Exploring Patterns of Language Variation on Social Media: A Case Study in Atlanta Online Communities

Hang Jiang,[1] **Brandon Roy,** [1] **Vivek Kulkarni,** [2] **Deb Roy** [1]
[1] Massachusetts Institute of Technology [2] Stanford University
hjian42@mit.edu, viveksck@stanford.edu, dkroy@media.mit.edu

Abstract

Recent research on characterizing language variation on social media has largely focused on detecting language variation across various dimensions like time, geography, domains. Several computational methods to detect language variation have been proposed which have now enabled sociolinguists to study language variation at scale by analyzing language use on the Internet and social media platforms. However, little work exists on investigating the social factors that relate to or lead to observed language variation in social media. In particular, how does language vary across online communities in social networks? How do the social network and its structural properties relate to language variation? How do social and demographic attributes of social network actors relate to observed language variation? Here, we attempt to investigate the role of some of these factors by focusing on a rich online social network of users based in a major US metropolitan area – Atlanta, GA – on Twitter that are segmented into 149 interest-based communities. We automatically identify a set of interest communities for a large population of Twitter users based in Atlanta. Interest communities are clusters of users who share common interests, such as sports, music, politics, etc., and tend to follow similar relevant popular accounts. Using this case study, we first show how to quantify community-specific shifts in lexical semantics using BERT. We then explore how such variation is related to community attributes, such as social identity and interaction. We find that communities with highly distinctive language are small to medium-sized and normally have a high proportion of young users. In addition, we find user activity, network centrality, and the interaction of organization status and social status have a negative effect on lexical innovation. At last, we leverage the language model to identify ideological concepts with significant shifts from the norm, providing a peek into the worldview of different groups. Our work thus yields an improved understanding of how social network structure and socio-demographic factors relate to language variation on social media platforms like Twitter.

Introduction

Language on the Internet is often used in non-standard and innovative ways (Androutsopoulos and Ziegler 2004). Many works in sociolinguistics have shown that social variables (e.g., ethnicity, age, gender, social status) are closely related to linguistic variation in online communities (Herring and Paolillo 2006; Eisenstein 2013; Nguyen et al. 2016). These communities tend to develop their sociolects, or social dialects, which is similar to the concept of regional dialects (Nguyen et al. 2016). Such linguistic variation includes orthographic variants (Stewart et al. 2017) as well as reductive/innovative meaning change (emergence/loss of a full-fledged meaning of a word) (Blank 2012; Schlechtweg, Walde, and Eckmann 2018). Therefore, one user's psychological encoding of a word also changes according to the community-specific context. For example, the connotative meaning of "police" or "school" may vary significantly for a speaker depending on their race, socioeconomic class, and geographical context (e.g, urban vs. rural). In this work, we hope to understand how "dialects" on social media vary from the norm across communities.

We focus our study on the city of Atlanta for several reasons. As background, our lab is involved in a collaborative project on the ground in Atlanta to leverage local listening and understanding of communities to inform public health communications. Thus, studying online communities based in Atlanta complements this ongoing work. Furthermore, although online interest communities need not have any grounding in geographic locale, location is a factor in defining linguistic communities. Atlanta, as a major metropolitan center of the Southeast, contains rich linguistic variation and change (Prichard 2010). This city is not only home to systematic changes of the Southern Vowel Shift, but also many recent immigrants from the North as well as sizeable populations of Hispanic, Jamaican, Korean, and Southeast Asian immigrants living within the metropolitan area (Roth and Ambrose 1996). In the past, researchers (Labov 2006; Prichard 2010) have studied linguistic variation in Atlanta with surveys and interviews, but none to our knowledge have used abundant social media data to characterize the linguistic variation in the city. The central goal of the paper is to leverage social media data and NLP tools to understand the relation of linguistic variation and social factors in a geographically rooted collection of communities.

Recently, many NLP approaches have been proposed to detect lexical variation across time and domains. Some of the latest methods (Hu, Li, and Liang 2019; Kutuzov and Giulianelli 2020) typically adopt dynamic contextual em-

beddings to represent a word based on its context instead of a static vector (Kulkarni et al. 2015; Hamilton, Leskovec, and Jurafsky 2016). In this work, BERT-based methods (Devlin et al. 2018; Nguyen, Vu, and Nguyen 2020) are an ideal choice to detect semantic changes between two Twitter communities since they adjust their linguistic representations to better capture subtle linguistic variation related to identity and context (Schlechtweg, Walde, and Eckmann 2018).

We begin by identifying a population of Twitter users based in Atlanta, infer a set of interest-based communities for this population, and construct the `TwitterATL` dataset. We validate our BERT-based method on a benchmark dataset called `TwitterAAE`, which contains ground truth lexical variation between African American Vernacular English and Standard American English. On `TwitterATL`, we quantify how each community-specific language deviates from the norm and detect words with significant changes using the BERTweet language model (Nguyen, Vu, and Nguyen 2020). After obtaining the semantic shifts, we study the relationship between language variation and community attributes and provide evidence to sociolinguistic theories related to language variation and social identity/interaction. Finally, we analyze ideological terms to probe the worldview of Atlanta communities.

Overall, our paper has three major contributions. First, we present an approach to study linguistic variation in a geographically bounded population "in the wild". Unlike previous studies on Reddit (Del Tredici and Fernández 2018b; Lucy and Bamman 2021), our study is grounded in Atlanta interest communities on Twitter. Second, we provide new empirical methods that might advance our theoretical understanding of language variation. Our work not only confirms previous findings related to community size and user activity generalize to the Twitter data but also explores the relationship between language variation and additional social factors, including social identity factors (e.g., age, gender, organization status) and social interaction factors (e.g., social status, closeness). At last, we detect significant semantic changes of ideological concepts in Atlanta communities and conduct qualitative analysis to understand their worldview.

