The University of Melbourne

COMP20008 A2

Predicting People's Subjectivity Using Their Posts On Twitter

Zhongyu(Andy) Xu

Zhen Liu

Yu Pin Gan

Malaila Safdar

October 21, 2022

Contents

1 Introduction

The aim of this project is to determine the subjectivity of a tweet.

Objective information is factual information whereas subjective information is biased on the basis of the conveyor's own opinions, prejudices and emotions about the topic at hand.

The ability to discern between objective and subjective tweets could be paramount to making a distinction between real and fake information being consumed by the general population. Subjectivity detection can be used by governmental agencies to find out which topics are polarizing the public so they can intervene early and attempt to prevent public discourse and further polarization. Similarly, businesses can benefit by keeping an ear out for opinionated posts to understand which of their products and/or services are causing a stir in the public domain and utilize the feedback constructively.

2 Pre-processing and Wrangling

2.1 Dataset Description

The Twitter dataset was initially in csv format, containing 71 features about tweets written by users (Horne et al., 2016). These features can be classified into three groups:

1. Behavioural Features: friends, followers, and percentage of messages with url, etc.

2. Linguistic Features: depth of syntax tree, and frequency of least common word used, etc.

3. Stylistic Features: punctuations, slang terms, and sentiment terms etc.

Our response (target) variable “Subjectivity Index” is synthesized from a set of stylistic features described in section 2.3.

2.2 Data Cleaning

To deal with missing values, all the rows containing at least one missing value(s) are deleted. In this case, only 1 out of 5282 rows are deleted.

Since all the user IDs are unique, the cleaned dataset contains 71 features from 5281 unique people after the data cleaning process.

2.3 Regression Target Variable: Subjectivity Index

We constructed a Subjectivity Index using some of the stylistic features provided in the Twitter dataset. The aim of this index is to measure the subjectivity in different domains of written language. The higher this index, the more subjective this person’s language is.

We define the subjectivity index as (see table 1 for variable definition)

$$SubjectivityIndex = exmark + first + interj + slang + senti + active\text{–}semi$$

The rationale behind this equation is: objective writings (e.g. academic papers) usually use less exclamation marks, first-person pronouns, interjections (e.g. yeah), slang, sentiment words and active words, but semi-colons are relatively more common in objective writing than subjective writing.

Note, all the features used in creating the subjectivity index have the same unit of measurement.[1] Therefore, the subjectivity index could be calculated by directly adding/subtracting these features. The index is positively skewed (Fig 1).

[1] number of words/punctuations used per tweet

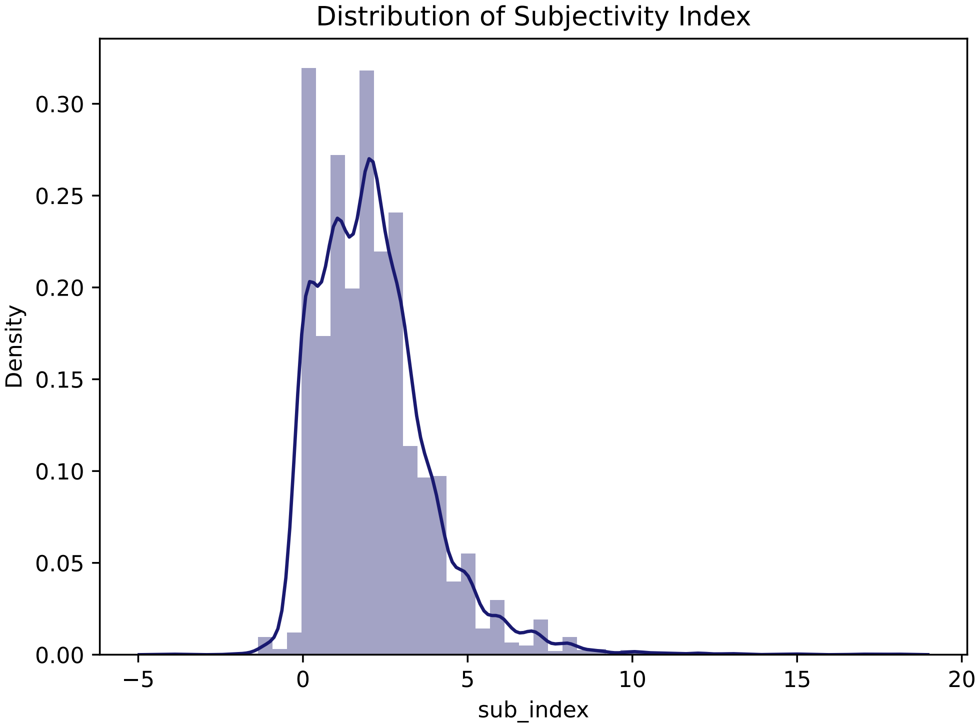


Figure 1: Density of Subjectivity Index

2.4 Combine Original Tweets and Re-Tweets For All Required Features

As we are not interested in the difference between original tweets and re-tweets, we combine the two features in the following way to gather an aggregate effect for each person:

$$combined\_feature = \begin{cases} retweetfeature, & \text{if } originalfeature = 0 \\ orginalfeature, & \text{if } retweetfeature = 0 \\ (original + retweet)/2, & \text{if otherwise} \end{cases} \tag{1}$$

2.5 Generating Indicator Variables For Categorical Variables

User types are encoded into 8 dummy variables for decision-tree analysis. For example:

$$expert = \begin{cases} 1, & \text{if } usertype = expert \\ 0, & \text{if } Otherwise \end{cases} \tag{2}$$

The domain variables are encoded into 5 dummy variables for decision-tree analysis. For example:

$$business = \begin{cases} 1, & \text{if } domain = business \\ 0, & \text{if } Otherwise \end{cases} \tag{3}$$

2.6 Dimensionality Reduction

From 71 features included in the Twitter dataset, we found only 24 features in table 1 are conceptually related to our research topic. Thus, 46 irrelevant features are deleted to reduce dimensionality.

