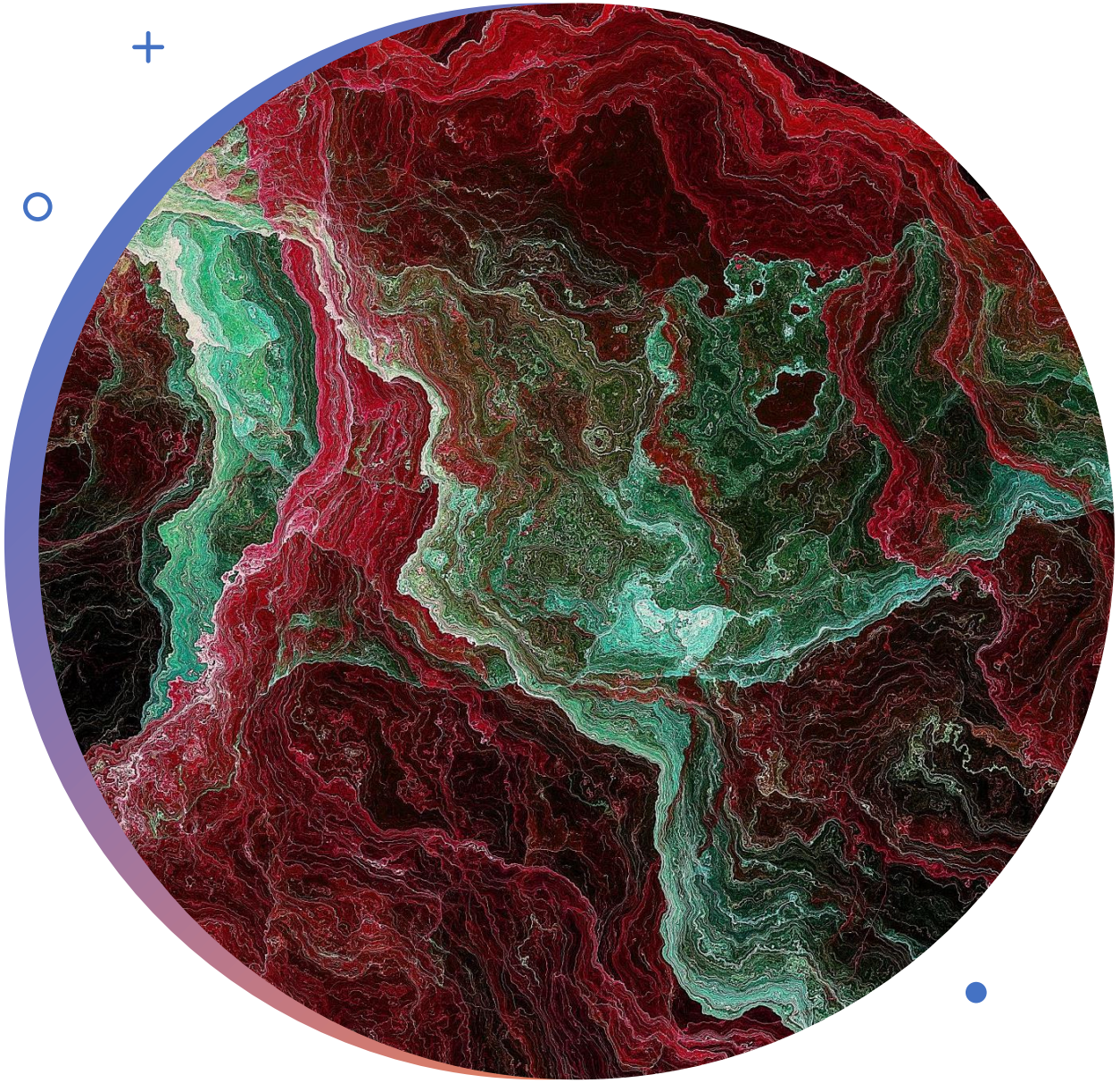
Web API and NLP Project 3

Vishnu Kodicherla

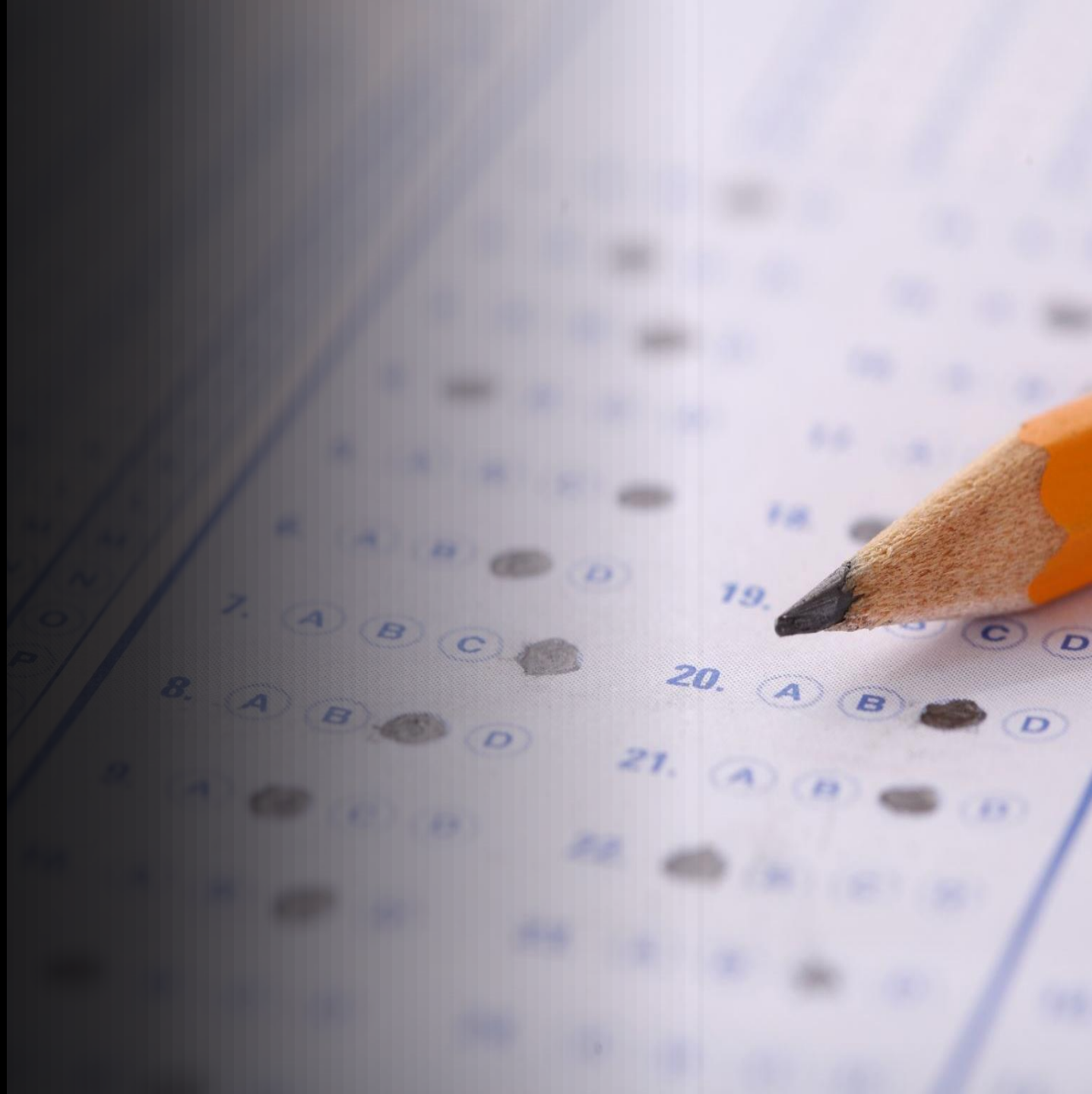


Background

- The two subreddits that I looked at were Nvidia and AMD
- Both these companies have become famous for developing graphics cards and improving performances for computers.

Problem Statement

- As a Data Scientist working at a startup company, I received a request from Reddit in which it informed me that there was a bug in their system that caused swithe tching of similar subreddits. Due to Reddit being busy with other issues, they assigned us to help with the problem. Our company agreed to work on the issue. To be able to analyze and figure out how to look at the subreddits, I started with a sample in which I chose Nvidia and AMD. To be able to distinguish these subreddits I used different models that include Logistic Regression, Random Forest, and decision tree. By using these methods I will make a confusion matrix in which I can look at true negatives, true negatives, false positives, and false negatives. From looking at the confusion matrix, I can see if the three models worked or not. A successful final model would have the lowest number of false positives and false negatives.



