Can Twitter be used to predict county excessive alcohol consumption rates?

Logan Ayliffe
ayliffe@seas

Joshua Gardos
gardos@sas

Audrey Yang
auyang@seas

Abstract

Inspired by work done at the World Well-Being Project, we investigate the correlation between Twitter language usage and excessive alcohol consumption rates. We replicate a study by Curtis et al. using updated county-tweet data. We successfully replicate the published baseline results as well as extensions to this baseline. We demonstrate that there is a strong correlation between the Twitter data features and drinking, stronger even than the socioeconomic and demographic features typically used to predict excessive alcohol consumption rates.

1 Introduction

The World Well-Being Project is an interdisciplinary effort among computer scientists, psychologists, and statisticians. Based out of the University of Pennsylvania's Positive Psychology Center and Stony Brook University's Human Language Analysis Lab, the WWBP develops scientific and programmatic methods for measuring physical and psychological health using social media data.

Among their many publications is the 2018 paper from Curtis et al. titled "Can Twitter be used to predict county excessive alcohol consumption rates?".[1] In this paper, the authors developed a model trained on a corpus of tweets that was able to predict excessive alcohol consumption more accurately than a baseline classifier using standard socioeconomic-demographic data.. Our project seeks to replicate a subset of this study and expand on the model developed by Curtis et al.

One motivation for this study is reducing barriers accessing the timely, accurate data required to effectively guide public health measures. Alcohol consumption data at the national level is typically collected via annual survey in the US by the Behavioral Risk Factor Surveillance System. As Curtis et al. note, there are several challenges with collecting this data via survey, including memory recall, socially desirable responding, and cost. On the other hand, tweets can be collected in near real-time and represent linguistic communication in a natural setting rather than a survey via phone call. Using tweets can therefore solve many of the problems associated with collecting alcohol consumption data via survey. This in turn can help guide resource allocation at all levels of government and among healthcare providers.

2 Literature Review

2.1 Mental Health

The World Well-Being Project extends their effort towards studying the correlation between social media activity and mental health. In their 2018 paper "Facebook language predicts depression in medical records," Eichstaedt, Smith, et al. analyzed the Facebook content data of 1,175 consenting emergency department (ED) patients, who also volunteered their electronic medical records (EMR), all over a 26-month span; this led to the construction of a dataset containing 949,530 status updates, as well as demographic information for each user.[2] From this information, qualities including post length, frequency, and temporal trends were featurized in an effort to predict the diagnosis of depression in medical records (11207). Specifically, to gauge the semantic content of the tweets, unigrams and

[1] Curtis, B., Giorgi, S., Buffone, A. E. K., Ungar, L. H., Ashford, R. D., Hemmons, J., Summers, D., Hamilton, C., Schwartz, H. A. (2018). Can twitter be used to predict county excessive alcohol consumption rates? PloS One, 13(4), e0194290-e0194290. https://doi.org/10.1371/journal.pone.0194290

[2] Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., Schwartz, H. A. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences - PNAS, 115(44), 11203-11208. https://doi.org/10.1073/pnas.1802331115

bigrams were extracted from the Facebook data, serving as inputs to an implementation of LDA, which in turn produced clusters of tokens used in similar contexts; 200 topics were generated, becoming features for each patient. These findings were then referenced with the Linguistic Inquiry and Word Count (LIWC) software to ascertain relative frequencies of word usage. Along with time of day and demographic characteristics obtained from EMRs, these features were used to learn a logistic regression equipped with a ridge penalty. The hyperparameters of the model were trained using 10-fold cross validation.

Since the predictive classes (i.e. whether or not depression was present on the EMRs) were unbalanced, area under the receiver operating characteristic (ROC) curve was used as the measurement of model performance. The researchers employed several models: one using just Facebook language (AUC of 0.69), another using post length and frequency (0.59), one using user demographics (0.57), one using temporal patterns (0.54), and one using a combination of all features (0.69). Interestingly, the model using only Facebook language scored just as well as the model using all features, suggestive of the strong predictive power language alone holds. Unsurprisingly, the found that higher frequencies in the PIWC dictionaries linked to depressed mood feeling, loneliness, hostility, somatic complaints, and medical referrals tended to indicate the diagnosis of depression. Furthermore, users diagnosed with depression also had the tendency to employ more first-person pronouns, perhaps suggestive of a preoccupation with the self (11205). These findings epitomize the relevance of social media as a tool to (ironically) prevent depression by unobtrusively recognizing warning signs and connecting potentially-depressed individuals to care resources more readily.

