MALIGNANT COMMENTS CLASSIFIER PROJECT

Submitted by:

Akriti Kakkar

ACKNOWLEDGMENT

I have referred to datatrained study material and python documentation for this model development.

INTRODUCTION

- Business Problem Framing

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments

uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

- # Conceptual Background of the Domain Problem

The data set includes:

Malignant: It is the label column, which indicates values 0 and 1, denoting if comment is malignant or not.

- **Highly Malignant:** It denotes comments that are highly malignant and hurtful.
- **Rude:** It denotes comments that are very rude and offensive.
- **Threat:** It contains indication of the comments that are giving any threat to someone.
- **Abuse:** It is for comments that are abusive in nature.
- **Loathe:** It describes the comments which are hateful and loathing in nature.
- **ID:** It includes unique Ids associated with each comment text given.
- **Comment text:** This column contains the comments extracted from various social media platforms.

- # Review of Literature
- # Follow the links to the dashboards:

1. https://docs.google.com/presentation/d/1DiZbjrNyupIfSy6BD9Z7Ly-y8jXmza6FDFjgFj9oAxA/edit?usp=sharing

2. https://docs.google.com/presentation/d/1-d5t6NuMkbTgQdJj4bz1rRniwxxx7SD5IPIMaCPrJ5o/edit?usp=sharing

3. https://docs.google.com/presentation/d/1AwzVWdihXRvpq8FR3Oh7Cd9Vf4jH9gAbnoVVDEn7dno/edit?usp=sharing

4. https://docs.google.com/presentation/d/1Hki6b-LDcIpBa6mlN_MCfnND2UN6XhuqfXndnzioPUg/edit?usp=sharing

5. https://docs.google.com/presentation/d/1Jnrkgl-flf61_neoaTYJ5B4UYfND5SYja9pkPgUNawM/edit?usp=sharing

6. https://docs.google.com/presentation/d/1cWBYgixIs2f2Wffow9UjOM-itoE2wpMur3iSaGvT6Sc/edit?usp=sharing

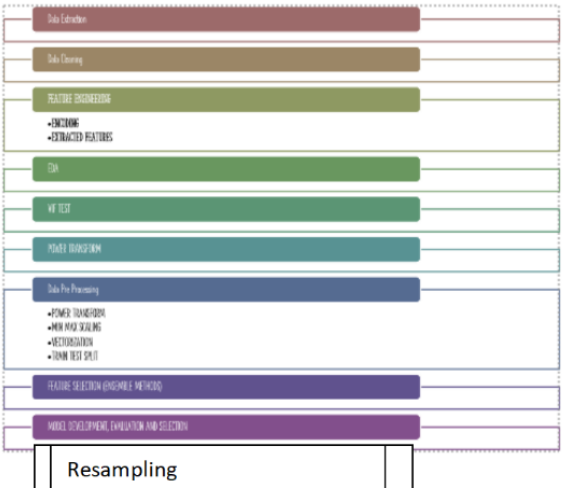
7. https://docs.google.com/presentation/d/1t38B5URUYV9QCBYBTQlDj_FdPE9a6L8iFVrmftPAM_Q/edit?usp=sharing

- Motivation for the Problem Undertaken

Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem



1. Data Extraction: Using read_ csv function of pandas library to read the data in tabulated format and analyze it.

2. Data Cleaning for missing values detection and its handiling.

3. Feature Engineering for encoding object format data and deriving more features.

4. EDA For data visualization and biasness detection:

· HEAD VIEW OF DATA

· TAIL VIEW OF DATA

· SAMPLE VIEW OF DATA

· GROUPBY EXPLORATION

· DESCRIPTIVE STATISTICS

· SCATTER PLOTS

· CORRELATION ANALYSIS

· BOX PLOTS EXPLORATION

· DESCRIPTIVE STATISTICS

· DISTRIBUTION PLOTS

5. VIF Test for multicollinearity reduction.

6. Power Transformation for standard scaling and outliers transformation.

7. Data PreProcessing for data transformation, scaling and vectorization.

8. Feature Selection (Ensemble Methods): ANOVA Test, p value, ftest, constant threshold filter to classify features based on relevance and biasness and select the most relevant features.

9. Model Development, Evaluation And Selection (Ensemble Methods and Grid Search CV) to do best hyper parameter tuning and develop low bias and low variance with right fit and minimal difference between test metrics and train metrics.

10. Resampling: To make 0 and 1 equal in number to remove biasness in learning of the models.

Data Sources and their formats

The data is provided in csv format.

Data Set Description

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes 'Id', 'Comments', 'Malignant', 'Highly malignant', 'Rude', 'Threat', 'Abuse' and 'Loathe'.

