2022

Element of Data Processing Assignment 2 Report

GROUP 62

JAHNAVI DATLA, DELPHINE DING, JINCHEN YUAN, ZOFIA WITKOWSKI-BLAKE

Contents

Aim

Our project aims to investigate whether the behavioural metrics or the language metrics of a Twitter user play an important role in determining if the user is an expert and identify which factors in the determinant metric have the highest impact.

Determining expertise is essential for applications such as user recommendation and talent seeking. It can be, determining which users to issue a blue tick to, recommending accounts to a user who has searched a term related to their area of expertise, recommendation of users to follow, as well as the management of misinformation, such as the warning label "soft moderation" that began in 2020 (1). Understanding demographics such as who uses Twitter and how they use it provides valuable insight for the company to improve its services and generate revenue through targeted advertisements.

Dataset

The dataset used in this investigation was sourced from https://github.com/rpitrust/expertisedataset (RPI_Expertise_2016_Features).

The dataset was originally in a CSV format. It contained data related to the behaviour metrics of a user such as number of followers and years on platform, and the language and linguistic style of their tweets.

The data was compiled for more than 5000 twitter users for the year 2016.
The user type was chosen as the response variable of our investigation.

Since the aim of our investigation was to find whether an expert can be identified using only their behavioural features or language features of their tweets, we split the cleaned dataset into two - a set of features describing the language and linguistic style adopted by a user in a tweet and the other describing their behaviour.

Additionally, the dataset was found to be imbalanced, ie, there was a high number of non-experts in comparison to experts. Keeping this characteristic in mind appropriate wrangling and analysis techniques were chosen.

The behavioural features used as explanatory variables in the investigation are as follows: 'utype', 'followers', 'friends', 'total_tweets', 'years', 'per_rt', 'tagpermsg', 'tagpermsg_rt', 'mentpermsg', 'mentpermsg_rt', 'urlpermsg', 'urlpermsg_rt', 'percent_msgwithtag'

The language features used as explanatory variables in the investigation are as follows: 'utype', 'chars', 'chars_rt', 'commas', 'commas_rt', 'entity','periods', 'periods_rt', 'quesmark', 'exmark', 'exmark_rt', 'colon', 'colon_rt', 'semi', 'punc', 'first', 'verb_phrase', 'noun_phrase', 'interj', 'stopwords', 'slang', 'sentiment', 'active_vb','modal_vb','negative_vb','dt', 'org_avg_d', 'pot_org','rt_avg_d', 'rt_avg_d_np', 'lexco', " topic_coherence"

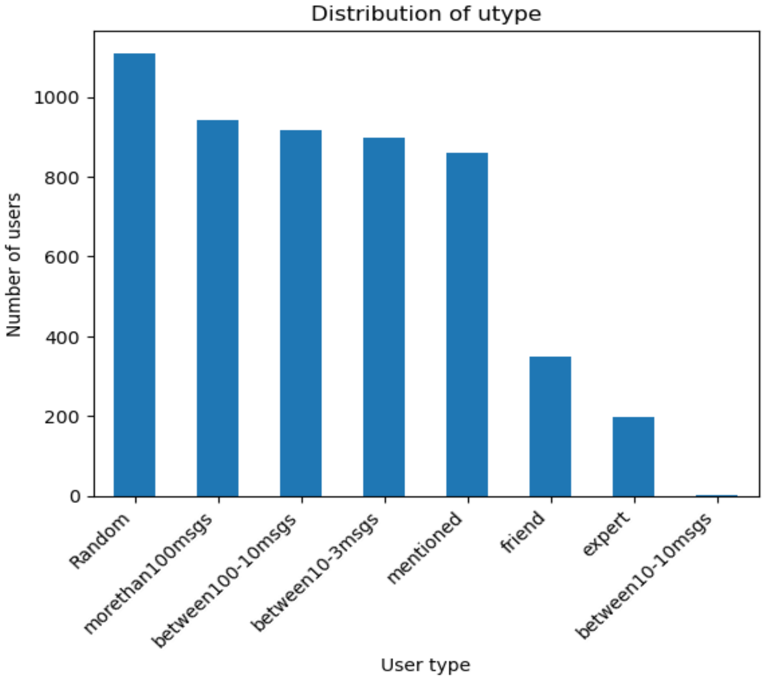
Pre-Processing and Wrangling:

Before proceeding with the analysis, the dataset was cleaned. Data cleaning is a process to fix corrupted and duplicate data within the given dataset. The original size of our dataset was 5282 rows and 158 columns and came with many unnamed columns. The last row in the dataset had a user ID but did not have any information regarding the attributes. Such rows and columns were dropped to save memory and to help us analyse the data more efficiently. After dropping all unnamed columns and rows containing NaN values the size of the dataset was reduced to 5278 rows and 71 columns.

Standardization of data was also performed to ensure the dataset was consistent. Any outliers were removed on applicable features to reduce variability bearing in mind their impact on the results.

Then, a heatmap was used to visualize the linear correlation between pairs of variables. This was done to reduce the collinearity and dimensionality of our data, any highly correlated features which were not significant were removed as they provide the same information, which makes them redundant, and can confuse our model thereby affecting its performance.

Figure 1.1



In Figure 1.1, we discovered that the distribution of expert within the whole dataset was imbalanced which leads to inaccurate prediction as the model would be biased towards the prediction of other non-expert utype.

In Figure 1.2, there was high correlation observed within the groups of variables related to depth of grammar components, sub-syntax trees in original tweets, and retweets. This is seen in the bottom right-hand corner of the matrix, with two boxes clustered in a lighter shade, with lighter colours indicating higher correlation. These features are markedly more correlated than the rest of the matrix. This means that there is correlation between these grammatical features, which makes sense as someone with complex grammatical usage may do this consistently across verb and noun phrases.

