Problem Statement 1: Tweet Classification

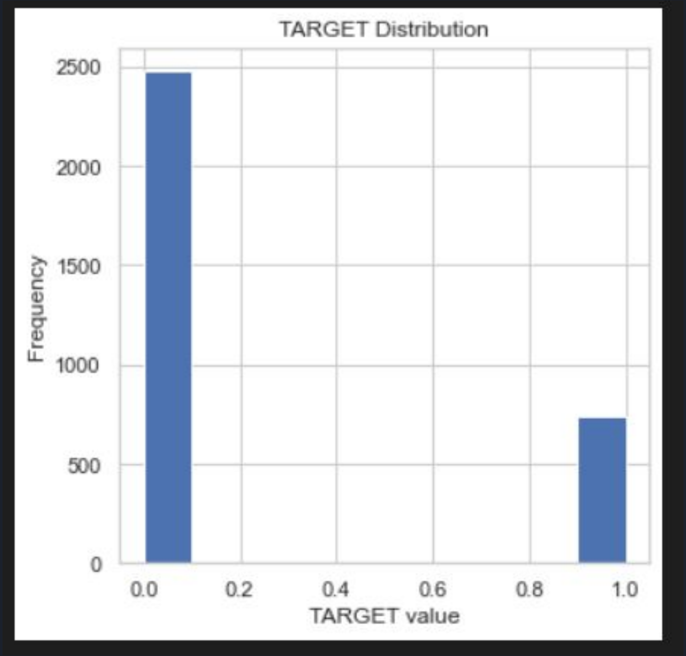
By,
TEAM: OkayXD
Prathmesh Deshmukh
Gavin Henry Lewis

Exploratory Data Analysis

1. Understanding the dataset.
2. Data Cleaning and removal of ‘NA’ values
3. Compressing the Variables and the Model:
   a. Reducing the memory impact of the variables by compressing and converting the int types in python
   b. This ‘memreduce’ function created reduces the memory used by the dataframe from 0.66 to 0.34 as seen in the notebook submitted.

Data Visualization

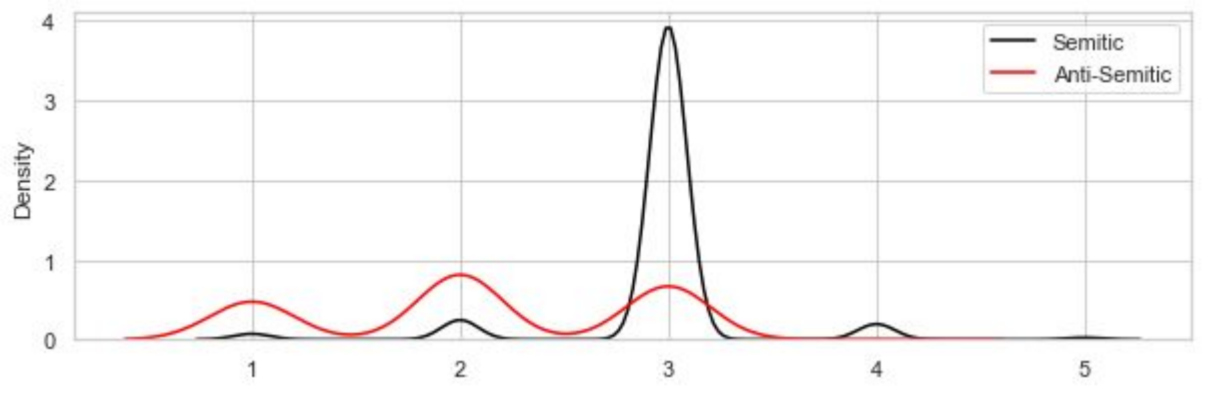
1. Understanding the skew in the Target Variable.
    a. Plotted the target variable to understand the distribution of target variable.
    b. This plot was done with respect to the 'Target' variable.