Data Cleaning

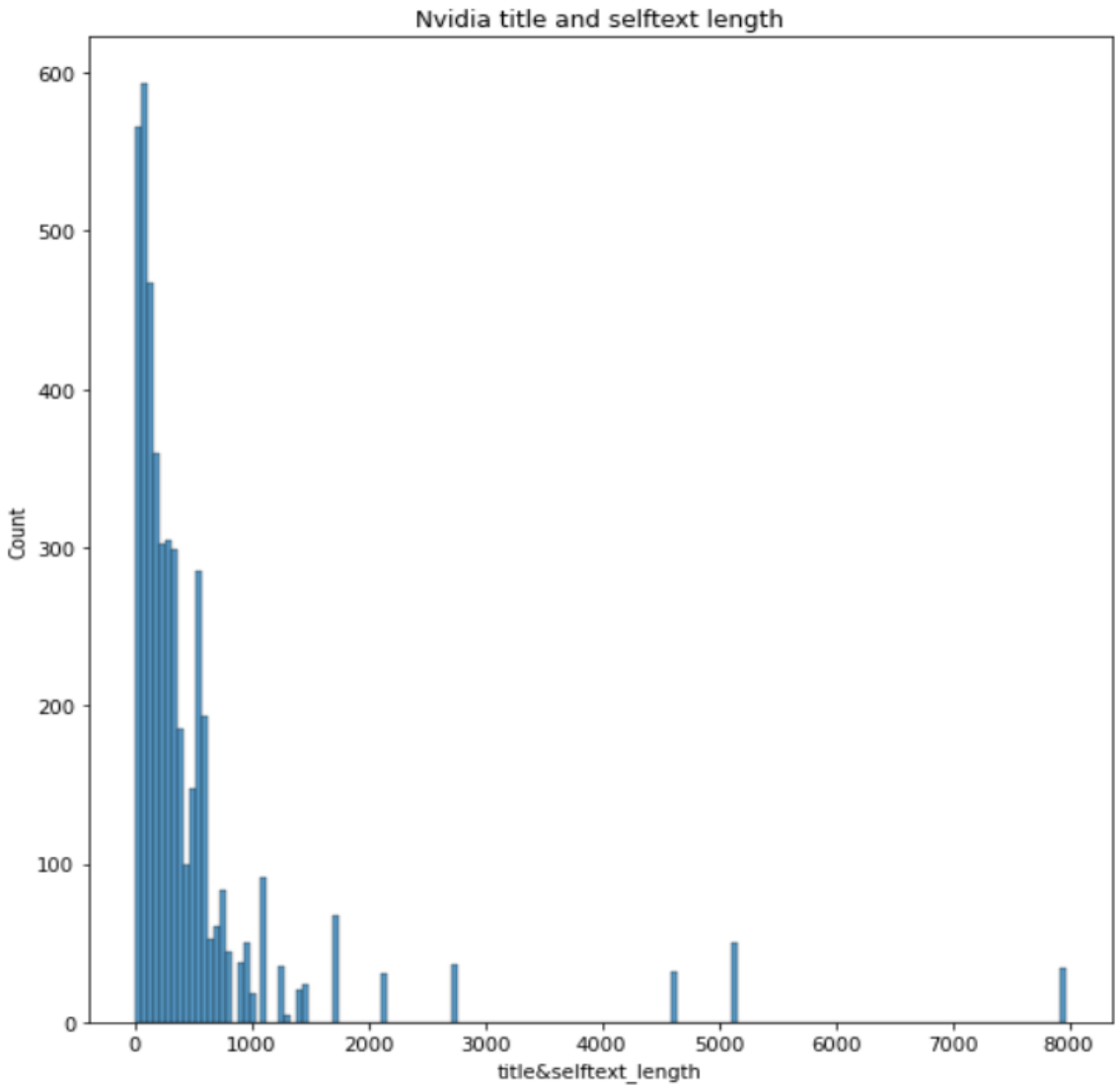
I chose these 7 columns for my dataset: subreddit, id, title, author, created_utc, self_text, is_self