Related Work

The growth of social media platforms has allowed sociolinguists to conduct large-scale studies in language variation (Nguyen et al. 2016), which is not practical using traditional methods such as surveys and interviews. Many related works have focused on how lexical variation from the norm is related to social and linguistic factors (Nguyen and Rose 2011; Danescu-Niculescu-Mizil et al. 2013; Eisenstein et al. 2014; Stewart and Eisenstein 2017). The community-specific linguistic norms and differences are often distinguished by which words are used. Recently, Zhang et al. (2017) proposed a PMI-based approach to identify the community-specific language use of Reddit communities. Later, Lucy and Bamman (2021) proposed a BERT-based approach to encode semantic variation on 474 Reddit communities.

Lexical semantic change detection is a rapidly growing field in NLP. Gulordava and Baroni (2011) first used a distribution similarity method to detect lexical changes. Eisenstein et al. (2014) applied latent vector models to study lexical shifts on social media. Later, researchers inspired by the skip-gram model (Mikolov et al. 2013) designed variants to detect semantic changes over time and across domains (Kim et al. 2014; Hamilton, Leskovec, and Jurafsky 2016; Kulkarni et al. 2015; Bamler and Mandt 2017; Dubossarsky et al. 2019). After the pre-trained language model BERT (Devlin et al. 2018) was released, Schlechtweg et al. (2020) adapts both BERT and ELMo (Peters et al. 2018) to study diachronic and synchronic lexical changes. Compared to the Word2vec approaches, BERT is easy to adapt to new domains (e.g., Twitter, scientific text) with finetuning (Beltagy, Lo, and Cohan 2019; Nguyen, Vu, and Nguyen 2020). Besides, a single BERT model is sufficient to study linguistic variation across communities whereas Word2vec-based approaches usually require training models on each community and aligning their vector spaces. The tweets from each community are not large enough to effectively train a separate Word2vec model but do not pose any challenges for finetuned BERTweet. At last, BERT's token representation varies depending on the token's context, allowing researchers to study subtle linguistic variation caused by context variation.

However, there are far fewer works that utilize state-of-art techniques to study community-specific language variation. Bamman, Dyer, and Smith (2014) first proposes dialect-aware skip-grams to study regional dialects and Del Tredici and Fernández (2018b) applies the method on Reddit sociolects. To extend the previous works to a larger set of communities, Lucy and Bamman (2021) employs BERT to characterize English variation in 474 Reddit communities, demonstrating promising directions to study community-specific language variation with pre-trained language models. Finally, Yang and Eisenstein (2017) proposes social attention to overcome language variation on Twitter for sentiment analysis, showing that attention-based approaches can be used to address language variation on social networks.

Previous works (Del Tredici and Fernández 2018b; Lucy and Bamman 2021) have analyzed language variation on Reddit communities and have also noted the role of community attributes (e.g., density, size) as factors influencing language variation. However a puzzle remains: Do these and other hypotheses generalize to other social media platforms like Twitter? Twitter is naturally structured as a social network, whereas Reddit is an online forum featuring aggregated content, news, and conversations. Morales, Monti, and Starnini (2021) have found that "echo chambers" are less prevalent on Reddit than Twitter, arguing this is due to structural differences between the platforms. To answer this question, we use a simple BERT-based approach to measure community-specific language variation. We validate our method by comparing with the baseline approach (Lucy and Bamman 2021) on the TwitterAAE benchmark dataset. Finally, we conduct analysis on the Atlanta data to cover social factors in both social identity and interaction. We observe similar patterns between linguistic variation and community attributes (e.g., community size, user activity) on Twitter as previous works found on Reddit (Del Tredici and Fernández 2018b; Lucy and Bamman 2021). In this study, we intro-

Corpus	TwitterAAE	TwitterATL
tweet count	557,128	7,602,853
token count	8,960,308	98,934,304
unique token count	234,691	750,033
# communities	2	149

Table 1: Statistics for TwitterAAE and TwitterATL datasets

duce additional social factors such as social identity (age, gender, organization status) from Twitter profiles and find both gender and organization status play a significant role in community-level linguistic variation.

Dataset

For our case study, we construct the TwitterATL dataset to study language variation across online communities in Atlanta and will explain our steps in the following paragraphs. Before the study, we validate our approach on a benchmark TwitterAAE dataset created by Blodgett, Green, and O'Connor (2016) and compare our approach with the baseline method in characterizing lexical variation between AAVE and SAE. We pick this benchmark dataset for validating our approach since this dataset presents two well-studied English dialects spoken in two communities and contains a wide range of linguistic changes (e.g., semantic, syntactic, and orthographic variation). Both datasets follow the same pre-processing process suggested by Nguyen, Vu, and Nguyen (2020). Table 1 shows the statistics of two datasets.

TwitterATL

The TwitterATL dataset consists of a population of 146,351 Twitter users believed to be located in Atlanta (assembled Aug-Sep 2020) who are grouped into interest clusters, and up to 100 of the recent tweets by each user (collected in mid-February 2021.)

Identifying the population We took two approaches to identifying an Atlanta-based population: first, by querying the followers of a set of "seed accounts" (e.g., "navicenthealth", "civicatlanta") that we expected to have a large Atlanta-based followership, fetching the user objects for those followers, and filtering this set. Second, by scanning the user objects embedded in tweets on the "decahose", Twitter's 10% sample of tweets, and filtering this set. In both cases, the large set of user objects was filtered using a simple rule-based classifier on the "location" field to match against various conventional identifiers for "Atlanta, GA" (e.g., "Atlanta, Georgia", "ATL", "atlanta, ga", etc.)

This yielded 174,911 users we believed to be located in Atlanta. For each of these users, we then queried for the user's "friends" (i.e. the set of users that the target user follows.) On Twitter, who you choose to follow is the primary driver of the user experience, largely determining the tweets you see in your timeline. As such, the users you follow may reflect both your interests and social connections.

Note that both the follow graph and decahose based methods for assembling the Atlanta population have some bias – the follow graph method reflects the audiences of the seed set and thus is best served by a broad and diverse set of seed accounts, while the decahose method only reflects users who tweet (including retweets and replies.) Finally, the population only includes users who indicate their location.

Interest clusters To study the linguistic variation of these Atlanta-based Twitter users, we first identify coherent interest clusters which reflect the myriad interests of this diverse population. Although there are many choices of who to follow on Twitter, users with similar interests may choose to follow the same or similar accounts. We use a decomposition method to uncover this latent structure and compute low dimensional vector embeddings for each user, such that users with similar interests will be near each other in the embedding space.