Further, the 7 features used in constructing the subjectivity index are excluded from the analysis to avoid perfect multicollinearity, further reducing the number of features to 17.

Table 1: Features Description

Variable Name	Description
Sub_index	subjectivity index
utype	category of user (expert or not)
domain	topic of user expertise or interest
followers	number of followers on Twitter
friends	number of friends on Twitter
total_tweets	total number of tweets during time frame
years	number of years the users Twitter account has existed
chars_combined	average number of characters per tweet
quesmark_combined	average number of question marks per tweet
exmark_combined	average number of exclaimation marks per tweet
semi_combined	average number of semi-colons per tweet
punc_combined	average number of punctuations per tweet
tagpermsg_combined	average number of hashtags per tweet
mentpermsg_combined	average number of mentions per tweet
percent_msgwithment_combined	percent of tweets that have at least 1 mention
percent_msgwithtag_combined	percent of tweets that have at least 1 hashtag
percent_msgwithurl_combined	percent of tweets that have at least 1 url
first	average number of first person terms per tweet
interj	average number of interjections per tweet
slang	average number of slang words per tweet
sentiment	average number of sentiment terms per tweet
active_vb	average number of active verbs per tweet
org_med	median depth of syntax tree for original tweets
org_med_np	median depth of noun phrase sub-syntax tree for original tweets
org_med_vp	median depth of verb phrase sub-syntax tree for original tweets

2.7 Removing Outliers

Outliers are defined as

$$outlier_i = \begin{cases} 1, & \text{if } i > (Q3 + 3 * IQR) \text{ or } i < (Q1 - 3 * IRQ) \\ 0, & \text{if } Otherwise \end{cases} \tag{4}$$

Where Q1 is the first quartile, Q3 is the third quartile, and IQR is Inter-Quartile-Range.

The outliers in the feature ‘followers’ and feature ‘friends’ are removed to ensure the results of this study could be generalised to average twitter users and are less affected by the social media influencers.

As a result, 778 people are excluded from our analysis. As an interesting side note, more than 20% of people who are identified as the outliers in ‘followers’ are also identified as the outliers in ‘friends.

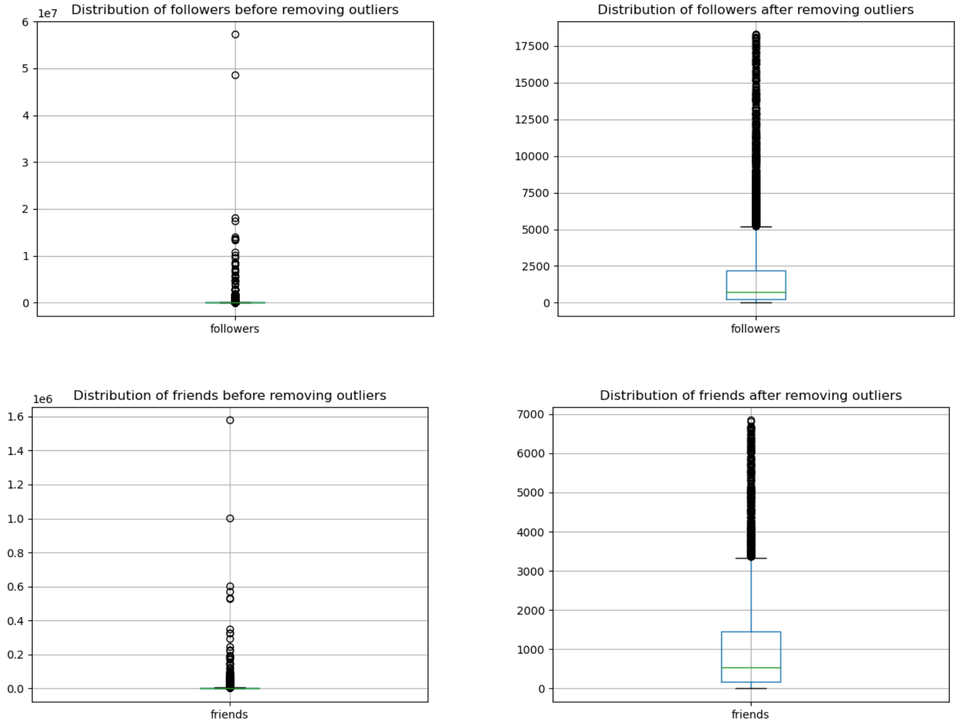


Figure 2: Followers and Friends Box Plot

As shown in Fig 2 and Fig 3, the distribution of ‘followers’ and ‘friends’ are less imbalanced after removing the outliers.

2.8 No Log Transformations

Although ‘Followers’, ‘friends’ and ‘total tweets’ are non-linearly correlated with the subjectivity index and heavily skewed (see Fig 3), no log transformations applied to these variables. This is because 1000+ people who have 0 followers or friends should not be excluded from the analysis.

