Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6287–6291
Marseille, 11–16 May 2020

Small Town or Metropolis? Analyzing the Relationship between Population Size and Language

Amy Rechkemmer♣, Steven R. Wilson♢, Rada Mihalcea♠
♣Purdue University ♢University of Edinburgh, ♠University of Michigan
arechke@purdue.edu, steven.wilson@ed.ac.uk, mihalcea@umich.edu

Abstract

The variance in language used by different cultures has been a topic of study for researchers in linguistics and psychology, but often times, language is compared across multiple countries in order to show a difference in culture. As a geographically large country that is diverse in population in terms of the background and experiences of its citizens, the U.S. also contains cultural differences within its own borders. Using a set of over 2 million posts from distinct Twitter users around the country dating back as far as 2014, we ask the following question: is there a difference in how Americans express themselves online depending on whether they reside in an urban or rural area? We categorize Twitter users as either urban or rural and identify ideas and language that are more commonly expressed in tweets written by one population over the other. We take this further by analyzing how the language from specific cities of the U.S. compares to the language of other cities and by training predictive models to predict whether a user is from an urban or rural area. We publicly release the tweet and user IDs that can be used to reconstruct the dataset for future studies in this direction.

Keywords: population, cities, culture, social media

1. Introduction

In recent years, online profiles and posts on social media platforms have provided a wealth of information for researchers to analyze, leading to discoveries about the relationships between the words that people write online and their emotions (Suttles and Ide, 2013), demographics (Rao et al., 2010), values (Boyd et al., 2015), geographic locations (Han et al., 2014), and more. Through these linguistic associations, we can not only predict certain characteristics about the authors of these messages, but also get a glimpse of how people of different demographics interact with their world and process information.

Prior studies have analyzed how location affects the type of language that people use, often looking at text written by authors from different countries when exploring cross-cultural differences (Poblete et al., 2011; Garcia-Gavilanes et al., 2013). However, it is not always necessary to look at multiple countries in order to view different cultures. Recent electoral results in the United States exemplify a divide in the political opinions between those living in densely populated areas and those living in rural parts of the country (Scala and Johnson, 2017). This urban-rural ideological divide is intensifying, yet the people living in these areas depend on one another more and more (Lichter and Ziliak, 2017), interacting at the geographical borders, but also online through platforms like Twitter. Can we leverage text analysis tools to investigate how people in these seemingly disparate areas are expressing themselves through the words that they write to each other and about themselves online?

In this study, we analyzed Twitter data collected from users that live in urban and rural areas of the United States and look for differences in what categories of words are more likely to be used in one area over the other. Our principal contributions are (1) a new dataset of twitter users from urban and rural areas as categorized by the U.S. Census Bureau along with tweets written by these users, (2) a linguistic analysis of this data, and (3) the proposal of city population size as an additional user-level variable to consider when performing social media studies or modeling user behavior online.

2. Data

We began with a set of tweets collected using the public Twitter API[1], sampled from the public Twitter stream over the course of five years from March 2014 to March 2019, with a sample of new tweets collected every four hours during this period. We added tweets to our dataset if they had been written in English, and if the user had provided a location of their account that matches a valid U.S. city as verified through the US census API.[2] We did not use geotagged locations from the tweets directly, instead relying on the locations provided explicitly in users' Twitter bios, meaning that mispelled or unconventional location names were discarded. In this way, these tweets are more likely to match the location of the user's place of residence rather than just a location they are currently visiting. This is important for our analysis, as we are interested in analyzing the content created by users who are actually *from* various cities, not tweets that happened to have been tweeted by a person visiting those cities. Using the census API, we also collected the populations of each valid location as of the most recent U.S. census in 2010. We consider a tweet to be from a rural area if the population of its author's location

[1] https://developer.twitter.com/content/developer-twitter/en.html
[2] https://www.census.gov/developers/

was less than 50,000 people.[3] Otherwise, we consider the tweet to be from an urban setting. This cutoff value comes directly from the U.S. government's definitions of urban and rural populations (Ratcliffe et al., 2016).

