Project 3, modeling AI

Problem statement

#1. From exploring the data what features best indicate AI?

#2. Do punctuation, abbreviations, and semicolons indicate AI?

#3. Which models should we use to help moderators automod posts?

Data

I crawled reddit for 6 subreddits, but only ended up using 1: r/nostupidquestions in my analysis because of the cost of OpenAIAPI. I did have to clear out a couple of automoderator bot posts from my human data.

Outliers and automoderator

Some reddit responses come from a bot called an automoderator which automatically flags as /u/ or u/ to call a new user I removed those and checked for mentions of chatgpt calling itself a language model and removed those.

Parameters

I looked at 3 grammatical syntaxes for my parameters.

Abbreviations

Semicolons

;

Punctuation

!?

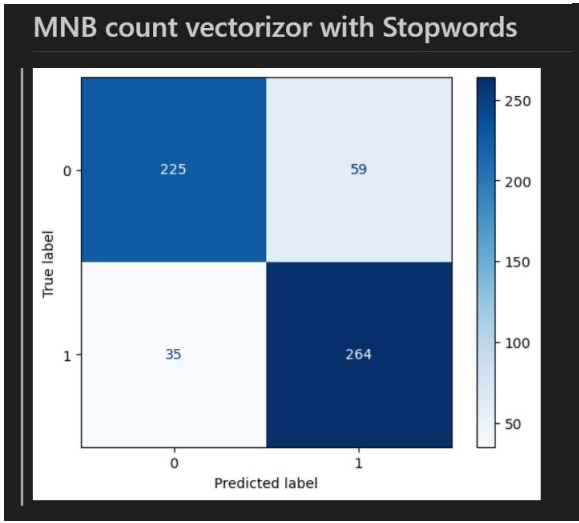
Models

So I used logistic regression and Multinomial naive bayes for my modeling because our target is a class of 1 and 0(AI written or not) and countvectorizer returns rows of integers for Multinomial naive bayes. This made punctuation a less powerful factor if I went forward I'd want to use it in a tree. Finally I used stemitization and that made my model slightly better at predicting,

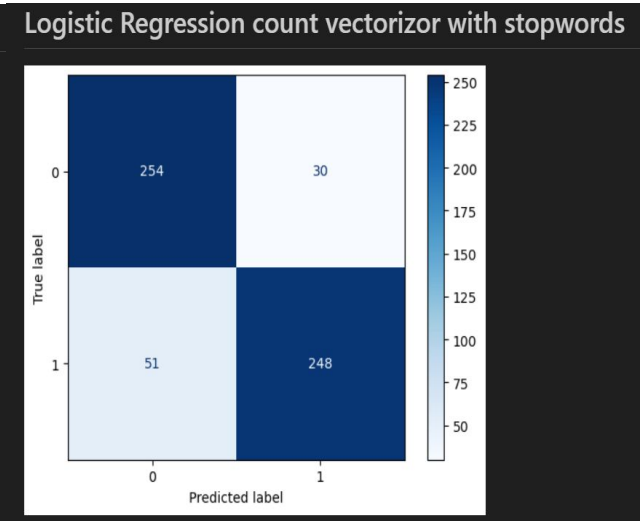
Baseline vs train/test and best_params

The baseline for my dataframe is 51.3%, roughly half human, half AI, because I dropped some responses recorded as human done by reddit's auto moderator.

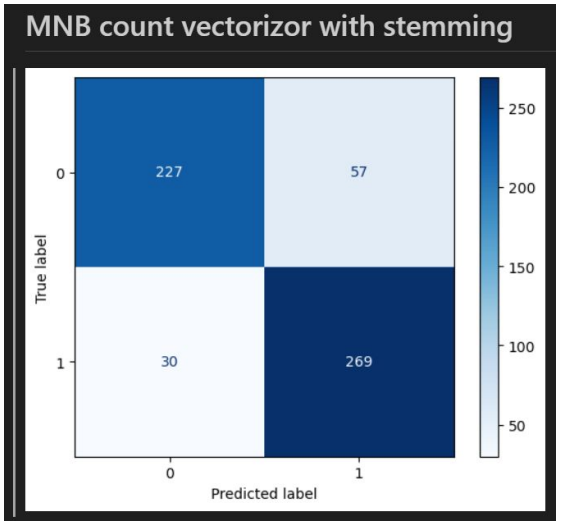
Model 1 MNB stop words



Model 2 Log stop words



Model 3 MNB stem



Model Evaluation

Model	MNB Stopwords	LR Stopwords	MNB stemming
Train score (%)	91.5	98.6	89.5
Best score(%)	84.8	89	87.4
Test score (%)	83.8	85.7	87.6
Over/under fit	Over	Over	Over

Conclusions

For now count vectorizors are still very good at catching AI generated posts and appear to be excellent tools, if I had more time, I would make a tree for punctuation.

Sources

Boston dynamics for image.