Figure 1.2

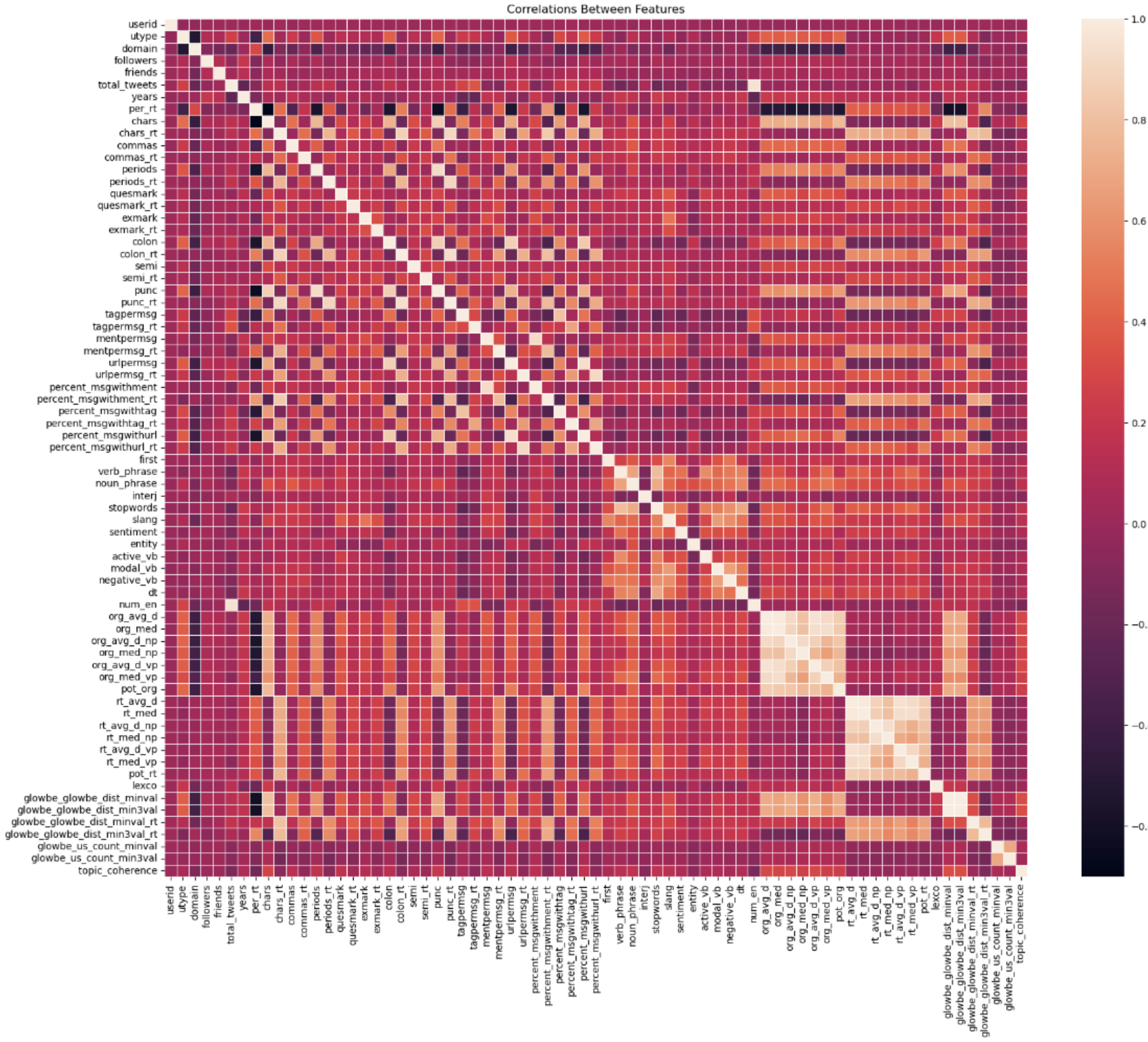
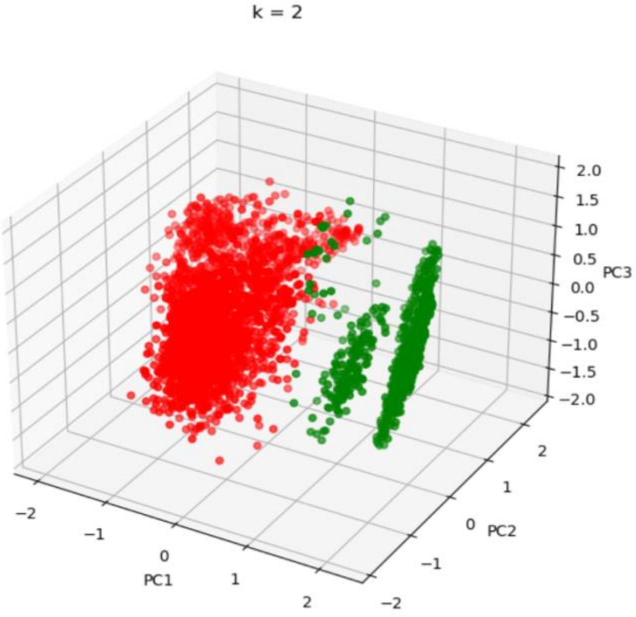


Figure 1.3



Our data is high dimensional and is an appropriate candidate for Principal Component Analysis, which is used to visualize clusters of related data. This is done by reducing dimensionality while retaining the maximum amount of variation. PCs are formed, which are directions that explains the maximum amount of variance.

In Figure 1.3, it is clear that 3 distinct clusters have formed in the 3-dimensional space formed by the 3 PCs, showing that there is structure within the data that must be accounted for the supervised learning model to proceed.

Methods:

Taking our chosen dataset and the aim of our project into consideration, we used analysis techniques which were more relevant and suitable to our requirements. Since our investigation is concerned with classifying whether a user is an expert or not, we found classification and clustering techniques to be more suitable (since the response variable user type is categorical). We chose not to use any linear models and regression techniques as they deal with data that has a linear relationship with the response variable and is preferred for continuous data.

