Geographic Wellbeing

Edit New page

Jump to bottom

tonyl-code edited this page 7 minutes ago · 5 revisions

Abstract

Policy makers are increasingly interested in pursuing societal wellbeing, not just GDP, as a policy priority (Graham 2021). However, traditional methods of obtaining this information involves surveying the population, which is expensive. Social media, where users often post their thoughts and feelings, can therefore serve as valuable indicators. In this project, we aim to improve upon the existing method of predicting wellbeing from social media posts established by Jaidka et al., 2020 by using word embeddings. We show that using embeddings as features moderately improves prediction accuracy.

Introduction

Existing methods of predicting wellbeing from social media posts uses dictionary methods and topic modelling to extract insight from text (Jaidka et al., 2020). There, the researchers obtained geolocated Twitter posts and the wellbeing index of every US county via national surveys. They conducted supervised learning on a county level, whereby for each country they combined Twitter posts located in that county, extracted the language features, and used ridge regression to predict that county's wellbeing index.

With the advancement of transformers, however, we wish to improve upon this method by using contextualized embeddings as features. As such, our overarching goal remains the same as Jaidka et al's (2020). To keep our methods comparable, I will be conducting my analysis at the county level and use supervised learning. However, due to data policies, the previous study did not post their data. I will be using a separate Twitter dataset and county level wellbeing indices from another national survey. Thus, I will first replicate Jaidka et al.'s method on our data set to establish a baseline performance, and then move on to the embeddings method.

Overview of Method

Main Approach

I will be conducting my analysis on the county level. Each county is labelled with a wellbeing index obtained from the University of Wisconsin Population Health Institute from 2010. I then obtained geolocated Twitter posts and users during the period of 2009 - 2010 (Cheng et al., 2010). I match each user to their respective counties, and for each user, I encode a separate sentence embedding for each of their posts via SBERT. Afterwards, I average the embeddings within each county to obtain a county level embedding. This 384 dimension embedding will serve as our features. Once we have the features and labels, we apply supervised learning methods to predict each county's wellbeing. We use a train-test split to obtain a set of test predictions which we will correlate with the true labels.

Baseline

For the scope of this study, we say that our method produces more accurate results if the correlation with the true labels is higher than the correlation produced by replicating the previous study's method. Note that, to show that our study is internally valid, the correlation which we will be using as a baseline is not the one reported in the reference paper, but rather the correlation we obtain by replicating the method on our new data set. Thus, a substantial part of this project will be on establishing this baseline.

Experiments

Data

The study by Jaidka et al. 2020 aggregated the wellbeing indices for each county from 2009 to 2015 obtained via the Gallup national survey and used a 10% Twitter sample from 2009-2015. Neither of these are publicly available, so we use the following two data sets as substitutes.

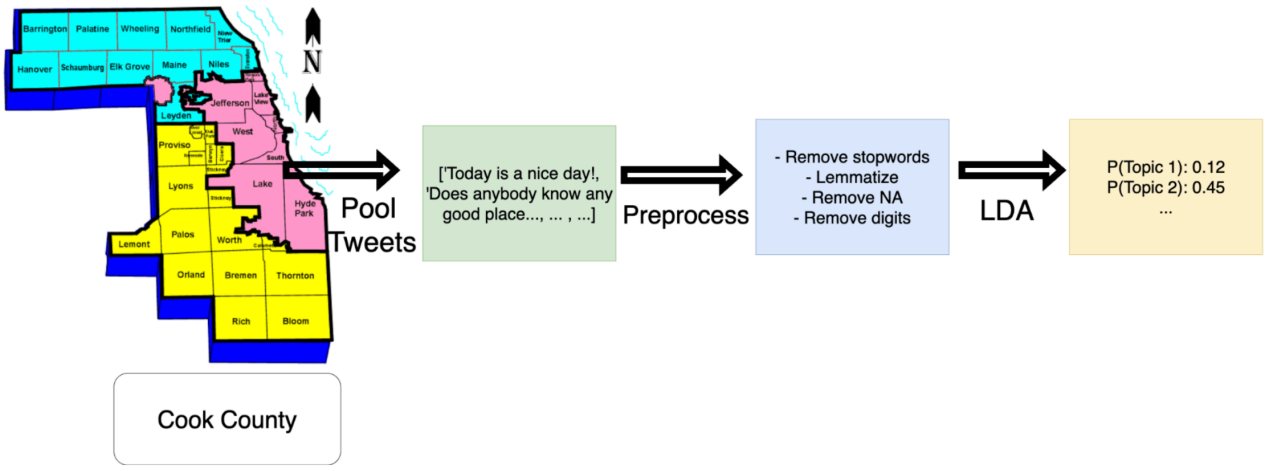
1. Geolocated twitter posts (Cheng et al., 2010). This data set contains 5,136 Twitter users and 5,156,047 posts. All posts were collected in the period of September 2009 to January 2010. This results in 682 unique counties being represented.

