Big Data Project Report

Fatemeh Kiaie

# Toxic Comment Classification Project

The aim of the project is to categorize the toxic comments based on the types of toxicity. Examples of toxicity types can be toxic, severely toxic, obscene, threat, insult, identity hate. Two machine learning techniques like Logistic Regression and Decision Tree are implemented to determine the 6 types of toxic comments

### Data Description:

The dataset we are using for toxic comment classification is taken from Kaggle competition which can be found at Kaggle. Dataset has a large number of comments from Wikipedia talk page edits. They have been labeled by human raters for toxic behavior. Each comment can be labeled as any of the toxicity labels. Therefore, we have multiple targets for each record and we are dealing with multi-labeled classification.

### Data Preprocessing:

1. Random sampling from data: due to the huge number of instances in dataset, a random sample of data containing 10000 instances are selected for data exploration and analysis. 2. Handling missing data: no missing value. 3. Encoding: not required because all features are numeric. 4. Dropping unnecessary feature: two columns including the id and the data type (either test or training) are removed. Moreover, Toxicity is the summation of the other toxic comments targets, so it is removed. 5. Scaling: not required since the data are almost in the same range of values. 6. Text Analysis: for the test analysis we need to follow the steps as follows, a. Lowerization: all texts are already in lower case and this step is not required, b.Tokenization: all texts have been tokenized using “nltk” package been installed, the output is an RDD format c. Stop word removal: the stop words are removed from the RDD of tokenized words using “nltk” package, d. Punctuation removal: the punctuation characters are removed from the words,

e. Lemmatization is done to get the data ready for the TF-IDF, finally TF-IDF is used to convert the results words to the numbers for being used by the

*Correlation Analysis:*

Since we are dealing with multi-label classification, we need to find the correlation among targets to remove the uncorrelated targets and the heat map for the target correlation is shown as below in Fig.1 (obtained from exporting the pyspark result into python heatmap). From Fig.1, there is relatively correlation among the targets and we cannot remove any targets as the feature of other targets.

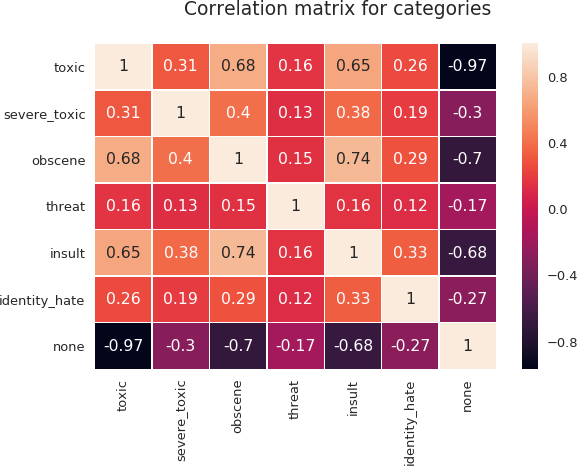
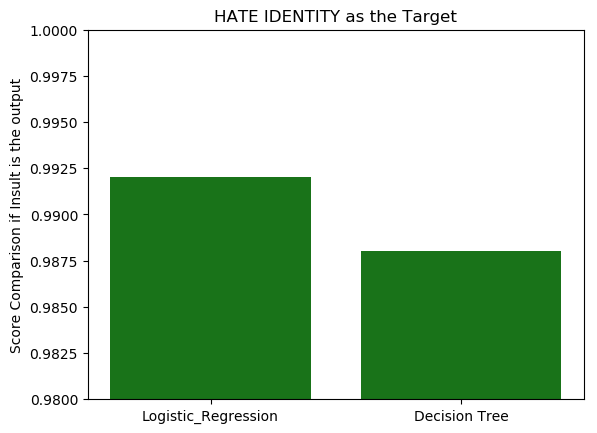


Fig. 1, the Heatmap of correlation matrix for categories

### Results

As can be seen in Fig. 2 for the HATE IDENTITY, logistic regression performs slightly better than the Decision Tree; however, slight amount of value for the overfitting comparison is shown in Fig.3. The term “overfitting” is not a proper term since the model has a good performance results.

