Reflection of Demographic Background on Word Usage

Aparna Garimella*
Adobe Research
Adobe Big Data Experience Lab
garimell@adobe.com

Carmen Banea*
University of Michigan
Computer Science and Engineering
carmen.banea@gmail.com

Rada Mihalcea
University of Michigan
Computer Science and Engineering
mihalcea@umich.edu

The availability of personal writings in electronic format provides researchers in the fields of linguistics, psychology, and computational linguistics with an unprecedented chance to study, on a large scale, the relationship between language use and the demographic background of writers, allowing us to better understand people across different demographics. In this article, we analyze the relation between language and demographics by developing cross-demographic word models to identify words with usage bias, or words that are used in significantly different ways by speakers of different demographics. Focusing on three demographic categories, namely, location, gender, and industry, we identify words with significant usage differences in each category and investigate various approaches of encoding a word's usage, allowing us to identify language aspects that contribute to the differences. Our word models using topic-based features achieve at least 20% *improvement in accuracy over the baseline for all demographic categories, even for scenarios with classification into 15 categories, illustrating the usefulness of topic-based features in identifying word usage differences. Further, we note that for location and industry, topics extracted from immediate context are the best predictors of word usages, hinting at the importance of word meaning and its grammatical function for these demographics, while for gender, topics obtained from longer contexts are better predictors for word usage.*

* This work was done while the authors were affiliated with the University of Michigan.

Action Editor: Mohit Bansal. Submission received: 18 February 2022; revised version received: 28 November 2022; accepted for publication: 19 December 2022.

https://doi.org/10.1162/coli_a_00475

1. Introduction

According to Shweder et al. (1998), "to be a member of a group is to think and act in a certain way, in the light of particular goals, values, pictures of the world; and to think and act so is to belong to a group." The demographics of people, such as geographic location, gender, industry, education, or age, can affect and shape their beliefs and behaviors, and are reflected in their everyday thoughts, ideas and actions (McCarty and Shrum 1993; Newcomer and Baldwin 1992). Examining what people say and write in their daily lives can help us identify ways in which language use is influenced by demographics. In this work, we focus on the geographic location, gender, and industry. We gather a large corpus of personal writings from a diverse set of online bloggers and explore differences in word usage across the various demographic groups.

We find inspiration in a line of research in psychology that posits that people from different demographic backgrounds and/or speaking different languages perceive the world around them differently, ranging from their perception of time and space (Kern 2003; Boroditsky 2001), body shapes (Furnham and Alibhai 1983), or surrounding objects (Boroditsky, Schmidt, and Phillips 2003). As an example, the study described by Boroditsky, Schmidt, and Phillips (2003) showed that the perception of objects in different languages can be affected by their assigned grammatical gender in those languages. For instance, one of the words used in the study is the word "bridge," which is masculine in Spanish and feminine in German. When asked about the descriptive properties of a bridge, Spanish speakers described them as being *big*, *dangerous*, *long*, *strong*, *sturdy*, and *towering*, while German speakers said they are *beautiful*, *elegant*, *fragile*, *peaceful*, *pretty*, and *slender*.

While this previous research has the benefit of careful in-lab studies, which explore differences in worldview with respect to one dimension (for example, *time*, *space*) or word (for example, *bridge*, *sun*) at a time, it also has limitations in terms of the number of experiments that can be run, as subjects have to be brought to the lab for every new aspect being examined. We aim to address this shortcoming by using large-scale computational linguistics to identify demographic differences in word usage in a data-driven bottom–up fashion. We hypothesize that these differences can be regarded as an approximation of a demographic group's distinctive worldview. Rather than starting with predetermined hypotheses (for example, that Spanish and German speakers may perceive bridges differently), computational linguistics methods allow us to run experiments on hundreds of words, and identify those where usage differences exist between the various demographic groups.

While previous studies in psychology have considered differences across speakers of different languages, in our work we choose to focus on differences across speakers of English from different demographics. We explicitly avoid the use of multiple languages, so that we can avoid the errors that may be introduced by the translation process.

We seek to answer two main research questions: First, given a word w, are there significant differences in its usage by demographic groups? At a high level, we use the phrase **word usage** to refer to the wordview of people while using the given word. That is, the manner in which a given word is used by people of a specific demographic group in their day-to-day life. In order to ground word usage to linguistic notions, we consider four aspects to capture word usage in this work: *(i)* the immediate words that are typically used while using the given word (that may indicate the meaning and/or grammatical functionality of the word); *(ii)* the words that are frequently present in the global context (100 words to the left and right); *(iii)* the sociolinguistic sense in which the word is used, as indicated by polarity, the sentiment/affect, or any morality terms

(using existing lexicons, such as LIWC [Pennebaker, Francis, and Booth 2001], WordNet Affect [Strapparava and Valitutti 2004], OpinionFinder [Wilson et al. 2005], and Morality dictionaries [Ignatow and Mihalcea 2012]); and finally, *(iv)* the topics that are generally spoken about while using the word. In other words, while "word usage" in this work is used to describe people's world views while talking about a specific word, we use the above four types of linguistic features to capture this worldview in a given word's usage by a specific audience. To answer the first research question, we build word models based on several classes of linguistic features and train classifiers to attempt to differentiate between usages of w across different demographic groups.

Second, if significant differences in word usage are identified, can we use feature analysis to understand the nature of these differences? We perform two analyses: (1) *feature ablation* that highlights the linguistic features contributing to these differences and (2) *topic modeling*, which is used to identify the dominant topic for each word in each demographic group and to measure the correlations between the topic distributions in the selected demographic groups.

2. Related Work

Most of the previous research work across cultures and demographics has been undertaken in fields such as sociology, psychology, or anthropology (de Secondat and de Montesquieu 1748; Shweder 1991; Cohen et al. 1996; Street 1993). For instance, Shweder (1991) examined the cross-cultural similarities and differences in the perceptions, emotions, and ideologies of people belonging to different cultures, while Pennebaker, Rimé, and Blankenship (1996) measured the emotional expressiveness among the northerners and southerners in their own countries, to test Montesquieu's geography hypothesis (de Secondat and de Montesquieu 1748), which states that residents of warmer climates are more emotionally expressive than those living in cooler ones. More recently, the findings of Boroditsky, Schmidt, and Phillips (2003) indicated that people's perception of some inanimate objects (such as a bridge) is influenced by the objects' grammatical genders in their native tongue.

To our knowledge, there is only limited work in computational linguistics that explored cross-cultural differences through language analysis. Paul and Girju (2009) identified cultural differences in people's experiences in various countries from the perspective of tourists and locals. Specifically, they analyzed forums and blogs written by tourists and locals about their experiences in Singapore, India, and United Kingdom, using an extension of LDA (Blei, Ng, and Jordan 2003). One of their findings is that while topic modeling on tourist forums offered an unsupervised aggregation of factual data specific to each country that would be important to travelers (such as destination's climate, law, and language), topic modeling on blogs authored by locals showed cultural differences between the three countries with respect to several topics (such as fashion, pets, religion, health).