2. County Health Rankings & Roadmaps (CHR&R) from the University of Wisconsin Population Health Institute. This report contains physical and mental health indices from each county in the periods 2010 - 2024. I use the 2010 data in this study so that the population we are estimating matches with the twitter users, and I calculate a general wellbeing index which averages the variables, “Number of Physically Unhealthy Days” and “ Number of Mentally Unhealthy Days”. This is so that we can construct a close estimate of the Gallup index, which is a composite of “five elements of well-being” – purpose, social, financial, community, and physical wellbeing (Gallup, n.d.). Indeed, when we average our scores for each state and compare them to the Gallup index at the state level (the index is publicly available at a state level), we obtain a correlation of r=0.62. Thus, we can reasonably assume that our indices can capture a number of dimensions of human wellbeing.

Evaluation method.

We will be evaluating the models based on performance on test/ hold-out data. Since we are using supervised learning, the usual MSE loss applies when training our models. However, our main metric is the correlation between the predicted labels and the true labels for each test county. We will be comparing the correlation obtained using our embedding method with the previous method to evaluate if the new method is an improvement. We use correlation rather than the typical validation loss metrics as it was used in the Jaidka et al. 2020 paper and as it seems to be the convention in the field; in the end, we are interested in whether or not our model has any explanatory power for wellbeing, and so correlation serves as a more intuitive metric.

Experiment 1: Topic Modelling



Features

In order to make meaningful comparisons between the previous study's method and the proposed BERT method, we must first replicate their method on our new data set. Note that the paper utilized several methods, many of which are closed dictionary based methods (i.e PERMA, LIWC) which consist of expert-annotated words rated according to psychological concepts. For the purposes of this project, we will focus on the Latent Dirichlet Allocation (LDA) method, which was used in their best performing model.

The previous study extracted 2000 topics from their data set, each of which was associated with several words and the weights of each word in determining the topic. Since the group published their data set of this, we can follow their procedure and calculate the probability of a topic for a given document as

$$p(topic \mid doc) = \sum_{i=0}^{N} \frac{freq(word_i)}{N} * weight(word_i),$$

where N is the total number of words in the document, freq(wordi) is the frequency of the word in the document, and weight(wordi) is the weight of the word relative to the topic. Note that this term would be zero if the word does not exist in either the dictionary or document.

For our purposes, all the tweets in one county would be pooled together to count as a single document. Thus, we apply this formula for each county and generate a matrix relating each county to their topic probability distribution (i.e a matrix with 648 counties as rows and 2000 probabilities associated with each topic as columns).

Training

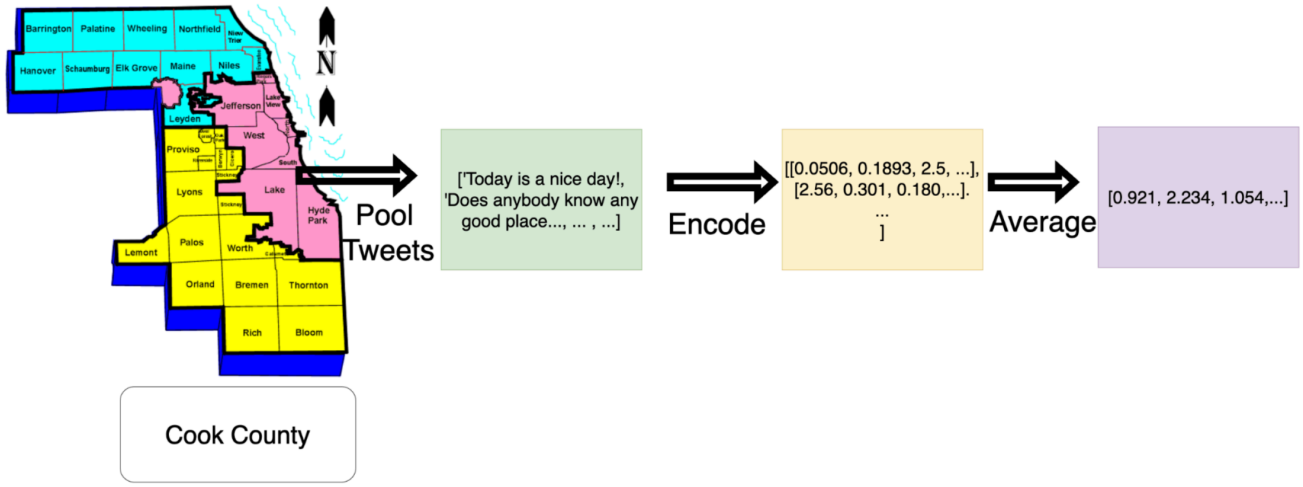
The above matrix will therefore serve as our feature matrix. We do a 80-20 train-test split and use ridge regression with penalty alpha = 1.0 and the default parameters. Following the procedures in Jaidka et al's (2020), we also first reduce the dimensions of the feature matrix using PCA (with 100 components) before inputting to the model.