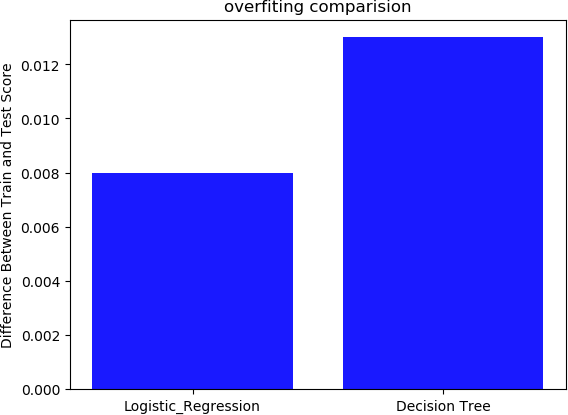


Fig. 2 Accuracy scores for HATE IDENTITY Fig. 3. Over-fitting values comparison for HATE IDENTITY

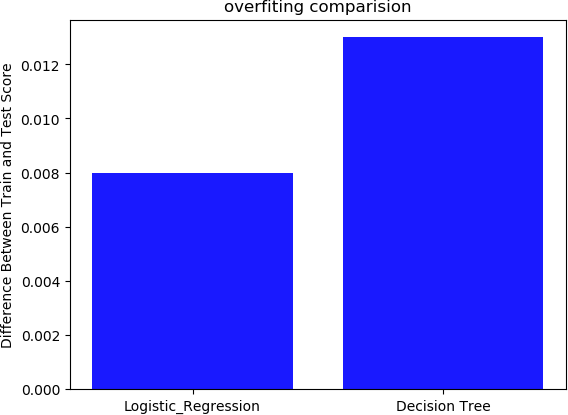


Fig. 4 Accuracy scores for INSULT Fig. 5. Over-fitting values comparison for HATE IDENTITY

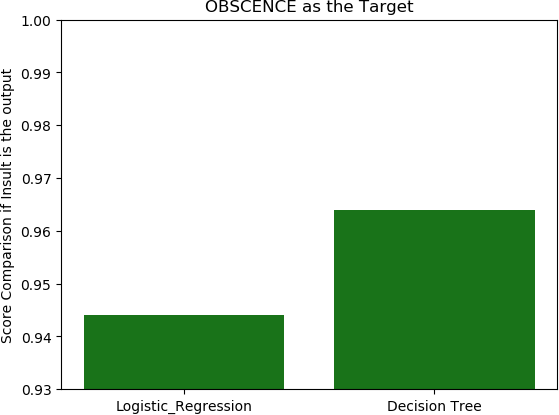
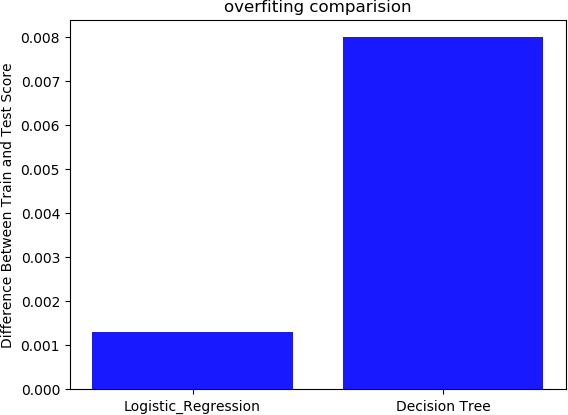
 

Fig. 6 Accuracy scores for OBSCENCE Fig. 7. Over-fitting values comparison for OBSENCE

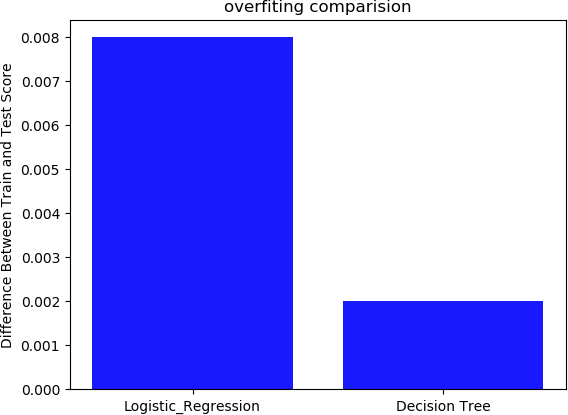
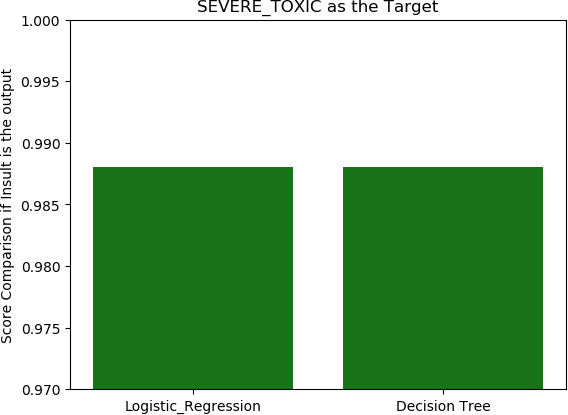