The identified Atlanta-based population consists of $174,911$ users who collectively follow $23,767,210$ distinct accounts. We first filter by considering only the subset of friends (i.e. the followed accounts) that have at least 1000 followers in this population, and then removing users (i.e. members of the population) who follow fewer than 10 of the remaining friends, yielding $147,049$ users and $5,603$ friends. We represent these follow relationships with a binary $n \times m$ adjacency matrix A_{ij}, with $A_{ij} = 1$ if user i follows user j, 0 otherwise. A is a large, sparse matrix, with $n = 147,049$ rows and $m = 5,603$ columns. We also rescale and regularize the adjacency matrix by the in- and out-degrees of the nodes, following the method in (Rohe, Qin, and Yu 2016) (except we use the adjacency matrix rather than the Laplacian).

With this filtered and rescaled adjacency matrix A, we perform the truncated singular value decomposition (SVD) to obtain matrices U, S, and V^t, truncating to only the top k singular values of S to obtain U, an $n \times k$ matrix, with each row corresponding to a k-dimensional embedding of a user. We use the method of Gavish and Donoho (2014), which yields a truncation value of $k = 583$.

To obtain interest clusters, we next cluster the row vectors of U using the HDBScan Python package (McInnes, Healy, and Astels 2017). Rather than directly apply HDBscan to the $k = 583$ dimensional row vectors (or reduce k further via the truncated SVD if needed), we use UMAP (McInnes, Healy, and Melville 2018) to reduce the embedding dimension following guidelines in the UMAP documentation.[1] Using a 10-dimensional embedding followed by HDBscan clustering, 60.1% of the population are clustered into 149 interest communities. Table 3 shows some examples of the identified communities, summarized by the top accounts followed by community members, and the top most distinctive words from the members' "bio text" fields. Although the interest communities are derived only from the follow graph, the coherence and interpretability of the bio text and top followed accounts supports our claim that we have identified meaningful interest structure for this population.

This method is perhaps more similar to Latent Semantic Analysis of document-term matrices (Deerwester et al. 1990) than spectral methods for network community detection, since the rows and columns effectively correspond to

[1] http://umap-learn.readthedocs.io/en/latest/clustering.html

different sets of items. Community detection methods, such as the Louvain algorithm (Blondel et al. 2008), typically focus on the link structure within a single set of nodes and may be better suited to revealing social rather than interest-based communities. In future work, we plan to investigate the linguistic features of such social communities.

Tweet history Finally, using the Twitter API we assembled a dataset of roughly 12.5 million tweets (including replies and retweets) from the filtered, clustered Atlanta user population. This dataset consists of up to 100 recent tweets from each user in the population, which we fetched in mid February 2021.

Methods

To compute significant semantic shifts per community, we follow three steps: (1) generate token-level embeddings; (2) calculate semantic changes; (3) detect significant changes.

Generate Token-level Embeddings

Both datasets are from the Twitter domain. Therefore, we use the pre-trained BERTweet model (Nguyen, Vu, and Nguyen 2020) to extract linguistic features. This model is pre-trained on 845M tweets and has a better capability of representing tweets than the original BERT-base model.

In our study, we focus on two sets of glossaries: `DialectAAE` and `IdeologyLex`. `DialectAAE` has 80 AAVE-specific terms and is based on related works (Pullum 1999; Hinton and Pollock 2000). Though not comprehensive, `DialectAAE` covers some reliable dialect variation in syntax, semantics, and orthography in AAVE. This word list enables us to validate and compare our method with other approaches on the TwitterAAE dataset. `IdeologyLex` is a list of ideological concepts and beliefs provided by Li, Schloss, and Follmer (2017)[2]. This list allows us to study how people from different communities in Atlanta understand the same ideological concepts differently.

Apart from the two glossaries, we also compute embeddings for 1000 words randomly sampled from the most frequent 5000 words from each dataset. These words are used to construct a semantic change distribution for each community, which is later used to detect significant changes.

Frequency can be a confounding factor for characterizing language variation. In order to reduce the effect of frequency, we sample up to 100 usage cases for each word per community and make sure to include the words used at least 5 times in the community. Following the previous work (Kutuzov and Giulianelli 2020), we extract token embeddings from the last layer of BERT on sampled occurrences.

Calculate Semantic Shifts

For each target word, we calculate its average token embedding on one community by averaging its token-level embeddings across all the usages. For TwitterAAE, we simply compare AAVE and SAE communities by computing the shifts from one to the other. On the contrary, there are 149 communities for TwitterATL and pairwise comparisons would be too computationally expensive. Therefore, to measure how each community-specific dialect shifts from the norm, we randomly sampled K tweets from all the tweets to create the norm community, where K is the mean number of tweets across communities. Afterward, we compared all the communities against the norm community. The semantic shift for each term is computed with the cosine distance between two average token embeddings (Giulianelli, Del Tredici, and Fernández 2020):

$$\mathcal{D}(U_w^{c_1}, V_w^{c_2}) = 1 - \cos\left(\frac{\sum_{u_{w_i} \in U_{w_i}^{c_1}} u_{w_i}}{n_w^{c_1}}, \frac{\sum_{v_{w_j} \in V_{w_j}^{c_2}} v_{w_j}}{n_w^{c_2}}\right)$$

where $U_w^{c_1}$ indicates all usage embeddings for a term on the corpus c_1 and n_w^{c1} indicates the number of occurrences for the term w in the corpus c_1. This score indicates the degree of semantic changes undergone by a word between corpus c_1 and c_2. As previous literature (Schlechtweg, Walde, and Eckmann 2018) suggests, this score does not distinguish a gain or loss of word meaning.

Robustness Check Using the same method for sampling the control group norm community, we sampled 10 additional norm communities. Comparing the 149 interest communities with each of these norm communities and repeating the analysis yielded similar patterns of results.