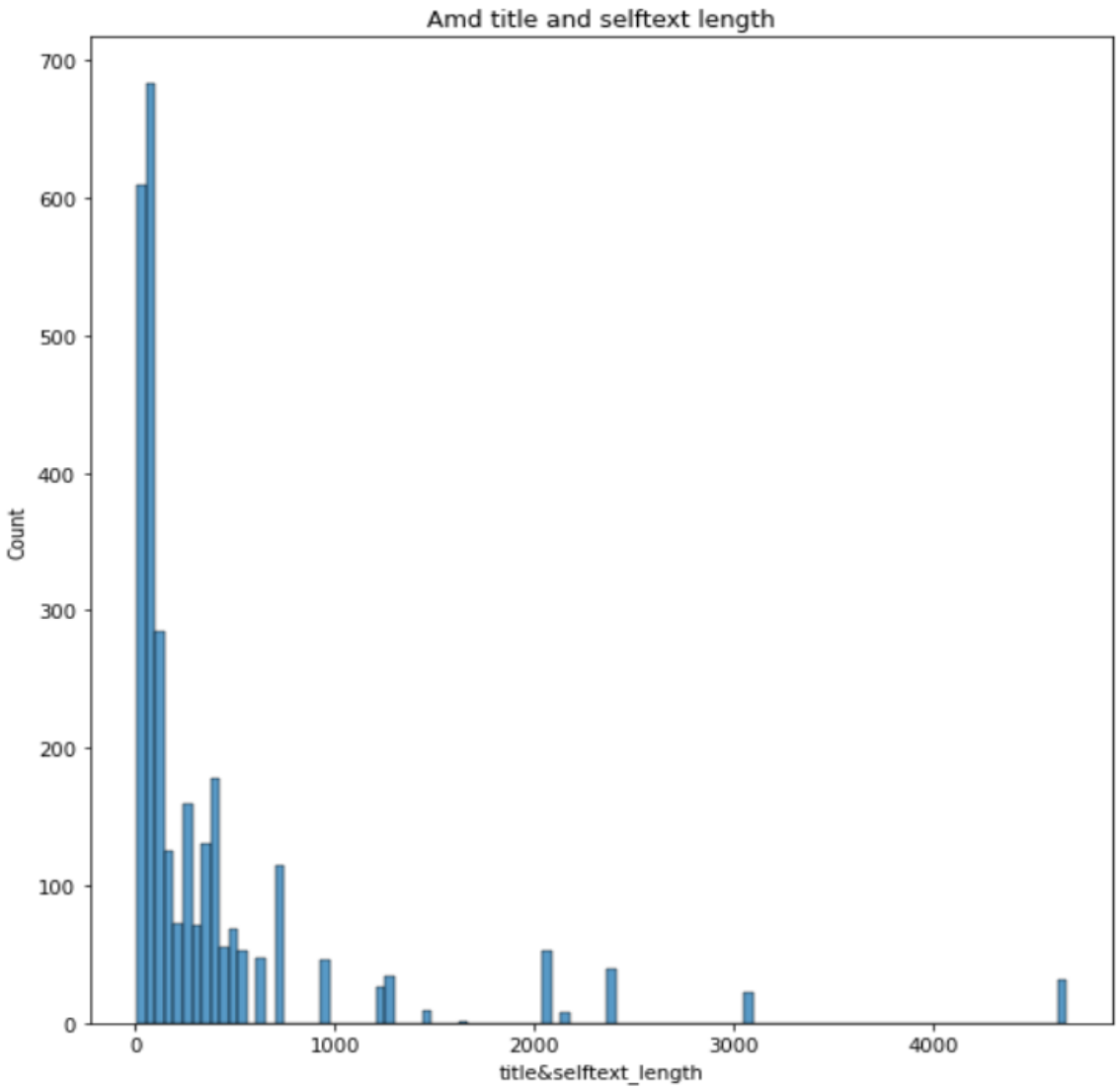
I combined the self_text column and title column together and made a new column