Data Visualization

2. Density Plots of the Numerical features:

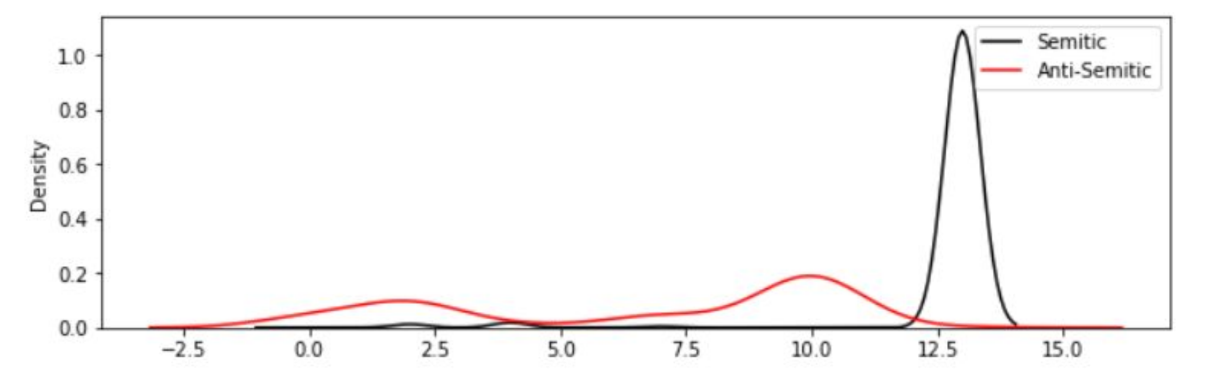
a. Sentiment Rating:



Data Visualization

2. Density Plots of the Numerical features:

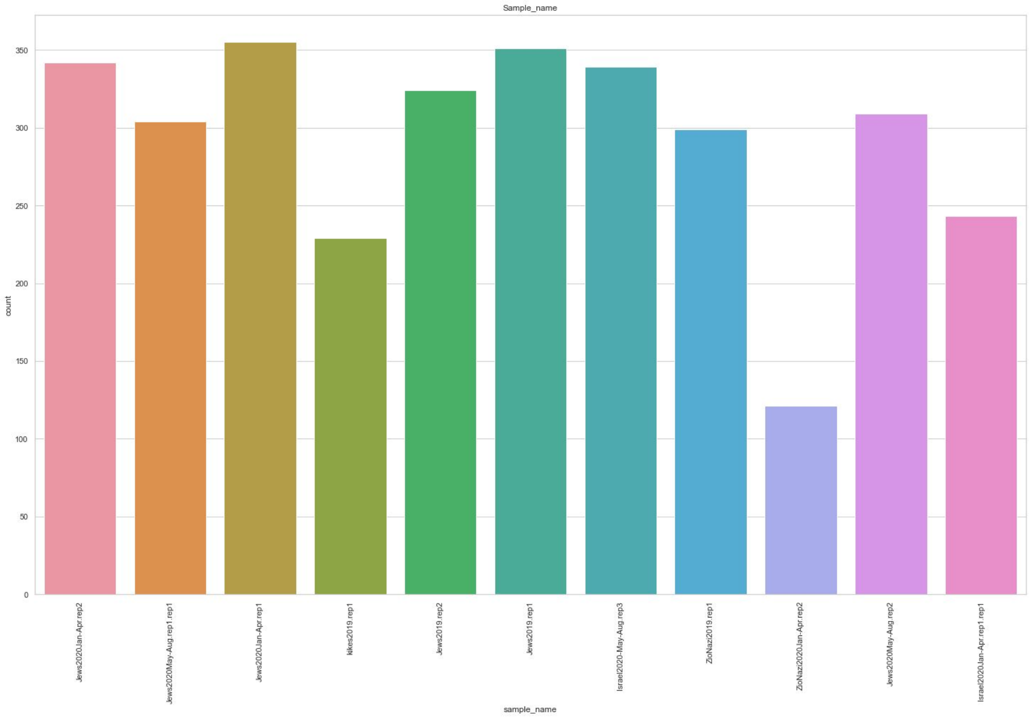
b. IHRA Section:



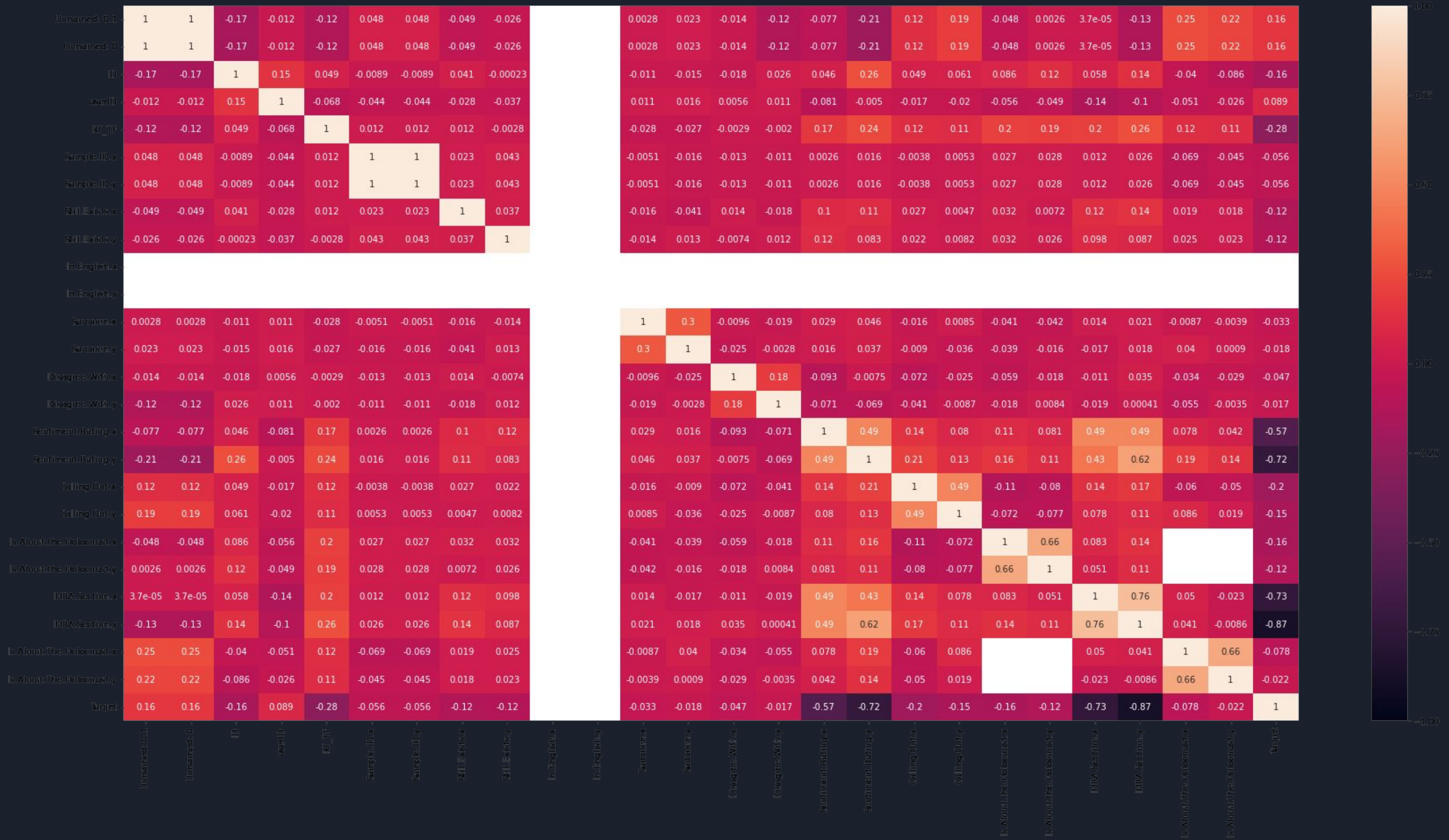
Data Visualization

3. Distribution of Categorical variables:

Distribution of Sample_name



Correlation Matrix



	Unnamed: 0.1	Unnamed: 0	ID	userID	RT_TF	Sample.ID.x	Sample.ID.y	Still.Exists.x	Still.Exists.y	In.English.x	In.English.y	Sarcasm.x	Sarcasm.y	Disagree.With.x	Disagree.With.y	Sentiment.Rating.x	Sentiment.Rating.y	Calling.Out.x	Calling.Out.y	Is.About.the.Holocaust.x	Is.About.the.Holocaust.y	IHRA.Section.x	IHRA.Section.y	Is.About.The.Holocaust.x	Is.About.The.Holocaust.y	Target
Unnamed: 0.1	1	1	-0.17	-0.012	-0.12	0.048	0.048	-0.049	-0.026			0.0028	0.023	-0.014	-0.12	-0.077	-0.21	0.12	0.19	-0.048	0.0026	3.7e-05	-0.13	0.25	0.22	0.16
Unnamed: 0	1	1	-0.17	-0.012	-0.12	0.048	0.048	-0.049	-0.026			0.0028	0.023	-0.014	-0.12	-0.077	-0.21	0.12	0.19	-0.048	0.0026	3.7e-05	-0.13	0.25	0.22	0.16
ID	-0.17	-0.17	1	0.15	0.049	-0.0089	-0.0089	0.041	-0.00023			-0.011	-0.015	-0.018	0.026	0.046	0.26	0.049	0.061	0.086	0.12	0.058	0.14	-0.04	-0.086	-0.16
userID	-0.012	-0.012	0.15	1	-0.068	-0.044	-0.044	-0.028	-0.037			0.011	0.016	0.0056	0.011	-0.081	-0.005	-0.017	-0.02	-0.056	-0.049	-0.14	-0.1	-0.051	-0.026	0.089
RT_TF	-0.12	-0.12	0.049	-0.068	1	0.012	0.012	0.012	-0.0028			-0.028	-0.027	-0.0029	-0.002	0.17	0.24	0.12	0.11	0.2	0.19	0.2	0.26	0.12	0.11	-0.28
Sample.ID.x	0.048	0.048	-0.0089	-0.044	0.012	1	1	0.023	0.043			-0.0051	-0.016	-0.013	-0.011	0.0026	0.016	-0.0038	0.0053	0.027	0.028	0.012	0.026	-0.069	-0.045	-0.056