2.2 Locus of Control

In their exploratory 2018 article "Modeling and Visualizing Locus of Control with Facebook Language," Jaidka, Buffone et al. explore how social media can be used to assess the construct of Locus of Control (LoC) through social media.[3] Rooted in psychology, LoC refers to the degree to which an individual believes they are in control of events in their lives. Generally, two attitudes surround the trait: externals tend to perceive events as out of their control, while internals ascribe control over their environment to themselves (1). The researchers administered a survey through Qualtrics asking participants to rate themselves given a variety of prompts (based on the MIDUS survey). Using standard psychological scales, an LoC score was then calculated and assigned a binary output. Facebook data - 1.2 million posts volunteered by the 2348 participants - was tokenized and featurized. Specifically, frequency distributions were obtained from the aforementioned LIWC database, as were the 1,000 most common uni-, bi-, and trigrams. Additionally, LDA was employed to generate 2,000 social-media-specific topics, with joint probabilities being used to determine a word's prevalence in a topic. After applying PCA and univariate regression, a gradient-boosted classifier was trained using 10-fold cross-validation (2).

Interestingly, linguistic data alone (specifically, N-grams and LDA topics) produced an F-score (0.82) much higher than that of the same model using census features such as age, race, gender, education, and income bracket (0.54), overall an impressive result. The researchers then shifted their attention towards exploring the correlation between LoC and the 'Big Five' taxonomical personality traits of conscientiousness and emotional stability. The researchers found that terms associated with internals often signified support, belonging, and well-being (such as 'fiance,' 'grateful,' or 'celebration') correlated with greater cognitive (conscientiousness) and emotional (stability) control. Conversely, words associated with externals often reflected greater self-focus and helplessness (such as 'confused' or 'unwanted'), and were linked to less cognitive and emotional control (3). The results of this study signify the ability of social media to predict LoC from posters. The importance of this extends beyond psychology, as locus of control is known to be linked to job performance and satisfaction; the ability to obtain and harness this data therefore has the potential to lead to monetary improvements from a commercial perspective.

In the 2019 article "Suicide Risk Assessment with Multi-level Dual-Context Language and BERT, " Matero et al. aimed to detect suicidal

[3] Jaidka, K., Buffone, A., Eichstaedt, J., Rouhizadeh, M., Ungar, L. (2018). Modeling and Visualizing Locus of Control with Facebook Language. Proceedings of the International AAAI Conference on Web and Social Media, 12(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15076

risk (low, medium, or high) of individuals using Reddit data. [4] Their goals were to a) develop a model with dual-context where language in a suicide-specific context is separate from other language, b) develop a deep learning architecture that applied dual-context modeling to GRU cells and attention layers and adds a user-factor adaptation layer, c) compare individual theoretically related linguistic assessments, and d) evaluate models based on theoretically-motivated features versus models based on open-vocabulary features. The data used in this paper came from the CLPsych 2019 Shared Task, which was split into three separate tasks. Task A included all user data from the subreddit r/SuicideWatch; Task B included all user data from Reddit; and Task C included user data from all of Reddit without r/Suicide Watch. They extracted three sets of linguistic features. The first were features on the theoretical dimensions of message level and the user level; the second was open-vocabulary features, including dimensionally reduced BERT embeddings and 25 LDA Topics trained using Gibb's Sampling over suicide watch posts; and meta-features, including post-statistics and 39 subreddit features derived from popular subreddits. They noted correlations and distributions that they found by exploring the data. For example, being female was found to be correlated with having a higher suicide risk (contradicting previous findings), age was found to have no significant effect, and agreeable, conscientious, and extroverted personality factors correlated with lower suicide risk.