Yin et al. (2011) used topic models along with geographical metadata in Flickr to analyze cultural differences in the tags used for specific image categories, such as cars, activities, or festivals. They performed a comparison over the topics across different geographical locations for each of the categories using three modeling strategies: location-driven, text-driven, and latent geographical topic analysis (LGTA) that combines location and text information; they found that the LGTA model worked well not only for finding regions of interest, but also for making effective comparisons between different topics across locations.

More recently, Vilares and Gómez-Rodríguez (2018) and Shwartz (2022) analyzed the differences in the interpretation of time expressions in different languages and cultures. Vilares and Gómez-Rodríguez (2018) studied the semantics of part-of-day nouns (such as "morning" or "night") by utilizing tweets containing time-specific greetings (such as "good morning") used in different cultures. They presented several interesting insights, such as that Asian, African, and American countries tend to begin the day earlier than European countries (with the exception of Germany), based on the language use of the people from these countries on Twitter. Shwartz (2022) proposed the task of mapping time expressions (such as "morning") used in different cultures in their corresponding languages to specific hours in the day, and further applied their methods to 23 additional unlabelled languages, and analyzed the differences predicted by the models.

3. Are There Significant Word Usage Differences by Demographic Groups?

We start out by exploring ways to model words that can highlight differences in usage within a demographic category. To achieve that, we look at words that occur with preponderance in each demographic, and refine candidate lists with seemingly differing usage. As mentioned before, we use "word usage" to refer to the worldview of people while using a given word. For each word w used by a given demographic, we generate a vector model that accounts for w's usage across all demographic slices. For instance, if we look at gender, the model will have samples of w from both male and female data. Each word is represented through several lenses, by looking at potential linguistic differences, accounting for social and psycholinguistic aspects, and modeling the topical content of the context in which the word appears.

In the rest of this article, **demographic** will refer to the large demographic categories (such as *gender* or *culture*) while **demographic slices** will refer to the sub-categories in each demographic (such as *male* or *female*).

We train machine learning models over the word representations, aiming to predict which slice a word instance belongs to. We take the model's performance to be an indicator of how much discriminating power w has within each slice. In this article, we focus on gender, location, and industry, but other demographic criteria can be explored as well.

3.1 Extracting Demographic-Biased Words

We start by collecting social media data from Google Blogger, which is a free online blog-publishing platform launched in August 1999, and it is used by people from numerous English-speaking countries. We use the blog genre because blog posts are lengthier than microblogs (posted on sites such as Twitter or Facebook), as well as richer in terms of vocabulary and topic usage. In addition, users complete their profile information using a drop-down list for countries, gender, and industry, ensuring a closed vocabulary that enables better matches across users. We were able to identify a large number of profiles and posts written in English, where users self-specified their country, gender, and industry. The blog posts were published between 1999 and 2016.

Country. We culled 5,527,606 blog posts associated with 20,533 user profiles from fifteen countries. From these, we retain countries with more than 700 profiles (to ensure user and language diversity), resulting in a set of twelve countries. Table 1 shows the profile

Table 1
Blog data statistics.

Category	Profiles	Posts		
		Total	Mean	Std. Dev.
Country				
United Kingdom	886	393,160	444	758
Australia	879	412,743	470	1,099
Canada	854	501,475	587	1,276
Nigeria	844	259,817	308	620
New Zealand	843	231,469	275	501
Ireland	831	215,702	260	612
South Africa	825	140,659	171	353
Philippines	808	243,194	301	553
United States	800	556,186	695	1,134
Singapore	748	262,006	350	556
India	719	210,456	293	797
Pakistan	707	100,755	143	688
Gender				
Female	4,683	1,256,161	385	943
Male	4,051	1,560,314	268	510
Industry				
Arts	670	221,083	330	626
Communications	499	217,032	435	900
Technology	330	96,500	293	858
Fashion	252	46,155	183	326
Internet	197	92,609	470	1,215
Business Svcs.	193	66,891	347	1,205
Publishing	190	93,671	493	748
Non-Profit	187	75,634	404	1,709
Engineering	176	56,514	321	803
Consulting	157	57,741	368	922
Science	138	38,734	281	498
Marketing	136	65,697	483	1,112
Religion	123	38,460	313	571
Tourism	110	28,680	261	585
Advertising	102	37,204	365	738

and post distribution for each country. The five continents represented encompass very different language use scenarios.

Gender. From the same data, we select the profiles which specified either male or female as the gender, resulting in approximately 9,000 users authoring approximately three million posts, where the data is roughly equally distributed between male and female users (Table 1). While we acknowledge the rich communities that form the other groups

Table 2
Sample list of words with the highest frequencies obtained for each demographic category.

LOCATION	GENDER	IND_{200}	IND_{100}
one	one	one	one
time	time	look	time
use	take	time	see
know	know	day	take
day	day	new	day
work	year	take	year
people	new	think	new
first	first	year	first

of gender, we limit our work to study the male and female groups, as most of the profiles provide one of these as their gender.

Industry. Google Blogger provides 39 industry options that users can specify in their profile. We consider those with over 100 user profiles, resulting in fifteen industries. The number of profiles is lower when compared to country and gender, as most users do not disclose their industry, and many industries do not meet the 100 profile threshold (Table 1). We refer to this data as $Industry_{100}$ (IND_{100}); $Industry_{200}$ (IND_{200}) represents a subset of this data with more than 200 profiles per slice.

Blog posts are preprocessed to remove HTML tags, email ids, urls, repeated characters, posts shorter than ten words or with more than 25% non-English words, and posts without published times or other demographic information. The content is lemmatized using Stanford CoreNLP (Manning et al. 2014) to generate a compact word index.

For each demographic category, we create a set of candidate target words by identifying the top 500 words satisfying the following constraints: they are the most frequent words within each demographic category, they are not stopwords, modal, or auxiliary verbs, they do not contain numerals or special characters, and they have at least three characters. We obtain four sets of 500 target words for location, gender, $Industry_{100}$, and $Industry_{200}$, respectively. For each demographic, these words are selected to have high frequencies across *all* demographic slices, enabling the corresponding word data set to contain diverse usage examples that are representative of all slices. Table 2 shows the top eight target words from each category; 319 words are common across them.

3.2 Encoding Word Usage

For each demographic–word tuple, we construct a **target word data set** by culling usage examples from blog posts, where the class label is the slice name. These posts are truncated to a maximum of 100 words to the left and right of the target word, disregarding sentence boundaries. We balance the target word data sets with respect to author and the time when the blog was posted. Specifically, we apply the following heuristics (Garimella, Banea, and Mihalcea 2017) to create our target word data sets for each demographic category: (1) Compute the minimum number of users n over all the demographic slices (e.g., female and male authors in the case of the gender). (2) From each slice, select the top n users based on the number of years they were blogging and the number of posts they wrote. This ensures that the maximum amount of data will

be available for the selected users. (3) For each of these n users, pick at most 100 posts in a round-robin fashion from the years in which they blogged. (4) Let M be the total number of posts collected in this manner from all the slices. In order to avoid having most of the posts coming from a small number of years, set a cutoff X as a fraction of M. For each year, a maximum of X posts will be chosen from the set of M posts (X = 0.15M). (5) To ensure that all the users get to contribute posts, and that the contribution of prolific writers is kept in check, maintain user participation scores:

$$p(user) = \frac{\text{posts collected from user}}{\text{total number of posts collected}} \quad (1)$$

These user participation scores are updated after each year is processed. (6) Sort the years in increasing order of number of posts and iterate through them; identify the lowest number of posts contributed by the least prolific writer, then collect the minimum number of posts from all users who published in that year in a round-robin manner. Then, select additional posts from users in increasing order of participation scores, until the number of posts for the year reaches the cutoff X. (7) After each year, update the user participation scores.