Unsupervised Significant Change Detection

To detect significant changes, we construct a distribution of semantic changes with a set of top frequent terms. Specifically, we sample 1000 words from the most frequent 5000 words, which do not overlap with the two glossaries presented above. Based on this null distribution, we can estimate the probability of any target word's semantic change on a community corpus and determine if the change is significant or not through a right-tailed p-value test.

Evaluation on TwitterAAE

We compare our approach, BERTweet + Avg Token Embeddings, against the baseline method proposed by Lucy and Bamman (2021) for an unsupervised lexical change detection task[3]. To construct the test set, we combine 80 positive AAVE words from `DialectAAE` with 80 sampled negative words. We compute the semantic changes on the test set and automatically detect words with significant changes at a significance level. Our models are evaluated with precision, recall, and F1 shown in Table 2. Based on the result, we decide to use our approach for the case study.

To adapt Lucy and Bamman (2021)'s model for the task, we cluster on the last four layers' concatenation of BERT and run k-means clustering algorithm. We use the criterion $k = argmin_k RSS(k) + \gamma k$ for cluster cardinality (Schütze, Manning, and Raghavan 2008) to determine the number of clusters for each word. To calculate the semantic shift of a word, we count the proportion of usages from one corpus for each cluster and use Jensen–Shannon divergence to calculate the distance between two communities.

[2] After removing phrases, there are 133 ideological terms.

[3] Only TwitterAAE is used for the task since TwitterATL does not have a target lexicon ground truth like DialectAAE.

Models	Precision	Recall	F1
Lucy and Bamman (2021)	0.6000	0.0750	0.1333
BERTweet + Avg Token Embeddings	0.6923	0.1125	0.1935

Table 2: Unsupervised change detection on TwitterAAE. Significance threshold is 0.05. Our approach consistently does better than the baseline in precision, recall, and F1.

Language Variation in the Atlanta Communities on Twitter

Online Communities in Atlanta

We select one Atlanta community from each category and show them in Table 3. Some communities are Atlanta-specific such as Emory Hospital, Gwinnett County, Historically Black Colleges and Universities (HBCU), Atlanta Falcons. The other communities also share Atlanta characteristics. For instance, the top followed accounts of the travel community contain "ajc" (Atlanta Journal Constitution) and "Delta" (Delta Airlines, based in Atlanta).

Category-level Variation We create ten categories shown in Figure 1 and annotate each community with a category by observing each community's topic words and top follow accounts. By averaging the semantic changes for each category, we calculate the average change on `IdeologyLex` and 1000 sampled frequent words. To show which category of communities use `IdeologyLex` in the most distinctive way, we compute the change of the ideological terms relative to topmost frequent words with $\delta_{1000}/\delta_{ideo}$, where δ_{ideo} and δ_{1000} represent the average change on `IdeologyLex` and 1000 sampled frequent words. We rank the categories the resulting values in Figure 1. This plot shows that communities belonging to the top four categories (religion, health, education, politics) tend to use ideological words in the most distinguished way from the 1000 words. For instance, the science community (education category) shows a distinctive use or understanding of ideological words. In the DJ/EDM and pop culture communities (entertainment community), there is a smaller gap between the two set of words.

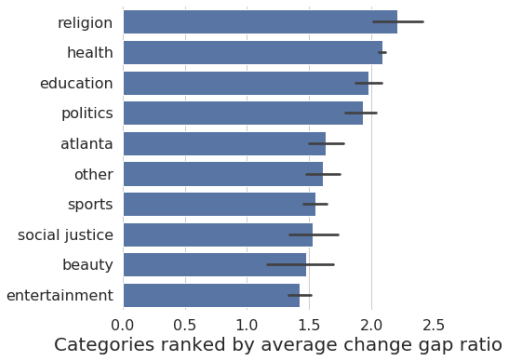


Figure 1: Categorical-level semantic change for both IdeologyLex and 1000 sampled frequent words.

Community Attributes and Language Variation

Both social interactions and social identity are key factors that shape language (Nguyen et al. 2016). In sociolinguistics, a major research direction is to find the relation between **social identity** (e.g., gender, age, ethnicity, status) and language use (Eckert 1997; Holmes and Meyerhoff 2008; Wagner 2012; Eckert and McConnell-Ginet 2013). Each speaker draws linguistic elements from their identity repertoire (Snow and Anderson 1987). In our work, we identify gender, age, and organization status information with M3[4], a demographic inference tool (Wang et al. 2019), to infer gender, age, and organization-status. This tool relies on the user name, screen name, description, profile image, and language to infer these three factors. We choose this tool because M3 has state-of-the-art performance on gender (F1 91.8%), age (F1 52.2%), and is-org (F1 89.8%) classification tasks while also reducing algorithmic bias (Wang et al. 2019). Wang et al. (2019) divides age into four groups (<=18, 19-29, 30-39, >=40), gender into two groups (male and female), and organization status into two groups (personal and organizational account). We predict the user's demographic information by picking the category with the max probability. Based on the demographic predictions, our data have 15.1% organizational accounts and 84.9% personal accounts. For gender, there are 42.4% female and 57.6% male. For each of the age groups, their proportions in the population are 12.0% (<=18), 32.2% (19-29), 25.0% (30-39), and 30.8% (>=40).

Apart from social identity, **social interaction** is closely tied to the norms of communities (Nguyen and Rose 2011; Danescu-Niculescu-Mizil et al. 2013; Lucy and Bamman 2021). Community members use these norms such as slang words and domain-specific jargon to signal their affiliations (Nguyen and Rose 2011). Language change is also observed on the community level since communities also keep evolving as members participate (Danescu-Niculescu-Mizil et al. 2013). In this work, we compute the following community-level behaviors to investigate: community size, social status, user tweet activity, user favorite activity, network centrality. The details are discussed in the corresponding section.

In the following paragraphs, we will first discuss the identity-based factors (age, gender, organization status) and then the community-level factors (community size, social status, user tweet activity, user favorite activity, network centrality). Specifically, we study how these attributes can be tied to unique language use, as shown in Figure 2. We use the lexical variation between each community and the norm community on the 1000 words to measure the level of lexical deviation from the norm.

