User Income Level Classification Using Twitter Data

Baiyue Cao, Chun-Chieh Tsai, Isha Chaturvedi

New York University Center for Urban Science and progress

ABSTRACT:

Real time census data has the potential to generate timely insights for urban policy makers, allowing them to capture important urban issues such as population displacement and neighborhood change. This study, building on top of the 2015 paper "*Studying User Income through Language, Behavior and Affect in Social Media*" by Preotiuc-Pietro et al. shows how twitter data can be used to predict user income level while using random forest selected top 20 features. In this study, Gaussian Process, Support Vector Machine and Random Forest models are trained for prediction, achieving 0.42 for highest 10 class income level prediction and 0.88 for highest 3 class income level prediction. In conclusion, this paper shows how using relatively few features one can predict twitter user income level, and provides a road map for policy makers to use twitter data to generate real time insights.
[Keywords: twitter, natural language processing, income prediction]

1. INTRODUCTION

In the last ten years, as mobile devices and social media services become increasingly popular among people of all age groups and profession, different social media platforms generated significant amount of rich data documenting every aspects of our lives, especially our preferences and our behaviors. In Cities, where these social media activities are concentrated, policy makers can utilize real time social media data to generate useful insights, responding promptly to the needs of urban residents. Traditionally, policy makers rely heavily on census data to gain insights about urban resident, although conclusive and relatively accurate, the time span between census taking place fails to enable us to study our rapidly changing urban space.

This paper explores the possibility to use twitter data to predict user economic well-being of through income level as an important indicator, showing that social media platforms can provide accurate and timely socioeconomic features such as income level, which unlocks the potential and provides new channels for studying recent urban phenomenons such as population displacement as well as neighborhood gentrification. This paper, building on top of "*Studying User Income through Language, Behavior and Affect in Social Media*" (Preotiuc-Pietro; Volkova; Lampos; Bachrach; Aletras, 2015), aims to improve the original model and to provide more interpretability into the prediction algorithms through random forest feature selection. Finally, this paper will show a less computationally expensive version of the original methodology can be adapted and implemented into everyday city operations, helping policy makers to be more responsive to the need of our ever changing urban landscape.

2. LITERATURE REVIEW

As social media become an important instrument for our social interactions, literature focusing on using social media data to characterize user attributes and preferences emerged to provide new perspectives into human behaviors. For example, by parsing twitter user handles and analyzing linguistic features generated by natural language processing, gender detection can be conducted with high accuracy using a simple linear model (Vicente; Batista; Carvalho, 2015). As political debates increasingly moving to social media platforms, twitter data enables us to classify users by political ideology (Preotiuc-Pietro; Hopkins et al. 2017). A recent study shows that one can use bag-of-words, Word2Vec topics combined with sentiment analysis to generate features that capture different tweeting habits, through which one can use to infer where the user stands in the political spectrum using a linear regression.

In terms of socioeconomic attributes, twitter data was also found to be helpful in extracting occupation and social class. Sloan, Morgan, Burnap et al. used the standard occupational classification code (SOC) published by the United Kingdom Office for National Statistics to match for mentioning of occupation in user description (2015). Different job titles revealed much about the position rank as well as social class of different users. Combining these techniques, Preotiuc-Pietro, Volkova, Lampos et al used methodologies mentioned above to conduct the first large scale income level prediction study using twitter data. Their paper constructed features such as word clusters as well as user sentiment by using natural language processing tools mentioned above; From user descriptions, SOC titles were generated for each user and then mapped with their corresponding Annual Survey of Hours and Earnings (2013) released by the United Kingdom Office for National Statistics as indicators of mean yearly income. Along with demographic features such as gender, ethnicity and behavioral features such as follower-friend ratio, number of tweets per day, the paper used support vector machine and gaussian process to achieve relatively high accuracy in income level prediction. Although this study collected more than 200 conclusive features to characterize each twitter user, the data collection process requires amazon mechanical turk service to crowd source annotation for perceived features such as gender and ethnicity.