We do not balance the target word data sets across topics, as we regard potentially different topic distributions as reflective of word usage variations across the slices within each demographic (for example, bloggers from India may be naturally more interested in cricket than bloggers from the United States are). For each target word, an equal number of posts is selected from each slice pertaining to each demographic category; hence, each target word data set is *class*-balanced.

Table 3 shows the average number of profiles and posts across the 500 word data sets retained after processing, and their standard deviation. For each word, we encode the word usage in each word context in terms of the linguistic choice in the context while using the word, the social sense in which the word is used, and the topics of discussion with respect to the word. For this purpose, we use the following four types of features to build the word usage representations from the corresponding word data sets for each demographic.

Local features (Loc). These consist of the target word itself, and five context words to the left and right of the target word; they capture the immediate surrounding words to the target word, which reflect the grammatical function of the word.

Contextual features (Con). To extract these features, we identify **contextual words**—ten most frequent words per demographic slice appearing more than 5 times in a target word's data set w; they are weighted by term frequency in each of w's contexts of ± 100 words. They capture the salient words in the wider context of the target word.

Table 3
Dataset statistics. First value: mean, second value: standard deviation.

CATEGORY	USERS	POSTS	POSTS/CATEGORY
LOCATION	6,684 ± 967	49,848 ± 24,583	4,151 ± 2,049
GENDER	5,820 ± 825	146,215 ± 50,239	54,969 ± 25,119
IND_{200}	1,145 ± 181	9,288 ± 5,541	2,322 ± 1,385
IND_{100}	2,180 ± 348	16,713 ± 7,665	1,114 ± 511

Sociolinguistic features (Soc). These include word classes from four sources capturing social and psycholinguistic insights as they pertain to bloggers, and are extracted from the full contexts (±100 words). They include (1) fractions of words that fall under each of the 70 Linguistic Inquiry and Word Count (LIWC) categories (Pennebaker, Francis, and Booth 2001) (the 2001 version of LIWC includes about 2,200 words and stems grouped into categories relevant to psychological processes, such as social, affective, cognitive, perceptual, biological); (2) fractions of words belonging to each of the five fine-grained polarity classes in OpinionFinder (Wilson et al. 2005), namely, strongly negative, weakly negative, neutral, weakly positive, and strongly positive; (3) fractions of words belonging to each of the ten Morality classes (Ignatow and Mihalcea 2012), such as authority, care, fairness, ingroup, sanctity; and (4) fractions of words belonging to each of the six WordNet Affect (Strapparava and Valitutti 2004) classes (anger, disgust, fear, joy, sadness, and surprise).

Topic features (Topic). These features capture the various topics that bloggers write about when using the target words. The features consist of the topic distributions learned over a word's data set when extracting latent topics using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). We use the LDA implementation included with the Gensim Python library (Řehůřek and Sojka 2010) and derive 10 topics. As typically done in topic modeling, we preprocess the target word's data set by removing (1) a standard list of stop words, (2) words with very high frequency ($> 0.25\times$ corpus_size [the total number of words in the word data set]), and (3) words with frequency less than 5.

3.3 Classifying Word Usage

We hypothesize that if a word w is used in highly varying ways by the different demographic slices (that is, has high usage differences), then a machine learning classifier trained to predict the demographic slice of any given context of w will have a high accuracy, as compared to a random chance or majority vote baseline. Conversely, if w is used in similar ways by the various demographic slices (that is, has very little usage differences), then the classifier will not be able to discriminate between its usages by the various slices (or will have a low accuracy in doing so). We utilize the vector spaces derived above to train classifiers to predict the demographic slice that a given word's context belongs to (or is authored by), and ultimately explore word usage differences across the various demographic slices. We use the WEKA machine learning toolkit (Hall et al. 2009) for all our experiments. We consider five multi-class classifiers: (1) Naive Bayes (NB), (2) Random Forest (RF), (3) Decision Tree (DT), (4) k-Nearest Neighbor (k-NN) with $k = 3$, and (5) AdaBoost (AB). In addition to these classifiers, we also use the majority class classifier (that is, one that predicts the majority class for every test example) as a baseline (BL). As the target word data sets are class-balanced, the BL has accuracies of around 8.34% for location (with 12 classes), 50% for gender (with 2 classes), 25% for Industry_{200} (with 4 classes), and 6.67% for Industry_{100} (with 15 classes). We use the classifier with the highest accuracy to identify words with high usage, as it has a higher discriminative ability between the word usages by the various demographic slices, for the rest of the article.

Table 4 shows the average accuracy over 50 randomly selected words obtained for the learners under consideration using all the features. The NB classifier performs consistently better than the others: for *location*, +4.09% compared to second best k-NN, for $Industry_{200}$, +2.07% compared to second best DT, and for $Industry_{100}$ +3.33%

Table 4
Ten-fold cross-validation accuracy over 50 words using all classifiers. Last row shows the difference between NB and the best other classifier. Best results are in bold, second best in italics.

LEARNER	LOCATION	GENDER	IND_{200}	IND_{100}
Avg NB	**17.18**	70.12	**40.28**	**14.06**
Avg DT	10.48	*70.69*	*38.21*	9.89
Avg RF	11.52	**71.06**	35.20	10.09
Avg *k*-NN	*13.08*	63.89	33.14	9.06
Avg AB	12.10	70.27	35.49	*10.71*
Avg BL	8.34	50.00	25.00	6.67
Difference	4.09	−0.94	2.07	3.33

compared to second best AB. For *gender*, RF exhibits a stronger predictive ability by 0.94%. As none of the other classifiers achieve consistent top results, we use the NB classifier for the remaining experiments in this article.

3.4 Results and Discussion

We compare the performance of the classifier against the majority class BL, which always predicts the class with the highest occurrence in the data. A significant increase in accuracy over the baseline for a given word suggests that we can automatically identify the demographic slice to which a writer belongs. This is taken as an indication that there exist usage differences for that word among the slices. The results reported in this article are obtained using ten-fold cross-validation on the word data sets. When creating the folds, we ensure that all posts authored by the same blogger are either in the test or the training set pertaining to a given fold, but never in both. This is important, as including a blogger's writings in both training and test can potentially tune the model to the writing styles of individual bloggers instead of learning the underlying demographic-based differences between the bloggers.

Garimella, Mihalcea, and Pennebaker (2016) focused on location, conducting experiments classifying words based on usage in Australia compared to the United States. The finding of that study, that word usage differences can be captured across countries, prompted us to extend the analysis by looking at three demographic dimensions (country, gender, and industry), while also significantly expanding diversity within each category, exploring differences across 12 countries, 2 genders, and 15 industries. While that preliminary study was using syntactic and part-of-speech features, due to their below-baseline performance in ablation studies, we focus our evaluations on the local, contextual, and sociolinguistic features introduced in Section 3.2, while also considering for the first time topic-based features.