Sample.ID.y	0.048	0.048	-0.0089	-0.044	0.012	1	1	0.023	0.043			-0.0051	-0.016	-0.013	-0.011	0.0026	0.016	-0.0038	0.0053	0.027	0.028	0.012	0.026	-0.069	-0.045	-0.056
Still.Exists.x	-0.049	-0.049	0.041	-0.028	0.012	0.023	0.023	1	0.037			-0.016	-0.041	0.014	-0.018	0.1	0.11	0.027	0.0047	0.032	0.0072	0.12	0.14	0.019	0.018	-0.12
Still.Exists.y	-0.026	-0.026	-0.00023	-0.037	-0.0028	0.043	0.043	0.037	1			-0.014	0.013	-0.0074	0.012	0.12	0.083	0.022	0.0082	0.032	0.026	0.098	0.087	0.025	0.023	-0.12
In.English.x																										
In.English.y																										
Sarcasm.x	0.0028	0.0028	-0.011	0.011	-0.028	-0.0051	-0.0051	-0.016	-0.014			1	0.3	-0.0096	-0.019	0.029	0.046	-0.016	0.0085	-0.041	-0.042	0.014	0.021	-0.0087	-0.0039	-0.033
Sarcasm.y	0.023	0.023	-0.015	0.016	-0.027	-0.016	-0.016	-0.041	0.013			0.3	1	-0.025	-0.0028	0.016	0.037	-0.009	-0.036	-0.039	-0.016	-0.017	0.018	0.04	0.0009	-0.018
Disagree.With.x	-0.014	-0.014	-0.018	0.0056	-0.0029	-0.013	-0.013	0.014	-0.0074			-0.0096	-0.025	1	0.18	-0.093	-0.0075	-0.072	-0.025	-0.059	-0.018	-0.011	0.035	-0.034	-0.029	-0.047
Disagree.With.y	-0.12	-0.12	0.026	0.011	-0.002	-0.011	-0.011	-0.018	0.012			-0.019	-0.0028	0.18	1	-0.071	-0.069	-0.041	-0.0087	-0.018	0.0084	-0.019	0.00041	-0.055	-0.0035	-0.017
Sentiment.Rating.x	-0.077	-0.077	0.046	-0.081	0.17	0.0026	0.0026	0.1	0.12			0.029	0.016	-0.093	-0.071	1	0.49	0.14	0.08	0.11	0.081	0.49	0.49	0.078	0.042	-0.57
Sentiment.Rating.y	-0.21	-0.21	0.26	-0.005	0.24	0.016	0.016	0.11	0.083			0.046	0.037	-0.0075	-0.069	0.49	1	0.21	0.13	0.16	0.11	0.43	0.62	0.19	0.14	-0.72