Training-Test Split:

Prior to our analysis, we applied a Stratified K Fold Cross on the data with 3 splits to reduce the imbalance factor in the dataset and reduce bias for the majority class. Although K-Fold is a good technique it is not appropriate to use it when the dataset is imbalanced. The data was then split into training and testing sets. About 70% of data was allocated for training and 30% for testing. The training set was used for preliminary analysis and model fitting while the test set was used to assess the performance of the models and their generalisability.

Preliminary-Analysis:

To understand how each indicator was related to user type, we tried to examine their relationship. We found that most of the indicators had a nonlinear relationship with respect to user type. So we decided to consider classification models such as Logistic Regression and embedded models such as Random Forest and Decision Tree Classifier which work well with both linear and non-linear data. Another reason for choosing embedded models, particularly Random Forest was the large number of features in the dataset. Appropriate feature selection methods were also used to investigate the relationship between each feature and user type and the most significant features were chosen as input for the models.

Feature Selection:

Chi square Test:

We used a Chi-square test for association, to test the probability of each feature being uncorrelated with the response variable (utype). For each feature, a hypothesis test was conducted with a standard alpha value of 0.05, which gives a 1/20 chance of falsely rejecting the null hypothesis.

We chose this method to select the features which are most highly correlated/dependent on the response variable. Some of the most highly correlated features were followers, and topic coherence, with p-values of 9.41965136e-28 and 2.58484779e-20. This is intriguing, as it indicates that experts may be more likely to have similar numbers of followers and be similarly coherent in how they tweet.

Mutual Information (MI Test):

We also chose to look at the mutual information of features to select their significance, which is a filtering method. Mutual information quantifies the amount of information about one feature such as “domain” and "quesmark_rt".

Using these two different statistical feature selection methods allowed us to get a broader understanding of what features may be most important.

Modelling

To gain a better understanding of how accurately the significant explanatory variables determine the expert, we fit our classification models.

Firstly, we fit a Logistic Regression model to the data as it is suitable for binary classification which meets our requirement for classifying a user as an expert or non-expert. (We converted the user type to a binary format, with expert as ‘1’ and non-expert as ‘0’. This was done during the Pre-processing stage keeping the possible models we may choose in mind) Performance metrics such as macro F1 Score and AUC-PR score were chosen instead of accuracy due to imbalanced data, as accuracy fails to indicate the actual performance of the model since it does not distinguish between the numbers of correctly classified examples of different classes. In such cases the macro F1 Score which is the average of F1 score acts a better determinant since it is the harmonic mean of precision and recall values. Since precision and recall don't consider true negatives, the PR curve is not affected by data imbalance.

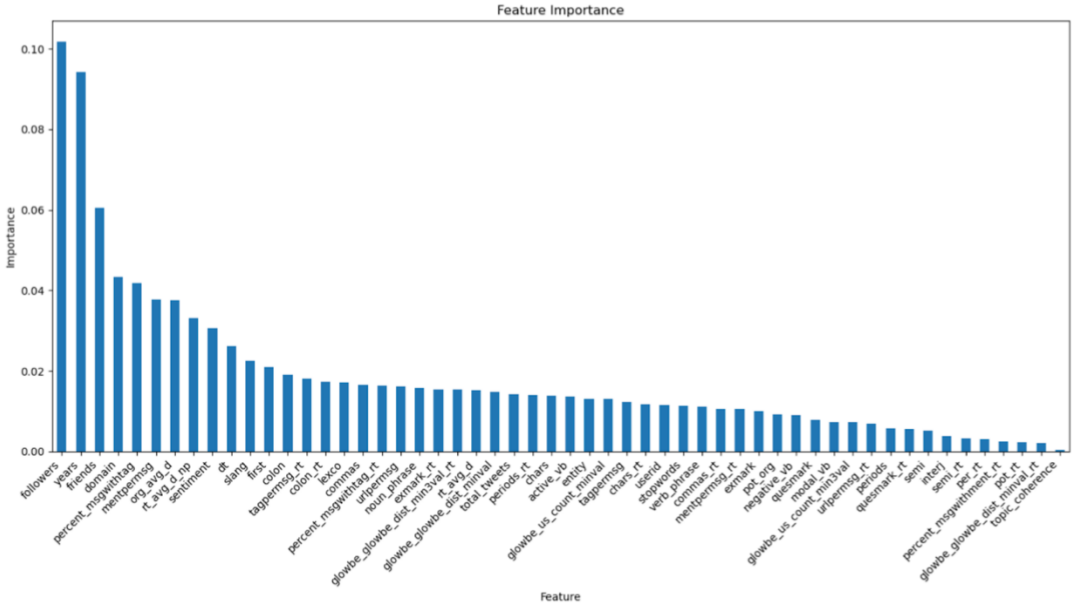
The Logistic Regression model gave an underwhelming performance, which can be due to collinearity between the independent variables or even the lack of a linear relation on the log odds of the independent variable which is a crucial assumption.