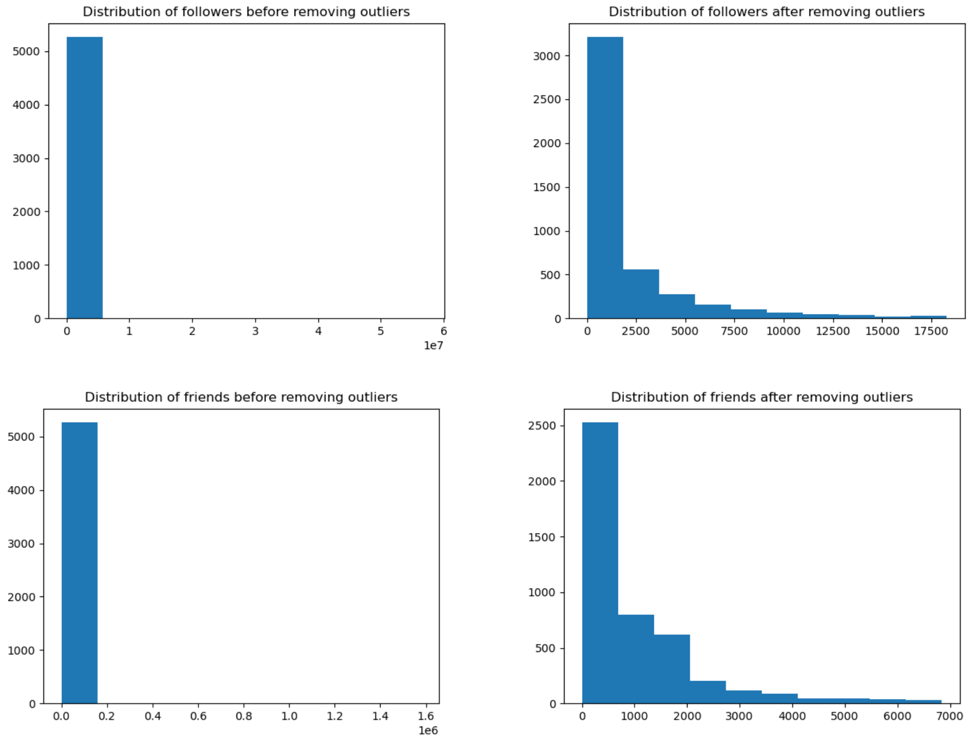


Figure 3: Followers and Friends Histogram

3 Analysis Methods and Train Test Split

As the project aim is to investigate the subjectivity level of a Twitter account based on its metrics, the data we are dealing with are primarily continuous data. Hence two analytical methods are chosen for the project aim: Simple Linear Regression and Decision Tree.

Before analysis was conducted, the data was split by an 80:20 ratio into training and testing sets. The training set was used for feature selection and model fitting, whereas the testing set was used for assessing the model's fit using the k-fold method and whether the model could make a generalizable prediction based on the training data.

4 Feature Selection

Feature selection aims to select the most significant features explaining the subjectivity index while reducing the dimensionality of the model to avoid over-fitting.

The decremental wrapper (greedy) approach is employed for both regression and decision-tree analysis.

4.1 Regression Feature Selection

User types and Domains cannot be transferred into dummy variables without losing their interpretation in the regression model due to the perfect multi-collinearity problem (Dummy Variable Trap). Therefore, we leave the feature analysis of User types and Domains to the decision tree because they are discrete categorical variables.

Starting with all 15 eligible features, we compute the model's R2 and MSE values using the validation dataset. Then, the feature with the highest p-value in this model will be dropped. We then estimate regression model again with 14 remaining features and record its R2 and MSE. This process continues until only 2 features are left. The recording results are in Fig 4.

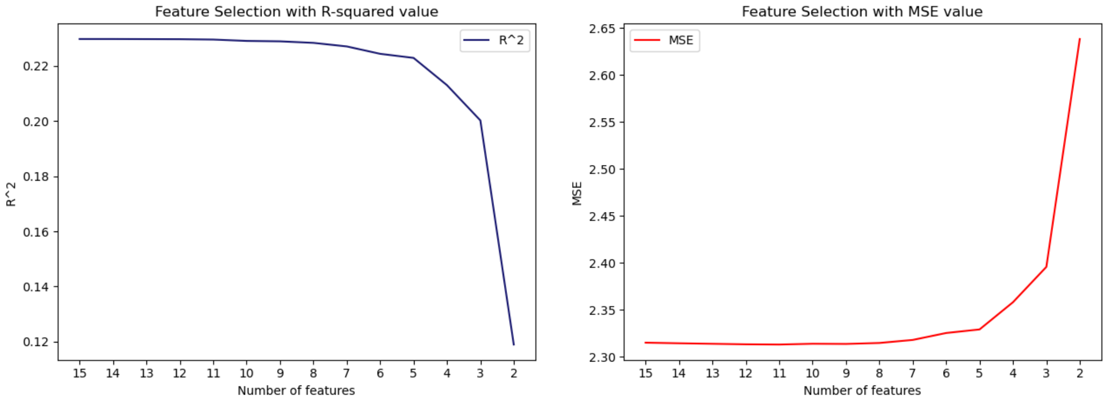


Figure 4: Regression Feature Selection using R2 and MSE

As we can see from Fig 4, when five features are left, a sharp decline in R2 and a sharp increase in MSE are observed. Therefore, we stop removing the features when only five features are preserved for fitting the linear regression model.

4.2 Decision-tree Target Variable: Subjective Language

The target variable of our study is 'Subjective_Language'.