The researchers worked on each task separately. For Task A, they found that a neural model that used LSTM with hierarchical post-level attention worked the best, with a 0.59 accuracy and 0.5 F1 score on the test set. One of the logistic regression models that they trained performed better on the training set, with a 0.57 accuracy and 0.46 F1 score. For Task B, they trained a logistic regression model that took in features of SuicideWatch and non-SuicideWatch language and processed them separately. They also trained a neural model with two separate GRU cells, one that took the same input features of the Task A model on SuicideWatch, and the other that took the subreddit info feature vector in addition to those same input features on non-SuicideWatch subreddits. They used separate attention weights for SuicideWatch, non-SuicideWatch and applied user-factor adaptation (a concatenation of the sum of hidden vectors with the attentions of the two GRU cells). Then, they concatenated user-level features with the factorized output vector. They found that there was a large improvement from the dual-context type approach. The dual-context BERT embeddings based logistic regression outperformed other approaches on the test set, with a 0.57 accuracy and 0.5 F1 score. For Task C, the authors trained logistic regression models using a) BERT embeddings alone, b) open vocabulary, theoretical dimensions, meta-features, and subreddit latent factors, and c) the same features as were used in Task B but without traits of personality, age/gender, and anxiety, anger, and depression scores. They found that a combination of open vocabulary and theoretical features outperformed the other approaches. The best had a 0.69 accuracy and 0.18 F1 score, and accounted for all user level traits, and mean aggregation of message-level open-vocabulary features.

2.3 Cardiovascular Health

In the 2021 article "Predicting Cardiovascular Risk Using Social Media Data: Performance Evaluation of Machine-Learning Models," the authors Andy et al. used social media data to predict Atherosclerotic Cardiovascular Disease (ASCVD) risk in an EMR and to character differences in posts relative to four categories based on 10-year primary risk from ASCVD risk scores. [5] The researchers obtained EMR data, demographic information, and Facebook statuses from up to five year in the past from patients who consented. They took two approaches to process language from social media posts to include in a regression model. The first was to use open vocabulary topics. They used latent Dirichlet allocation (LDA) to obtain 20 topics generated using Facebook posts. The second was dictionary-based psycholinguistic features. They used language from posts to identify prevalence of predefined word categories represented in the Linguistic Inquiry and Word Count (LIWC) dictionary,

[4] Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., Schwartz, H. A. (2019, June). Suicide risk assessment with multi-level dual-context language and BERT. Paper presented at the Sixth Workshop on Computational Linguistics and Clinical Psychology, Minneapolis, Minnesota. Retrieved from http://wwbp.org/papers/matero2019suicide.pdf.

[5] Andy AU, Guntuku SC, Adusumalli S, Asch DA, Groeneveld PW, Ungar LH, Merchant RM Predicting Cardiovascular Risk Using Social Media Data: Performance Evaluation of Machine-Learning Models JMIR Cardio 2021;5(1):e24473 doi: 10.2196/24473 PMID: 33605888

which has 73 categories.

The researchers trained three models. Model 1 was a multiclass logistic regression model for four categories of risk scores: $< 5\%, 5\% - 7.4\%, 7.5\% - 9.9\%$, and $\geq 10\%$. Using language (text data) only, the $< 5\%$ category could be extracted with 0.78 AUC (area under the curve). In Model 2, a linear regression model, the risk score was treated as a continuous variable (to contrast with the categorical variable of Model 1), and had a Pearson correlation coefficient of $r = 0.26$. In Model 3, risk was treated as a dichotomous variable. It used logistic regression to distinguish the high-risk category using language correlated with low ASCVD scores. The researchers also used LIWC to distinguish different features associated with high-risk by correlating category features. Model 3 had an AUC of 0.69. From the study, the researchers concluded that this method probably will not replace traditional methods but is still helpful in providing supplemental information about individual's lifestyle and behavior. They also found that a high ASCVD risk score was associated with "sad" language (language that fell in the sadness category of LIWC). The results showed low accuracy from data, so the authors hypothesized that that could have been because the patients were between 40 and 79 year old and may not use social media very often.

3 Experimental Design

3.1 Data

We are working with three main datasets. For features, we use Socioconomic and Demographic Data as well as Twitter data. For labels, we use excessive alcohol consumption data. Each dataset is described below.

3.1.1 Drinking Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a population-based cross-sectional telephone and cell phone health survey of U.S. adults, aged 18 years, conducted by state and territorial health departments in conjunction with the Centers for Disease Control and Prevention.