We then fit embedded models such as Decision Tree Classifier and Random Forest Classifier. Hyperparameter tuning was also done to make sure that the models worked optimally. The Decision Tree Classifier exceeded the performance of the Logistic Regression model, with Random Forest outperforming both of them.

Random Forest was considered since the algorithm internally uses bootstrapping on the training set which is preferred for the dataset chosen. The Random Forest showed improved performance with a higher macro F1 Score and AUC-PR score. Random Forest can be a powerful classification technique especially when dealing with a larger number of features.

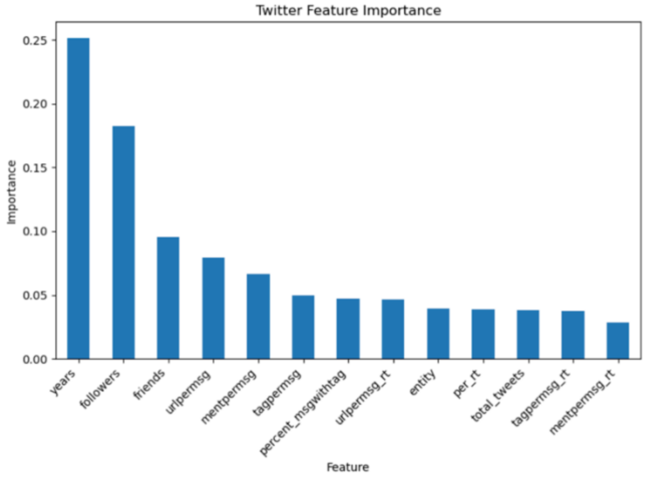
The optimum model was then fitted to our split behavioural and language feature datasets to see which type of metric was a better indicator of an expert. The behavioural data showed greater accuracy in determining the experts when compared to language data. To answer the second part of our aim, we chose to use the feature importance method of Random Forest Classifier to find the most significant behavioural features. We found that the number of years a user spent on the platform and the number of friends and followers they have are some of the most important indicators.

Figure 2.1



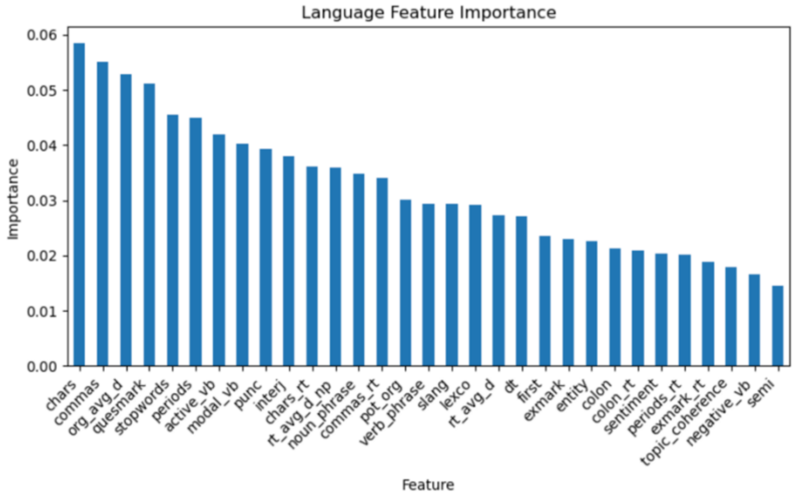
In Figure 2.1, across the whole data frame, followers, years and friends are outliers in feature importance, while the remaining are approximately consistent.

