***# Data exploration***

*I firstly did a basic filter of the dataset, removing columns with greater than half of the empty values, and then removing rows with null values.*

*It was also found that 'post\_retweets', 'post\_quotes' both had more than 90% 0 values, so they were removed.*

*A point plot was observed for 'user\_mentions', and it was found that there was an almost perfect average distribution from the minimum to the maximum values (0-12), so it was determined that the data in this column did not affect the results, therefore 'user\_mentions' column had been removed.*

*Point plot observations were made on the other metadata and found that extreme values existed in most of the metadata, so 5% of the outlier was dropped. Ather this operation, the data became more evenly distributed and the trend of the dot plot has become more stable.*

*It was also found that the remaining two columns of the post metadata() were positively correlated with misinformation, so we were prepared to extract the values of 0 and non-0 respectively, and designed a special function to facilitate the subsequent operation.*

***# Text preparation***

*For the Dataframes that were already preprocessed, the necessary text processing was performed: tokenlization, lowercasing, symbol removal, stopwords removal, and also lemmatisation.*

*After the first-round check, wordcloud was used to find the keywords where the misinformation was true or false separately, no significant differences were found. However words that were found as main components but not meaningful were also iterated into the list of stopwords.*

*A tf-idf calculation and dimensionality reduction process was performed and also analysed the sentiment values for each textual content.*

*Topic probabilities were predicted for the textual content. Based on the partitioning of tweets, the number of topics was designed as 7.*

***# Feature engineering***

*Using the SVD model, the dimensionality reduced Tf-idf vectors were added to df6. After several model tests, it was found that retaining only compound socre to represent the sentiment of the text had higher accuracy. Only the Topics that the text is most likely to belong to are also selected to streamline the number of columns of feature variables.*

*For the post metadata type, a self-written function: get\_dummies\_special\_values() was used to specially dummilised the metadata to distinguish whether it was zero or not.*

*Use the K-Mean model to convert the post metadata into a column of feature variables, which are markers for clusters. The elbow method was used to determine that K in K-Mean is best taken as 3.*

*For the User metadata type, principal component analysis(PCA) was used to de-quantify this data and downscale it to 3 columns.*

***#Model building & evaluation***

*I used an ensemble classifier, which is divided into two sub-models: for all feature variables, versus just for NLP feature variables.*

*For sub-model 1, I used Random Forest and a separate decision tree. This is because they are good at dealing with large data and have the ability to avoid overfitting.*

*For sub-model2, I use a Naive Bayes model with 5-Nearest Neighbors to analyse the NLP variables. Experiments proved that analysing only NLP variables resulted in much greater accuracy than global analysis, also because both models performed better in predicting NLP features.*

*I have dropped the use of Support Vector Machines in this model due to excessive training time, black-box effects & overfitting issues.*

*Since the two sub-models took different feature datasets for analysis (because of the different number of columns of X's selected), I ran each classifier individually and stored the predictions and composed new Dataframes with a dataset full of 0's and 1's that were used as predictions.*

*They will be used in a new voting classifier. This solves the problem that traditional voting classifier can only fit one feature dataset. The quadratic voting mechanism also avoids the risk of overfitting. However, due to the use of hard voting in this model, it may aggravate the black box effect.*

*This model de-predicted the results for the training data and generated the confusion matrix. The data from the matrix indicates that this model is well capable and possesses 77% accuracy and recall. Although the individual performance is not as good as the individuals in the aggregated classification, the voting mechanism stabilises the model's performance.*