Table 5 shows the average accuracies of the NB learner and the BL over 500 target words in each demographic using all the feature types. The highest average improvement in accuracy with respect to the baseline is observed for gender (2 classes) at 20.07%, while for location (12 classes) it is 8.67%. In the case of industry, for $Industry_{200}$ (4 classes), the improvement is 15.18%, while for $Industry_{100}$ (15 classes), it is 7.15%.

Not only do we notice an *average* improvement, but *every individual* target word exhibits a higher NB accuracy than the majority class baseline. For location, these range

Table 5
Ten-fold cross-validation accuracies averaged over 500 target words for NB and BL classifiers. Last row shows the difference between them for each demographic.

LEARNER	LOCATION	GENDER	IND_{200}	IND_{100}
Avg NB	17.00	70.12	40.38	13.93
Avg BL	8.33	49.96	25.06	6.78
Difference	4.09	20.16	15.32	7.15

from 11.71% for *Monday* to 4.1% for *hand*, for industry, from 9.18% for *create* to 3.69% for *movie*, and for gender, from 26.3% for *product* to 8.45% for *government*.

These results indicate that there are indeed differences in the ways bloggers from the various demographic slices use the target words.

Performance for Generic Compared with Specific Target Words. While Table 2 shows a few sample target words that are most frequently occurring (and hence are overly generic, such as *time*, *one*, *day*), we would like to point out that our target word list also includes several specific words, such as *bank*, *attack*, *business*, and *development* for location, *daughter*, *relationship*, *happy*, and *weekend* for gender, and *job*, *product*, *idea*, and *company* for industry. Our target word list is a mix of generic and specific words, and thus we believe that the results from our current set of experiments also reflect the word usage differences for demographic-specific words. As a sanity check, we compare the accuracies averaged over generic and specific target words (that is, those that have occurrences $\geq$ and $<$ 0.5 times the maximum occurrence for any target word, respectively) for each demographic category. There are around 100 generic words and 400 specific words for each demographic category. We observe the following performance: For location, the generic target word accuracy is 17.07%, while the specific target word accuracy is 16.98%; for gender, the generic target word accuracy is 70.87%, and the specific target word accuracy is 70.01%. These results suggest that the performance for specific target words does not vary too much from those for generic target words.

We perform an error analysis by examining the confusion matrices constructed using the results obtained from the NB classifiers using all the four feature types. A confusion matrix illustrates the performance of a given classifier by presenting the counts of actual and predicted values in the form of a table, which enables us to view which class labels are predicted incorrectly. For each demographic category, we first obtain the target word-wise confusion matrices with normalized values instead of absolute counts. We then obtain aggregate confusion matrices for each demographic category by averaging the matrices for all the 500 target words (Figures 1, 2, 3). Based on these matrices, we can cluster the various slices within each demographic category based on which locations, industries, or genders are confused with each other by the classification model. In that respect, each demographic category has specific slices that the model often defaults to, or, in other words, there are generic patterns based on the four feature types that cause the model to get confused.

Figure 1 shows the confusion matrix for location obtained by aggregating the confusion matrices of all the corresponding target words. From the figure, we notice that Australia, Canada, New Zealand, Philippines, South Africa, and United States often get confused with Singapore, while India gets confused with Pakistan, and Ireland with United Kingdom. While Singapore is over-predicted for six slices, the United States is

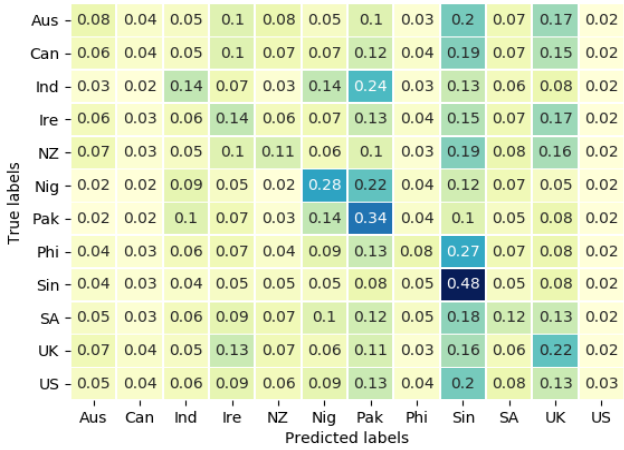

Figure 1
Confusion matrix for locations.

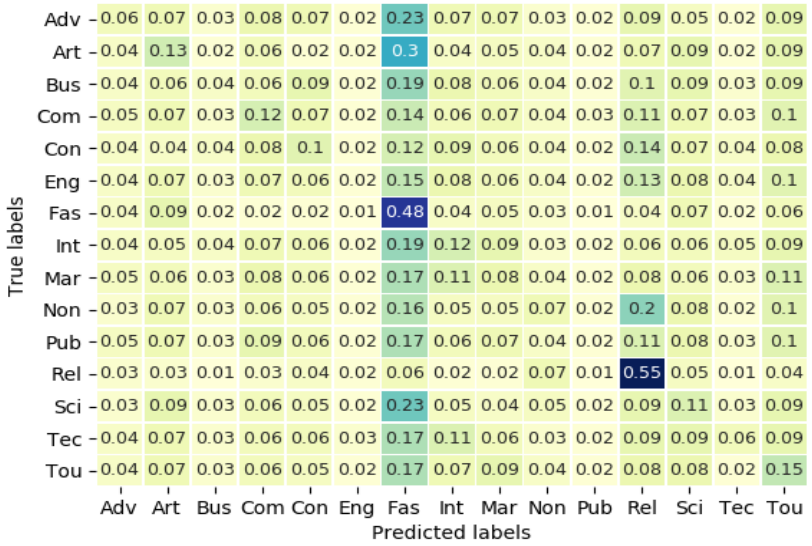

Figure 2
Confusion matrix for industries.

predicted the minimum number of times. In other words, six of the twelve countries are confused as Singapore. However, in comparison to the other countries, Singapore has the highest one-versus-all classification accuracy (48%), making it the country with the most distinct word usage, and United States the country with the most generic word usage (3%).

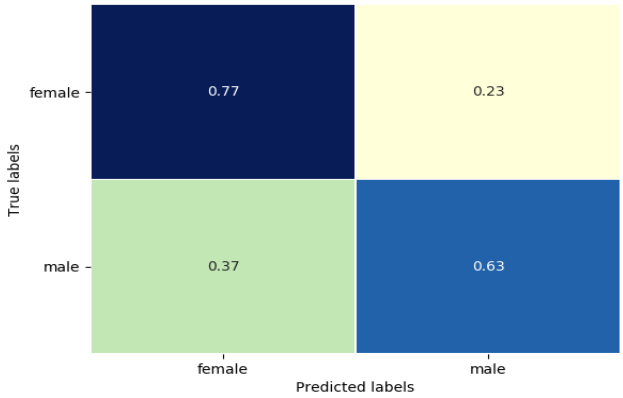

Figure 3
Confusion matrix for genders.