Exploring the possibility of feature selection without losing a considerable amount of prediction accuracy became the priority for real world adoption of income prediction with twitter data. Prediction models such as Random Forest shows its advantage. Relatively robust to outliers and noises, Random Forest also gives useful estimate indicators for feature importance (Breiman, 2001), which allows us to identify features that are the most relevant. Building on top of selected features, this study will then run Support Vector Machine (SVM), Gaussian Process (GP) as the original paper, however, with income level simplified into 3 and 10 even income brackets. Alongside, a Random Forest model is added for comparison. Combining Random Forest (RF) feature selection with SVM, GP and RF is found to improve computational performance as well as prediction accuracy (Adrias; Kotsiantis, 2015). Building on top of previous literature and methodologies, this study will show how a prediction model can benefit from feature selection and how a simplified model can accelerate the process of data collection, as well as income prediction, while achieving a reasonable accuracy.

3. DATA

The dataset in this study, is obtained from the research data sharing repository Figshare, published along the paper "*Studying User Income through Language, Behaviour and Affect in Social Media*"(Preotiuc-Pietro; Volkova; Lampos; Bachrach; Aletras, 2015). The raw dataset used in the above paper consists of 10,796,836 tweets from August 2014 to September 2015. After processing, the authors above arrived at a dataset that contains 5191 data entries, with twitter user individuals as twitter entry, and 285 features. Four over all categories of user features can be found in the dataset, including profiles, psycho-demographic, emotion and textual features. In the profile category, there are eight features, such as number of friends, number of tweets and follower friend ratio. In Inferred Perceived Psycho-Demographic category, features such as occupation SOC number, age and gender are included. These perceived user attributes were annotated by crowdsourcing using internet marketplace Amazon Mechanical Turk. Emotion features were generated by using text based predictive model, which gives each tweet scores across emotions such as joy, sadness, fear, disgust, surprise and anger. In aggregation, each user will have proportions of these emotions throughout the scraped tweet history as their emotion features. The textual features, are composed of 200 topic clusters, generated using Word2Vec algorithms (Section 5). As mentioned previously, the dependent variable -- user income level was created using SOC title as a proxy, mapped with their corresponding 2013 Annual Survey of Hours and Earnings. As can be seen from above, the data set contains a rich selection of features that characterizes each twitter user from all different aspects. However, the data acquisition process, which involves in human annotation, was complex and labor intensive. To better adapt this methodology for everyday city operations, this paper will show the advantage of using feature selection to pinpoint relevant user attribute, which in a large extent, would simplify the data acquisition process, as well as modeling performance.

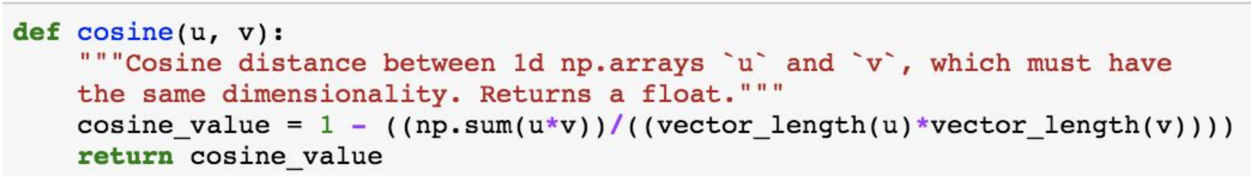
4. METHODOLOGY

The words in the tweets are converted to vectors using Word2Vec algorithm. Word2Vec aids in converting word to vector representations, also called word embeddings. After the words are converted to corresponding vector representation, a word by word matrix is constructed representing the semantic similarity between the words. The semantic similarity is calculated using cosine distance method. Cosine distance method takes overall length of the word vector in account. The cosine distance between two vectors u and v of dimension n is

$$1 - \left(\frac{\sum_{i=1}^{n} u_i \cdot v_i}{\|u\| \cdot \|v\|} \right)$$

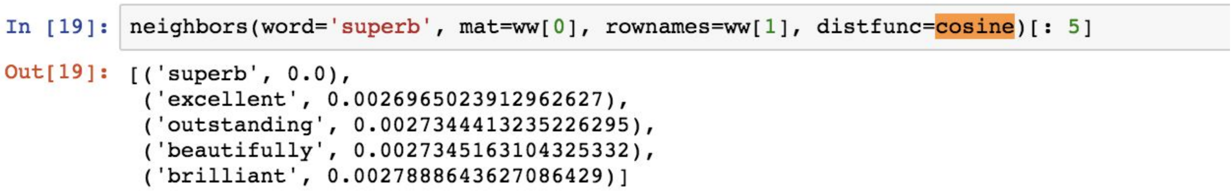
Here, the similarity part is the measuring the angle between the two vectors. The result would be same if one first normalize the two vectors using the length of the vector, and then calculate the Euclidean distance.

