Measuring Regional Variation in Culture Through Embedding-Based Lexica

Keywords: NLP, Culture, Word Embeddings, Regional Variance, Twitter

Extended Abstract

Introduction: This study builds a new measure of regional variation in culture using social media language. Historically, measuring how cultural constructs[1] vary across regions has been mostly done through questionnaires, such as the World Values Survey[2]. However, questionnaires are time-consuming and heavily restricted in scope – the most recent wave of the WVS took over 4 years to complete and only averages 52 participants per US state. One solution that is more scalable is using data-driven lexica to count words on geolocated social media, but this requires lexica that are statistically validated and linked to cultural theory.

We propose a method that leverages FastText word embeddings[4] to efficiently create lexica based on a small set of seed words. This method enables researchers to measure regional variation in any cultural construct using language from those regions. We put this into practice by building lexica to study individualism and collectivism[1] and measure their variation across US counties using Twitter. We validate our lexica results against prior indicators from the Global Collectivism Index (GCI)[3], gathered using state-level Census data and responses from the WVS. Results indicate our embedding-based lexica can accurately measure cultural constructs through social media language. The results also provide the first-ever county-level map of individualism and collectivism across the US.

Lexica Creation: Our embedding-based lexica creation method has two components: *Expansion* and *Purification*. **Figure 1** details this two-stage approach.

Expansion: Given a small set of seed words curated by a psychologist with expertise in collectivism/individualism research, we utilize word embeddings to expand the set of seed words in two ways: we locate all words that are similar to each individual seed word (i.e. synonym expansion), as well as locate the words that are similar to the overall construct described by the complete set of seed words (i.e. concept expansion). First, we find the nearest neighbors for each individual seed word in FastText embedding space[4] – we use FastText instead of context-based embeddings due to their superior ability to find synonyms in context-free scenarios. Next, we average all of the seed words to find the centroid of the seed word cluster. We then find all of the nearest neighbors of this centroid. The number of nearest neighbors returned by either of these expansions are tunable parameters, and can be adjusted based on the desired length of the final lexicon.

Purification: Upon aggregating the words returned from both expansion types, we want to ensure that the resulting lexica are both pertinent and internally correlated. First, we filter out rare words, i.e. any words below a given usage frequency. We then apply the lexica to our US Twitter Corpus to get overall lexica scores (the total word frequencies for words in the individualism lexicon and the collectivism lexicon) for every state and county. Finally, we remove any low validity words by measuring the correlation of each individual word frequency with the overall lexica scores and removing all words that do not correlate positively (Pearson $r < 0.15$).

Validation: Upon expanding and purifying the lexica, we validate our results against known indicators of collectivism from the GCI[3]. All six variables in the GCI – total fertility rate, living arrangements (% households with people over 60 and children under 14), stability of marriage (divorce rate to marriage rate ratio), religiosity, collective transportation, and

ingroup bias (approximated by compatriotism due to lack of state-level data) – are replicable at the state-level using US census data and WVS data. Note that when aggregating US census data from county-level to state-level, we treat each county as being weighted equally, due to disproportionate amounts of data coming from big cities. In order to determine which of these six replicated variables also measure collectivism within the United States, we sample subsets of the six variables and use Cronbach's alpha to measure internal consistency. We limit the subsets to size three or larger, following Pelham and colleagues'[3] validation of three collectivism indicators per nation. The set of living arrangements, religiosity, and compatriotism yielded the highest Cronbach's alpha (0.702), so we chose these three variables as a validation metric.

Table 1 shows the correlations between each of the three validation variables, the collectivism lexicon score, and the individualism lexicon score for US states. Collectivism word use positively correlates with all validation outcomes, and individualism word use correlates negatively. We further validate against median income at the state-level. Prior research has found that income is negatively correlated with collectivism.[3] Similarly, income was negatively correlated with our collectivism lexicon scores (-0.273) and positively with our individualism lexicon scores (0.424). We also validate against Vandello and Cohen's collectivism scores[5]. We see a positive correlation (0.388) with our collectivism lexicon scores and a negative correlation (-0.374) with our individualism lexicon scores. This suggests that our lexica measurements are indeed tapping into real cultural differences.