Calling.Out.x	0.12	0.12	0.049	-0.017	0.12	-0.0038	-0.0038	0.027	0.022			-0.016	-0.009	-0.072	-0.041	0.14	0.21	1	0.49	-0.11	-0.08	0.14	0.17	-0.06	-0.05	-0.2
Calling.Out.y	0.19	0.19	0.061	-0.02	0.11	0.0053	0.0053	0.0047	0.0082			0.0085	-0.036	-0.025	-0.0087	0.08	0.13	0.49	1	-0.072	-0.077	0.078	0.11	0.086	0.019	-0.15
Is.About.the.Holocaust.x	-0.048	-0.048	0.086	-0.056	0.2	0.027	0.027	0.032	0.032			-0.041	-0.039	-0.059	-0.018	0.11	0.16	-0.11	-0.072	1	0.66	0.083	0.14			-0.16
Is.About.the.Holocaust.y	0.0026	0.0026	0.12	-0.049	0.19	0.028	0.028	0.0072	0.026			-0.042	-0.016	-0.018	0.0084	0.081	0.11	-0.08	-0.077	0.66	1	0.051	0.11			-0.12
IHRA.Section.x	3.7e-05	3.7e-05	0.058	-0.14	0.2	0.012	0.012	0.12	0.098			0.014	-0.017	-0.011	-0.019	0.49	0.43	0.14	0.078	0.083	0.051	1	0.76	0.05	-0.023	-0.73
IHRA.Section.y	-0.13	-0.13	0.14	-0.1	0.26	0.026	0.026	0.14	0.087			0.021	0.018	0.035	0.00041	0.49	0.62	0.17	0.11	0.14	0.11	0.76	1	0.041	-0.0086	-0.87
Is.About.The.Holocaust.x	0.25	0.25	-0.04	-0.051	0.12	-0.069	-0.069	0.019	0.025			-0.0087	0.04	-0.034	-0.055	0.078	0.19	-0.06	0.086			0.05	0.041	1	0.66	-0.078
Is.About.The.Holocaust.y	0.22	0.22	-0.086	-0.026	0.11	-0.045	-0.045	0.018	0.023			-0.0039	0.0009	-0.029	-0.0035	0.042	0.14	-0.05	0.019			-0.023	-0.0086	0.66	1	-0.022
Target	0.16	0.16	-0.16	0.089	-0.28	-0.056	-0.056	-0.12	-0.12			-0.033	-0.018	-0.047	-0.017	-0.57	-0.72	-0.2	-0.15	-0.16	-0.12	-0.73	-0.87	-0.078	-0.022	1