The function for calculating the cosine distance is :



```
def cosine(u, v):
    """Cosine distance between 1d np.arrays `u` and `v`, which must have
    the same dimensionality. Returns a float."""
    cosine_value = 1 - ((np.sum(u*v))/((vector_length(u)*vector_length(v))))
    return cosine_value
```

Finally cosine distance is used to rank all the words according to their distance from a specific word.

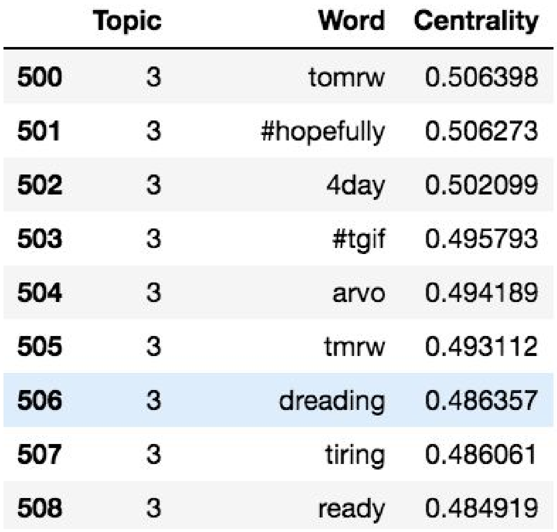


```
In [19]: neighbors(word='superb', mat=ww[0], rownames=ww[1], distfunc=cosine)[: 5]

Out[19]: [('superb', 0.0),
          ('excellent', 0.0026965023912962627),
          ('outstanding', 0.0027344413235226295),
          ('beautifully', 0.0027345163104325332),
          ('brilliant', 0.0027888643627086429)]
```

However this method is inefficient as one has to go over every word to find nearest neighbours of a particular word and some words have the potential to be clustered. Hence, spectral clustering (sklearn implementation of spectral clustering[1]) is applied on the similarity matrix to obtain 200 distinct word clusters of topics (Figure 1).

[1] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html



	Topic	Word	Centrality
500	3	tomrw	0.506398
501	3	#hopefully	0.506273
502	3	4day	0.502099
503	3	#tgif	0.495793
504	3	arvo	0.494189
505	3	tmrw	0.493112
506	3	dreading	0.486357
507	3	tiring	0.486061
508	3	ready	0.484919

Figure 1: Topic 3 containing some of the words and their significant similarities (centrality - calculated to form the clusters)

4.1. Feature Selection

As mentioned above, using 285 features to predict income increases the difficulty level of interpretability and makes the model computationally inexpensive. Hence, there is a need to conduct dimensionality reduction on the dataset. Initially, methods like principal component analysis and t-SNE methods are used to perform dimensionality reduction (Figure 2, 3). The results are absurd and are of no use. One of the reasons for this is that these methods are reducing the data from 285 features to 2 or 3 features. t-SNE can reduce to maximum of 3 features and even If PCA is used to reduce the dimensionality to more than 3 features, the results will be still absurd, as the data is of non-linear shape. Because of these reasons, dimensionality reductions method for non-linear data is required. Both Decision Trees and Random Forest (RF) are good models to know important features in a dataset. As RF generally has a higher stability and accuracy as compared to that of Decision Tree, it is used to perform feature selection with 80:20 split for training and testing data. Here, the hyperparameters of RF like “max_depth” are tuned using GridSearchCV to achieve a higher accuracy on the training data. Using these hyperparameters, RF is iterated multiple times until the set of top 20 features became constant. These top 20 features are used to perform modelling and build a dictionary.