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

The data set includes:

- **Malignant:** It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
- **Highly Malignant:** It denotes comments that are highly malignant and hurtful.
- **Rude:** It denotes comments that are very rude and offensive.
- **Threat:** It contains indication of the comments that are giving any threat to someone.
- **Abuse:** It is for comments that are abusive in nature.
- **Loathe:** It describes the comments which are hateful and loathing in nature.

- **ID:** It includes unique Ids associated with each comment text given.
- **Comment text:** This column contains the comments extracted from various social media platforms.

• Data Preprocessing Done

· · Data Pre Processing

· · POWER TRANSFORM

· · MIN MAX SCALING

· · VECTORIZATION

· · TRAIN TEST SPLIT

· The data was further used for feature selection.

· Assumption made:

· • Acceptable Skewness Range Is +/-0.65

· • Acceptable VIF Score Is Than 6

· • Acceptable P Value Is Less Than 0.05

· • Variance Threshold Is 0.01

• Data Inputs- Logic- Output Relationships

Data inputs include:

comment_text_encoded comment_text_encoded 0.0

· Data Input Type: float; Min Max Scaling in the range of 0 to 1.

· Impact On Output: highly malignant, rude, threat, abuse, loathe, comment text encoded are positively correlated with label.

	malignant	highly_malignant	rude	threat	abuse	loathe	comment_text_encoded
malignant	1.000000	0.308619	0.676515	0.157058	0.647518	0.266009	0.132016

• State the set of assumptions (if any) related to the problem under consideration

· · Acceptable Skewness Range Is +/-0.65

· · Acceptable VIF Score Is Than 6

· · Acceptable P Value Is Less Than 0.05

· · Variance Threshold Is 0.01

- Hardware and Software Requirements and Tools Used

· Installation Of Anaconda Community.

· Required Installations:

· · Pandas (Within environment)

· · Numpy (Within environment)

· · Seaborn (Within environment)

· · Matplotlib (Within environment)

· · Cufflinks

· · Plotly Express

· · Sklearn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

1. Data Extraction: Using read_ csv function of pandas library to read the data in tabulated format and analyze it.

2. Data Cleaning for missing values detection and its handiling.

3. Feature Engineering for encoding object format data and deriving more features.

4. EDA For data visualization and biasness detection:

· HEAD VIEW OF DATA

· TAIL VIEW OF DATA

· SAMPLE VIEW OF DATA

· GROUPBY EXPLORATION

· DESCRIPTIVE STATISTICS

· SCATTER PLOTS

· CORRELATION ANALYSIS

· BOX PLOTS EXPLORATION

· DESCRIPTIVE STATISTICS

· DISTRIBUTION PLOTS

5. VIF Test for multicollinearity reduction.

6. Power Transformation for standard scaling and outliers transformation.

7. Data PreProcessing for data transformation, scaling and vectorization.

8. Feature Selection (Ensemble Methods): ANOVA Test, p value, ftest, constant threshold filter to classify features based on relevance and biasness and select the most relevant features.

9. Model Development, Evaluation And Selection (Ensemble Methods and Grid Search CV) to do best hyper parameter tuning and develop low bias and low variance with right fit and minimal difference between test metrics and train metrics.

10. Resampling: To make 0 and 1 equal in number to remove biasness in learning of the models.

- Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing (Total Model=5):

Model 1:

Random Forest Classifier With Intuitional Hyper Parameter Tuning

Model 2:

Random Forest Classifier With Default Hyper Parameter Tuning

Model 3:

Decision Tree Regressor With Default Hyper Parameter Tuning

Model 4:

RFC On Resampled Data With Intuitional Hyper Parameter Tuning

Model 5:

RFC On Resampled Data With Default Hyper Parameter Tuning

Run and Evaluate selected models

Follow the link to the dashboard:

https://docs.google.com/presentation/d/1EM_uwcOdGyUzsvdwFH0xh4MR0SW-_yaV2Gl3aumG2Kw/edit?usp=sharing

- Key Metrics for success in solving problem under consideration

· 1. Power Transform: To remove outliers from extremely spread out data.

· 2. VIF Scores: To reduce multicollinearity from a highly biased dataset.

· 3. Ensemble Methods: To remove over fitting in a complex dataset and finding maximum explanatory power

· 4. Resampling: To remove biasness and optimize models by improving the way of learning patterns.