Age. We encode four age groups <=18, 19-29, 30-39, >=40 into 0-3 and use the average age of community members to represent the community-level age. Figure 2 show that age is a strong indicator of language variation, confirming previous findings that **most linguistic changes happen in young ages** (Wagner 2012; Nguyen et al. 2013).

Gender Diversity. We measure the *gender diversity* with the female ratio in each community. Gender diversity does

[4]https://github.com/euagendas/m3inference

Community Name	Comm. size	Top 10 (stemmed) Biography Words (1st row) / Followed Accounts (2nd row)
Emory Hospital (Health)	514	emori,health,@emorymedicin,research,medicin,@emoryunivers,fellow,care,diseas,professor,medic,md,resid,@winshipatemori,physician EmoryMedicine,emoryhealthcare,EmoryUniversity,emoryhealthsci,CarlosdelRio7,GradyHealth,CDCgov,ajc,EmoryDeptofMed,BarackObama
Gwinnett County (Atlanta)	355	gwinnett,mom,consult,realtor,duluth,clean,metro,pilat,profession,develop,help,atlanta,web,famili,counti GwinnettDaily,ajc,GwinnettNewsNow,wsbtv,GwinnettSchools,GwinnettChamber,GwinnettMag,GDPsports,GwinnettEvents,FOX5Atlanta
HBCU (Education)	417	morehous,colleg,spelman,student,educ,cau,@morehous,clark,communiti,univers,hbcu,offici,career,@cau,alumna Morehouse,SpelmanCollege,ajc,CAU,BarackObama,CityofAtlanta,HowardU,HBCUBuzz,wsbtv,CNN
Travel (Other)	511	travel,hotel,airport,trip,luxuri,experi,pilot,wanderlust,cruis,adventur,servic,compani,bed,airlin,food ajc,TravelLeisure,travelchannel,NatGeoTravel,Delta,CNTraveler,lonelyplanet,SouthwestAir,TravelMagazine,wsbtv
Democrat (Politics)	1883	#resist,polit,trump,liber,democrat,resist,vote,justic,#bidenharris2020,#fbr,#blm,retir,#theresist,wife,lover BarackObama,JoeBiden,KamalaHarris,maddow,ProjectLincoln,HillaryClinton,SpeakerPelosi,gtconway3d,AOC,JoyAnnReid
Christian religious (Religion)	1353	jesus,christ,husband,church,pastor,follow,father,love,christian,wife,god,grace,ministri,author,discipl AndyStanley,louiegiglio,JohnPiper,CSLewisDaily,ajc,lecrae,RickWarren,BethMooreLPM,plattdavid,christomlin
LGBT (Social Justice)	671	gay,lgbtq,lgbt,hiv,atlanta,communiti,activist,lesbian,queer,transgend,support,tran,aid,pride atlantapride,ProjectQAtlanta,theGAVoice,HRC,GAEquality,NOH8Campaign,glaad,BarackObama,TheEllenShow,HRCATL
Atlanta Falcons (Sports)	292	falcon,#riseup,sport,@atlantafalcon,fan,#falcon,atlanta,#brave,#truetoatlanta,nfl,brave,#hawk,#atlantafalcon,justwaldrop,#inbrotherhood AtlantaFalcons,M_Ryan02,debo,juliojones_11,devontafreeman,GradyJarrett,Keanu_Neal,VicBeasley3,TheFalcoholic,roddywhiteTV

Table 3: Examples of Atlanta online communities. Each has its category in parentheses. On the right side, the 1st row contains the top 10 biography words and the 2nd row contains the top 10 followed accounts among members of a community.

not have a clear correlation with distinctive language use. This is not surprising since previous studies (Herring and Paolillo 2006; Gianfortoni, Adamson, and Rose 2011) have suggested that **the relation between gender and linguistic features is not significant when controlling for some factors (genre and occupation)** and each interest-based community has a relatively different topic.

Organization Status. We represent *organization status* with the proportion of organizational accounts in each community. We observe that **communities with a larger proportion of organizational accounts tend to use a more standard language**. As previous works suggest (McCorriston, Jurgens, and Ruths 2015; Alzahrani et al. 2018), personal and organizational accounts have significant differences in terms of intentions and linguistic styles.

Community Size. We observe a similar pattern (Del Tredici and Fernández 2018a; Lucy and Bamman 2021) that **"small to medium" size communities tend to produce more lexical innovations than large communities**. We define *community size* as the number of unique users in a Twitter community. When communities become as large as the Atlanta food community, the social interactions tend to be simpler with shorter tweets and users do not have enough opportunities to form their group-specific language (Jones, Ravid, and Rafaeli 2004).