From the BRFSS (2006–2012), we use the county-level prevalence of self-reported binge drinking and heavy drinking. The full dataset is available here.

Excessive alcohol consumption was defined as having drunk more than two drinks per day on average (for men) or more than one drink per day on average (for women) or having drunk 5 or more drinks during a single occasion (for men) or 4 or more drinks during a single occasion (for women).

3.1.2 Socioeconomic and Demographic Data

We use percentages of female, foreign born, African-American and Hispanic as well as age percentages (19 bins) from the U.S. Census Bureau. County-level measures of log income, percentage of married residents, high school and college graduation rates, and unemployment were obtained from the American Community Survey. Because income and education are highly correlated, we also use a composite county-level socioeconomic index by averaging standardized log income and standardized high school graduation rates. Such averaging often results in a stronger single predictor.[6] The full dataset is available here.

In line with the original authors, we split the Demographic and Socioeconomic data into four feature sets:

- Demographic features only: percentage of female, African American, Hispanic, foreign born and married residents as well as four age bins: 1–14, 15–29, 30–60 and 60+
- Socioeconomic features only: log income, high school and college graduation rates, and unemployment
- Single composite socioeconomic index: log income + HS grad rate
- All of the above combined: all demographic, socioeconomic, and the composite index.

3.1.3 Twitter Data

The linguistic features are derived from a large corpus of tweets collected by Curtis et al.The original paper collected a random 1% of twitter data between 10/2011 and 12/2013.These tweets were then geolocated using latlong and self-reported location information in the user's profile. Of the original set of 2.24 billion tweets, 138 million were mapped to counties. However, in conversation with

[6]Gelman, A., Hill, J. (2007). Data analysis using regression and Multilevel/Hierarchical models Analytical Methods for Social Research series. Cambridge and New York: Cambridge University Press. Retrieved from https://proxy.library.upenn.edu/login?url=https://www-proquest-com.proxy.library.upenn.edu/books/data-analysis-using-regression-multilevel/docview/56615274/se-2?accountid=14707

one of the authors Salvatore Giorgio, we learned that the group has since collected and validated a better set of county level Twitter lexical data. This data set is derived from a 10% Twitter sample from 2009-2015, and uses over 1.5 billion tweets that were successfully mapped to counties. Here we use this updated dataset.

The Twitter data from the original paper is here, and the updated data can be found here. To understand more about how these features are extracted, please see the detailed explanations in our Jupyter notebook.

3.2 Evaluation Metric

In line with Curtis et al, we report the accuracy of our models using the Pearson Product-Moment Correlation Coefficient between the prediction of the model and the excess drinking data. We will also consider the mean absolute error produced from each of our models. We will compare models built on the Twitter data alone, demographics, socioeconomics, and combinations of same.

3.3 Simple Baseline

For our simple baseline we limited our feature space to only include demographic and socioeconomic features (i.e. feature set 1). We trained a simple linear regression model on these raw features without performing any further feature selection nor dimensionality reduction. This simple baseline produced a Pearson-r value of 0.577 and a mean absolute error of 3.055. We'll keep these values in mind as we consider more complex models

3.4 Published Baseline

For our published baseline, we sought to emulate the linguistic model employed by Curtis et al. More involved than our simple baseline, we begin with our dataset of linguistic topics, and - after dropping null values, which reduced the number of counties by roughly 7% (from 1604 to 1488 counties) - removed low variance features (i.e. those whose family-wise error falls below a threshold with $\alpha = 60$). We then split the data into train/dev/test sets using an 80%/10%/10% demarcation. To reduce dimensionality further, we apply principal component analysis (automatically discovering the n components explaining 95% of the total variance) before learning a ridge regression model.

4 Experimental Results

4.1 Published Baseline

The result of our published baseline was a Pearson-r value of 0.702, which not only exceeds our simple baseline performance by a considerable margin (an improvement of 0.125), but even out-performs the correlation coefficient 0.65 achieved by Curtis et al. This result speaks to the power linguistic data alone holds in its predictive ability.

4.2 Extensions

We implemented two extensions to our baseline model. These are run on the two Twitter datasets (impute and dropna).

