A Predictive Approach to Help Online Discussions Become More Productive and Respectful

Danielle Bresich, Lawrence Baryoh, Libby Gonzalez, Maddie McDonough, Matthew A. Lanham

Purdue University Krannert School of Management

dbresich@purdue.edu; lbaryoh@purdue.edu; gonza255@purdue.edu; mcdonom@purdue.edu; lanhamm@purdue.edu

Abstract

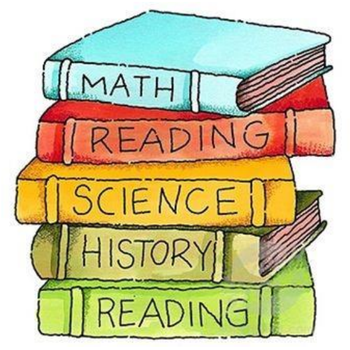
This study focuses on how to make online discussions more productive using predictive analytics. The motivation for this investigation is: since more people interact more via online discussions, is there a way to make these interactions better for the users? With the rise in big data analytics improving all areas of life, we believe a predictive model could be developed to eliminate this problem. In our study, we investigated Wikipedia comments that have been labeled as toxic behavior by human raters. We developed a classification predictive model using R that would allow future comments to be flagged for removal based on their content. We posit that our proposed solution could be extended to other online discussion forums to improve interactions by making them more productive for the community.

Introduction

Online discussion forums are powerful tools used in our society. Universities, work-places, and social sites all use these platforms as a way to communicate and share ideas with others across the world. This project is relatable to society's today since discussion forums are found in several environments. The goal of this project is to be able to identify toxic behavior that may results in unproductive discussion forums. Examples of toxic behavior includes things such as: insults, threats, obscene language, and hate speech.

Discussion Forum Environments:

1. Educational
2. Business reviews
3. Community Engagement



Research Question:
Through the use of data analytics and machine learning, how can a predictive model identify and potentially delete toxic language from online discussion forums?

Literature Review

Social: Gay Bullying Online Opinion Expression

- Different opinions displayed on social media were tested and surveys were given based on hypothetical situations based on hostile and friendly environments.

Academic: Exploring the Use of Discussion Strategies & University of Missouri

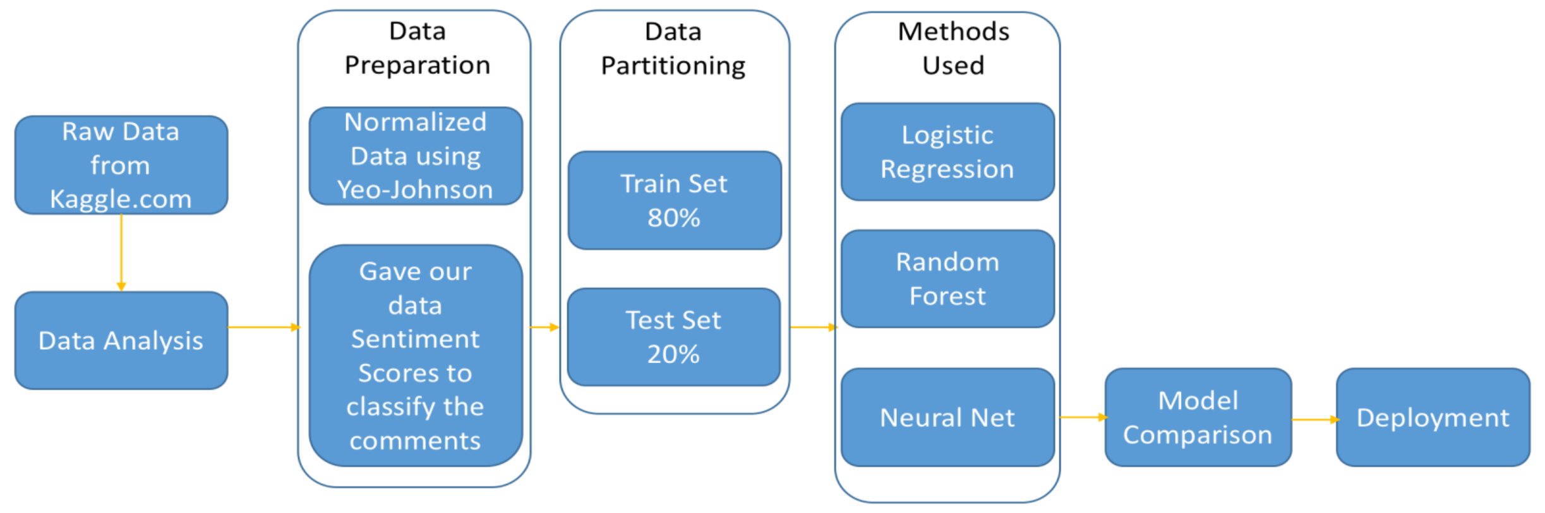
- Researchers gave participants, discussion strategies to develop productive discussions and aid teachers contribute to productive discussions.
- Participants received different levels of structure to improve the quality of learning