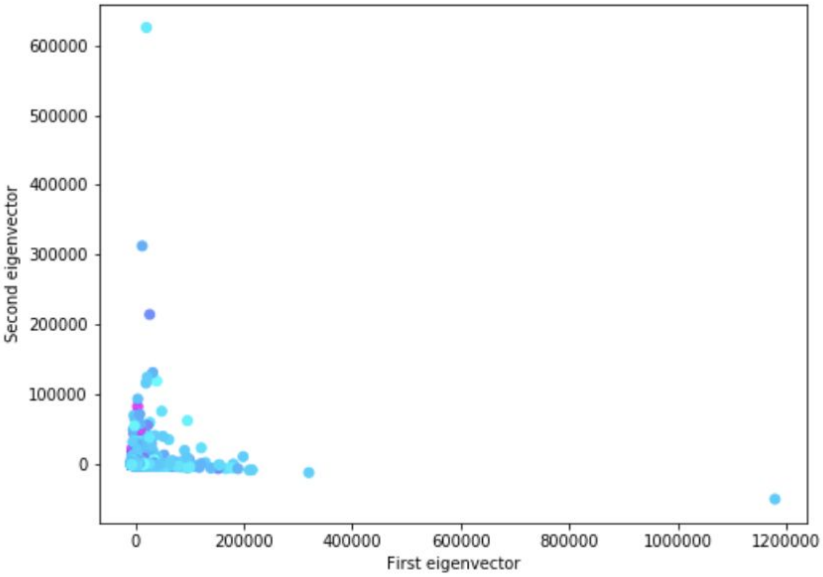


Figure 2: PCA results (2 dimensions)

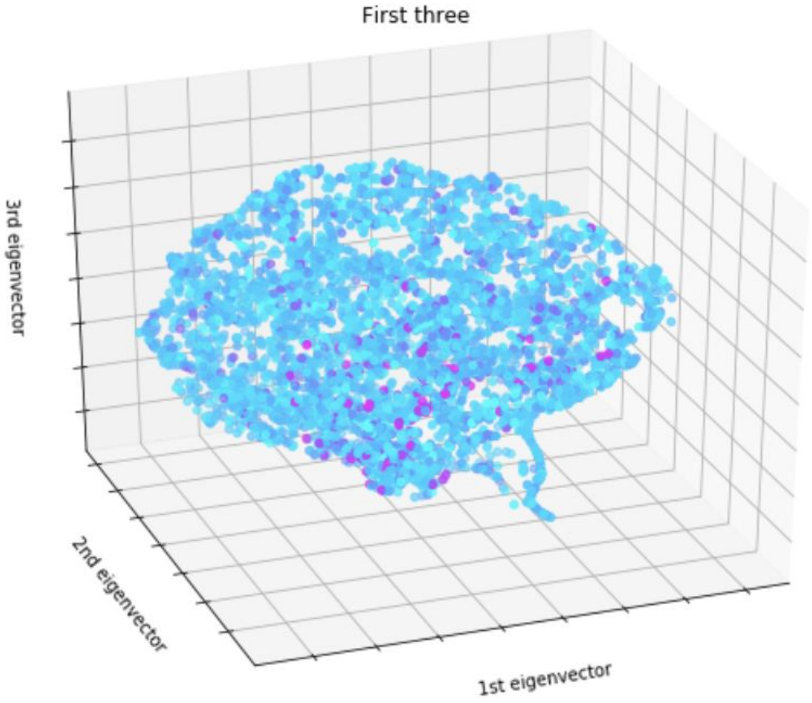


Figure 3: t-SNE results (3 dimensions)

4.2. Modeling

As mentioned in Section 5.1, feature selection resulted in a set of stable top features, which are then used in running the models. The income is classified in two ways, 3 classes and 10 classes. The three-class multi classification divided the income distribution (Figure 4) into 3 brackets: high, medium and low income. The ten-class multi-classification divided the income into 10 equal brackets. The models are run for the two cases separately.