4.2.1 Extension 1: Dimensionality Reduction

In our first extension, we experimented with different methods of dimensionality reduction other than PCA. This is meant to search for a replacement for PCA – we still retained all other preprocessing methods, which was only the low variance and family-wise error filter applied before PCA initially. The variance threshold was 0 and family-wise error threshold was 60 as per the published baseline. We also continued to use a ridge regression model with parameter 1,000 as our model. We tested four different methods for dimensionality reduction on our two Twitter datasets (impute and dropna). The results are reported in Table 1.

The first method we tested was another low variance and family-wise error filter. This is the same type of filter we used earlier, but applied again only to the training data. We used a GridSearch over the range of 1 to 2000 with a step size of 25 to search for the optimal parameters. These came out to be 0 for the low variance filter and 601 in the impute set and 1376 in the dropna set for the family-wise error threshold filter. This does not work as well as PCA.

The second method we tested was random projection, which is a simple and efficient way to reduce dimensionality. The function trades a controlled amount of accuracy (as an additional variance) for a faster processing time and a smaller model size. Scikit-learn uses the Johnson-Lindenstrauss lemma to give an estimate of the minimal size of the random subspace to guarantee a bounded distortion from the random projection. However, when we tried to get an automatic discovery of the number of components, the estimate was too conservative (it actually was more than our

Dataset	Method	Dimension	MAE	Pearson R
IMPUTE	Low Variance + FWE	1471	3.344	0.488
DROPNA	Low Variance + FWE	1464	3.438	0.553
IMPUTE	Random Projection	201	3.400	0.451
DROPNA	Random Projection	100	3.432	0.525
IMPUTE	Feature Agglomeration / Distance Threshold	45	3.710	0.311
DROPNA	Feature Agglomeration / Distance Threshold	155	3.824	0.470
IMPUTE	Feature Agglomeration / GridSearch	997	3.400	0.463
DROPNA	Feature Agglomeration / GridSearch	997	3.520	0.529
IMPUTE	LDA	100	3.718	0.320
DROPNA	LDA	50	3.964	0.368

Table 1: Extension 1 Results

original number). So, we used GridSearch over the range of 1 to 300 instead to tune the number of components.

Our GridSearch results were 225 components for the impute set and 203 components for the dropna set.

The third method we tested was feature agglomeration. In Scikit-learn, the function uses hierarchichal clustering to group features that behave similarly. We searched for the optimal number of clusters in two ways – by finding a good distance threshold and through GridSearch.

With the distance threshold, we plotted dendrograms with 3 levels (see Section 8 – Appendix for the diagrams) and took the value at the lowest cluster for the distance threshold. Again, this method did not perform as well as PCA. For the imputed data, we had a threshold of 8, which gave a dimension of 45. For the dropna data, we had a threshold of 5, which gave a dimension of 155.

With the GridSearched cluster value over the range of 1 to 1000 with a step size of 4, we had 997 clusters for both sets of data.

The fourth method we tested was Latent Dirichlet Allocation (LDA). LDA is a method is usually used for topic modeling. In fact, it was used to obtain the topics from Facebook that was a part of our Twitter data. We use it on our data to group our features into a certain number of components. We ran GridSearch for the optimal number of components. However, since it takes a long time to run, we did not make many guesses (only the values of 50, 100, and 200 for each). We chose 100 as our number of components for the impute data and 50 for dropna data. Though this method did not yield great results, there is a chance that we could improve it by doing a more exhaustive search, given more computing power.

Overall, these methods of dimensionality reduction do not work very well, and none even beat the baseline. PCA outperformed all of our tests, so we continue to use it moving forwards.

4.2.2 Extension 2: Neural Network

Our second extension experimented with a neural network instead of ridge regression. We thought this might work better than regularized linear regression because neural networks can better capture the complexities that exist in the data.

We continued to use PCA as in the baseline. Since the dropna set worked better in the baseline, we tuned the model on it. We then defined a model and tested out a variety of hyperparameters. Our training function takes in the training data and labels and converts them into tensors. We then create a DataLoader with batch size of 16 (chosen over the other tested options of 8, 32, and 64). The method then trains the model over a certain number of epochs, iterating through the dataloader object each time. Our chosen optimizer is Adam, and we optimize with the mean squared error loss function. Though our results are reported as a mean absolute error, the training loss with that function did not converge. Our test function runs the model on the validation set and reports the MAE.