Figure 2 shows the confusion matrix for industry obtained by aggregating the confusion matrices of all the corresponding target words. Most industries (12 out of 15) often get confused with Fashion, with the exception of Consulting, Religion, and Non-Profit. While Fashion is predicted as the target class for most of the industries, Engineering is the industry that is predicted the minimum number of times, despite not being the most sparsely represented. However, in comparison to the other countries, Fashion and Religion have the highest one-versus-all classification accuracy (48% and 55%, respectively), making them the industries with the most distinct word usage, while Engineering and Publishing have the most generic word usage (2% and 2%, respectively).

In the case of gender (Figure 3), since the classifier accuracies are greater than 50% and there are only two classes, the majority of instances are classified correctly. One interesting note is that female word usage offers a stronger signal compared with male word usage (that males are more likely to be confused by the classifier with females [37%], than females with males [23%]).

As seen from Figures 1 through 3, words do exhibit differences in their usage across various countries, genders, and industries, with some demographic slices being more similar to each other, while other slices exhibiting stronger differences. In the next section, we examine which factors contribute the most to these usage differences by focusing on the various features used, both from a qualitative and quantitative perspective.

4. What Factors Best Encode Usage Differences?

We now take a closer look at each demographic to analyze the extent to which the linguistic features contribute to word usage differences. For each target word, we also seek to understand differing usage patterns based on topics and word classes across the various slices in a demographic category. For this, we perform (1) feature analysis and (2) qualitative and quantitative analysis of the various topics and word classes appearing in the surrounding contexts of the target words.

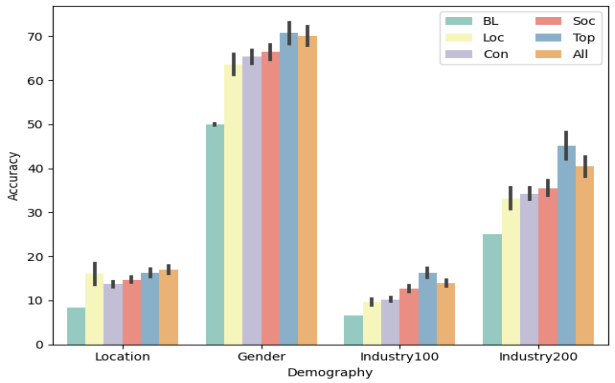

Figure 4
NB and BL accuracies averaged over 500 target words. Features: local, contextual, sociolinguistic, topic, and combined.

4.1 Feature Analysis

We study the role played by the various linguistic features in differentiating word usage among authors of different demographics. We explore how the feature types perform individually, as well as jointly, and we identify which of the sociolinguistic signals are more suited to encode usage differences. Then, we analyze the effect of context size on classification performance.

4.1.1 Feature Type Performance. To assess the encoding ability of the feature types introduced in Section 3.2, we retrain our word models using one feature type at a time. Figure 4 shows the average accuracies obtained by training Naive Bayes classifiers on the data set of each target word across the 500 in each demographic. BL represents the majority class baseline, which is surpassed by every single one of our feature types. Interestingly, the information that each of these feature types encodes does not seem to be orthogonal, as when combining them, we get lower accuracies for all demographics with the exception of *location*; the latter seems to be boosted by the interaction between local and topic based features, attaining an accuracy that is higher than achievable by topic features alone by 0.7%. Local features perform well for the location demographic, probably because they are able to capture the regional usage of words, while for gender and industry the discriminating power (at least over a context of ± 5 words) is not sufficient to surpass features extracted from contexts of ± 100 words.

The performance of topic-based features exceeds the baseline: for location +7.97%, for gender +20.72%, for Industry_{200} +20.07%, and for Industry_{100} +9.44%. Similarly, they surpass the performance of other feature types: for location, the strongest contender is the local features type (+0.07%), while for the other demographics, the strongest contender is the sociolinguistic features type (+4.3% for gender, +3.6% for Industry_{200}, and +9.58% for Industry_{100}).

4.1.2 Sociolinguistic Features. Figure 5 shows the classification accuracy for each lexicon included under sociolinguistic features, namely LIWC, OpinionFinder (OF), Morality (ML), and WordNet Affect (WNA), and all lexicons combined (Soc). The accuracies in the figure suggest that LIWC features contribute the most to the overall sociolinguistic

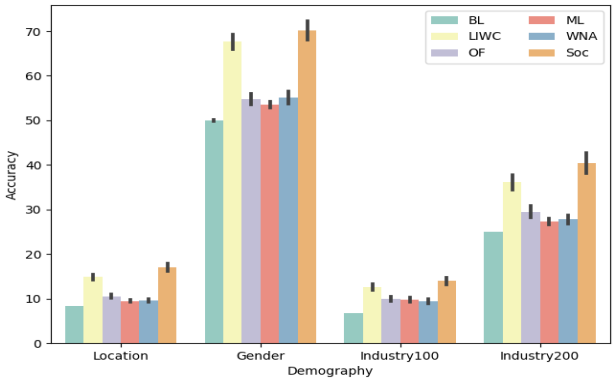

Figure 5
NB accuracy averaged over 500 target words using sociolingistic features.

classification performance, driving the accuracy when classifying over the entire feature set. OF, ML, and WNA all have significantly lower accuracies (with $p < 0.01$ using the t-test). Accordingly, we reassess the overall classification performance using all feature types, but replacing sociolinguistic features with LIWC; we note no change in performance, signaling that LIWC features are sufficient to model sociolinguistic aspects.

4.1.3 Context Size and Performance. To examine the effect of context size on feature type performance, we train NB classifiers with context lengths ranging from ± 5 to ± 100 words. Figure 6 shows the results for each demographic. Overall, increasing the context length results in an improved performance for local, contextual, and sociolinguistic

(a) Location. (b) Gender.

(c) Industry$_{200}$ with 4 categories. (d) Industry$_{100}$ with 15 categories.

Figure 6
NB accuracy averaged over 500 target words with varying context sizes: 5, 20, 50, and 100.

features for all the demographics. However, in the case of topic features, performance is highest when only five context words are used for *location* and *industry*. This generates instances in the data set that are sparse (as the topics are extracted cumulatively from a narrow context), and are easy for the classifier to assign to a given class, even for Industry_{100} with 15 categories.

For *gender*, despite the binary classification, sparsely populated topic-based instances are not sufficiently discriminative; gender seems to be best represented by the interplay of topic features extracted from lengthy contexts; an increase in window size for topic extraction from 5 to 100 results in a jump in accuracy from 53% to 71%.