- Visualizations
- Follow the links to the dashboards:

1. https://docs.google.com/presentation/d/1DiZbjrNyupIfSy6BD9Z7Ly-y8jXmza6FDFjgFj9oAxA/edit?usp=sharing

2. https://docs.google.com/presentation/d/1-d5t6NuMkbTgQdJj4bz1rRniwxxx7SD5IPIMaCPrJ5o/edit?usp=sharing

3. https://docs.google.com/presentation/d/1AwzVWdihXRvpq8FR3Oh7Cd9Vf4jH9gAbnoVVDEn7dno/edit?usp=sharing

4. https://docs.google.com/presentation/d/1Hki6b-LDcIpBa6mlN_MCfnND2UN6XhuqfXndnzioPUg/edit?usp=sharing

5. https://docs.google.com/presentation/d/1Jnrkgl-flf61_neoaTYJ5B4UYfND5SYja9pkPgUNawM/edit?usp=sharing

6. https://docs.google.com/presentation/d/1cWBYgixIs2f2Wffow9UjOM-itoE2wpMur3iSaGvT6Sc/edit?usp=sharing

7. https://docs.google.com/presentation/d/1t38B5URUYV9QCBYBTQlDj_FdPE9a6L8iFVrmftPAM_Q/edit?usp=sharing

Interpretation of the Results

Based On EDA done above in two parts, it can be concluded that all the features are relevant for making prediction because of moderate to strong correlation with the label. Even the outliers don't seem to be very high. However, the pattern in all the columns is that skewness and kurtosis are not normal. To overcome this problem and to arrive at good prediction, I will following steps:

1. Feature Scaling
2. Outliers Transformation.
3. ANOVA Test.
4. Resampling

For detailed interpretation, follow the link to the dashboards:

1. https://docs.google.com/presentation/d/1DiZbjrNyupIfSy6BD9Z7Ly-y8jXmza6FDFjgFj9oAxA/edit?usp=sharing

2. https://docs.google.com/presentation/d/1-d5t6NuMkbTgQdJj4bz1rRniwxxx7SD5IPIMaCPrJ5o/edit?usp=sharing

3. https://docs.google.com/presentation/d/1AwzVWdihXRvpq8FR3Oh7Cd9Vf4jH9gAbnoVVDEn7dno/edit?usp=sharing

4. https://docs.google.com/presentation/d/1Hki6b-LDcIpBa6mlN_MCfnND2UN6XhuqfXndnzioPUg/edit?usp=sharing

5. https://docs.google.com/presentation/d/1Jnrkgl-flf61_neoaTYJ5B4UYfND5SYja9pkPgUNawM/edit?usp=sharing

6. https://docs.google.com/presentation/d/1cWBYgixls2f2Wffow9UjOM-itoE2wpMur3iSaGvT6Sc/edit?usp=sharing

7. https://docs.google.com/presentation/d/1t38B5URUYV9QCBYBTQlDj_FdPE9a6L8iFVrmftPAM_Q/edit?usp=sharing

CONCLUSION

- Key Findings and Conclusions of the Study

Based On EDA done above in two parts, it can be concluded that all the features are relevant for making prediction because of moderate to strong correlation with the label. Even the outliers don't seem to be very high. However, the pattern in all the columns is that skewness and kurtosis are not normal. To overcome this problem and to arrive at good prediction, I will following steps:

5. Feature Scaling
6. Outliers Transformation.
7. ANOVA Test.
8. Resampling

For detailed interpretation, follow the link to the dashboards:

1. https://docs.google.com/presentation/d/1DiZbjrNyupIfSy6BD9Z7Ly-y8jXmza6FDFjgFj9oAxA/edit?usp=sharing

2. https://docs.google.com/presentation/d/1-d5t6NuMkbTgQdJj4bz1rRniwxxx7SD5IPIMaCPrJ5o/edit?usp=sharing

3. https://docs.google.com/presentation/d/1AwzVWdihXRvpq8FR3Oh7Cd9Vf4jH9gAbnoVVDEn7dno/edit?usp=sharing

4. https://docs.google.com/presentation/d/1Hki6b-LDcIpBa6mlN_MCfnND2UN6XhuqfXndnzioPUg/edit?usp=sharing

5. https://docs.google.com/presentation/d/1Jnrkgl-flf61_neoaTYJ5B4UYfND5SYja9pkPgUNawM/edit?usp=sharing

6. https://docs.google.com/presentation/d/1cWBYgixIs2f2Wffow9UjOM-itoE2wpMur3iSaGvT6Sc/edit?usp=sharing

7. https://docs.google.com/presentation/d/1t38B5URUYV9QCBYBTQlDj_FdPE9a6L8iFVrmftPAM_Q/edit?usp=sharing

- Learning Outcomes of the Study in respect of Data Science

· Visualizations and data cleaning convert a whole complex and messy dataset into insightful and interesting representation, which make it easier to reach the core of the problem and solve it.

·· The best model is RFC With Default Hyper Parameter tuning on resampled data, the most

challenging part in models development process was to reduce overfitting and biasness in the data, that is why I have applied ensemble methods on base estimators... RFC provided the best framework to reduce overfitting. By running it on resampled data, I achieved an accuracy score of 1.0.

- Limitations of this work and Scope for Future Work

Further optimatization can be obtained by applying deep learning solutions. Since, it requires very high RAM capacity, it could not be displayed in jupyter notebook... I would like to update Google Colab Notebook for future projects, if acceptable... That can help me to submit a completely optimized model.