Result

Here, we report the MSE and correlation between the predicted wellbeing index and actual wellbeing index (from the UW-Madison study specified above) on the test set. We see that the version without dimensionality reduction performed better. One potential explanation is that since our data set is much smaller than the one used in the previous study, we end up with less information and so any additional dimension could be informative.

	Ridge	Ridge (With PCA)
MSE	0.414	0.414
Correlation	0.116	0.041

Experiment 2: BERT



Features

To obtain the sentence embeddings, I used the SBERT python module and the mpnet-base model from Microsoft. This model creates embeddings with 384 dimensions. To construct the features, I produce a BERT for each tweet and average all embeddings within a county to obtain county-level embeddings. This county to embeddings matrix will be our feature matrix.

Training

Like before, we do a 80-20 train-test split. For ridge regression, I use a penalty term alpha=1.0 with default parameters. I first run it on the training set, and then use the test set to obtain a set of predictions which I will correlate with the true labels. Since our matrix is likely less sparse than before in the topic modelling case, I also use Random Forest regression with 100 trees (default parameters in sklearn) instead of just ridge regression. I apply both models with and without dimensionality reduction.

Result

As we see, the BERT embeddings perform ~1.5 to 2 times better with Ridge and Random Forest regressions. However, the models still perform much worse with PCA.

Model	Ridge	Forest	Ridge PCA	Forest PCA
MSE	0.449	0.436	0.467	0.470
Correlation	0.178	0.213	-0.214	0.001

Discussion

While our study focused on the most relevant method for prediction, the original study we replicated employed a broader range of techniques. Notably, it incorporated post-stratification, which adjusts for demographic biases in social media data by weighting users based on the population distribution of their county. This approach helps mitigate the issue that Twitter users are not representative of the U.S. population as a whole. However, implementing post-stratification requires demographic information about individual users, which we lack in our dataset. Future work could investigate alternative strategies to account for this bias, such as leveraging external demographic data or estimating user demographics through auxiliary models.

Beyond improving predictive accuracy, future research could also focus on enhancing interpretability. One potential avenue is the use of sparse autoencoders to identify the most salient features within BERT embeddings that contribute to wellbeing predictions. This could provide deeper insights into the linguistic and thematic signals associated with county-level wellbeing, offering a more interpretable framework for understanding the connection between social media discourse and public health metrics.

Additionally, expanding the data sources beyond Twitter could improve model robustness and generalizability. Platforms like Reddit, which host long-form discussions and diverse topic-specific communities, may provide complementary signals that are not captured in Twitter's predominantly short-text format. Comparing wellbeing predictions across multiple social media sources could help assess the consistency of our findings and identify platform-specific biases. Integrating multiple data sources could also enable a richer, more holistic understanding of how online discourse reflects and potentially influences community wellbeing.

Conclusion

Our findings suggest that using contextualized word embeddings from BERT improves the prediction of county-level wellbeing from social media posts compared to traditional topic modeling methods. Ridge regression and Random Forest models trained on BERT embeddings achieved higher correlations with true wellbeing scores, indicating that embeddings capture richer semantic information than topic distributions. However, we observed that dimensionality reduction via PCA consistently degraded performance, likely due to the loss of crucial information in our relatively small dataset. These results highlight the potential of transformer-based embeddings for social media-based wellbeing prediction and suggest that future work, in addition to the ones mentioned above, could explore optimizing feature selection and model architectures to further refine predictive accuracy.

References

SBERT. https://sbert.net/index.html

Graham, Carol (2021). Making well-being a policy priority: Lessons from the 2021 World Happiness Report. https://www.brookings.edu/articles/making-well-being-a-policy-priority-lessons-from-the-2021-world-happiness-report/

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. Proceedings of the National Academy of Sciences, 117(19), 10165-10171.

Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010.

County Health Rankings. https://www.countyhealthrankings.org/

Gallup. (n.d.). Gallup-Healthways Well-Being Index: How does it work? Gallup. Retrieved March 11, 2025, from https://news.gallup.com/poll/128186/gallup-healthways-index-work.aspx

+ Add a custom sidebar

Clone this wiki locally

```
https://github.com/minalee-research/cs257-students.wiki.git
```