In conclusion, for *location* and *industry*, the topics extracted from the immediate context surrounding a target word are the best predictors of word usage classification performance. Lengthier contexts lead to signal degradation; the steepest drop in performance is noted when increasing the window from 5 to 20 words. For *gender*, the trend is opposite, observing the steepest increase in performance between 5 and 20 context words, and reaching the peak at 100 words. This implies that location and industry are best predicted by word *meaning*, namely, the sense that a word takes given the topics that co-occur with it. A word's neighboring context (when we look at ± 5 words) is mostly able to capture the *grammatical function* of the word, and not its meaning. For this reason, for *location*, the prediction accuracy increases from 16% for local features to 27% for topic features for the same context of 5. For *industry*, the trend is even steeper: For Industry_{200}, local features have a prediction accuracy of 33% compared with 57% for topic features, and for Industry_{100}, local features have a prediction accuracy of 9% compared with 32% for topic features. For *gender*, the word itself is not used with a markedly different sense by male and female bloggers; rather what generally male and female individuals talk about allows us to identify whether the word appears in male or female writing. For this reason, local features at a context of 100 display the same performance as topic features extracted from the same context length (with an accuracy of 71%); while the large number of features obtained from a vocabulary of $\pm$ 100 context words, in the first case, generates sparse instances, for topic features, we achieve a dense 10 feature space for every instance.

By pairing the context length findings (Figure 6) with the accuracies observed per feature type and overall (Figure 4), we conclude that the strongest signal in differentiating word usage across demographics comes from topic modeling. A short window of ± 5 words from which to extract latent topics is optimal for differentiating across *location* (accuracy of 28% compared to BL of 7%) and *industry* (for Industry_{200}, accuracy of 57% compared to BL of 25%, and for Industry_{100}, accuracy of 33% compared to BL of 8%). A wider window is necessary to extract topics that encode usage differences for gender, with ± 100 words being optimal (accuracy of 70% compared to BL of 50%).

Figure 7 shows the classification results (NB_{Best}) using features extracted from optimal context sizes for each feature type (as seen in Figure 6). Comparing NB_{Best} with the original algorithm (NB) trained over all feature types, we note a significant increase in performance ($+3.35\%$ for location, $+6.23\%$ Industry_{200}, and $+3.98\%$ for Industry_{100}). Despite these improvements, leveraging only topic features on the optimal context size (Best_feature) results in the best performance (by an average of $+8.15\%$ across all demographics). Since location and industry are best represented through 10 topic features, additional features degrade the signal, and cause a performance drop. Gender is the only demographic that is resilient, and maintains a strong performance, whether using the best feature type, or overall.

To conclude, we emphasize that for every demographic we are able to achieve at least a 20% improvement in accuracy compared to the baseline, which is a notable

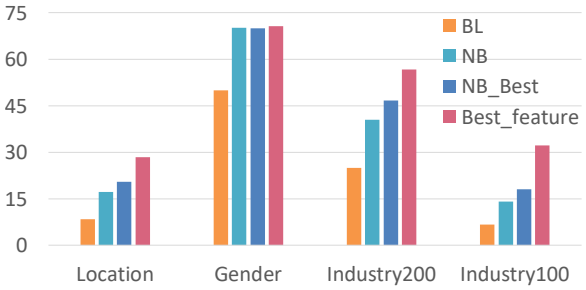

Figure 7
Average ten-fold cross-validation accuracy over 500 target words when using all features (NB, NB_Best) vs. topic features (Best_feature).

performance in itself, and this is even more commendable considering that this improvement is achieved over demographic categories such as location (12 classes) or industry (15 classes for Industry_{100}).

4.2 Quantitative and Qualitative Analyses

As topic features contribute the most to identifying usage differences, we investigate their modeling ability by analyzing the various latent topics that emerge from the contexts associated with a word's occurrences in the writings of people from various demographic slices. Since examining the cross-demographic topics for all the 500 target word data sets qualitatively is tedious, we consider a subset of 30 words that have the highest improvements over the baseline for each demographic (that is, with *strong demographic bias* in their use). Because our focus here is to aggregate the latent topics and to explore differences over the various demographic slices, for each instance in the word data set, we determine the topic that has the highest preponderance in the topic distribution predicted by LDA. We call this **the dominating topic**. We then tally the instances and their dominating topic across each slice, to obtain a word-centered topic distribution per slice. Using this experimental set-up, we perform the quantitative and qualitative analyses below. Since these analyses are carried out in view of the dominating topic, we use the original 100 word context.

4.2.1 Quantitative Analysis. To gauge how similar or different the various demographic slices are with respect to word usage, we compute the Pearson correlation between the topic distributions for each pair of demographic slices (that is, combinations of slices in a given demographic taken two at a time) for the top 30 target words. Hence, when two slices have a high correlation, they are more similar, while when they have a low correlation, they are more dissimilar in terms of the primary topic with which a word is associated.

Gender. The pair-wise gender correlation for the top 30 words is -0.75, meaning that the extracted topics are very different. This is in line with our findings from Section 4.1.3, where we also saw that the combinations of topics extracted from lengthier contexts are better predictors of gender.

Table 6
Pearson correlation for the top 30 words with demographic bias for location. Values below 0.6 are in light gray; in each row the highest values are in bold and the lowest in italics.

	AU	CA	IN	IE	NZ	NG	PK	PH	SG	ZA	GB	US
AU	1	0.77	−0.50	0.53	**0.89**	*−0.55*	−0.52	0.54	0.40	0.63	0.81	0.55
CA	0.77	1	−0.45	0.69	**0.80**	−0.39	*−0.47*	0.51	0.26	0.52	0.73	0.77
IN	*−0.50*	−0.45	1	−0.49	−0.47	**0.66**	0.63	−0.17	−0.28	−0.06	*−0.50*	−0.20
IE	0.53	0.69	−0.49	1	0.60	−0.27	−0.35	0.39	0.02	0.37	**0.74**	0.57
NZ	**0.89**	0.80	−0.47	0.60	1	−0.53	*−0.57*	0.51	0.35	0.69	0.75	0.65
NG	*−0.55*	−0.39	**0.66**	−0.27	−0.53	1	0.61	−0.29	−0.32	−0.23	−0.43	−0.19
PK	−0.52	−0.47	**0.63**	−0.35	*−0.57*	0.61	1	−0.31	−0.40	−0.38	−0.43	−0.38
PH	0.54	0.51	−0.17	0.39	0.51	−0.29	*−0.31*	1	**0.60**	0.42	0.36	0.48
SG	0.40	0.26	−0.28	0.02	0.35	−0.32	*−0.40*	**0.60**	1	0.18	0.13	0.33
ZA	**0.63**	0.52	−0.06	0.37	0.69	−0.23	*−0.38*	0.42	0.18	1	0.49	0.44
GB	**0.81**	0.73	*−0.50*	0.74	0.75	−0.43	−0.43	0.36	0.13	0.49	1	0.50
US	0.55	**0.77**	−0.20	0.57	0.65	−0.19	*−0.38*	0.48	0.33	0.44	0.50	1

Location. Table 6 shows the pair-wise country correlations.[1] Australia and Nigeria are the least correlated, while the highest correlation is observed between Australia and New Zealand. We note that this trend also holds among the 30 target words that have the lowest NB improvements over the baseline. While a rushed conclusion may be that the disparities are caused by differences in topics covered by a demographic slice, and not differences in topics as they are associated with word usage, we show in Section 4.2.2 that the words we randomly sampled are associated with a plethora of dominant topics for the same slice. Also, from our earlier analysis regarding context length (Section 4.1.3), we know that the shorter the window from which we extract topics for *location* and *industry*, the stronger the discriminative power of the model. India, Nigeria, and Pakistan all have negative correlations with 9 other countries, and positive correlations with each other, indicating that word usage differences are quite minor among them.