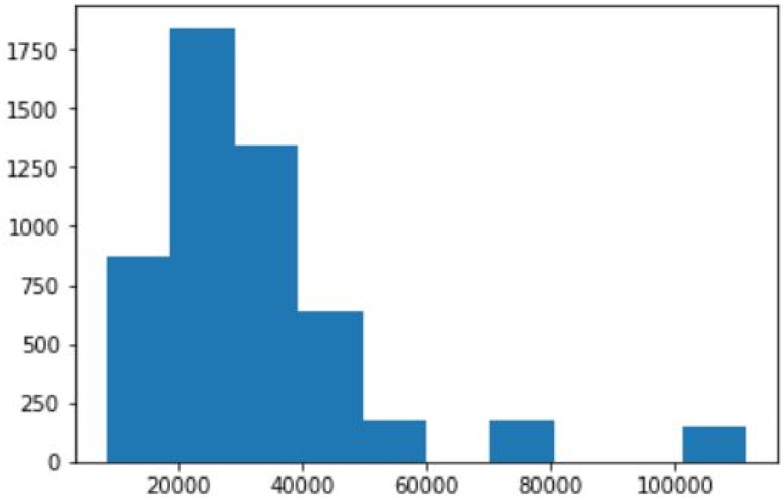


Figure 4: Histogram showing the Income Distribution

Since the data is non-linear, non-linear classification methods like Gaussian Process (GP), Support Vector Machine (SVM) and Random Forest are used for the modelling. The random state is kept same for all the three models (random_state = 999). Radial based function (RBF) kernel is used for both GP and SVM as its an effective kernel for classifying non-linear data. RBF kernel for GP is further tuned to improve the accuracy.

The Kernel for GP is: 1.* RBF(length_scale=1., length_scale_bounds=(1e-5, 1e5))

The Random Forest is tuned to take the max depth = 7 for modelling. The performance accuracy, known as "out of sample accuracy" is ran on 20% test data. Later cross-validation (cross_val_score) is run with cross validation generator or iterator = 10, on the entire dataset, to check the accuracy scores of the model on different training and test split. The mean of these accuracy scores is used for comparing the performance accuracies of the three models (GP, RF, SVM).

5. RESULTS

5.1. Feature Selection

As mentioned in Section 5.1, the RF is iterated numerous times to make sure it starts to give stable set of top 20 features. The Table 1 below shows the top 5 important features that are obtained by RF model. We further gave the features a "Category", according to the type of words they contained in their topic bucket (shown as Examples in Table 1). This tables shows that categories Hair/Cosmetics, Corporate/Business professions etc., are important features to predict income. It is interesting to see that the dependence of these features on income, holds true even in real world scenario. People in corporate and legal professions often have income. Similarly, people who have the rich accessibility of cosmetics, tend to belong to high-income classes. However, to understand the relation between each individual topic and income, more information about each category is required. Nevertheless, these categories definitely aid in interpreting the used twitter data.

Feature	Category	Examples
Topic 173	Hair/Cosmetics	Hair, comb, bleached, combed, slick
Topic 116	Arts and crafts	Archival, stencil canvas, minimalist, illustration
Topic 124	Corporate/Business professions	Consortium, institutional, firm's, acquisition, enterprises
Topic 107	Criminal Justice	Allegations, prosecution, indictment, alleged, convicted
Topic 139	Politics/Elections	Republican, democratic, gop, congressional, judiciary

Table 1: Set of top 5 features and their respective categories

5.2. Modeling

The cross-validation scores (or accuracy) for the three models are shown in Table 2 below. It can be seen that 3 class multi-classification performs better than the 10 class multi-classification. This is obvious because it is harder to predict accurately as the number of classes increases. Here, RF performs the best out of the three models for both the cases. The accuracy of GP is slightly higher (18% for 10 class division and 1.2% for 3 class division approximately) than that of the SVM. However, GP was far more computationally expensive than SVM, specifically for 10 class multiclassifcation (SVM took 2 hours to run, whereas GP took 6 hours for the 10 class).

	GP	**RF**	**SVM**
10 Class Accuracy	0.39	0.42	0.33
3 Class Accuracy	0.87	0.88	0.86

Table 2: Accuracies of different models for 10 and 3 class multi classification

5.3. Dictionary

Using the feature selection method in Section 5.1, and categorization approach in Section 6.1, following dictionary was produced to query on Twitter. This is specifically important because often in Natural Language Processing (NLP), finding right words to scrape social media is one of the most difficult tasks. This dictionary can be used as a reference for the social scientists to fetch relevant income related tweets for income predictions. The dictionary has word examples, which can be easily be expanded by the user as long as the added word examples falls in a specific category in the dictionary.