The number of epochs we train on is 150, since it seems like the loss has converged at that point. This is what we saw from graphing the loss per epoch.

We tested a few combinations of hyperparameters; the results of these are listed in Table 2.

We experimented with different combinations of hyperparameters. The activation functions tested are Sigmoid, ReLU, and LeakyReLU. The number

Dataset	Layers	Activation	Dropout	MAE	Pearson R
IMPUTE	1	Sigmoid	0	2.812	0.669
DROPNA	1	Sigmoid	0	2.763	0.696
IMPUTE	1	Sigmoid	0.1	2.827	0.669
DROPNA	1	Sigmoid	0.1	2.812	0.698
IMPUTE	1	ReLU	0	3.000	0.653
DROPNA	1	ReLU	0	3.052	0.632
IMPUTE	1	LeakyReLU	0	2.964	0.661
DROPNA	1	LeakyReLU	0	3.050	0.627
IMPUTE	2	Sigmoid	0	2.849	0.656
DROPNA	2	Sigmoid	0	2.776	0.685
IMPUTE	2	Sigmoid	0.1	2.859	0.648
DROPNA	2	Sigmoid	0.1	2.771	0.696
IMPUTE	2	LeakyReLU	0	3.265	0.583
DROPNA	2	LeakyReLU	0	3.072	0.643
IMPUTE	3	Sigmoid	0	2.915	0.611
DROPNA	3	Sigmoid	0	2.894	0.674
IMPUTE	3	Sigmoid	0.1	3.149	0.544
DROPNA	3	Sigmoid	0.1	3.384	0.597
IMPUTE	3	LeakyReLU	0	3.545	0.514
DROPNA	3	LeakyReLU	0	3.331	0.581
IMPUTE	3	LeakyReLU	0.1	2.947	0.618
DROPNA	3	LeakyReLU	0.1	2.926	0.678
IMPUTE	3	LeakyReLU	0.25	2.873	0.641
DROPNA	3	LeakyReLU	0.25	2.904	0.673
IMPUTE	4	Sigmoid	0	3.752	0.467
DROPNA	4	Sigmoid	0	4.024	N/A
IMPUTE	4	Sigmoid	0.25	3.715	0.475
DROPNA	4	Sigmoid	0.25	4.032	0.412
IMPUTE	4	LeakyReLU	0	3.545	0.514
DROPNA	4	LeakyReLU	0	3.331	0.581
IMPUTE	4	LeakyReLU	0.1	2.963	0.615
DROPNA	4	LeakyReLU	0.1	2.860	0.686
IMPUTE	4	LeakyReLU	0.25	2.890	0.638
DROPNA	4	LeakyReLU	0.25	2.924	0.686

Table 2: Extension 2 Results

of hidden layers tested are in the range of 1 to 4. The dropout probabilities tested are 0, 0.1, and 0.25.

The combinations tested reflect general directions of where we thought the model could improve. For example, the LeakyReLU function worked better with droupout, so we tested that with more layers and did not test Sigmoid the same way. We also preferred LeakyReLU over ReLU based on its performance on this dataset and its improvement over ReLU in general.

Overall, we found that the neural network with one hidden layer and a sigmoid activation and no dropout worked the best. This was somewhat surprising because we expected a more complex model to fit this complex data better.

4.3 Test Results

We have so far tested everything on the validation set. We now evaluate our published baseline and our best performing extension model on the test set. Though we did not tune any of our extensions on the combined dataset (sociodemographic + linguistic data), we report the results of the combined dataset run on the baseline and our best-performing model. The results are shown in Table 3.

On the test set, we beat the published baseline for all datasets in both mean absolute error and Pearson R. This is to be expected because neural networks have the capacity to capture a lot more than simple linear regression. The models on combined dataset performed the same or better than on the linguistic data alone. The only exception to this is the the DROPNA combination run on the neural network, in which the MAE is greater than that of the DROPNA set on its own, but the Pearson R is still higher. There were also two different train/test set splits between combination and linguistic only. So, we probably conclude that the models perform somewhat similarly, though perhaps better, on the combined than linguistic only data.

