifying review groups by incorporating them into classical ML classifiers and transformer-based language models. As expected, we observe that transformer-based models, when utilizing lexical features, outperform both types of feature-based classical ML classifiers.

A. Contributions

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

1) We identify differences in stylistic, syntactic, and statistical features among reviews from diverse demographic groups.
2) We show that it is possible to distinguish demographically diverse informal social media reviews by incorporating various linguistic features into the machine learning classifiers.
3) Finally, we compare the differences in the performance of linguistic and lexical n-gram-based features for distinguishing diverse review groups.

II. DEMOGRAPHY PREDICTION TASK

A. Dataset

As mentioned earlier, in this study, we investigate restaurant reviews written in English from two different demographic groups. The first group, which we refer to as *Demography-1*, consists of reviews collected from restaurants located in Bangladesh [3]. In Bangladesh, the majority of people (98%) speak standard Bengali or one of its many dialects as their first language, while the remaining population speaks regional or minority languages. English proficiency in Bangladesh is generally categorized as low among secondary speakers [4]. The second review group, referred to as *Demography-2*, comprises a subset of the Yelp restaurant review dataset written by primarily English native speakers residing in the USA.

To mitigate potential domain bias in the evaluation process, the textual content used in both groups was sourced

---

[1] https://www.ef.com/wwen/epi/
[2] https://www.worldatlas.com/articles/what-languages-are-spoken-in-bangladesh.html
[3] https://www.kaggle.com/tuxboy/restaurant-reviews-in-dhaka-bangladesh
[4] https://www.ef.com/wwen/epi/

TABLE I: Sample reviews from *Demography-1* and *Demography-2*

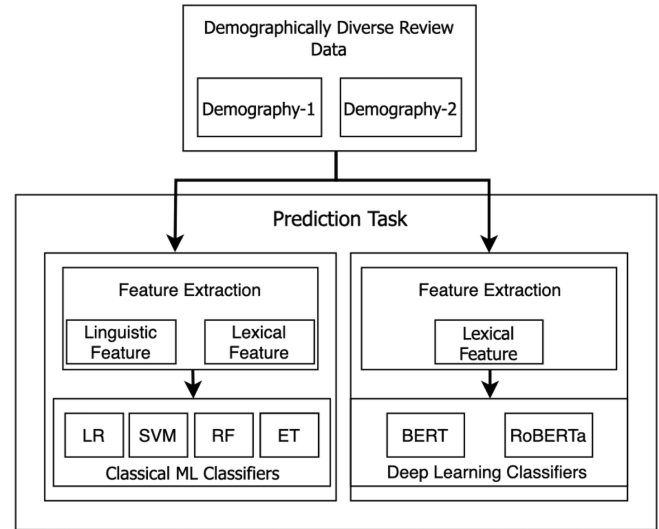| Demography-1 |
|---|
| **1.** The food and environments was sooo good . But stuff behavior was not up to mark...It was worst .... |
| **2.** It is a very nice place for eating. Last night we arranged our daughter's birthday party. It was excellent. All guests are happy with their food & services. It was really good & more than our expectation. We are fully satisfied with their food & services. Thanks to cafe Rio. We really loved it. Keep it up... |
| **3.** Worst buffet ever!!The service sucks,the food sucks....basically everything sucks about this place.And that one plate policy is bullshit." |
| **Demography-2** |
| **1.** Great place for lunch, transport yourself back in time to this quaint farm grill which offers high quality food, fresh ingredients, killer burgers and an atmosphere that is one of a kind. Good place for breakfast as well. Family owned restaurant goodness. Only drawback is the line, however good things come to those who wait. |
| **2.** I usually do not complain about bad food but the fish and chips gave my wife a serious case of food poisoning. So if some manager reads this please check out your fish supply or change the cooking oil. She never gets sick and has been sick all day (December 29, 2012) from the meal eaten December 28, 2012. |
| **3.** Slices Pizza is pretty good, not the best though. But if you are on Mill Ave its the best Pizza option. Uno's is around the corner and probably the worst pizza in Phoenix. If the main guy that works there was not so rude probably would have given it 3 stars. If you are in Tempe a much better Pizza place is Red Devil Pizza about 5 minutes South. |

Fig. 1: Overall framework of the study

exclusively from the same domain, namely, restaurant reviews. Additionally, an equal number of reviews were selected for each group to prevent the influence of class imbalance on the classification outcomes. The final dataset comprises a total of 9974 reviews, where each group includes 4987 reviews.