We define the subjective_language as an indicator variable

$$subjective\_language = \begin{cases} 1, & \text{if } subjectivity\_index > m \\ 0, & \text{if } Otherwise \end{cases} \tag{5}$$

where m is the median of the subjectivity index

Our target variable is to use the median of the subjectivity index to divide the people in the Twitter dataset into two equally sized groups. People with a subjectivity index higher than half of their peers are classified as preferring subjective written language, whereas those with a subjectivity index lower than 50% of their peers are classified as preferring objective written language.

4.3 Decision-Tree Feature Selection

The decremental wrapper (greedy) approach is achieved by first constructing the decision tree using all 17 features described in section 2.6 where user types and domains were encoded into 13 distinct dummy variables described in section 2.5.

Then, each of the features was iteratively and individually removed from the decision tree and the resulting model accuracy was recorded. For example, 19 models with 18 features are estimated and the one with the highest accuracy is retained as the best 18 features model.

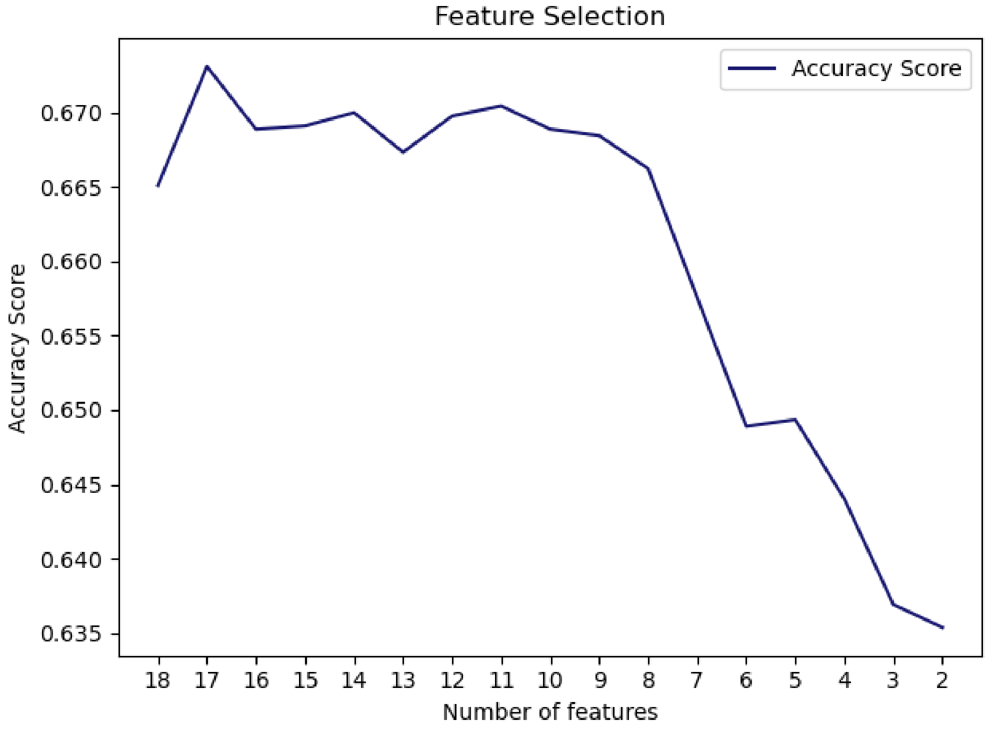


Figure 5: Decision-Tree Feature Selection using Accuracy Score

The accuracy score is picked as the evaluation metric because the target variable 'subjective_index' is balanced by definition.

We keep dropping features until the accuracy of the model decreased significantly (See figure 5). As a result, 8 features were retained in the model.

4.4 Decision-Tree Hyper Parameter Selection

Sensitivity analysis was done on the 'maximum depth of the tree' and the 'minimum sample for the tree to split'.

This was achieved by looping through possible values and getting the model with the highest accuracy score similar to section 4.3.

Then, the best maximum depth (6) and minimum split (212 samples) were used along with the selected features to construct the 'optimal' decision tree in figure 9.

Note, we also built a simpler tree to enhance interpretability, where the max depth of the tree is 3 and accuracy was compromised (see figure 10).

5 Visualisation

5.1 Scatter Plot

Following the feature selection process in section 4, we need to confirm the relationship between explanatory variables and the dependent variable. At a first glance of figure 6, all features seem linear but weakly correlated with the dependent variable. This is especially true after removing the outliers.

Among all features, the median depth of the syntax tree appears to have the strongest correlation with the subjectivity index.

These linear relationships support the usage of Pearson correlation and regression analysis.