I got rid of punctuation by using regex

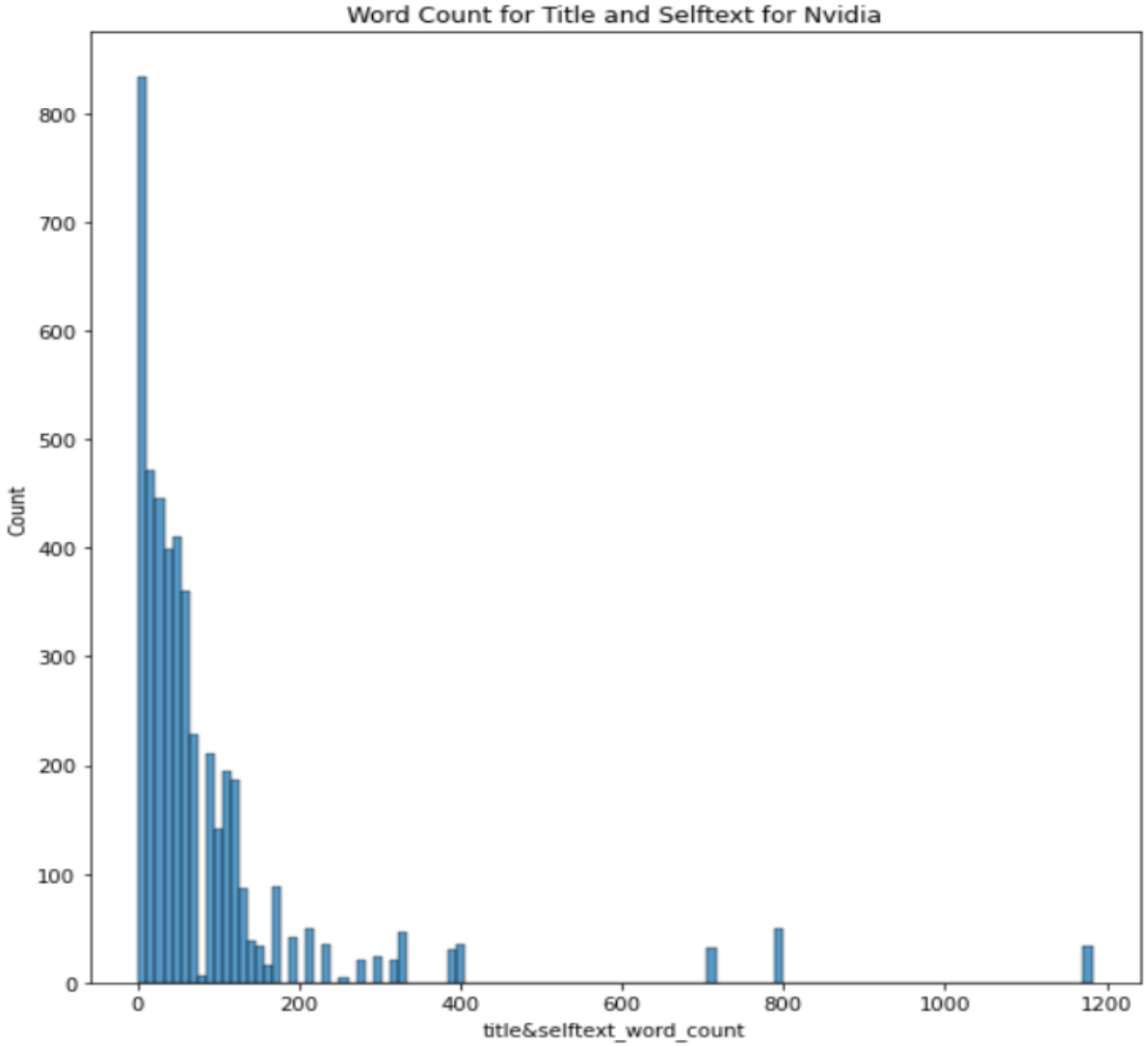
Then I made a title word count and length count