Category	**Word examples (for querying on Twitter)**
Hair/Cosmetics	Hair, comb, bleached, combed, slick
Arts and crafts	Archival, stencil canvas, minimalist, illustration
Corporate/ Business professions	Consortium, institutional, firm's, acquisition, enterprises
Criminal Justice	Allegations, prosecution, indictment, alleged, convicted
Politics/Elections	Republican, democratic, gop, congressional, judiciary',
Activism/Volunteership/Charity	Advocacy, organization, organizations, advocates, disadvantaged
Health and medical care	Physician, neonatal, dialysis, Topic, inpatient, medical
Disease	Chemotherapy, diagnosis, disease, inflammation, diseases
Food and cuisine	Buffet, sushi, seafood, deli, diner
Government/Public policy	Privatisation, bureaucrats, draconian, safeguards
Industrial/Environmental	Hydraulic, valves, sensors, halogen, voltage
Jewellery	Rhinestone, beaded, amethyst, bead, silver
Consumer electronics	Smartphones, zdnet, cnet, eweek, next-gen

Table 3: Dictionary of important words to query twitter data for income prediction

6. CONCLUSIONS

From the results shown above, it can seen that for 10 class income level classification, random forest out perform SVM and GP, with an accuracy of 0.42. For 3 class income level classification, accuracy scores are significantly higher because of the less granular income brackets, and the three models used showed similar accuracy scores. In conclusion, this study shows that even after feature selection, the twitter user income level can be predicted with reasonable accuracy. More importantly, feature importance shows that topic clusters are more relevant for predicting user income, especially with topics such as corporate, art, politics and philanthropy, which encodes much information about level of language capability, professional status, as well as education level. Using these word topics as a dictionary, policy makers will identify target keywords needed for income prediction and have more guidance in vectorizing tweets. Along with key topic clusters, only user age, gender, and follower friend ratio need to be generated to complete the model. Instead of using human annotation for these perceived user attributes, these features can be generated by using automated gender detection (Vicente; Batista; Carvalho; 2015) and age detection (Chamberlain; Humby; Deisenroth, 2017), which saves computational and labor resources. As a result, this study shows that of using relatively few twitter user features, which can be automatically generated without human annotation, one can make generic yet accurate predictions about twitter user income levels, providing valuable real time socioeconomic insights for policy makers.

7. LIMITATIONS AND FUTURE WORK

Although reasonable prediction accuracy for the used models is achieved, there are some limitations to our approach for income prediction using twitter data. First of all, the dataset used in this study is small, and because of the difficulties to obtain accurate twitter user income data, which is usually considered private information, the obtained income distribution from using SOC titles and annual income survey shows few data points with income from 60,000£ to 70,000£ and from 80,000£ to 100,000£. As a result, the model can predict lower income brackets much better than higher income brackets. Although 10 fold cross validation is done, the relatively small data set can also cause overfitting. Using the methodology in this paper, feature extraction can be narrowed down to the topic clusters mentioned in the dictionary (Table 3) in the future studies, as well as the topic clusters can be paired with simple user attributes including gender, age and follower friend ratio, hence more computation power can be devoted to mine more twitter user profiles.

Also, since the SOC titles and annual income survey that corresponds to these titles are published by the government of United Kingdom, whether the predicted income level can be applied to other regions of the world still need to be confirmed. For future studies, the same process will be applied on geotagged New York City based twitter users and prediction result with American

Community Survey data on census block level will be compared to explore the possibility of applying scaling factors to make prediction result applicable elsewhere.

The timely nature of the twitter dataset also poses some constraints to our methodology. Since the twitter data used in this study was collected from 2014 to 2015, whether topic clusters that characterize user socioeconomic attributes can change overtime would need further studies. However, the important topics derived in this study such as politics, charity and art, have been status symbols throughout the history. In future studies, as well as in city operations, it is important to regularly update the topic cluster dictionary to make sure prediction reflects current reality.

Although feature selection saved some computation power, this research progress is slowed down by using kernel algorithms such as Gaussian Process and Support Vector Machine. In the future, parallelized computed will be implemented in order to derive timely results and to update our models responsively.