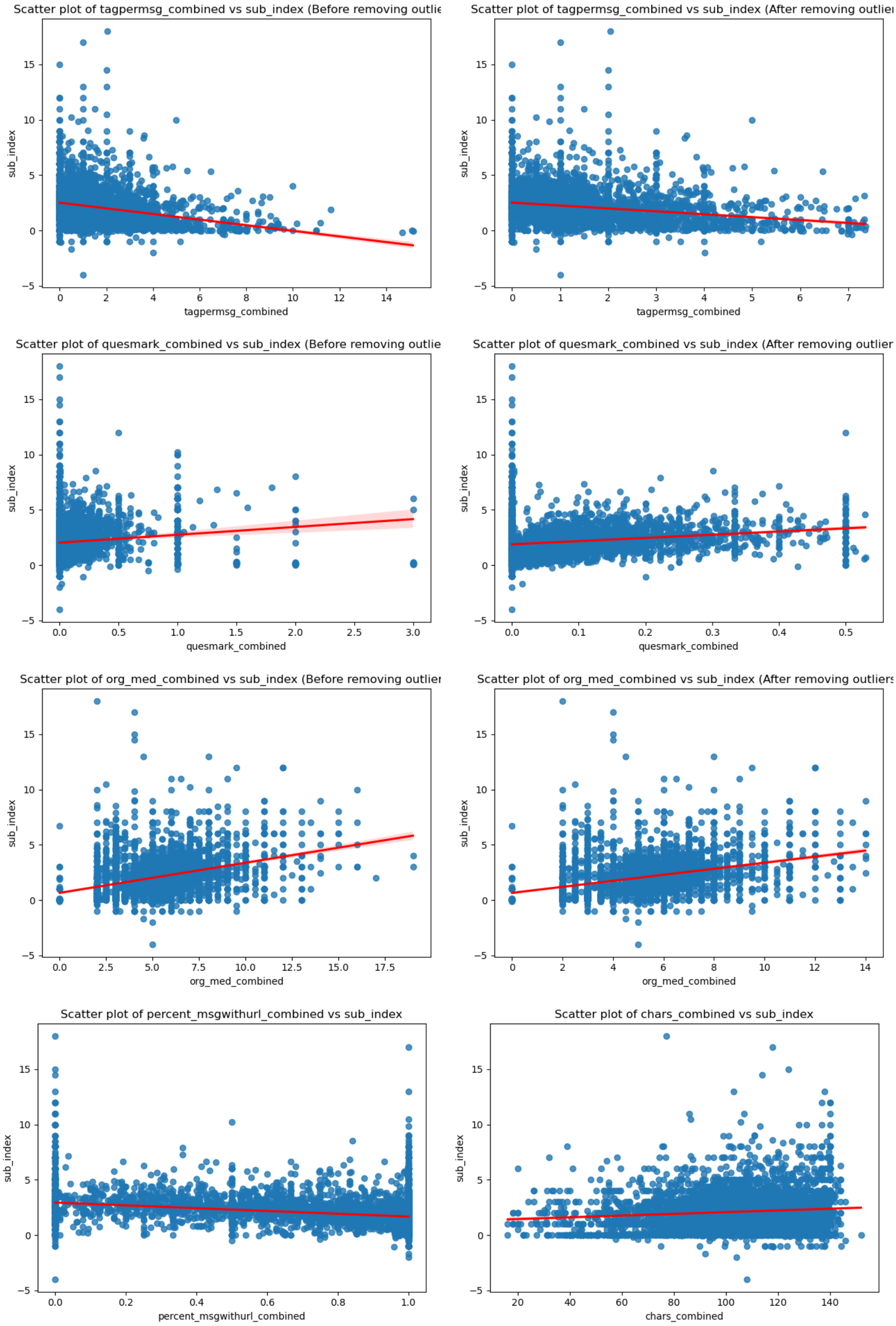


Figure 6: Scatter Plots of Selected Features

5.2 Bar Plot

To investigate the significance of user type and domain, a bar plot (figure 7) of the average subjectivity index within each group is investigated.

Users who have more than 100 messages in the technology domain are the most objective tweet writer whereas people who have exactly 10 messages in the unknown domain are the most subjective.

Note, the subjectivity of experts and non-experts is not obvious.