Figure 2.2



In Figure 2.2, the twitter feature dataset subset demonstrates that followers and years remain the standout predictors, more than double the importance of the remaining features.

Figure 2.3



In Figure 2.3, among the language features, characters, commas and average depth of syntax tree for original tweets were the top three predictors of an expert.

Discussion

In this section we discuss the significance of our results by interpreting our models. Recognising that we have a large amount of features, we only chose to use the most significant features that can determine an expert through feature selection. The fitting of our models revealed that while most of the features were correlated with user type, behavioural metrics showed high positive correlation, indicating that an increase in these metrics leads to an increase in the chances of the user being an expert. Across all behavioural metrics, the feature 'Followers' and 'Years' are the most crucial feature in determining if a user was an expert in their domain. Our observations were only reinforced as the model fitting over the behavioural metrics provided us similar results. These results are valuable as it is significant for both Twitter and its users.

For Twitter, this result reduces expenditure on resources such as time and other additional costs for information collection and pre-processing their datasets. It would be efficient in identifying and tracking expert users in their database. Twitter also has its own Twitter analytics system which users can access to track their statistics such as the number of tweets posted, followers and Twitter cards, which helps users boost their impact on Twitter. From the user's perspective, the Twitter Analytics will be a beneficial tool that tracks the most important metrics on their Dashboard and assists them in recognising the areas to improve, enabling them to increase their likelihood of becoming an expert in their domain if they wish to.

However, it is to be noted that there is only a slight difference between the performance of the model fitted over the language metrics and twitter metrics (0.97 and 0.99 respectively) in terms of accuracy. From this observation it can be inferred that the combination of the most important language and behavioural features can also be used as predictors.

Evaluation

In this section we examine the limitations in our findings and propose appropriate improvements and alternatives.

Firstly, as mentioned earlier the dataset is imbalanced (see Figure 1.1), as there are over 3 experts for every 100 non-experts. This means the models would be inclined to predict the majority non-expert class. Although we used methods such as stratified k-fold and bootstrapping along with changing our performance metrics, there still might be some bias existing in the models. One way to correct this is by adding more experts in the dataset, or by oversampling the minority class. Secondly, our Logistic Regression model also showed poor performance which can be due to some collinearity and non-existence of a linear log odds relation for the independent variables. Thirdly, we removed all outliers to reduce high variance in the dataset (which is considerably large), which might increase the statistical significance of the models. Also, since we used more generalized techniques of normalization for the data, these techniques might give us different results due to the imbalance factor, so we might want to consider more appropriate techniques.

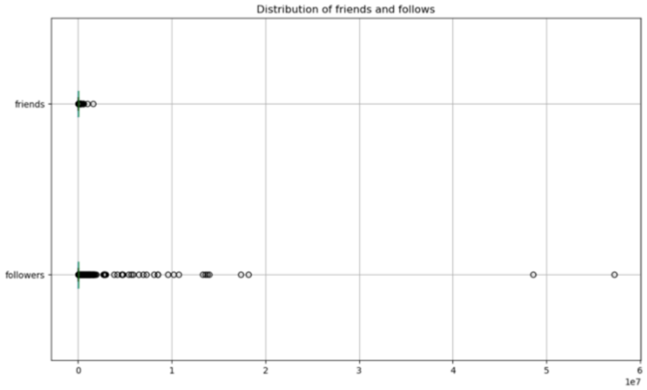
Performance for Classification Models:

Classification Models considered	Macro F1 score	AUC-PR
Logistic Regression on Full Dataset	0.62	0.510
Decision Tree on Full Dataset	0.74	0.560
Random Forest Classifier (with Bootstrapping) on Twitter dataset	0.76	0.982

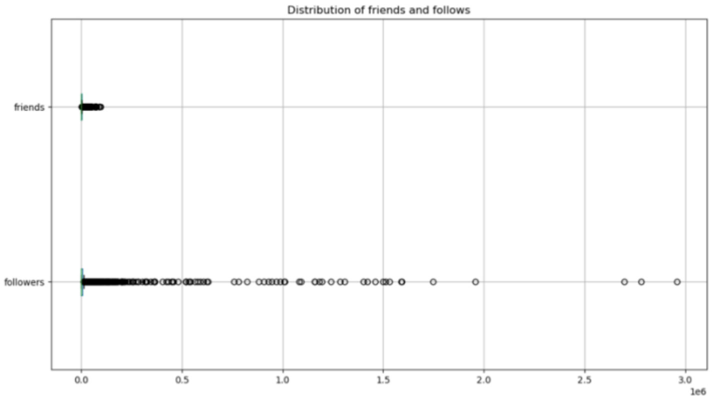
Random Forest Classifier (with Bootstrapping)	Macro F1 score	AUC-PR
On Behavioural metrics	0.94	0.999
On Language metrics	0.78	0.995

Appendix

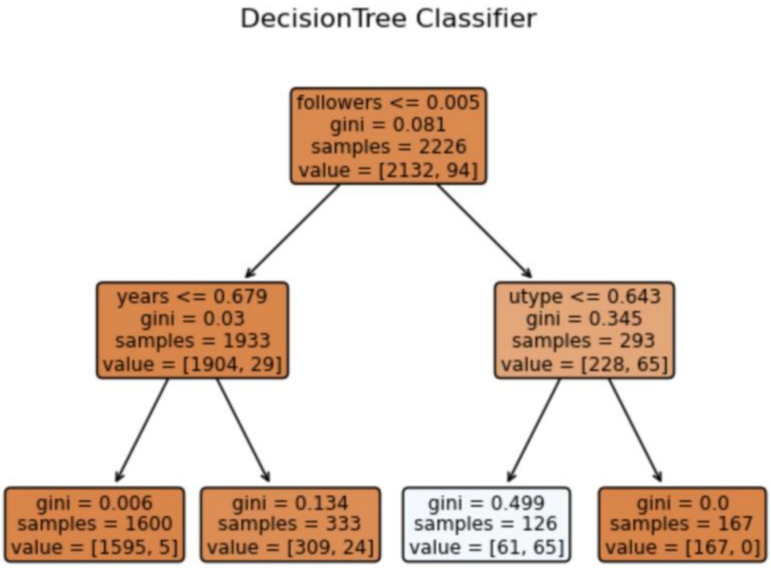
Before removal of outliers:



After removal of outliers:



Visualisation of the decision tree:



Reference:

Benjamin D. Horne, Dorit Nevo, Jesse Freitas, Heng Ji & Sibel Adali, ICWSM 2016.
Expertise in Social Networks: How Do Experts Differ From Other Users?
https://github.com/rpitrust/expertisedataset

Xu, Y., Zhou, D., & Lawless, S. (2017). User Expertise Inference on Twitter. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*.
https://doi.org/10.1145/3079628.3079646

Xu, Y., Zhou, D., & Lawless, S. (2017). Inferring your expertise from Twitter. *Proceedings of the International Conference on Web Intelligence*.
https://doi.org/10.1145/3106426.3106468

Stephen Allwright. (2022, August 9). *Metrics for imbalanced data (simply explained)*.
https://stephenallwright.com/imbalanced-data-metric/

Scikit-learn. (2018). *3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation*. Scikit-Learn.org.
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

sklearn.feature_selection.chi2 — scikit-learn 0.23.2 documentation. (n.d.). Scikit-Learn.org.
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html