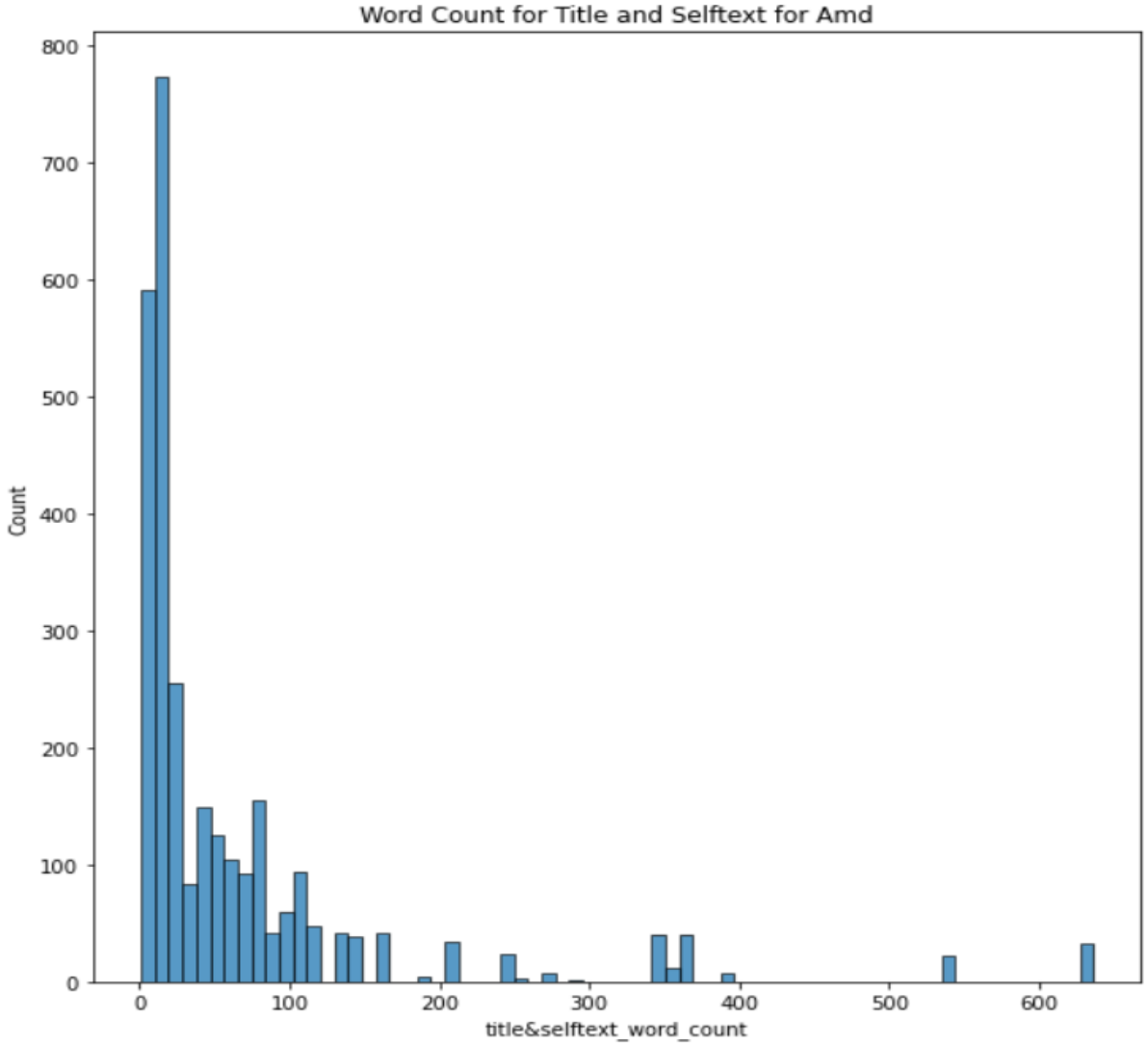
EDA Analysis





EDA Analysis continued





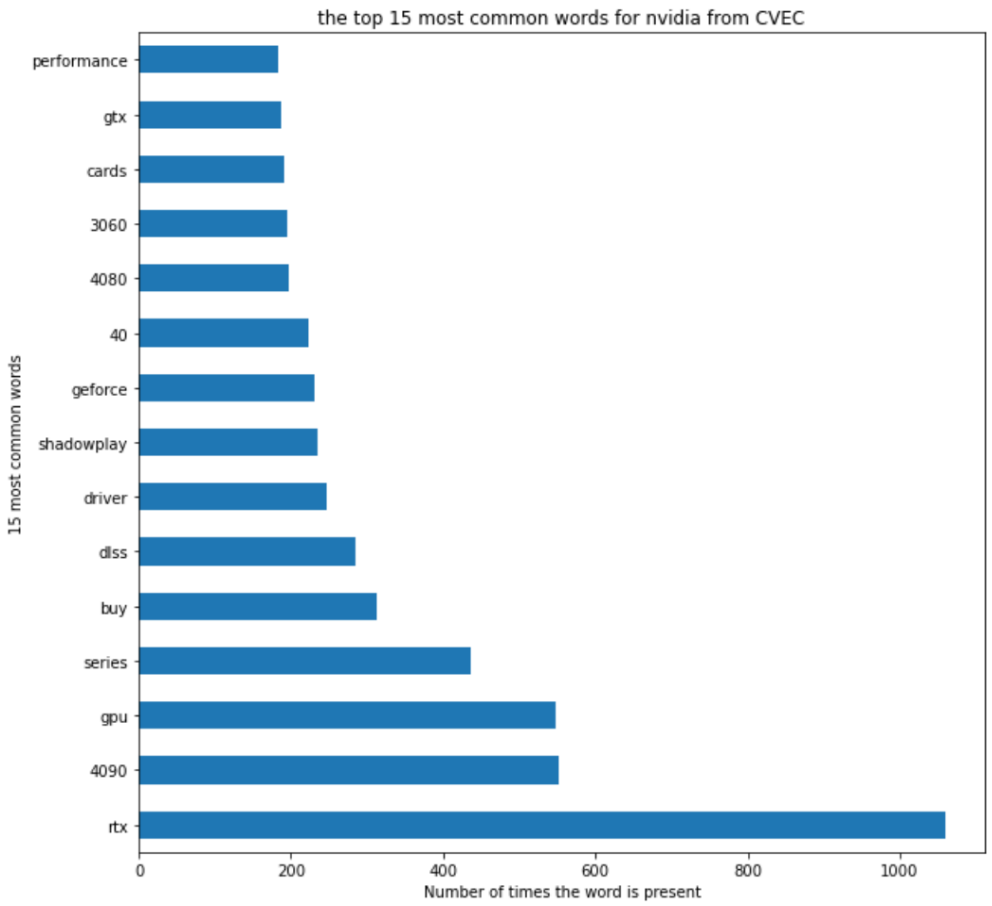