Industry. Table 7 shows the same analysis as it pertains to industry. Religion and Non-Profit have the highest pair-wise correlation, while Arts and Consulting are the least correlated. Religion is negatively correlated with 9 other industries making it the most "different," while Publishing is positively correlated with 13 other industries, making it the most "similar" to the others.

4.2.2 Qualitative Analysis. For a qualitative overview of usage differences in terms of the overall topics, we analyze the dominating topics in each demographic slice for the top 30 words with strong demographic bias. For ease of comparison, we associate labels to the hidden topics. These are either picked from the words associated with the topics, or, where intuitive, they are given based on the overall themes in the topics. Table 8 shows sample words for each demographic, accompanied by 3 sample words for each latent topic (with the exception of the target word itself). The two countries and two industries shown in the Table 8 are among the pairs that have least confusion score in the confusion matrices.

1 The country codes can be found at: `http://www.nationsonline.org/oneworld/country_code_list.htm`.

Table 7
Pearson correlation for the top 30 words with demographic bias for Industry_{100}. Values below 0.2 are light gray; in each row the highest values are in bold and the lowest in italics.

	Adv	Art	Bus	Com	Con	Eng	Fas	Int	Mar	Non	Pub	Rel	Sci	Tec	Tou
Adv	1	0.37	0.14	**0.51**	0.07	−0.07	0.45	0.09	0.32	−0.15	0.46	*−0.18*	0.08	−0.05	−0.12
Art	0.37	1	*−0.34*	0	−0.7	*−0.34*	**0.72**	−0.16	0.09	−0.17	0.18	−0.18	0.48	−0.26	0.05
Bus	0.14	*−0.34*	1	−0.01	**0.67**	0.22	−0.04	0.13	0.06	−0.03	−0.13	−0.24	0.1	0.23	−0.09
Com	0.51	0	−0.01	1	0.28	0.24	*−0.26*	−0.01	0.28	0.24	**0.85**	−0.06	−0.2	−0.1	0
Con	0.07	*−0.7*	**0.67**	0.28	1	0.43	−0.5	0.13	0.07	0.19	0.11	0.11	−0.34	0.27	−0.18
Eng	−0.07	−0.34	0.22	0.24	0.43	1	*−0.48*	0.34	0.09	0.39	0.17	0.23	−0.24	**0.56**	0.21
Fas	0.45	**0.72**	−0.04	−0.26	−0.5	−0.48	1	−0.11	−0.04	*−0.32*	−0.19	−0.21	0.53	−0.22	−0.08
Int	0.09	−0.16	0.13	−0.01	0.13	0.34	−0.11	1	0.6	*−0.52*	−0.02	−0.4	−0.26	**0.87**	−0.04
Mar	0.32	0.09	0.06	0.28	0.07	0.09	−0.04	**0.6**	1	−0.4	0.32	*−0.44*	−0.18	0.37	0.26
Non	−0.15	−0.17	−0.03	0.24	0.19	0.39	−0.32	*−0.52*	−0.4	1	0.23	**0.67**	0.05	−0.35	0.16
Pub	0.46	0.18	−0.13	**0.85**	0.11	0.17	*−0.19*	−0.02	0.32	0.23	1	−0.01	−0.04	−0.1	−0.03
Rel	−0.18	−0.18	−0.24	−0.06	0.11	0.23	−0.21	−0.4	−0.44	**0.67**	−0.01	1	−0.2	−0.22	*−0.24*
Sci	0.08	0.48	0.1	−0.2	*−0.34*	−0.24	**0.53**	−0.26	−0.18	0.05	−0.04	−0.2	1	−0.26	0.11
Tec	−0.05	−0.26	0.23	−0.1	0.27	**0.56**	−0.22	0.87	0.37	*−0.35*	−0.1	−0.22	−0.26	1	−0.08
Tou	−0.12	0.05	−0.09	0	−0.18	0.21	−0.08	−0.04	**0.26**	0.16	−0.03	*−0.24*	0.11	−0.08	1

Table 8
Seven sample words with significant usage difference among demographic slices.

	COUNTRY	
WORD	IN	US
group	BUSINESS (company, business, service)	(make, food, day, love)
national	HEALTH (bank, health, system)	(day, time, good)
wednesday	MARKET (company, market, share)	FRIEND (good, friend, night)
south	WAR (war, india, pakistan)	WORK (work, school, live)
town	PROJECT (city, business, project)	BOOKS (book, story, read)
market	STOCK (price, stock, bank)	FILM (film, family, play)
store	APP (phone, app, free)	(place, house, visit)
	GENDER	
WORD	FEMALE	MALE
product	BEAUTY (skin, face, beauty, eye)	TRADE (market, increase, trade)
oil	COOKING (olive, onion, garlic, cook)	POLITICAL (government, national, war)
card	DESIGN (stamp, cut, image)	ELECTION (report, party, office)
cut	PAPER (card, paper, piece)	TAX (pay, tax, price)
party	CELEBRATION (birthday, family, christmas)	POLITICAL (state, political, election)
box	GIFT (card, paper, gift)	GAME (film, game, play)
purchase	SHOP (shop, wear, buy)	STOCK (market, company, stock)
	INDUSTRY	
WORD	BUSINESS	ENGINEERING
create	COMPANY (company, business, service)	ONLINE (blog, follow, website)
daily	MONEY (money, company, service)	REPORT (news, report, national)
remove	GROUP (member, report, police, party)	CLEARING (house, car, tree)
follow	PLAN (plan, company, money)	ONLINE (post, link, twitter)
service	CUSTOMER (company, customer, product)	ACCESS (account, free, online, website)
provide	SUPPORT (support, project, customer)	INFORMATION (free, information, site)
result	RESEARCH (process, research, project)	ELECTION (govt, million, election, report)

For **country**, the word *national* predominantly describes (1) nature parks by the bloggers from Australia, New Zealand, and South Africa, (2) live shows by those from Canada, (3) health systems in India, (4) sports teams in Ireland and Nigeria, (5) art museums in Philippines, and (6) security in Pakistan. Another interesting example is the word *Wednesday*; Australians and Canadians talk about Wednesdays to post blogs and photos, while Indians think about market shares and companies in relation to Wednesday. It is used to describe events by those in Ireland, school work by those in New Zealand, politics by those in Nigeria and Pakistan, respectively, friends by those in Philippines, Singapore, United Kingdom, and United States, and God by South Africans.
For **gender**, female writers predominantly use the word *party* for celebratory events, while male writers use it to describe the political parties. Similarly, the word *oil* is used by women in cooking-focused posts, while males mention it in geo-political contexts.
For **industry**, the word *result* is used in terms of (1) research by writers from Business Services and Consulting, (2) elections by those from Communications or Media, Engineering, and Publishing, (3) search by those from Internet and Technology, and (4) medical test by those from Marketing and Tourism. Looking at the word *follow*, it describes (1) game shows and films by bloggers from Advertising and Marketing, (2) company plans and money by those from Business Services and Consulting, (3) political parties and members by those from Communications or Media and Publishing, (4) online posts and tweets by those from Engineering, Internet, and Technology, and (5) God by those from Religion and Non-Profit.