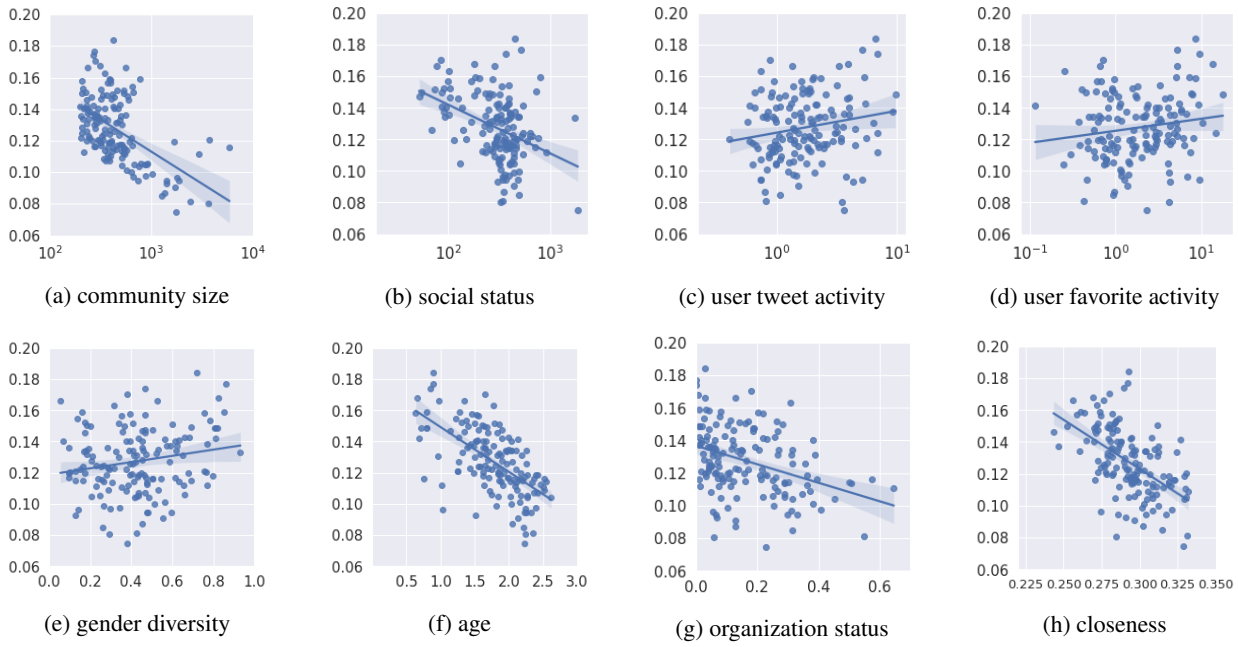
Social Status. Communities having more high-status users tend to use a language closer to the norm. According to Bakshy et al. (2011); Danescu-Niculescu-Mizil, Gamon, and Dumais (2011), we measure social status by the median number of friends per user in a Twitter community[5]. Communities with less distinctive language use tend to have more users with high social status on Twitter, which supports Labov's theory of social stratification in which people in a higher social class tend to use a more standard language (Labov 2006; Mallinson 2007). For instance, both entrepreneurship and artist Atlanta Twitter communities have many friends but lower lexical innovation.

[5]We try the mean and median of followers per user and its effect on linguistic change is not as significant as that of friends.

User Tweet/Favorite Activity. Regular social interaction is also necessary for a community of practice and can be vital to lexical innovation (Holmes and Meyerhoff 1999; Blackmore 2010). In general, to write a post is a higher level of involvement in social media than to like a post since it requires more input from the user. Therefore, we measure user activity from two perspectives: *user tweet activity* and *user favorite activity*. User tweet activity is the average number of tweets per user in the community, whereas user favorite activity is measured by the average number of favorites given out per user in a community. Interestingly, *user tweet activity* has a positive effect on linguistic variation, whereas *user favorite activity* has a negative effect (Table 4). This result indicates that **a high proportion of content creators in a community tend to produce lexical innovations, whereas a high proportion of active content endorsers do not**.

Network Centrality. We represent *network centrality* of a community with the average closeness score per user. This closeness score of a user measures its average inverse distance to all other users (Cohen et al. 2014). Users with a high closeness score have the shortest distances to others and is therefore a way of detecting users who can spread information very efficiently through a social network. According to Guinote and Vescio (2010), closeness can serve as a proxy for relative power (e.g. positive regard, trust, and commitment) specific to communities. Therefore, it is expected that **users who are close to everyone in the Atlanta communities to have a closer language to the norm**. As a result, communities with more high-closeness users show a less distinctive use of language relative to the norm community. We also tested betweenness centrality and found that closeness has a more significant effect in relation to linguistic change. Since two factors are dependent on each other, we pick closeness in our study.

Modeling Variation Based on the analysis above, we build an ordinary least squares (OLS) regression model (Montgomery and Elizabeth 1991) and a linear mixed (LM) regression model to study the effect of social factors on linguistic variation in Atlanta communities in Table 4 and 5.



(a) community size (b) social status (c) user tweet activity (d) user favorite activity

(e) gender diversity (f) age (g) organization status (h) closeness

Figure 2: Scatter plots between language variation (y-axis) and community-level social factors (x-axis).

OLS Model We observe that community size, user favorite activity, and age have the most significant effects on the variation in all the models ($p < 0.001$). Among these three factors, age has the biggest impact, followed by community size and user favorite activity. Besides, user tweet activity and network centrality show a lower level of significance ($p < 0.01$). Organization status is the least significant ($p < 0.05$). Two features (gender diversity and social status) are not significant in relation to linguistic variation in our study. Similar to previous findings (Zhang et al. 2018; Lucy and Bamman 2021), we show that the category has no additional significant impact on distinctive language use beyond the user-based attributes. This suggest that *who* is involved in a community matters more than *what* they are discussing in a community (Lucy and Bamman 2021).

Linear Mixed Model Based on the OLS analysis, we develop the mixed-effects model (Lindstrom and Bates 1988): `lexical_change ~ community_size + age + network_centrality + user_fav_activity + (organization_status*social_status) + (1|category)` on selected factors. Specifically, we allow organization status and social status to interact with each other because personal and organizational Twitter accounts with many friends (high social status) are different users and should be distinguished from each other. Shown in Table 5, we demonstrate that **interaction between organization status and social status is significant** even though the social status alone is not significant and organization status is also not the most significant factor ($p < 0.01$).

Variable	Estimate	Std
(intercept)	$1.012e{-}15$	0.048
community size	-0.4529^{***}	0.050
social status	-0.0237	0.063
user tweet activity	0.2983^{**}	0.091
user favorite activity	-0.3721^{***}	0.097
gender diversity	0.0486	0.054
age	-0.7046^{***}	0.100
organization status	0.1871^{*}	0.073
network centrality	-0.2339^{**}	0.075
category	0.0416	0.061
N	149	
R^2	0.685	
Adjusted R^2	0.665	

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 4: OLS regression results for the effect of community attributes on the fraction of words used in each community.