Social Welfare: Individual and Social Benefits of Online Discussions

- Anonymous engagements between individuals were tested and found people who participate in more meaningful online communities are more likely to have higher civic engagement and overall meaningful experience

Methodology

Figure 1 outlines our study design, starting from data collection, data cleaning, data pre-processing, feature creation and selection, model/approach selection, cross-validation design, and model assessment/performance measures.



Raw Data from Kaggle.com

Data Analysis

Data Preparation

Normalized Data using Yeo-Johnson

Gave our data Sentiment Scores to classify the comments

Data Partitioning

Train Set 80%

Test Set 20%

Methods Used

Logistic Regression

Random Forest

Neural Net

Model Comparison

Deployment

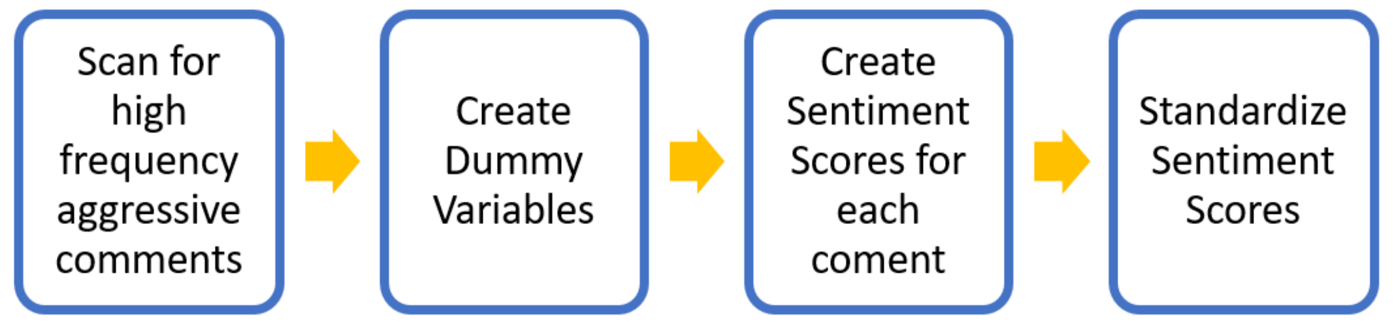
Figure 1. Study Design

Data

A large dataset of Wikipedia comments was obtained from Kaggle. These comments were categorized by human raters based on toxic behavior, including toxic, severe toxic, obscene, threat, insult, and identity hate.

Feature Creation

Our raw data was text and we processed it as follows:



Scan for high frequency aggressive comments

Create Dummy Variables

Create Sentiment Scores for each comment

Standardize Sentiment Scores

Figure 2. Feature Creation

Data Cleaning & Pre-Processing

Numeric features such as sentiment scores (anger, joy, etc.) were transformed using a min-max normalization that kept the values between 0 and 1, as well as Yeo-Johnson transformed to make them more normally distributed. These features helped to detect the overall vibe of the sentence. Most of the features were dummy (0/1) variables that indicated if a certain word was found within the comment or not.