CONTRIBUTIONS

The team most commonly met and worked in collaborative coding sessions, therefore it is difficult to assign roles specifically.

- *Baiyue: data cleaning, Gaussian Process, report*
- *Chun-Chieh Tsai: data cleaning, Support Vector Machine, Gaussian Process, report*
- *Isha Chaturvedi: feature selection, Random Forest, modeling, report*

REFERENCES

Aridas, C., & Kotsiantis, S. (n.d.). Combining Random Forest and Support Vector Machines for Semi-Supervised Learning.

Breiman, L. (2001). *Machine Learning.*

Lampos, V., Aletras, N., Geyti, J., Zou, B., & Cox, I. (2016). Inferring the Socioeconomic Status of Social Media Users Based on Behaviour and Language. In V. Lampos, N. Aletras, J. Geyti, B. Zou, & I. Cox. Springer, Cham.

Preoiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (n.d.). Beyond Binary Labels: Political Ideology Prediction of Twitter Users.

Preoțiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015, 9 22). Studying User Income through Language, Behaviour and Affect in Social Media. (L. Braunstein, Ed.) *PLOS ONE, 10*(9), e0138717.

Preotiuc-Pietro, Daniel; Volkova, Svitlana; Lampos, Vasileios; Bachrach, Yoram; Aletras, Nikolaos (2015): Twitter User Income Dataset. figshare. Dataset.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015, 3 2). Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. (T. Preis, Ed.) *PLOS ONE, 10*(3), e0115545.

Vicente, M., Batista, F., & Carvalho, J. (2015). Twitter gender classification using user unstructured information. *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-7). IEEE.

Chamberlain, B., Humby, C., & Deisenroth, M. (n.d.). Probabilistic Inference of Twitter Users' Age based on What They Follow.

APPENDIX

1. (Code)Feature Selection: Random Forest: https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/RandomForestRegression_FeatureSelection.ipynb
2. (Code)Feature Selection: PCA: https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/data_PCA.ipynb
3. (Code)Linear Regression:https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/Linear%20Regression.ipynb
4. (Code)Gaussian Processes(GP) for 3 classes: https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/GP%20-%20Class%203_0429_Jack.ipynb
5. Gaussian Processes(GP) for 10 classes : https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/GP%20-%20Class%2010_0430_Jack.ipynb
6. (Code)SVM for 3 classes: https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/SVM_Class%203_0430_Jack.ipynb
7. (Code)SVM for 10 classes: https://github.com/DishT/Machine_Learning_City/blob/master/4_CODE/distributed-word-reps-Isha%20Chaturvedi.ipynb
8. Top 20 features (in order of feature importance):

Feature	Category	Examples
Topic 173	Hair/Cosmetics	Hair, comb, bleached, combed, slick
Topic 116	Arts and crafts	Archival, stencil canvas, minimalist, illustration
Topic 124	Corporate/ Business professions	Consortium, institutional, firm's, acquisition, enterprises
Topic 107	Criminal Justice	Allegations, prosecution, indictment, alleged, convicted
Topic 139	Politics/Elections	Republican, democratic, gop, congressional, judiciary',
Topic 163	Activism/Volunteership/Charity	Advocacy, organization, organizations, advocates, disadvantaged
age	NA	NA
Topic 11	Health and medical care	Physician, neonatal, dialysis, Topic, inpatient, medical
Topic 105	Disease	Chemotherapy, diagnosis, disease,

		inflammation, diseases
Topic 29	Food and cuisine	Buffet, sushi, seafood, deli, diner
Topic 180	Government/Public policy	"Govt's", privatisation, bureaucrats, draconian, safeguards
Topic 66	Viral food tags	#foodtweet, yummm, #nomnom, spaghetti, sandwich
Topic 49	Animals	Lizards, pigeons, insects, rabbits, reptiles
Follower friend ratio	NA	NA
Topic 76	Industrial/Environmental	Hydraulic, valves, sensors, halogen, voltage
Topic 160	Jewellery	Rhinestone, beaded, amethyst, bead, silver
Gender: female	NA	NA
Topic 95	Consumer electronics	Smartphones, zdnet, cnet, eweek, next-gen