Semantic Shifts in Ideological Terms

Following the significant change detection method, we identify ideological words (Li, Schloss, and Follmer 2017) with significant changes for each Atlanta community. We select some words and their usages in Table 6. The first 4 words ("defense", "class", "race", "trust") are used to refer to different senses between the selected community and the norm community. For instance, the word "defense" is primarily used as "the action or role of defending one's goal against the opposition" in the Atlanta Falcon community, but is used for defending a stance in the norm community. Similarly, "race" refers to a grouping of humankind in the science community, and is used to mean "a competition between individuals or groups" in the norm community. On the con-

Variable	Estimate	Std
(intercept)	$-0.003***$	0.048
community size	$-0.442***$	0.042
user favorite activity	$-0.170***$	0.011
age	$-0.769***$	0.079
network centrality	$-0.178***$	0.047
organization status	$0.166**$	0.061
social status	0.012	0.012
organization status:social status	$-0.217***$	0.020
N	149	
R^2	0.668	

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 5: LM regression results for the effect of community attributes on the fraction of words used in each community.

trary, the lexical meanings do not change for the last three words ("president" and "media"). However, their connotation changes significantly with contexts. The word "president" is deeply bounded with "the elected head of a republic" in the democrat community but is used more flexibly to mean "the head of an organization or company" in the norm community. In the eyes of Trump community members, "media" is portrayed as misleading and fake, whereas the norm community members see it in a neutral view. These examples show that our approach is capable of detecting not only lexical changes in denotation but also subtle social meaning shifts in connotation, revealing the worldview of different communities.

PCA Visualization To further investigate whether BERT embeddings can distinguish the subtle linguistic meaning variation, we extract the usage representations of selected ideological words from the BERT's last layer, compress the vectors with PCA, and visualize them in Fig. 3. The two lexical meanings of "rule" are captured in two clusters by the BERTweet model in the first subplot. As we mentioned earlier, the Atlanta Falcon community predominantly uses this word to refer to "what is allowed or not allowed on and off the court", but the Emory hospital community uses it in a more flexible manner to indicate "regulations or principles governing conduct within a particular activity or sphere". The "president" has a stable literal meaning in both HBCU (Historically Black College or University) and Trump communities, but its connotation is different. In HBCU, the word is mainly used to describe both the head of their schools or other organizations, whereas the Trump community primarily uses it to refer to the president of the United States. Overall, our PCA visualization shows that BERT embeddings are sensitive to both lexical and semantic variation, allowing us to identify community-specific language use.

Conclusion

In this paper, we study linguistic variation in a geographically grounded population. We identify 149 interest communities of Twitter users in Atlanta and apply a BERT-based approach to quantify community-specific semantic variation. After validating our approach on a benchmark dataset, we explore how such variation is related to social identity and interaction. Our results support previous observations on Reddit (Lucy and Bamman 2021) that "small to medium-sized" communities tend to use language in a highly distinctive manner. Besides, we find that communities with lexical innovations normally have a high proportion of young users, content creators, or users who are socially distant from others. At last, we identify ideological words with significant semantic changes from the norm (e.g., "defense", "president", and "media") and probe their ideological attitude through qualitative analysis. Overall, our work provides both evidence to socio-linguistic theories and an improved understanding of how network structure and sociodemographic factors relate to linguistic variation on social media platforms like Twitter.

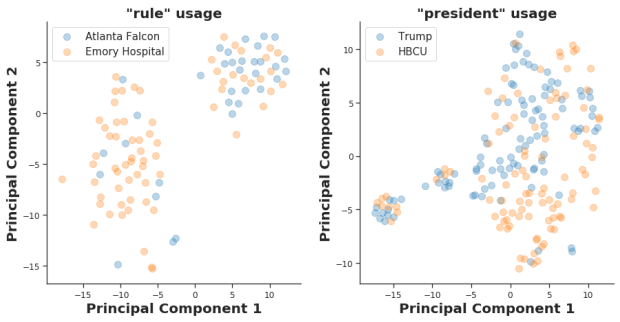


Figure 3: PCA visualization for BERTweet representations of selected words between pairs of Atlanta communities.

References

Alzahrani, S.; Gore, C.; Salehi, A.; and Davulcu, H. 2018. Finding organizational accounts based on structural and behavioral factors on twitter. In *SocialCom*, 164–175.

Androutsopoulos, J.; and Ziegler, E. 2004. Exploring language variation on the Internet: Regional speech in a chat community. In *ICLaVE*, volume 2, 99–111.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, 65–74.

Bamler, R.; and Mandt, S. 2017. Dynamic word embeddings. *ICML* .

Bamman, D.; Dyer, C.; and Smith, N. A. 2014. Distributed representations of geographically situated language. In *ACL*, 828–834.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *EMNLP* .

Blackmore, C. 2010. *Social learning systems and communities of practice*. Springer.

Blank, A. 2012. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*, volume 285. Walter de Gruyter.

Word	Community Name	Examples
defense	Atlanta Falcons	matt ryan can slice most **defenses** when he has better play calling on his side and almost every drive ends up in the red zone
	Norm	but in my **defense** you see me twice a year only at night time and only at a bar
class	Car	the e **class** initially stood for einspritzmotor german for fuel injection engine
	Norm	one of my teachers from animation mentor pete paquette was at blue sky when i was in his **class**
race	Science	shes talking about cultural differences not physical **race** differences
	Norm	i am from athens georgia for the **race** . we have participants from as far away as upstate new york
trust	Finance	if you have a **trust** did you actually fund it is your plan ready for the new secure act
	Norm	if you **trust** them i trust them
president	Democrat	when you took over **president** obama had done the hard work
	Norm	the **president** of stonehenge consulting group joined us today to share key success strategies
media	Trump	**media** to be so flagrant with their lies and misdirection
	Norm	open up a lucid **media** studio where ill be taking pictures and also booking gigs and helping other creators

Table 6: Ideological terms with examples used for different meanings in the selected and norm communities.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic dialectal variation in social media: A case study of African-American English. *ACL* .

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *JSTAT* 2008(10): P10008.

Cohen, E.; Delling, D.; Pajor, T.; and Werneck, R. F. 2014. Computing classic closeness centrality, at scale. In *ACM WOSN*, 37–50.

Danescu-Niculescu-Mizil, C.; Gamon, M.; and Dumais, S. 2011. Mark my words! Linguistic style accommodation in social media. In *WWW*, 745–754.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*, 307–318.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6): 391–407.

Del Tredici, M.; and Fernández, R. 2018a. The road to success: Assessing the fate of linguistic innovations in online communities. *COLING* .

Del Tredici, M.; and Fernández, R. 2018b. Semantic variation in online communities of practice. *IWCS* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL* .