B. Classical Machine Learning-based Prediction

To distinguish the two demographic groups, we employ several classical ML classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Extra Trees (ET). For all classical ML classifiers, the default parameter settings of the scikit-learn library [17] are used.

As an input of classical ML classifiers, we utilize two distinct groups of features: i) linguistic features and ii) word unigram and bigram-based lexical features, separately. The lin-

guistic features further encompass several sub-types, including grammatical and sentiment features.

1) Linguistic Features: We extract three different types of linguistic features from the reviews, which are utilized as input for the classical ML classifiers. Text statistics such as review length in terms of words and sentences have been employed in related work such as language variety identification task [18]. Grammatical features such as the usage of PoS tags have been studied for language nativeness identification tasks in the earlier works as these stylistic features reflect user communication behavior and interaction style [19]. In addition, lexicon coverage, a feature that refers to the usage of opinion or sentiment words, is considered.

We employ the Mann-Whitney U test to determine which linguistic features show significant differences in the two groups of reviews. The Mann-Whitney U test is a non-parametric test of the null hypothesis, which is often used to test the differences in the distributions of two sets of values. We utilize the Mann-Whitney U test between two groups for all the linguistic features. The null hypothesis states that the distribution of a specific attribute in *Demography-1* is the same as the underlying distribution of the same attribute in *Demography-2*, while the alternative hypothesis suggests the opposite.

TABLE II: The statistics of various linguistic features in the reviews belong to two groups, *Demography-1* and *Demography-2*

Type	Feature	Demo-1	Demo-2
Text Statistics	Total words (in corpus)	147401	656672
	Total sentences (in corpus)	16272	47922
	Mean review length (#words)	29.56	131.70
	Mean review length (#sentences)	3.32	9.611
	Mean sentence length (#words)	8.89	13.70
Lexical Diversity	Total unique words (in corpus)	10473	28297
Negation	Total negation words (in corpus)	1585	5287
	(%) of negation words (in corpus)	1.0%	0.80%
	Negation words per review (mean)	0.32	1.06
Grammatical Feature	Total articles (in corpus)	6553	45918
	(%) of article in corpus (word)	4.44%	6.99%
	Articles/review (words)	1.31	9.20
	Total adjectives (in corpus)	15632	57561
	(%) of the adjective (in corpus)	10.60%	8.76%
	Number of adjectives/review	3.13	11.54
	Total verbs (in corpus)	14754	74881
	(%) of the verb in corpus	10.01%	11.40%
	Number of verbs/review	2.95	15.01
	Total prepositions (in corpus)	8466	52637
	(%) of the preposition (in corpus)	5.79%	8.00%
	Number of prepositions/review	1.69	10.55
	Total SC(in corpus)	2960	19402
	(%)of the SC(in corpus)	2.0%	2.95%
	SC per reviews	0.59	3.89
Sentiment Lexicon	Total opinion words (Hu &Liu)	8799	33691
	Lexicon coverage (Hu & Liu)	6.22%	5.14%
	Total opinion words (VADER)	9516	36390
	Lexicon coverage (VADER)	6.51%	5.56%

a) Text Length Features: These features represent the length of the reviews based on word and sentence level: i) Number of words per review, ii) Number of sentences per review, iii) Number of words per sentence. As earlier

studies suggested that English texts written by non-native English speakers are usually simpler than those of natives [20], the sentence length could be a distinguishing factor for the demography prediction task.

b) Grammatical and Negation Features: We consider a set of grammatical attributes that may provide signals to discern the diverse demographic groups.