We filtered the set of tweets so that only one tweet per user appeared in our dataset to prevent having certain accounts with many tweets bias the analyses. For the dominance score analysis, tweets and user descriptions were preprocessed by expanding contractions, removing symbols, and tokenizing text. Stopwords were kept in on account of some of the lexical resources that include these words in their word classes. City-level analysis and classification of users included the same preprocessing in addition to stopword removal. In total, the dataset contains over 2.6 million unique users, each with a tweet and self-written description for their account.[4]

3. Methodology

We use a set of six lexical resources containing meaningful word classes in order to identify trends in the language produced by users in areas classified as urban and rural. The frequency that a person uses words from each of these classes allows us to calculate a score that indicates the prevalence of a word class in one group compared to the other.

3.1. Word Classes

The six lexicons we consider are the Linguistic Inquiry and Word Count, General Inquirer, Roget, Morality, Values, and Opinion Finder.

LIWC The Linguistic Inquiry and Word Count dictionary was constructed in order to identify groups of words related to important psychological phenomena that may be present in writing samples. We utilize the 2015 edition of the LIWC dictionary (Pennebaker et al., 2015), which contains over 80 categories comprised of a total of nearly 6,400 words, word stems, and emoticons.

General Inquirer (GI) As a resource for automated content analysis, this dictionary (Stone et al., 1966) provides categories of words related to emotional and cognitive states, as well as semantic dimensions and words from the categories outline in the Laswell Dictionary (Namenwirth and Weber, 2016).

Roget's Thesaurus (R) This widely popular lexical resource (Roget, 1883) has been used as a resource for natural language processing (Jarmasz, 2012) due to its hierarchical nature and ability to categorize words into broad classes. We use the Open Roget resource (Kennedy and Szpakowicz, 2014) as another source of categories of words to measure.

Morality (M) Based on Moral Foundations Theory (Graham et al., 2013), this dictionary allows for the measurement of words that either align with or oppose five main moral foundations, which include care, fairness, in-group loyalty, sanctity, and authority (Garten et al., 2016).

Values (V) The hierarchical lexicon for personal values (Wilson et al., 2018) was created by sorting and expanding a set of seed terms collected from surveys of the values of people from around the world. The words in these classes describe the types of things that are important to people in their everyday lives. While the tool allows for the generation of various word classes from the tree structure of the lexicon, we use the author's originally recommended set of 50 value categories.

Opinion Finder (OF) This lexicon (Wilson et al., 2005) is used specifically to search for subjectivity in language. All words in the lexicon are grouped into either the Positive class or the Negative class depending on the word's connotation.

3.2. Dominance Scores

In order to compare texts written by users from areas classified as urban to text written by users from rural areas, we calculate dominance scores (Mihalcea and Pulman, 2009) for each lexicon category. This allows us to compare the relative difference in usage of words from the lexicon class between the two groups.

For a class of words $C = W_1, W_2, ..., W_N$, the coverage of the class in a corpus X is the percentage of words in X that belong to class C, normalized by the number of words in X:

$$Coverage_X(C) = \frac{\sum_{W_i \in C} Frequency_X(W_i)}{Size_X}$$

Consider two corpora: the foreground corpus, F, and the background corpus, B. The dominance score for any class C of F with respect to B is the ratio between the coverage of the corpus F and the coverage of the corpus B.

$$Dominance_F(C) = \frac{Coverage_F(C)}{Coverage_B(C)}$$

A score of 1 indicates that both groups use words from the class C at equal rates, while scores higher than 1 indicate a greater usage of the words in F, and scores less than 1 indicate the opposite. In the following sections, we will consider the text written by users from either rural or urban areas as F and the text written by the other users as B.

4. Results and Analysis

Given the data and approach described above, we analyze the differences in content written by users from urban and rural areas.

4.1. Word Class Dominance Scores

We calculate dominance scores of the urban and rural corpus' for each class represented among the six lexicons. Tables 1-4 show the classes with the top dominance scores for each lexicon for both user descriptions and tweets when

[3] Technically, this group also contains "urban clusters", but for the purposes of this study, we group these together with areas categorized as rural.