Fig. 8 Accuracy scores for SEVERE\_TOXIC Fig. 9. Over-fitting values comparison for SEVERE\_TOXIC

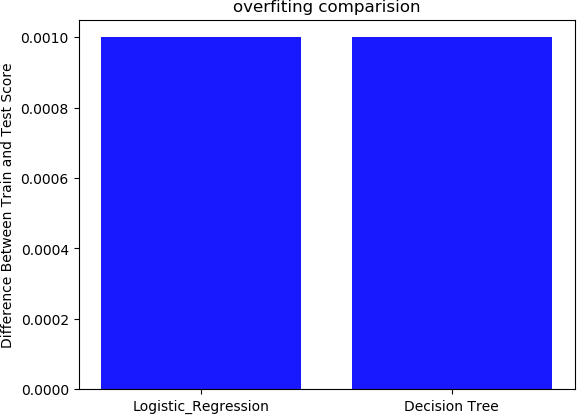
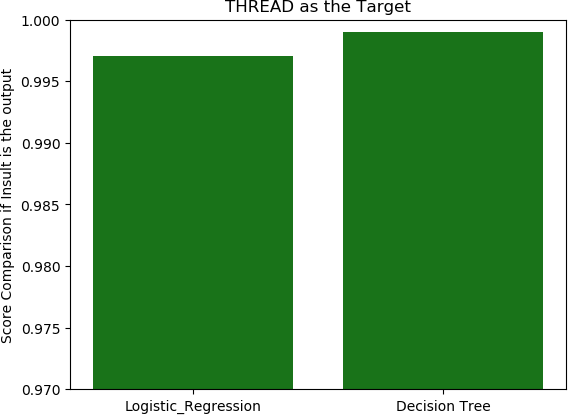


Fig. 10 Accuracy scores for THREAT Fig. 11. Over-fitting values comparison for THREAT

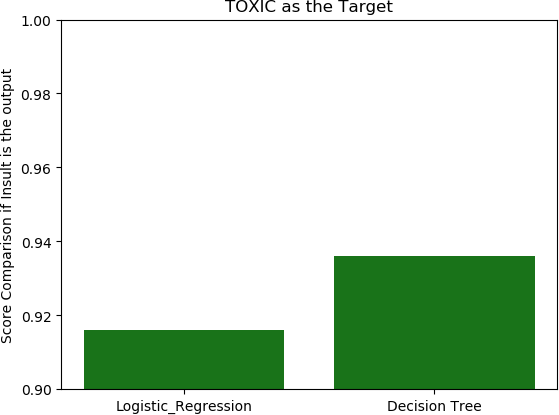
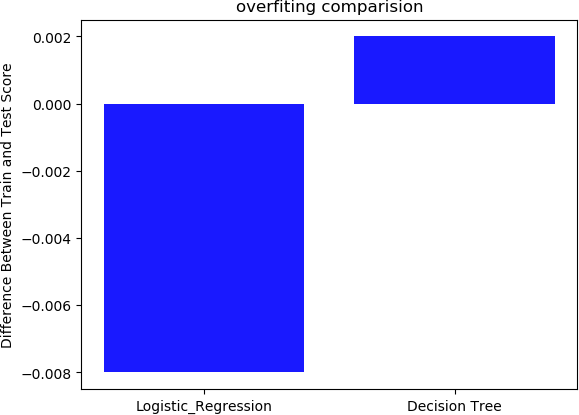
 

Fig. 12 Accuracy scores for TOXIC Fig. 13. Over-fitting values comparison for TOXIC

Similarly, as can be seen in Fig. 6, 8, 10, and 12, decision tree performs slightly better than the logistic regression in the targets of INSULT, OBSCENCE, SEVERE\_TOXIC, THREAT, and TOXIC. Comparing overfitting values for different labels in Fig. 3, 5, 7, 9, 11, and 13 show that the overfitting comparison is negligible value in the results.

### Conclusion

The results proving that both algorithms are performing pretty good and decision tree performs slightly better than logistic regression in general. However, overfitting value is a little higher in decision tree compared to the logistic regression. Moreover, models are performing best in detecting THREAT compare to the other target labels.