- Articles: The number of articles (i.e., *a, an, the*) present in a review is computed.
- Adjectives: The number of adjectives present in a review is computed. The spaCy [21] library is used to identify adjectives in a text.
- Verbs: The usage of verbs in both corpora is provided. Similar to adjective identification, spaCy [21] library is used for verb identification.
- Prepositions: We calculate the number of prepositions present in the reviews of two groups. A list of commonly used prepositions is considered (details can be found here [5]).
- Subordinating conjunctions (SC): Additionally, we take into account the presence of subordinating conjunctions that indicate complex sentences. A complex sentence typically consists of one or more dependent (subordinate) clauses and one or more independent clauses. Subordinating conjunctions are words or phrases that connect dependent clauses to independent clauses. Examples of subordinating conjunctions include "although," "as," "because," "before," "how," "if," "once," "since," and so on. We examined the occurrence of 50 commonly used subordinating conjunctions in each review [6].
- Negative words: The VADER [22] negative word list is used as a reference to find the number of negative words in each group [7].

c) Sentiment Lexicon Coverage: The coverage of two popular English sentiment lexicons, Opinion Lexicon [23] and VADER [22], is computed for each of the reviews from both groups.

2) Lexical n-gram Features: In addition, as lexical features, we extract word n-grams from the reviews of both groups. An n-gram denotes a consecutive sequence of n items within a given textual context. Particularly, we extract both unigrams (individual words) and bigrams (two-word combinations) from the review texts. Following this extraction, we calculate their respective term frequency-inverse document frequency (tf-idf) scores and incorporate these computed scores as input features into the CML classifiers.

C. Deep Learning-based Classification

The transformer-based pre-trained language models, such as BERT [24] and RoBERTa [25], have shown state-of-the-art results in various text classification tasks with limited labeled data. We fine-tune both transformer-based pre-trained models

---

[5]https://github.com/sazzadcsedu/LinguisticAnalysis
[6]https://github.com/sazzadcsedu/LinguisticAnalysis
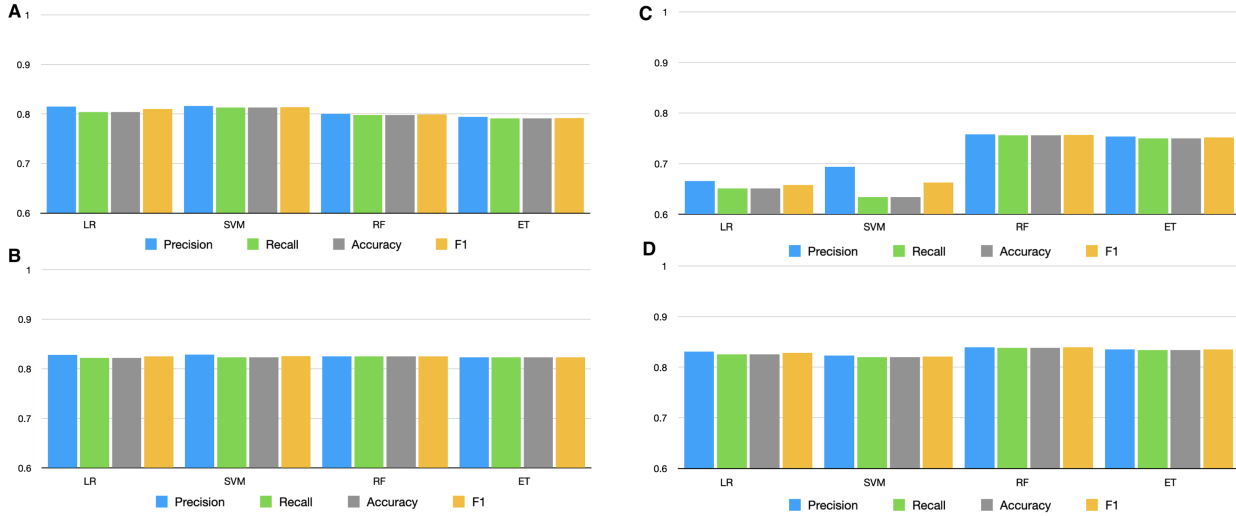[7]https://github.com/cjhutto/vaderSentiment