Down-sampling

In order to focus on each category of non-productive comment individually, we down-sampled the entire dataset into smaller sets, comprised of only one category.

Model Design & Methodology (Approach) Selection

We trained and evaluated a random forest, logistic regression, and neural net models. Each of these models learn in different ways.

Statistical & Business Performance Measures

We evaluated the models based on AUC. AUC strikes a balance among sensitivity (correctly predicting non-productive comments) versus specificity (correctly predicting productive comments).

Results

The VarImpPlot function was used to determine which variables were the best predictors for the toxicity classifications. For example, the graph for toxic is posted below:

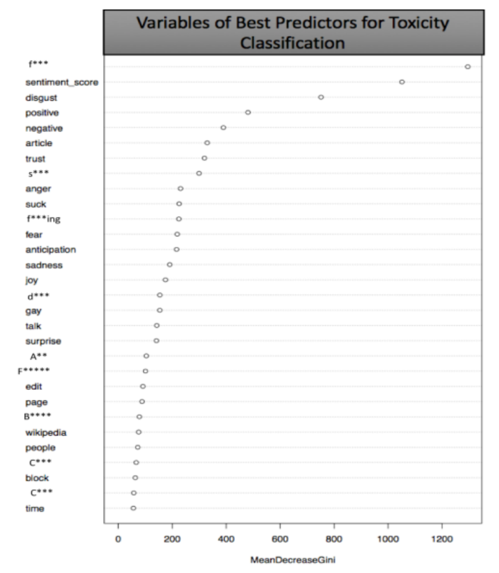


Figure 3. Toxic Predictors

Variable	#1 Best Variable Predictor	#2 Best Variable Predictor	#3 Best Variable Predictor
Toxic	F***	Sentiment Score	Disgust
Severe Toxic	F***	Sentiment Score	F***ing
Obscene	F***	Sentiment Score	S***
Threat	Sentiment Score	Fear	Sadness
Insult	F***	Sentiment Score	Disgust
Hate	F***	Sentiment Score	Gay

Figure 4. Classification Predictors

The table above summarizes the best variable predictors for each variable. Each of these were found by using the VarImpPlot function in R.

The plots below show the predictive performance of each model and the optimal model's performance for each type of non-productive type of comment:

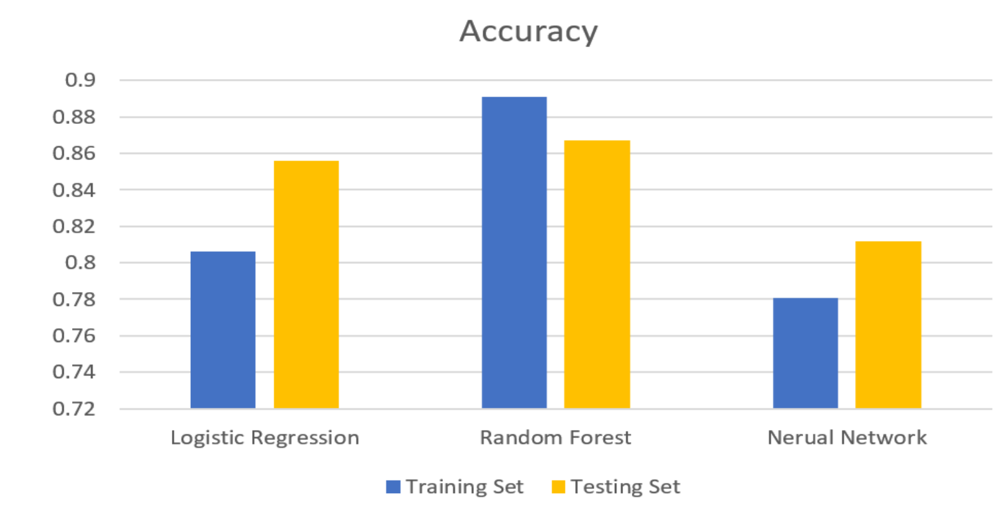


Figure 5. Model Accuracies

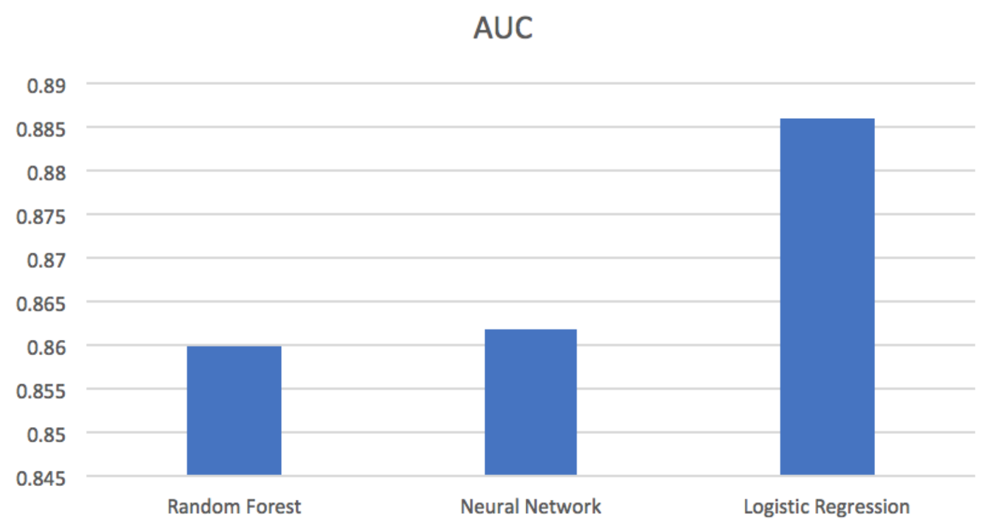


Figure 6. Model AUCs

According to the Accuracy chart, Random Forest performed the best. Although it was the most accurate, our focus was on AUC and Logistic Regression proved to be the best performing model.

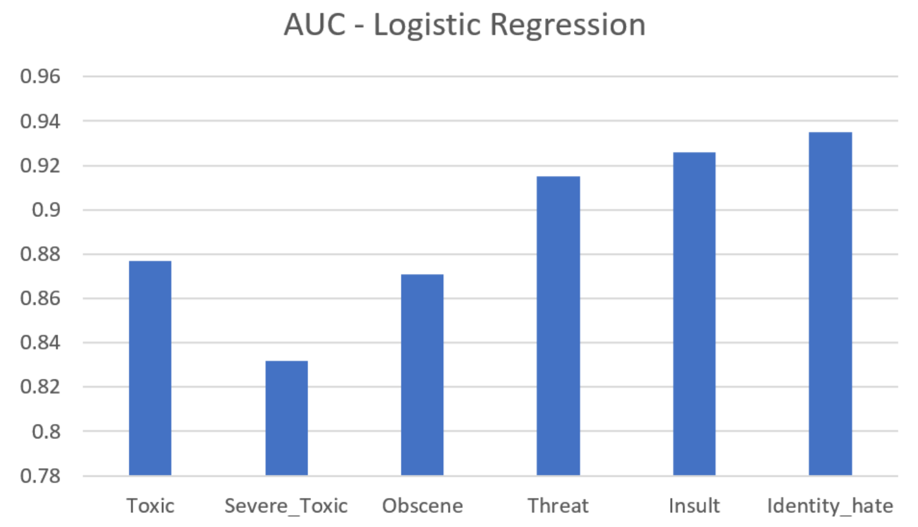


Figure 7. L.R. Category AUCs

Conclusions

As technology becomes more prevalent in our everyday lives, the threat of unproductive and disrespectful online forums has been at an all-time high. Our study aimed to create a predictive approach to make online discussions more productive and beneficial to the society. Based on our research, a firm could expect to obtain a 89% accuracy using a logistic regression model at flagging and removing non-productive or disrespectful comments. In conclusion, all three discussion forum environments could benefit from our study and eliminate almost 90% of toxic comments after implementing this model.

Acknowledgements

We thank Professor Matthew Lanham & Josh McKenney for their guidance.