Preprocessing Analysis

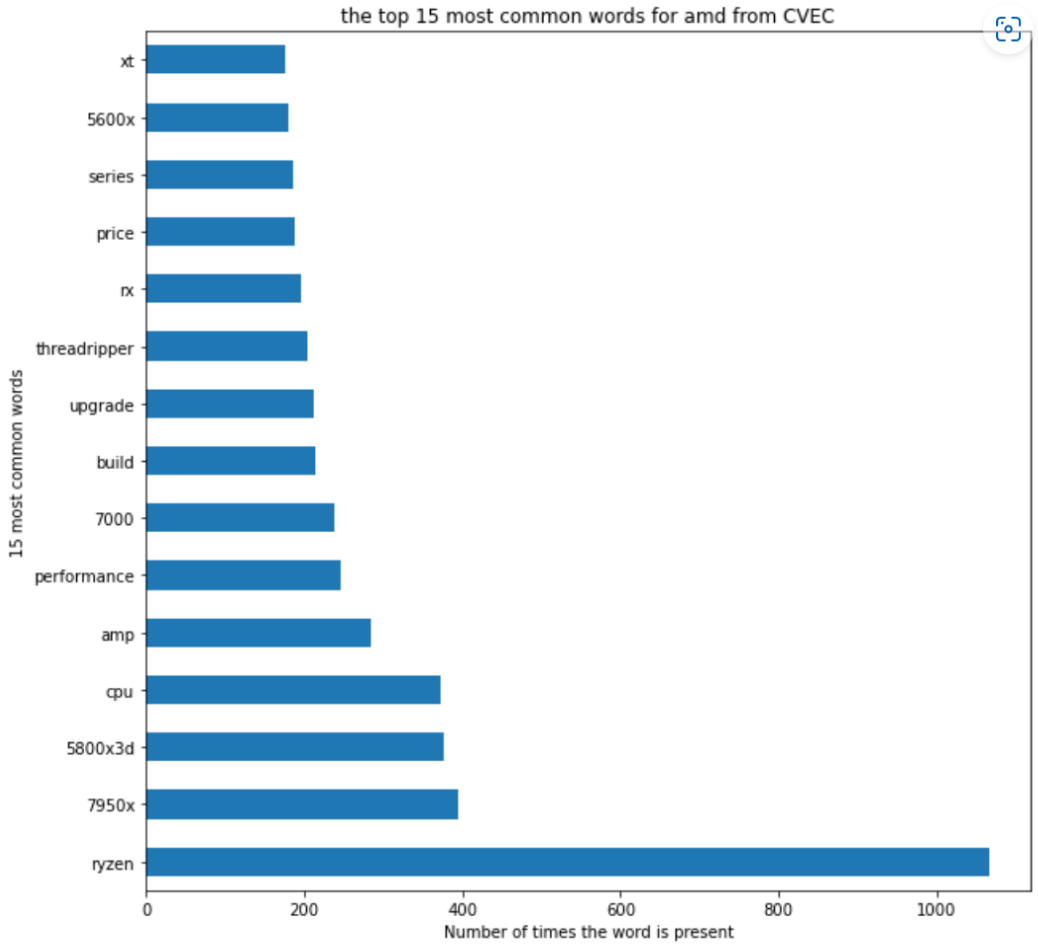
Did Stemming and Lemmitization