Fig. 2: Performances of classical ML classifiers using various types of linguistic features, (A) text statistics features, (B) grammatical features, (C) sentiment features, and (D) using features A, B, and C altogether
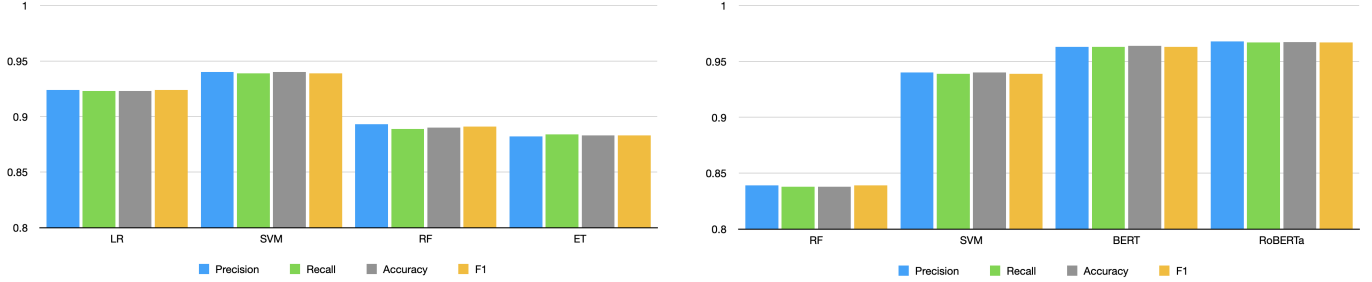

Fig. 3: Performances of classical ML classifiers using lexical features (i.e., word unigrams and bigrams)

Fig. 4: The comparisons between the best results achieved by classical ML classifiers (i.e., RF with linguistic features and SVM with lexical features), and two transformer-based language models, for the demographic group prediction task

for categorizing reviews into two demographically diverse groups. Since this is a binary-level classification task, we utilize the classification module of these pre-trained models. The Hugging face library [26] is used for fine-tuning all the pre-trained models. As the initial layers of pre-trained models primarily learn general features, they are left unchanged during fine-tuning process. Only the last layer of the pre-trained model is trained using new data specifically for the binary-level classification task.

We tokenize the input data for fine-tuning the language model. As pre-trained models typically support texts with a maximum of 512 tokens, we divide reviews longer than 512 words into 512-word chunks. During training, all the 512-token chunks are assigned the same class as the original review. During testing, the final class of the review is determined by majority voting. In the event of a tie, we consider the word length of the chunks to decide the final class label. For training, we use a mini-batch size of 8 and a learning rate of $4*10^{-5}$. During the training process, 20% of the samples are dedicated to validation. We optimize the pre-trained models using the Adam optimizer, with the loss parameter set to

categorical cross-entropy. The training procedure is carried out for four epochs with an early stopping criterion set. Note that all hyperparameters of deep learning models are determined based on empirical evaluation.

III. RESULTS AND DISCUSSION

A. Evaluation Settings

To compare different classification methods for the demography prediction task, we utilize 5-fold cross-validation. Several key metrics, namely precision, recall, macro F1 score, and accuracy, are employed to assess the performance of classifiers. These metrics provide a comprehensive assessment of the performance of the different classifiers.

B. Performance of classical ML Classifiers using Linguistic and Lexical Features

The linguistic analysis reveals that several attributes of reviews, such as review length, range of vocabulary used, coverage of opinion lexicon, and usage of part-of-speech (POS)

TABLE III: Top adjectives(#occurrences) and verbs (#occurrences)