Table 9
Few other words with significant usage difference among various demographic slices with least confusion scores for location.

	COUNTRY		
WORD	AUSTRALIA	CANADA	US
national	NATURE (park, walk, drive)	SHOW (show, live, film, music)	DAY (day, time, good)
buy	HOUSE (house, home, place)	DRESS (wear, shop, photo)	BOOK (book, life, read)
service	CAR (city, car, water)	ENTERTAINMENT (book, play, show)	GOD (life, church, god)

Table 10
Few other words with significant usage difference among various demographic slices with least confusion scores for industry.

	INDUSTRY			
WORD	FASHION	ENGINEERING	BUSINESS	RELIGION
personal	(style, design)	INFORMATION (information, service)	EXPERIENCE (work, school)	LIFE (life, god)
natural	LOOK (hair, light, eye)	PLACE (area, view)	RESOURCE (gas, company)	LIFE (life, god)
save	WATER (water, food)	FILE (post, blog)	COMPANY (company, service)	WORLD (life, world)
share	PHOTO (picture, photo)	SITE (blog, post)	MARKET (company, market)	LOVE (god, love)

Table 9 shows a few more words with differences in their topics of discussion with respect to the given words, for three other countries with the least pair-wise confusion scores (of 0.02). Table 10 shows similar analysis for industry, with three industry groups with least pair-wise confusion scores.

5. Conclusions

In this article, we performed an extensive analysis of demographic differences in word usage between people belonging to various demographics; we analyzed blog posts written by people from twelve countries, two genders, and fifteen industries. We selected 500 target words for each of the demographics based on occurrence frequency in blog posts, and studied how these words are utilized. For that, we built classifiers for each target word based on linguistic and psycholinguistic features, and topic-based word classes. We noted that the classification performance for all demographics surpasses the baseline by at least 20%, and that topic features are best in predicting how words are used by the various demographic slices. We also showed that the immediate context ($\pm$5 words) is best for extracting topics that differentiate between demographic slices for *location* and *industry*, as word meaning can be directly encoded by a few topics, while for gender, the interplay of topics extracted from a wider context ($\pm$100 words) is the best predictor.

Looking at the topic-based differences for the top words with demographic bias, we examined the dominating topics in each demographic group using topic modeling, both qualitatively and quantitatively. The correlations between the topic distributions suggest that some demographic slices are closer to each other in terms of word usage, while others are further apart. Further, in the case of gender, different groups tend to write about different topics while using a certain word, and so by recognizing the topics surrounding a target word, it is very likely to recognize the gender of the authors. However, in the case of location and industry, the group is best predicted by recognizing the word meaning, assuming that topics from a context of five neighboring words indicate the target word's meaning only, and not the general topic of discussion. While we acknowledge that some of the identified differences may be due to classifier errors, we believe that there do exist differences in the usage for words in general between different demographic groups, as indicated in our qualitative analysis, and further research can aid in uncovering these differences further.

Acknowledgments

This material is based in part upon work supported by the John Templeton Foundation (#48503, #62256). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Boroditsky, Lera. 2001. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1):1–22. https://doi.org/10.1006/cogp.2001.0748, PubMed: 11487292

Boroditsky, Lera, Lauren A. Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. In Dedre Gentner and S. Goldin-Meadow, editors, *Language in Mind: Advances in the Study of Language and Thought*. The MIT Press, Cambridge, Massachusetts, chapter 4, pages 61–79.

Cohen, Dov, Richard E. Nisbett, Brian F. Bowdle, and Norbert Schwarz. 1996. Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and*

Social Psychology, 70(5):945. https://doi.org/10.1037/0022-3514.70.5.945, PubMed: 8656339

Furnham, Adrian and Naznin Alibhai. 1983. Cross-cultural differences in the perception of female body shapes. *Psychological Medicine*, 13(04):829–837. https://doi.org/10.1017/S0033291700051540, PubMed: 6665099

Garimella, Aparna, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 2275–2285.

Garimella, Aparna, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, COLING 2016, pages 674–683.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18. https://doi.org/10.1145/1656274.1656278

Ignatow, Gabe and Rada Mihalcea. 2012. Injustice frames in social media. In *American Sociological Association Annual Meeting*.

Kern, Stephen. 2003. *The Culture of Time and Space, 1880-1918: With a New Preface*. Harvard University Press.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics System Demonstrations (ACL 2014)*, pages 55–60.

McCarty, John A. and L. J. Shrum. 1993. The role of personal values and demographics in predicting television viewing behavior: Implications for theory and application. *Journal of Advertising*, 22(4):77–101. https://doi.org/10.1080/00913367.1993.10673420

Newcomer, Susan and Wendy Baldwin. 1992. Demographics of adolescent sexual behavior, contraception, pregnancy, and STDs. *Journal of School Health*, 62(7):265–270. https://doi.org/10.1111/j.1746-1561.1992.tb01242.x

Paul, Michael and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3 of *EMNLP 2009*, pages 1408–1417.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71:2001. https://www.researchgate.net/publication/239667728_Linguistic_Inquiry_and_Word_Count_LIWC_LIWC2001

Pennebaker, James W., Bernard Rimé, and Virginia E. Blankenship. 1996. Stereotypes of emotional expressiveness of northerners and southerners: A cross-cultural test of Montesquieu's hypotheses. *Journal of Personality and Social Psychology*, 70(2):372–380. https://doi.org/10.1037/0022-3514.70.2.372, PubMed: 8636889

Řehůřek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, LREC 2010, pages 45–50.

de Secondat, Charles and Baron de Montesquieu. 1748. *The Spirit of Laws*. Hayes Barton Press.

Shwartz, Vered. 2022. Good night at 4 pm?! Time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853. https://doi.org/10.18653/v1/2022.findings-acl.224

Shweder, Richard A. 1991. *Thinking Through Cultures: Expeditions in Cultural Psychology*. Harvard University Press.

Shweder, Richard A., Jacqueline J. Goodnow, Giyoo Hatano, Robert A. LeVine, Hazel R. Markus, and Peggy J. Miller. 1998. The cultural psychology of development: One mind, many mentalities. *Handbook of Child Psychology*, pages 865–937.

Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet Affect: An affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC 2004, pages 1083–1086.

Street, Brian. 1993. Culture is a verb: Anthropological aspects of language and cultural process. In D. Graddol, L. Thompson, and M. Byram, editors, *Language and Culture*. Clevedon: Multilingual Matters, pages 23–43.

Vilares, David and Carlos Gómez-Rodríguez. 2018. Grounding the semantics of part-of-day nouns worldwide using Twitter. In *Proceedings of the Second*

Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 123–128. https://doi.org/10.18653/v1/W18-1116

Wilson, Theresa, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. https://doi.org/10.3115/1225733.1225751

Yin, Zhijun, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW 2011, pages 247–256. https://doi.org/10.1145/1963405.1963443