[4] We make the set of user ids, tweet ids, and associated cities and populations available at `http://lit.eecs.umich.edu/downloads.html`.

	Users in Urban Areas			*Users in Rural Areas*		
Lexicon	Category	Score	Example Words	Category	Score	Example Words
LIWC	Assent	2.573	awesome, cool, ok	She_He	2.573	she, him, oneself
	Discrep	1.617	could, hope, need	Home	2.105	bedroom, door, window
	Insight	1.591	consider, feel, know	Death	1.700	alive, coffin, kill
	Friend	1.514	bf, buddy, date	Relig	1.607	belief, god, pray
	Hear	1.470	concert, listen, said	Focus_Past	1.413	asked, gave, ran
GI	Place_Route	8.234	road, highway, street	Infants	2.915	baby, kid, young
	Ought	4.632	deservedly, piety, rightful	Natural_Objects	2.429	coal, ice, plant
	Racial	4.117	black, ethnic, native	Increase	2.332	accumulate, expand, overflow
	Yes	3.088	agree, right, yeah	Religion	2.105	christ, holy, sin
	Building	2.058	apartment, deck, wall	Pain	1.666	ache, despair, fury
R	Deafness	7.205	hearing, alphabet, shut	Angularity	6.801	angle, point, bend
	Vice	6.175	evil, moral, error	Ceramics	6.801	cement, clay, enamel
	Defective_Vision	6.175	visual, eyes, sight	Circularity	5.830	ring, cycle, wheel
	Stench	6.175	stink, odor, repulsive	Religious_Bldgs	5.830	choir, shrine, mosque
	Eccentricity	6.175	strangeness, kooky, freak	Alarm	4.858	alert, siren, startle
M	Care_Neg	2.058	cruel, hurt, pain	In_Group_Pos	2.429	attachment, earnest, loyal
	Sanctity_Pos	1.235	blessed, faith, religious	Care_Pos	0.995	concern, patience, tolerance
	Fairness_Pos	1.166	equal, honest, right	Authority_Pos	0.972	follow, command, tradition
	Authority_Pos	1.029	follow, command, tradition	Fairness_Pos	0.857	equal, honest, right
	Care_Pos	1.005	concern, patience, tolerance	Sanctity_Pos	0.810	blessed, faith, religious
V	Creativity	1.687	inventive, novelty, curiosity	Forgiving	1.626	pardon, acceptance, grace
	Justice	1.579	fair, equal, law	Religion	1.532	god, heavenly, church
	Advice	1.529	opinions, views, counsel	Significant-Other	1.445	married, companion, fiance
	Moral	1.363	ethical, moral, ethos	Marriage	1.431	partners, wedding, wife
	Art	1.352	music, painter, artistic	Family	1.363	son, parent, grandparents
OF	Negative	1.213	doubt, lull, stupid OF	Positive	1.006	awe, dream, happy

Table 1: Top Classes and dominance scores for user self-descriptions from Authors in Urban (left) and Rural (right) areas.

the urban corpus is in the foreground and the rural corpus is in the background, and when the rural corpus is in the foreground and the urban corpus is in the background.

Twitter users from urban areas used more words in both their descriptions and tweets that come from the Creativity and Art categories of the Values lexicon. They also included words from the Moral and Advice Values classes in their descriptions more often, suggesting that these users are more likely to share their opinions and moral stances on Twitter than their rural counterparts. They also talk about friends more often than the rural users, who talked more about familial relationships and various members of their families. The users from rural areas were more likely to use words from the LIWC Home class, the Values Family class, and the Values Significant-Other class, suggesting that in addition to family, users from rural areas are more likely to discuss their home life and significant others in their descriptions. The rural population represented in our dataset also uses words from the LIWC, General Inquirer, and Values Religion classes more often in their user descriptions than our urban population, suggesting that Twitter users from rural communities are more likely to identify themselves using religious terms, and are also more likely to tweet about theological concepts.