Results & Future Work: We apply the validated lexica on the county-level Twitter language to gain a more fine-grained understanding of how these cultural constructs vary regionally. **Figure 2** illustrates this variation. We see that the deep south shows high levels of collectivism (dark red) and low levels of individualism (light blue). Conversely, we see that the west coast and the northeast show low levels of collectivism (light red) and high levels of individualism (dark blue). Note that counties without any attributed individualism or collectivism scores, due to lack of Twitter data from those counties, are colored in gray. We further scope by ACP communities and find that College Towns and Big Cities have the highest levels of individualism, while Evangelical Hubs and the African American South have the highest levels of collectivism, as shown in **Figure 3**.

In conclusion, our results show that it is possible to efficiently create accurate lexica by leveraging word embeddings as shown in **Figure 1**. Using this method, we can create lexica to measure cultural constructs and apply the lexica to social media language in order to get labels at fine-grained geographic levels, such as states, counties, and Census tracts. Researchers can use these scores to get deeper insight into communities and cultures. Researchers can also extend this method to other languages by leveraging a multilingual embedding, allowing them to efficiently measure cultural constructs on a global scale.

References

[1] Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov. *Cultures and organizations: Software of the mind*. Vol. 2. New York: Mcgraw-hill, 2005. **[2]** Haerpfer, C., Inglehart, R., et al. (eds.). 2020. World Values Survey: Round Seven – Country-Pooled Datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. **[3]** Pelham, Brett, et al. "A truly global, non-WEIRD examination of collectivism: The global collectivism index (GCI)." *Current Research in Ecological and Social Psychology* 3 (2022). **[4]** Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the association for computational linguistics* 5 (2017): 135-146. **[5]** Vandello, Joseph A., and Dov Cohen. "Patterns of individualism and collectivism across the United States." *Journal of personality and social psychology* 77.2 (1999).

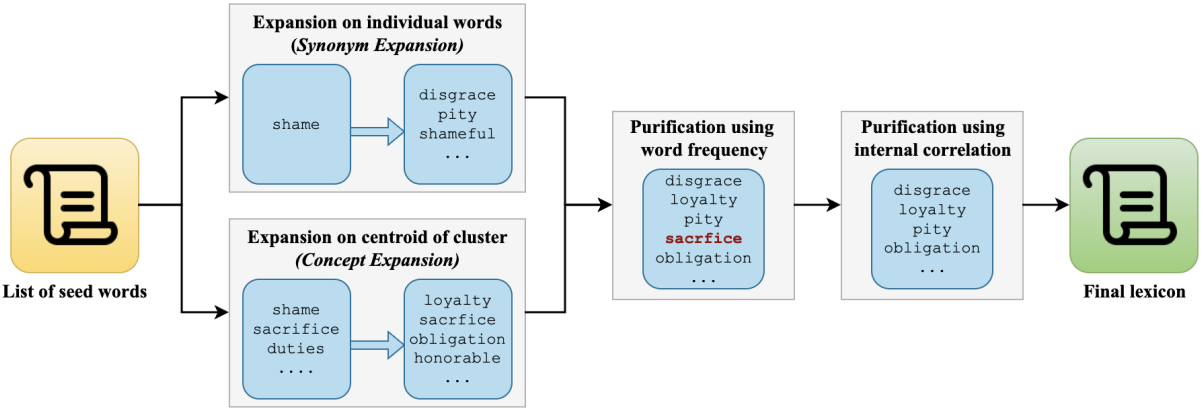

Figure 1: Our embedding-based lexica creation method. The first stage, *Expansion,* consists of synonym expansion and concept expansion, done in parallel. The second stage, *Purification,* includes frequency-based and correlation-based purification, done sequentially.

	Individualism Lexicon Score	Collectivism Lexicon Score	Living Arrangements	Religiosity
Collectivism Lexicon Score	**-0.470**			
Living Arrangements	**-0.291**	**0.200**		
Religiosity	**-0.658**	**0.400**	**0.344**	
Ingroup Bias (Compatriotism)	**-0.513**	**0.464**	**0.169**	**0.750**

Table 1: Pairwise Pearson correlations between the results of our individualism and collectivism lexica applied to our Twitter Corpus and the GCI validation variables, at the US state-level.