4.4 Error Analysis

We keep on using the same best extension model for error analysis. This model does work slightly better than the baseline.

The error from our neural net could be from a variety of factors. We could have more exhaustively tried hyperparameter options as well as combinations of hyperparameters to better fit the model. We also only tuned on the linguistic data, and not on any of the sociodemographic data.

Dataset	Model	MAE	R
IMPUTE	Base	2.966	0.624
DROPNA	Base	2.520	0.726
IMPUTE COMBO	Base	2.966	0.624
DROPNA COMBO	Base	2.520	0.726
IMPUTE	NN	2.888	0.635
DROPNA	NN	2.410	0.754
IMPUTE COMBO	NN	2.823	0.669
DROPNA COMBO	NN	2.432	0.759

Table 3: Final Test Results

We would also like to note that since the data are shuffled, we trained and tested on different sets every time. This leads to some difficulties, including that if someone ran our notebook, they will most likely would obtain different values. This might lead to larger differences for the second extension, since our GridSearch is not completely exhaustive (meaning we do take steps bigger than 1 in our search space). However, since this is the same large dataset, the MAE and Pearson R values should not be so far off.

5 Conclusion

In our final project, we replicated and attempted to improve an earlier project from the World Well-Being Project to answer the question – can we use Twitter data to better predict excessive alcohol consumption rates in US counties? We worked with socioeconomic, demographic, excessive alcohol consumption, and Twitter data, and used the mean absolute error and Pearson r to evaluate our models.

The simple baseline was a model run on demographic and socioeconomic features. The published baseline (from Curtis et al.) was a ridge regression model run on features that had undergone feature selection through a low variance and family-wise error filters and dimensionality reduction through PCA. The best results of the published baseline came from a combination of sociodemographic and linguistic features.

We attempted to beat the baseline in two ways. First, we experimented with different methods of dimensionality reduction. However, we found the PCA still worked the best. Second, we experimented with a neural network in place of ridge regression. We obtained the most improvement over the baseline by using a neural network with one hidden layer, a sigmoid activation function, and

no dropout. This was our best performing model.

We tested both the published baseline and our best performing neural network on the test set. Our neural network showed improvements in both the mean absolute error and Pearson r on both linguistic-only and combined data.

Some possible next steps we could take would be to change our features. We used the features straight from the article our baseline comes from, and we could perhaps do more topic modeling to extract better data, since it has been a while since the initial study.

Another is to tune the neural network even more and to the combined dataset. We also never tried dimensionality reduction with the combined dataset, but we expect that the results will be similar.

6 Ethics

It is worth noting here that there are ethical concerns in this space. Most importantly, it must be noted that not all people have the access or inclination to use Twitter specifically or social media in general. As such, using solely social media data to allocate public health resources runs the risk of allowing higher-status, privileged individuals to dictate scare resources. As such, it is important to consider social media data as a supplementary tool to be used in context and conjunction with other sources.

It is also important to consider that participants in studies which use social media data are rarely given informed consent by the researchers. Instead, the terms of service for a platform like Twitter make it clear that academic research is often performed using their data. This means that future work should ensure that users identities are protected by not releasing full JSON objects into public, only the tweet IDs.

7 Acknowledgements

We are grateful to the original authors of the inspirational paper, in particular to Salvatore Giorgio for answering questions and guiding us to the updated county tweet lexical bank.