4.2. City-Level Analysis

In order to perform an analysis at the city-level, we split our dataset up into the tweets and corresponding user descriptions from users from each city represented in our data. For this analysis, we only included cities in our analysis represented by at least 50 unique users, leaving us with 1,985 rural areas and 716 urban cities. To generate lexicon scores for a city, we average the scores of all users from that city. Further, we computed an average rural area vector and an average urban city vector by averaging the city-level scores from all cities falling into either classification.

Next, we used cosine similarity to compare each city's vector with the average vector representing all cities in either the urban or rural grouping to find the "most similar" city to these average vectors (Table 3). Interestingly, the rural area most similar to the average is the same for both the average urban city and the average rural area for descriptions and tweets. We suspect that this indicates that the average urban city vector and average rural area vector are very similar to one another, and the vectors for each rural area are less similar to one another than the vectors for each urban city are to each other. Additionally, the urban cities for both tweets and descriptions that are most similar to rural areas are large Midwestern cities that are surrounded by more rural areas.

	Users in Urban Areas			*Users in Rural Areas*		
Lexicon	Category	Score	Example Words	Category	Score	Example Words
LIWC	Death	2.942	alive, coffin, kill	Anx	3.371	anxiety, fears, nervous
	They	1.557	they, they'll they'd	Home	1.751	bedroom, door, window
	She_He	1.490	she, him, oneself	Sad	1.686	cry, grim, tragic
	Netspeak	1.483	:), idk, lol	Sexual	1.541	sex, lover, naked
	Male	1.446	he, husband, nephew	Compare	1.445	best, easier, before
GI	Stay	2.192	hang, locate, remain	Building	4.816	apartment, deck, wall
	Animal	2.076	chicken, insect, tiger	Academic	2.793	campus, degree, research
	Fall	1.817	drop, sink, tumble	Place_Aquatic	1.926	ocean, river, swamp
	Judgment	1.817	adorn, idiot, madness	Place_Route	1.685	road, highway, street
	Military	1.780	army, march, sword	Exchange	1.651	cash, owe, spend
R	Architecture_Design	7.267	style, city, classical	Art_Criticism	5.779	aesthetic, critique, analyze
	Misbehavior	7.267	disorderly, rascal, rough	Land	5.779	dirt, territory, bay
	Resonance	5.191	boom, echo, clang	Inorganic_Matter	4.816	inanimate, mineral, unfeeling
	Marsh	5.191	swamp, meadow, bog	Religions_Cults_Sects	3.853	faith, Islamism, Confucianism
	Other_Sports	5.191	billiards, rowing, judo	Dryness	3.853	thirst, desert, shrivel
M	Sanctity_Neg	3.115	detest, repel, loathe	Sanctity_Pos	3.853	blessed, faith, religious
	Authority_Pos	1.187	follow, command, traditions	Care_Pos	1.127	concern, patience, tolerance
	Fairness_Pos	1.083	equal, honest, right	Care_Neg	0.963	cruel, hurt, pain
	Care_Neg	1.038	cruel, hurt, pain	In_Group_Pos	0.963	attachment, earnest, loyal
	In_Group_Pos	1.038	attachment, earnest, loyal	Fairness_Pos	0.923	equal, honest, right
V	Creativity	1.429	inventive, novelty, curiosity	Feeling-Good	1.201	happiness, joys, glee
	Moral	1.291	ethical, moral, ethos	Religion	1.193	god, heavenly, church
	Art	1.270	music, painter, artistic	Perseverance	1.185	stamina, vigor, endurance
	Wealth	1.225	money, funding, income	Animals	1.143	livestock, cats, hamsters
	Achievement	1.210	success, progress, revenue	Friends	1.132	buddies, mates, friend
OF	Negative	1.027	doubt, lull, stupid	Positive	1.021	awe, dream, happy

Table 2: Top Classes and dominance scores for tweets written by users from Urban areas (left) and Rural areas (right).