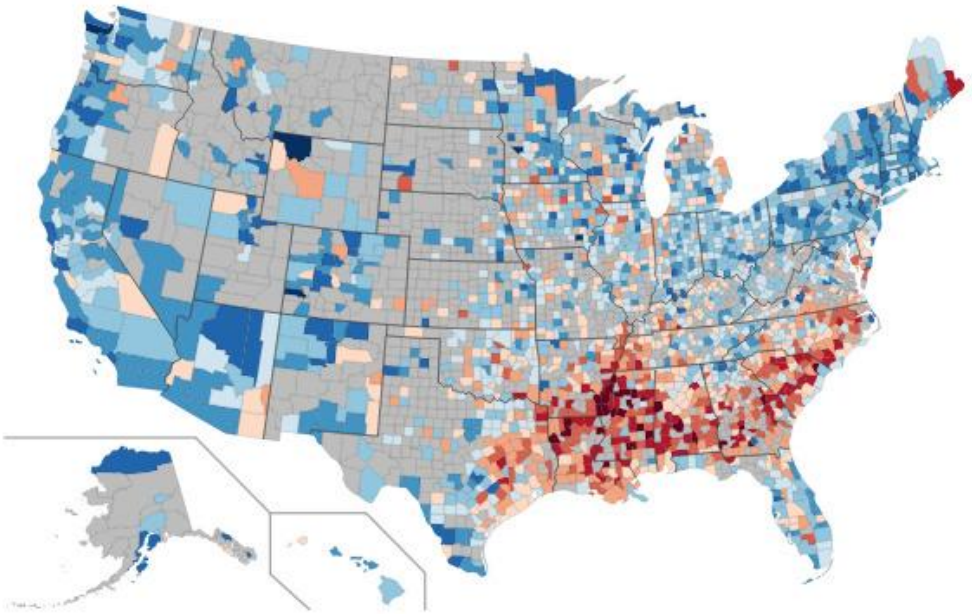

Figure 2: Collectivism (red) and individualism (blue) across US counties. Dark red represents higher collectivism, and dark blue represents higher individualism. Gray counties do not have enough Twitter data to estimate a score.

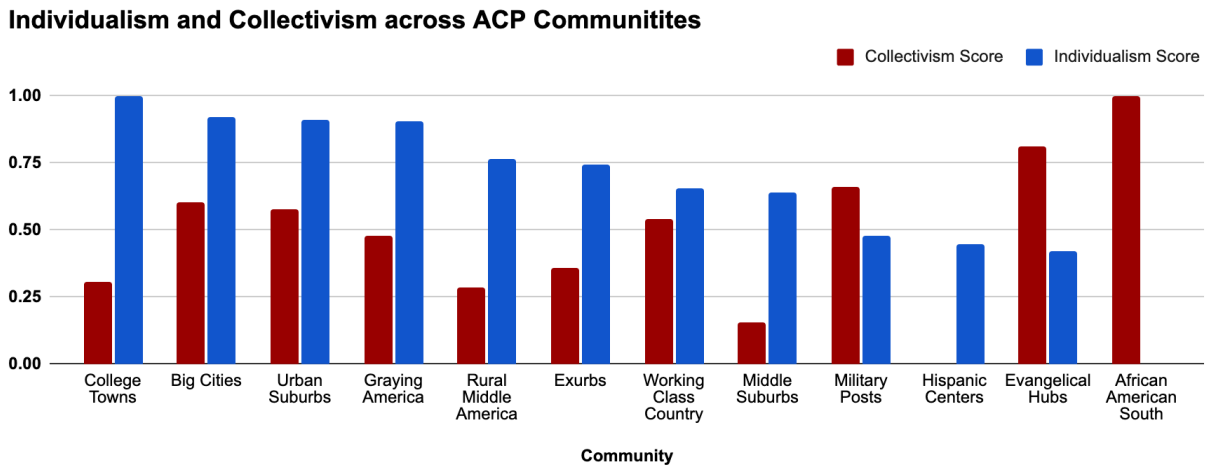

Figure 3: A comparison of collectivism lexica (red) and individualism lexica (blue) scores across communities defined by the American Communities project (ACP). We 0-1 normalize both sets of scores and order the communities from most individualistic (left) to least individualistic (right). The results suggest convergent validity of our lexica with the American Communities Project. For example, it makes sense that college towns and big cities (left) score high on individualism and that military posts and evangelical communities score low on individualism (right).