8 Appendix: Extension Visuals

Figure 1: Dendrogram for IMPUTE Data
Figure 2: Dendrogram for DROPNA Data

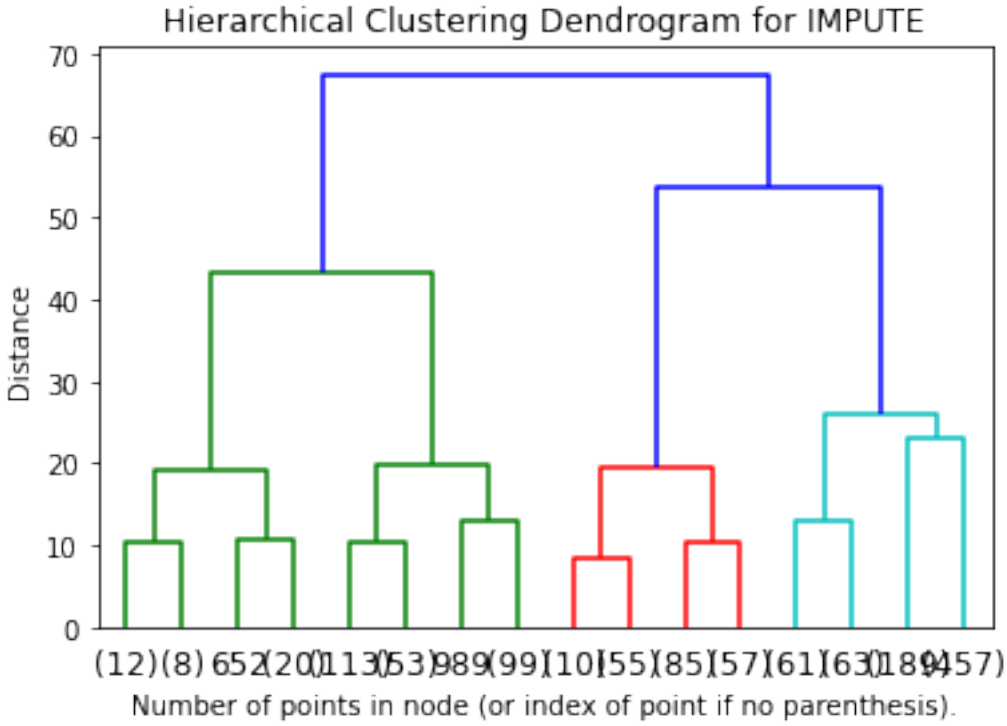

Figure 1: Dendrogram for IMPUTE Data

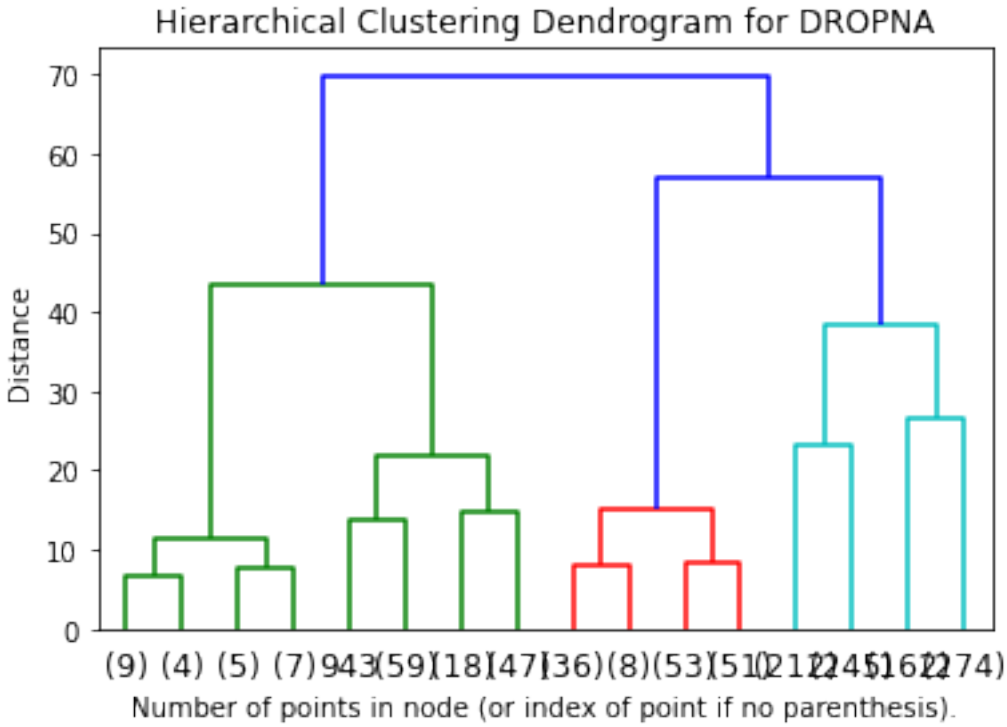

Figure 2: Dendrogram for DROPNA Data