Text	**Type**	**Average**	**City**
Tweets	Rural	Rural	Myrtle Beach, South Carolina
		Urban	Myrtle Beach, South Carolina
	Urban	Rural	Columbus, Ohio
		Urban	Orlando, Florida
Descriptions	Rural	Rural	York, Pennsylvania
		Urban	York, Pennsylvania
	Urban	Rural	Omaha, Nebraska
		Urban	Phoenix, Arizona

Table 3: Cities with tweets and descriptions most similar to the average tweets and descriptions from Urban and rural areas. For example, the first row shows the rural area that is closest to the average of all rural areas, while the third row shows the urban city that is most similar to the average of all rural areas.

4.3. Classification of Urban/Rural Users

Lastly, we sought to determine whether or not, using linguistic features alone, we could predict whether a user claims to be located in an urban or rural area. To do this, we created a dataset balanced equally between both categories, and we used the FastText classifier (Joulin et al., 2016) with default parameters, training new word embeddings and subwords to be used as features for the classification with cross-entropy loss.

Using only users' tweets, we were able to achieve an accuracy of 55.7%, while using users' descriptions allowed the model to make the prediction with 62.6% accuracy (compared to a baseline of 50%). This indicates that there are identifiable linguistic differences between the groups, but the performance is not strong enough to suggest that this model could be used to automatically tag users by their city's population size.

5. Conclusion

In this paper, we introduced a new set of Twitter data tied to U.S. cities and their corresponding population sizes, we compared dominance scores across a series of lexicons to explore the difference in tweets and self-descriptions written

by urban and rural users, and we measured the similarity of individual cities to the average urban and rural areas.

We also trained classification models that are able to capture a relationship between population size classification and users' language patterns. Through our analyses, we provided insights into how users present themselves and communicate online based on the size of the cities in which they reside.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (grant #1815291) and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation.

6. Bibliographical References

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., and Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In Ninth International AAAI Conference on Web and Social Media.

Garcia-Gavilanes, R., Quercia, D., and Jaimes, A. (2013). Cultural dimensions in twitter: Time, individualism and power. In Seventh International AAAI Conference on Weblogs and Social Media.

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., and Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47. Elsevier, pp. 55–130.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Jarmasz, M. (2012). Roget's thesaurus as a lexical resource for natural language processing. *arXiv preprint arXiv:1204.0140*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kennedy, A. and Szpakowicz, S. (2014). Evaluation of automatic updates of roget's thesaurus. *Journal of Language Modelling*, 2(1):1–49.

Lichter, D. T. and Ziliak, J. P. (2017). The rural-urban interface: New patterns of spatial interdependence and inequality in america. *The ANNALS of the American Academy of Political and Social Science*, 672(1):6–25.

Mihalcea, R. and Pulman, S. (2009). Linguistic ethnography: Identifying dominant word classes in text. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 594–602. Springer.

Namenwirth, J. Z. and Weber, R. P. (2016). Dynamics of culture. Routledge.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). Do all birds tweet the same?: characterizing twitter around the world. In Proceedings of the 20th ACM international conference on Information and knowledge management, pages 1025–1030. ACM.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents, pages 37–44. ACM.

Ratcliffe, M., Burd, C., Holder, K., and Fields, A. (2016). Defining rural at the us census bureau. *American community survey and geography brief*, pages 1–8.

Roget, P. M. (1883). Thesaurus of english words and phrases. Avenel Books.

Scala, D. J. and Johnson, K. M. (2017). Political polarization along the rural-urban continuum? the geography of the presidential vote, 2000–2016. *The ANNALS of the American Academy of Political and Social Science*, 672(1):162–184.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Suttles, J. and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 121–136. Springer.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Opinionfinder: A system for subjectivity analysis. In Proceedings of HLT/EMNLP 2005 Interactive Demonstrations.

Wilson, S. R., Shen, Y., and Mihalcea, R. (2018). Building and validating hierarchical lexicons with a case study on personal values. In International Conference on Social Informatics, pages 455–470. Springer.