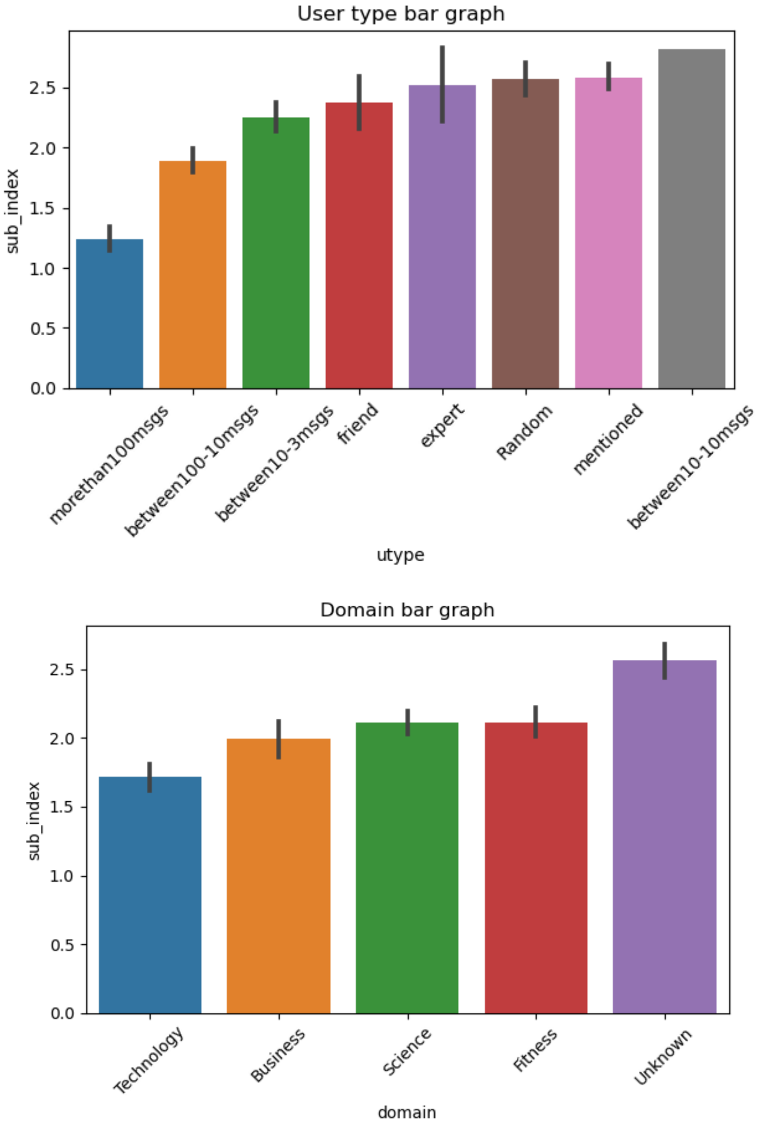


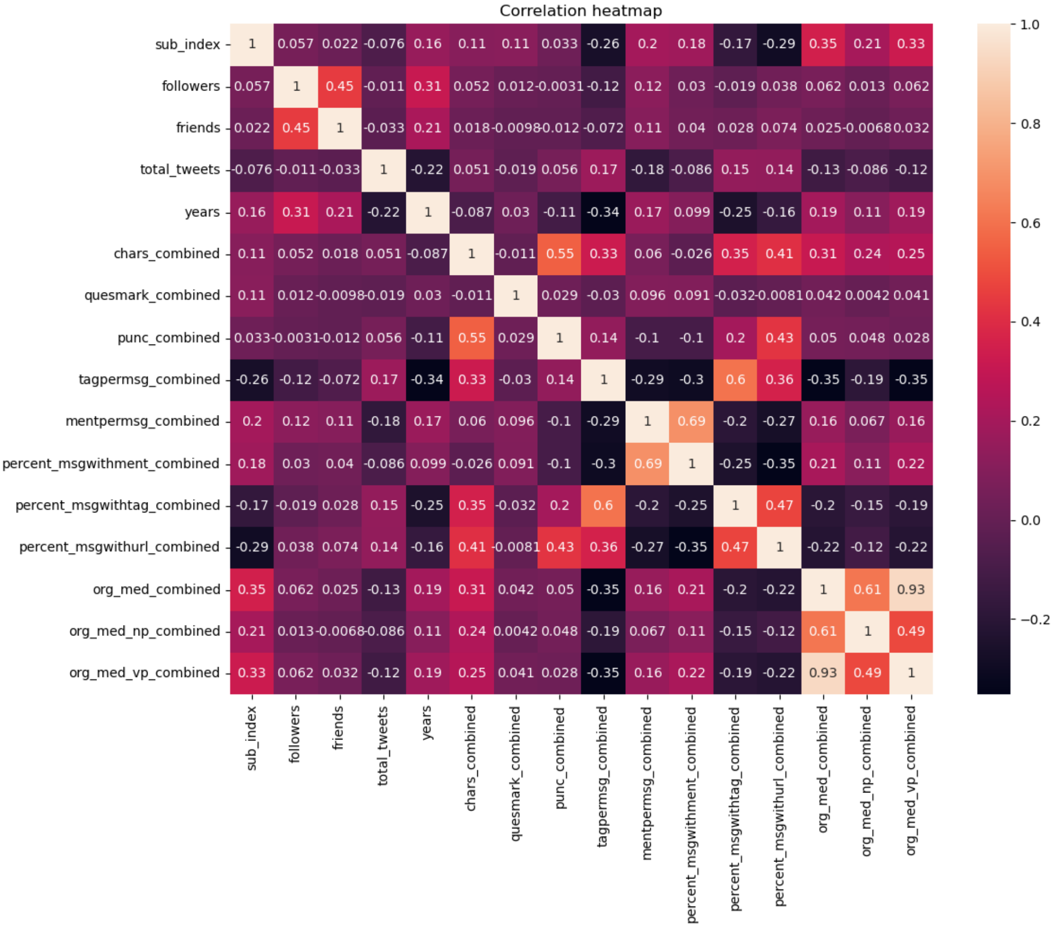
Figure 7: Average Subjectivity Index Vs Different Groups of People

5.3 Heat Map

A heatmap of continuous variables (Fig 8) is generated to get an overview of the Pearson correlation between features.

Five features have a relatively high magnitude of correlation: average tags per message ($\rho = -0.26$), percentage of messages with URL ($\rho = -0.29$), median depth of syntax tree($\rho = 0.35$), median depth of noun tree($\rho = 0.21$), and median depth of verb tree ($\rho = 0.33$). So, we pay more attention to these features.

The annotated heat map also provides an overview of the multi-collinearity between features (See table 2). Therefore, we ideally should avoid having them simultaneously in our regression model and decision-tree models.