Top adjectives				Top verbs			
Demography-1		Demography-2		Demography-1		Demography-2	
good	1609	good	2854	had	1609	had	2987
bad	407	great	1925	have	363	have	2489
worst	331	other	1073	go	361	get	1818
best	327	little	944	ordered	272	go	1609
great	274	nice	884	serve	242	got	888
worst	264	more	822	went	218	ordered	848
nice	250	best	802	love	198	know	829
poor	232	few	654	served	196	make	767
awesome	214	friendly	645	visit	160	going	743

vary across two demographic groups that can be attributed to the reviewer's English language nativeness or proficiency level and socio-culture (shown in Table II). It is observed that the usage of common opinion terms is more apparent (i.e., high coverage of sentiment lexicon) in the writing of non-native speakers (Demography-1). Figure 2 provides the precision, recall, F1 score, and accuracy of various classical ML-based methods utilizing linguistic features. We can see that when all the linguistic features (i.e., length, grammatical, and sentiment) are utilized, all of the four classical ML classifiers LR, SVM, RF, and ET perform similarly; They achieve F1 scores of around 0.833 and accuracy around 83%.

We apply the Mann-Whitney test to find whether any of the stylistic features are significantly different in the reviews of *Demography-1* and *Demography-2*. A *p*-value of 0.05 is used for the significance test. The Mann-Whitney test indicates many of the linguistic features are significantly different in the two groups; however, we observe they do not provide very high distinguishing power when incorporated into the classical ML classifiers. For example, the text statistic features, such as review length and sentence length, are significantly different (p-value less than 0.05) in both groups; however, when incorporated into the classical ML, they obtain a much lower F1 score of 0.813 than the lexical features. The high standard deviations (std.) of the review length, with respect to both word and sentence, indicate that many reviews in each group spread far away from its group mean value, which may induce the classifier to yield wrong predictions.

Classical ML classifiers yield F1 scores between 0.89 and 0.94 when word unigram and bigram lexical features are utilized (see Fig. 3). The better performance of ML classifiers with lexical features indicates the presence of some distinguishing socio-culture-specific words and named entities in the reviews of both groups, which generate effective signals to discern the two groups. Nevertheless, we find that reviews of both groups share some common adjectives, such as *good*, *great*, *nice*, and *best* (Table III), which are among the most frequently occurring adjectives in both groups. A few other adjectives are also common to both groups.

This observation suggests that both demographic groups, irrespective of English language nativeness, tend to use simple and commonly used adjectives to express opinions and feelings. One dissimilarity we observe is that in reviews of *Demography-2* (written mostly by native speakers), adjectives of quantity such as *more*, *little*, and *few* are more frequent than in the *Demography-1* review group. None of these adjectives of quantity appear among the top 10 adjectives in *Demography-1*, and none of them occur more than 214 times. When examining the verbs used in the two groups of reviews, we observe that the most frequent verbs are also very similar (see Table III). Although there is a high presence of overlapping words in both groups, there are certainly distinguishing words, such as named entities, that help the classifiers discern between the two groups.

C. Performance of Deep Learning based Classifiers

The fine-tuned transformer-based language models yield impressive results utilizing unigram and bigram-based lexical features; they attain almost perfect accuracy by correctly classifying around 97% instances (Fig 4). The pre-trained language models are generated based on an enormous amount of textual content, which helps to capture the implicit pattern of the reviews and can effectively identify the language nativeness of reviewers. Also, the high efficacy of lexical feature-based classical ML classifiers and transformer-based models can be attributed to the English language proficiency level of the people of Bangladesh, who are not known as very fluent English speakers. Additionally, the presence of named entities with specific meanings and socio-cultural terms in the reviews of both groups supports the lexical n-gram-based approach, leading to better performance.

D. Implications of the Study

This study provides insights into various perspectives by analyzing data from two demographic groups, including the followings-

a) Language Landscapes and Identity: The study contributes to understanding how English language variation in social media usage is linked to different demographic groups. It can shed light on how individuals from diverse demographic backgrounds use language to express their cultural and social identities, providing insights into the complex relationship between language, ethnicity, and nationality.

b) Lexical Diversity in Online Communication: Examining the lexical diversity in online social media review data can shed light on the richness and variety of language used by individuals. This analysis can help researchers understand the level of vocabulary sophistication, linguistic creativity, or the influence of cultural factors in online communication.

c) Sociolinguistic Research and Language Policy: The study's findings can contribute to sociolinguistic research and language policy discussions. Understanding the relationship between language nativeness, demographic diversity, and social media usage can inform discussions on language rights, language maintenance, and linguistic identity in digital spaces.

