Detecting Insults In Social Commentary

Insights and Conclusions

During the development of the project, our goal was to implement an effective insult detection model. Three different models were trained and tested: logistic regression, random forest, and linear support vector classification. By comparing the metrics report of all models, and by manually exploring the prediction results, we concluded that linear SVC was the best model implemented in terms of accuracy, F1 score, and respecting the comment's context.

The model scored a respectable accuracy of 86% and an F1 score of 85%

Business Recommendations

We recommend integrating the model developed during this project into moderation systems for online platforms, forums, or social media channels, contributing significantly to maintaining a positive and respectful online environment. Continuous monitoring, periodic updates, and the establishment of a user feedback mechanism are suggested to ensure the model remains effective against new or tricky linguistic patterns and trends.

We also recommend using the model during online games and live streams that offer a chatting system to their customers. The messages can be run through the model for verification first, and only healthy messages would be allowed to pass. An extension can be implemented to punish biewers or players that constantly show signs of cyber-bullying and misbehaviour.

Future Work and Improvements

Exploration into more advanced feature engineering techniques, such as word embeddings, ensemble learning strategies, and multilingual support, could potentially enhance the model's insult detection capabilities.

Moreover, some comments intentionally included grammatically incorrect slurs and insulting words to avoid detection. The model can be improved by also considering these variations of words.