Dubossarsky, H.; Hengchen, S.; Tahmasebi, N.; and Schlechtweg, D. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. *ACL* .

Eckert, P. 1997. Age as a sociolinguistic variable. InF. Coulmas (Ed.), Handbook of sociolinguistics (pp. 151–167).

Eckert, P.; and McConnell-Ginet, S. 2013. *Language and gender*. Cambridge University Press.

Eisenstein, J. 2013. What to do about bad language on the internet. In *NAACL*, 359–369.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2014. Diffusion of lexical change in social media. *PloS one* 9(11): e113114.

Gavish, M.; and Donoho, D. L. 2014. The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory* 60(8): 5040–5053.

Gianfortoni, P.; Adamson, D.; and Rose, C. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *DIALECTS*, 49–59.

Giulianelli, M.; Del Tredici, M.; and Fernández, R. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. *ACL* .

Guinote, A. E.; and Vescio, T. K. 2010. *The social psychology of power.* Guilford Press.

Gulordava, K.; and Baroni, M. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *GEMS SIGLEX*, 67–71.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *ACL* .

Herring, S. C.; and Paolillo, J. C. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10(4): 439–459.

Hinton, L. N.; and Pollock, K. E. 2000. Regional variations in the phonological characteristics of African American Vernacular English. *World Englishes* 19(1): 59–71.

Holmes, J.; and Meyerhoff, M. 1999. The community of practice: Theories and methodologies in language and gender research. *Language in society* 28(2): 173–183.

Holmes, J.; and Meyerhoff, M. 2008. *The handbook of language and gender*, volume 25. John Wiley & Sons.

Hu, R.; Li, S.; and Liang, S. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *ACL*, 3899–3908.

Jones, Q.; Ravid, G.; and Rafaeli, S. 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research* 15(2): 194–210.

Kim, Y.; Chiu, Y.-I.; Hanaki, K.; Hegde, D.; and Petrov, S. 2014. Temporal analysis of language through neural language models. *ACL* .

Kulkarni, V.; Al-Rfou, R.; Perozzi, B.; and Skiena, S. 2015. Statistically significant detection of linguistic change. In *WWW*, 625–635.

Kutuzov, A.; and Giulianelli, M. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. *SemEval* .

Labov, W. 2006. *The social stratification of English in New York city*. Cambridge University Press.

Li, P.; Schloss, B.; and Follmer, D. J. 2017. Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior research methods* 49(5): 1668–1685.

Lindstrom, M. J.; and Bates, D. M. 1988. Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 83(404): 1014–1022.

Lucy, L.; and Bamman, D. 2021. Characterizing English Variation across Social Media Communities with BERT. *TACL* .

Mallinson, C. 2007. Social class, social status and stratification: Revisiting familiar concepts in sociolinguistics. *University of Penn Working Papers in Linguistics* 13(2): 12.

McCorriston, J.; Jurgens, D.; and Ruths, D. 2015. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. In *ICWSM*, volume 9.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2(11).

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software* .

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 3111–3119.

Montgomery, D. C.; and Elizabeth, A. 1991. PECK Introduction to linear regression analysis.

Morales, G. D. F.; Monti, C.; and Starnini, M. 2021. No echo in the chambers of political interactions on Reddit. *Scientific Reports* 11(1): 1–12.

Nguyen, D.; Doğruöz, A. S.; Rosé, C. P.; and de Jong, F. 2016. Computational sociolinguistics: A survey. *Computational linguistics* 42(3): 537–593.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. " How Old Do You Think I Am?" A Study of Language and Age in Twitter. In *ICWSM*, volume 7.

Nguyen, D.; and Rose, C. 2011. Language use as a reflection of socialization in online communities. In *LSM*, 76–85.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. *EMNLP: System Demonstrations* .

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *NAACL* .

Prichard, H. 2010. Linguistic variation and change in Atlanta, Georgia. *University of Pennsylvania Working Papers in Linguistics* 16(2): 17.

Pullum, G. K. 1999. African American Vernacular English is not standard English with mistakes. *The workings of language: From prescriptions to perspectives* 59–66.

Rohe, K.; Qin, T.; and Yu, B. 2016. Co-clustering directed graphs to discover asymmetries and directional communities. *PNAS* 113(45): 12679–12684.

Roth, D. R.; and Ambrose, A. 1996. *Metropolitan Frontiers: A Short History of Atlanta*. Longstreet Press.

Schlechtweg, D.; McGillivray, B.; Hengchen, S.; Dubossarsky, H.; and Tahmasebi, N. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *SemEval@COLING2020* .

Schlechtweg, D.; Walde, S. S. i.; and Eckmann, S. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. *NAACL* .

Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Snow, D. A.; and Anderson, L. 1987. Identity work among the homeless: The verbal construction and avowal of personal identities. *American journal of sociology* 92(6): 1336–1371.

Stewart, I.; Chancellor, S.; De Choudhury, M.; and Eisenstein, J. 2017. # Anorexia,# anarexia,# anarexyia: Characterizing online community practices with orthographic variation. In *Big Data*, 4353–4361. IEEE.

Stewart, I.; and Eisenstein, J. 2017. Making" fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. *EMNLP* .

Wagner, S. E. 2012. Age grading in sociolinguistic theory. *Language and Linguistics Compass* 6(6): 371–382.

Wang, Z.; Hale, S.; Adelani, D. I.; Grabowicz, P.; Hartman, T.; Flöck, F.; and Jurgens, D. 2019. Demographic inference and representative population estimates from multilingual social media data. In *WWW*, 2056–2067. ACM.

Yang, Y.; and Eisenstein, J. 2017. Overcoming language variation in sentiment analysis with social attention. *TACL* 5: 295–307.

Zhang, J.; Danescu-Niculescu-Mizil, C.; Sauper, C.; and Taylor, S. J. 2018. Characterizing online public discussions through patterns of participant interactions. *PACM HCI* 2(CSCW): 1–27.

Zhang, J.; Hamilton, W.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Community identity and user engagement in a multi-community landscape. In *ICWSM*, volume 11.