	sub_index	followers	friends	total_tweets	years	chars_combined	quesmark_combined	punc_combined	tagpermsg_combined	mentpermsg_combined	percent_msgwithment_combined	percent_msgwithtag_combined	percent_msgwithurl_combined	org_med_combined	org_med_np_combined	org_med_vp_combined
sub_index	1	0.057	0.022	-0.076	0.16	0.11	0.11	0.033	-0.26	0.2	0.18	-0.17	-0.29	0.35	0.21	0.33
followers	0.057	1	0.45	-0.011	0.31	0.052	0.012	-0.0031	-0.12	0.12	0.03	-0.019	0.038	0.062	0.013	0.062
friends	0.022	0.45	1	-0.033	0.21	0.018	-0.0098	-0.012	-0.072	0.11	0.04	0.028	0.074	0.025	-0.0068	0.032
total_tweets	-0.076	-0.011	-0.033	1	-0.22	0.051	-0.019	0.056	0.17	-0.18	-0.086	0.15	0.14	-0.13	-0.086	-0.12
years	0.16	0.31	0.21	-0.22	1	-0.087	0.03	-0.11	-0.34	0.17	0.099	-0.25	-0.16	0.19	0.11	0.19
chars_combined	0.11	0.052	0.018	0.051	-0.087	1	-0.011	0.55	0.33	0.06	-0.026	0.35	0.41	0.31	0.24	0.25
quesmark_combined	0.11	0.012	-0.0098	-0.019	0.03	-0.011	1	0.029	-0.03	0.096	0.091	-0.032	-0.0081	0.042	0.0042	0.041
punc_combined	0.033	-0.0031	-0.012	0.056	-0.11	0.55	0.029	1	0.14	-0.1	-0.1	0.2	0.43	0.05	0.048	0.028
tagpermsg_combined	-0.26	-0.12	-0.072	0.17	-0.34	0.33	-0.03	0.14	1	-0.29	-0.3	0.6	0.36	-0.35	-0.19	-0.35
mentpermsg_combined	0.2	0.12	0.11	-0.18	0.17	0.06	0.096	-0.1	-0.29	1	0.69	-0.2	-0.27	0.16	0.067	0.16
percent_msgwithment_combined	0.18	0.03	0.04	-0.086	0.099	-0.026	0.091	-0.1	-0.3	0.69	1	-0.25	-0.35	0.21	0.11	0.22
percent_msgwithtag_combined	-0.17	-0.019	0.028	0.15	-0.25	0.35	-0.032	0.2	0.6	-0.2	-0.25	1	0.47	-0.2	-0.15	-0.19
percent_msgwithurl_combined	-0.29	0.038	0.074	0.14	-0.16	0.41	-0.0081	0.43	0.36	-0.27	-0.35	0.47	1	-0.22	-0.12	-0.22
org_med_combined	0.35	0.062	0.025	-0.13	0.19	0.31	0.042	0.05	-0.35	0.16	0.21	-0.2	-0.22	1	0.61	0.93
org_med_np_combined	0.21	0.013	-0.0068	-0.086	0.11	0.24	0.0042	0.048	-0.19	0.067	0.11	-0.15	-0.12	0.61	1	0.49
org_med_vp_combined	0.33	0.062	0.032	-0.12	0.19	0.25	0.041	0.028	-0.35	0.16	0.22	-0.19	-0.22	0.93	0.49	1

Figure 8: Heat Map of Pearson Correlation

Table 2: Selected Features that Exhibits High Correlation

Variables	Chars combined	quesmark combined	tagpermsg combined	percent msgwithurl combined	org med combined
chars_combined	1	−0.011	0.33	0.41	0.31
quesmark_combined	−0.011	1	−0.03	−0.0081	0.042
tagpermsg_combined	0.33	−0.03	1	0.36	−0.35
percent_msgwithurl_combined	0.41	−0.0081	0.36	1	−0.22
org_med_combined	0.31	0.042	−0.35	−0.22	1

6 Model Fitting and Evaluations

6.1 Linear Regression

As discussed in feature selection (section 4), five selected features are fitted into the training set's regression model in table 3. All five features are significant at 1% significant level.

On average, longer tweets, deeper syntax trees, more question marks per tweet, fewer tags and URLs per message are associated with more subjective tweets, holding else equal.

However, the correlation coefficient cannot determine the causal relationship between these variables due to the omitted variable bias and other biases. For example, education level could be a confounding factor correlated with both the subjectivity index and tweet length.

As explained in the heatmap section 5.3, the number of characters per tweet exhibits some high degrees of collinearity with other explanatory variables, inflating the standard error of the parameter of the multiple regression model.

6.2 Decision Tree

Based on the feature selection process in section 4, domain, number of characters, question marks, hashtags, URLs, depth of syntax trees and noun trees have the most explanatory power on people's subjectivity.

Dep. Variable:	sub_index	**R-squared:**	0.223
Model:	OLS	**Adj. R-squared:**	0.222
Method:	Least Squares	**F-statistic:**	172.6
Date:	Thu, 20 Oct 2022	**Prob (F-statistic):**	5.79e-165
Time:	16:42:46	**Log-Likelihood:**	-6630.6
No. Observations:	3602	**AIC:**	1.327e+04
Df Residuals:	3596	**BIC:**	1.331e+04
Df Model:	5		
Covariance Type:	HC1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.6663	0.123	5.398	0.000	0.424	0.908
chars_combined	0.0181	0.002	11.345	0.000	0.015	0.021
quesmark_combined	0.6801	0.121	5.605	0.000	0.442	0.918
tagpermsg_combined	-0.1898	0.015	-12.423	0.000	-0.220	-0.160
percent_msgwithurl_combined	-1.2572	0.095	-13.192	0.000	-1.444	-1.070
org_med_combined	0.1001	0.019	5.184	0.000	0.062	0.138

Omnibus:	1472.017	**Durbin-Watson:**	1.986
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	12544.387
Skew:	1.718	**Prob(JB):**	0.00
Kurtosis:	11.472	**Cond. No.**	504.

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Table 3: Linear Regression Model

We firstly build the optimal model using selected features and optimal model hyper-parameters in section 4. As seen in figure 9, the minimum entropy we could achieve is 0 with 38 people being classified as not subjective and 0 people being classified as subjective. Another notable achievement from the tree is entropy of 0.214 with 142 people being classified as not subjective and 5 people being classified as subjective.