IV. SUMMARY, LIMITATIONS AND FUTURE WORK

This study aims to distinguish reviews of two different socio-demographic groups leveraging various linguistic and

lexical features, language models, and ML classifiers. From two demographically distinct review groups (*Demography-1* and *Demography-2*), various linguistic features are extracted to train ML classifiers for the demography prediction task. In addition, two state-of-the-art pre-trained language models, BERT and RoBERTa, are fine-tuned with the n-gram-based features for the prediction task. We observe that linguistic features are capable of distinguishing demographically diverse reviews; when they are fed into classical ML classifiers, an F1 score (best result) close to 0.85 is obtained. The pre-trained models exhibit very high efficacy for distinguishing reviews using lexical features, which can be attributed to the presence of name-entity and sociocultural-specific features in the two review groups. Our analysis reveals the contrast and similarity of implicit characteristics of reviews written by two demographically diverse review groups. We present multiple approaches for the demography prediction task that can help a diverse set of downstream decision-making tasks.

One of the limitations of this work is that it only considers two demographically distinct groups. Since English proficiency levels (a predominant factor affecting linguistic characteristics) among non-native speakers may vary across demographics, such as geography, cultures, and language families, it is worthwhile to analyze non-native English review data from multiple demographics. Furthermore, it is worth exploring whether linguistic features exhibit similar distinguishing signals when demographics representing people with similar English language proficiency levels (e.g., native and highly proficient non-native speakers) are considered. In our future work, we will focus on collecting, annotating, and analyzing review data from diverse demographics, including variations in English fluency, geographical regions, native language families, and socio-cultures.

References

[1] Z. Wood-Doughty, M. Smith, D. Broniatowski, and M. Dredze, "How does twitter user behavior vary across demographic groups?," in *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 83–89, 2017.

[2] R. G. Guimaraes, R. L. Rosa, D. De Gaetano, D. Z. Rodriguez, and G. Bressan, "Age groups classification in social network using deep learning," *IEEE Access*, vol. 5, pp. 10805–10816, 2017.

[3] X. Chen, Y. Wang, E. Agichtein, and F. Wang, "A comparative study of demographic attribute inference in twitter," in *Proceedings of the international AAAI conference on web and social media*, vol. 9, pp. 590–593, 2015.

[4] F. Hsieh, R. Dias, and I. Paraboni, "Author profiling from facebook corpora," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[5] R. Dias and I. Paraboni, "Cross-domain author gender classification in brazilian portuguese," in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1227–1234, 2020.

[6] 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)*, pp. 1950–1961, Dublin City University and Association for Computational Linguistics, 2014.

[7] 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*, pp. 763–772, 2011.

[8] 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, pp. 199–205, 2006.

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

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

[11] 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.

[12] 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.

[13] R. Sarkar, S. Mahinder, and A. KhudaBukhsh, "The non-native speaker aspect: Indian english in social media," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pp. 61–70, 2020.

[14] S. Sazzed, "Impact of demography on linguistic aspects and readability of reviews and performances of sentiment classifiers," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100135, 2022.

[15] 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)*, pp. 1281–1288, 2021.

[16] S. Sazzed, "Stylometric and semantic analysis of demographically diverse non-native english review data," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 470–476, IEEE, 2022.

[17] 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.

[18] C. van der Lee and A. van den Bosch, "Exploring lexical and syntactic features for language variety identification," in *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pp. 190–199, 2017.

[19] S. Volkova and E. Bell, "Identifying effective signals to predict deleted and suspended accounts on twitter across languages," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, pp. 290–298, 2017.

[20] 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*, pp. 3591–3601, 2018.

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

[22] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.

[23] 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*, KDD '04, (New York, NY, USA), pp. 168–177, ACM, 2004.

[24] 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.

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

[26] 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), pp. 38–45, Association for Computational Linguistics, Oct. 2020.