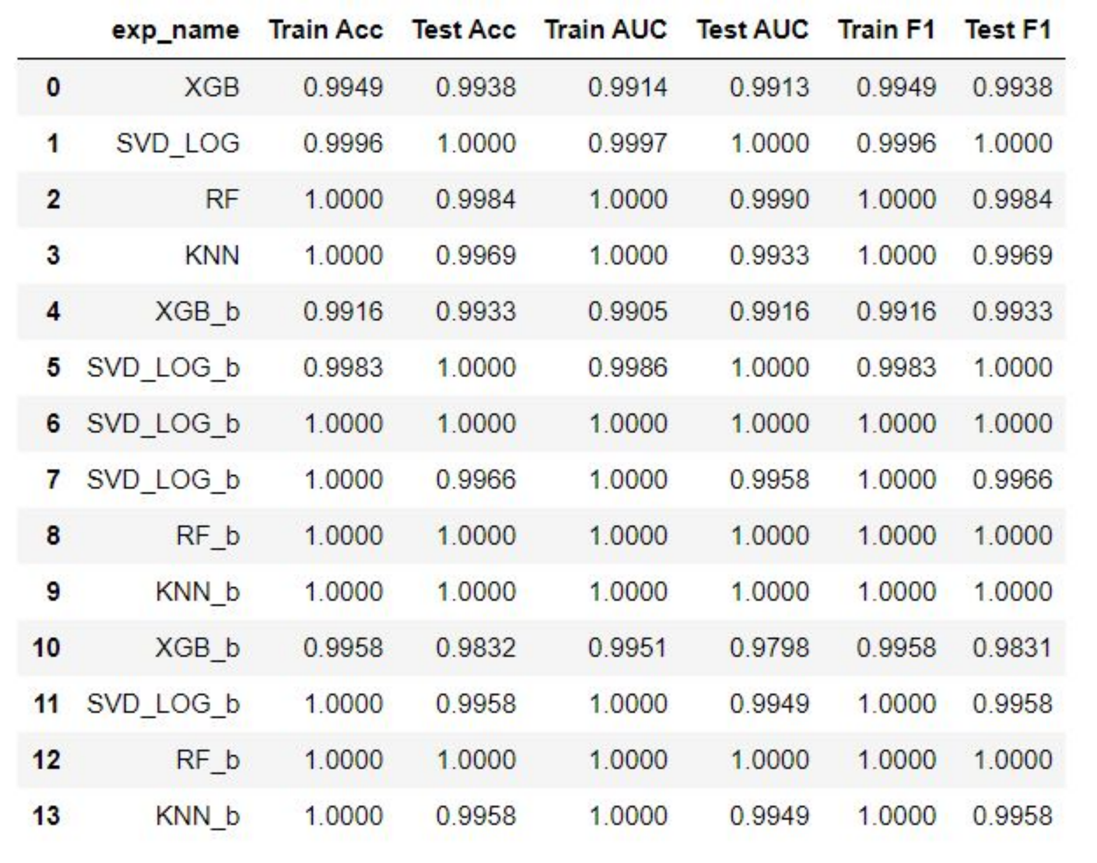
Models Used

Conducted model training with the given processed data setwith the models:

1. XGBoost
2. SVD
3. Random Forest
4. KNN

The results are stored in the expLog variable in the code.

Experiment Log



	exp_name	Train Acc	Test Acc	Train AUC	Test AUC	Train F1	Test F1
0	XGB	0.9949	0.9938	0.9914	0.9913	0.9949	0.9938
1	SVD_LOG	0.9996	1.0000	0.9997	1.0000	0.9996	1.0000
2	RF	1.0000	0.9984	1.0000	0.9990	1.0000	0.9984
3	KNN	1.0000	0.9969	1.0000	0.9933	1.0000	0.9969
4	XGB_b	0.9916	0.9933	0.9905	0.9916	0.9916	0.9933
5	SVD_LOG_b	0.9983	1.0000	0.9986	1.0000	0.9983	1.0000
6	SVD_LOG_b	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	SVD_LOG_b	1.0000	0.9966	1.0000	0.9958	1.0000	0.9966
8	RF_b	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	KNN_b	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	XGB_b	0.9958	0.9832	0.9951	0.9798	0.9958	0.9831
11	SVD_LOG_b	1.0000	0.9958	1.0000	0.9949	1.0000	0.9958
12	RF_b	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	KNN_b	1.0000	0.9958	1.0000	0.9949	1.0000	0.9958

Conclusion

From the experiment log we can conclude that Random Forest model performs better than other models considered in testing.

This is because he test AUC of this model is higher than all others on which tests were conducted.