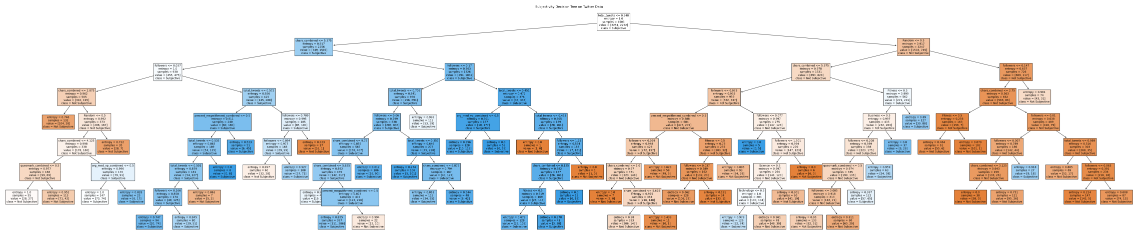


Figure 9: Optimal Decision Tree with max depth 6

Then, we build the simpler tree setting the maximum tree depth to 3 (figure 10). The minimum entropy in this tree is 0.472 where 38 people are classified as not subjective and 338 people are classified as subjective. This is saying people with fewer URLs, deeper syntax trees, and more question marks are more likely to be subjective. Another notable finding is people with more URLs, writing more than 100 messages, and fewer question marks are more likely to be objective (entropy 0.563 with 566 objective people and 86 subjective people).

These results are consistent with the regression method.

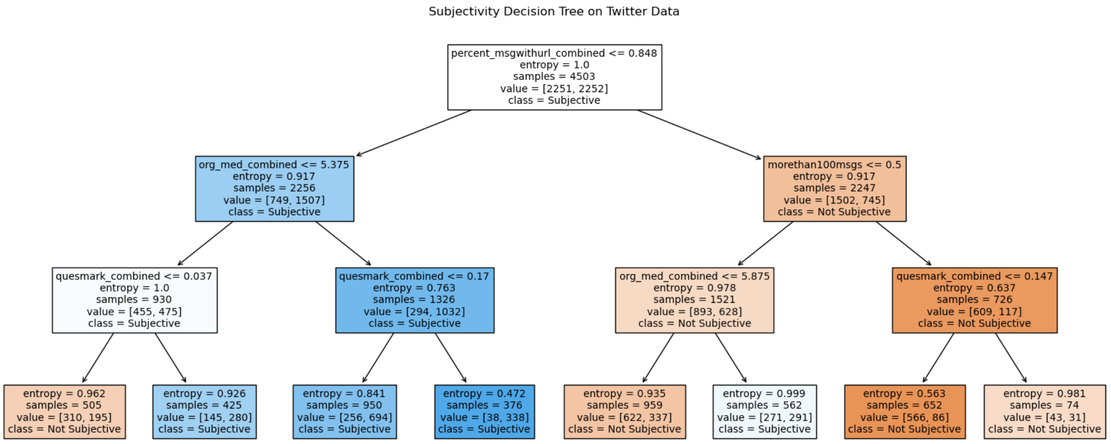


Figure 10: Simpler Decision Tree with max depth 3

Notably, as presented in figure 11, the predictions are balanced and false negatives are slightly more in the simpler model than in the optimal one.

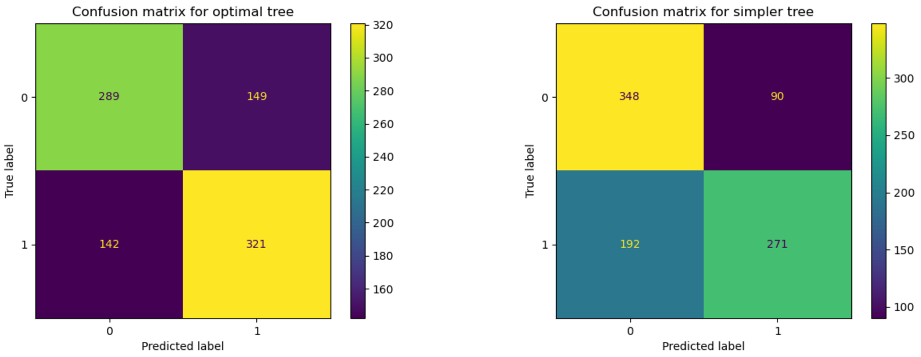


Figure 11: Confusion Matrix in these two models

7 K-Fold Cross Validation

10-Fold cross-validation is used to assess the model fit to reduce the randomness of model performance.

7.1 Regression

After 10-fold validation, approximately 22.4% (R2) of variations in the subjectivity index are explained by the features and MSE is approximately 2.34. The given 0.224 R-squared value indicates that the linear relationship between the selected features and the subjectivity index is relatively weak.

7.2 Decision Tree

This optimal tree yielded a mean accuracy of 0.703 and a standard deviation of 0.019.

The simpler tree achieved a mean accuracy of 0.692 and a standard deviation of 0.018.

The performance of the decision tree seems better than regression because of the inclusion of domain variables and the discretization of the subjectivity index.

8 Limitations and Suggestions For Future Improvements

Although a thorough investigation has been done on the dataset, the training and tests are far from exhaustive.

To start off with, the subjectivity index is calculated purely based on the syntax of sentences, which leaves out other features, such as the semantics and pragmatics which could be analyzed using more sophisticated machine learning techniques such as natural language processing that would allow us to achieve more distinctive results in terms of whether the subjectivity is positive or negative as well.

Furthermore, the correlation between individual features are high. This could be solved by using a dataset with more features, so there are more options to be considered when doing regression model fitting.

Last but not least, A tweet that is subjective to a certain extent might still have some objective factual information as well. We would need a more sophisticated approach to differentiate between the exact portions of a tweet that are objective vs those that are subjective.

References

Horne, B., Nevo, D., Freitas, J., Ji, H., & Adali, S. (2016). Expertise in social networks: How do experts differ from other users? *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 583–586.