Included some words in my list of stopwords

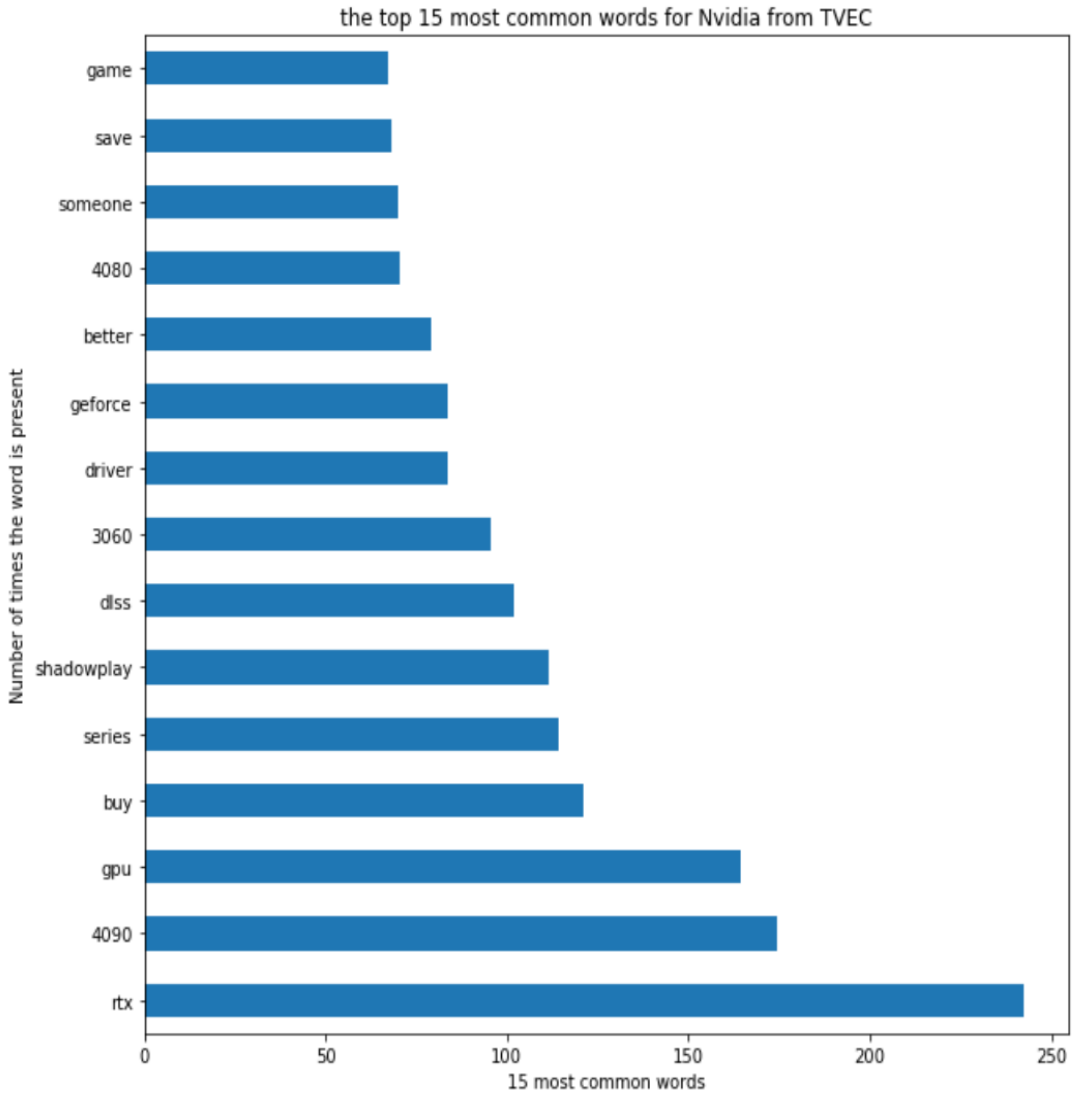
Chose lemmitization for my model analysis

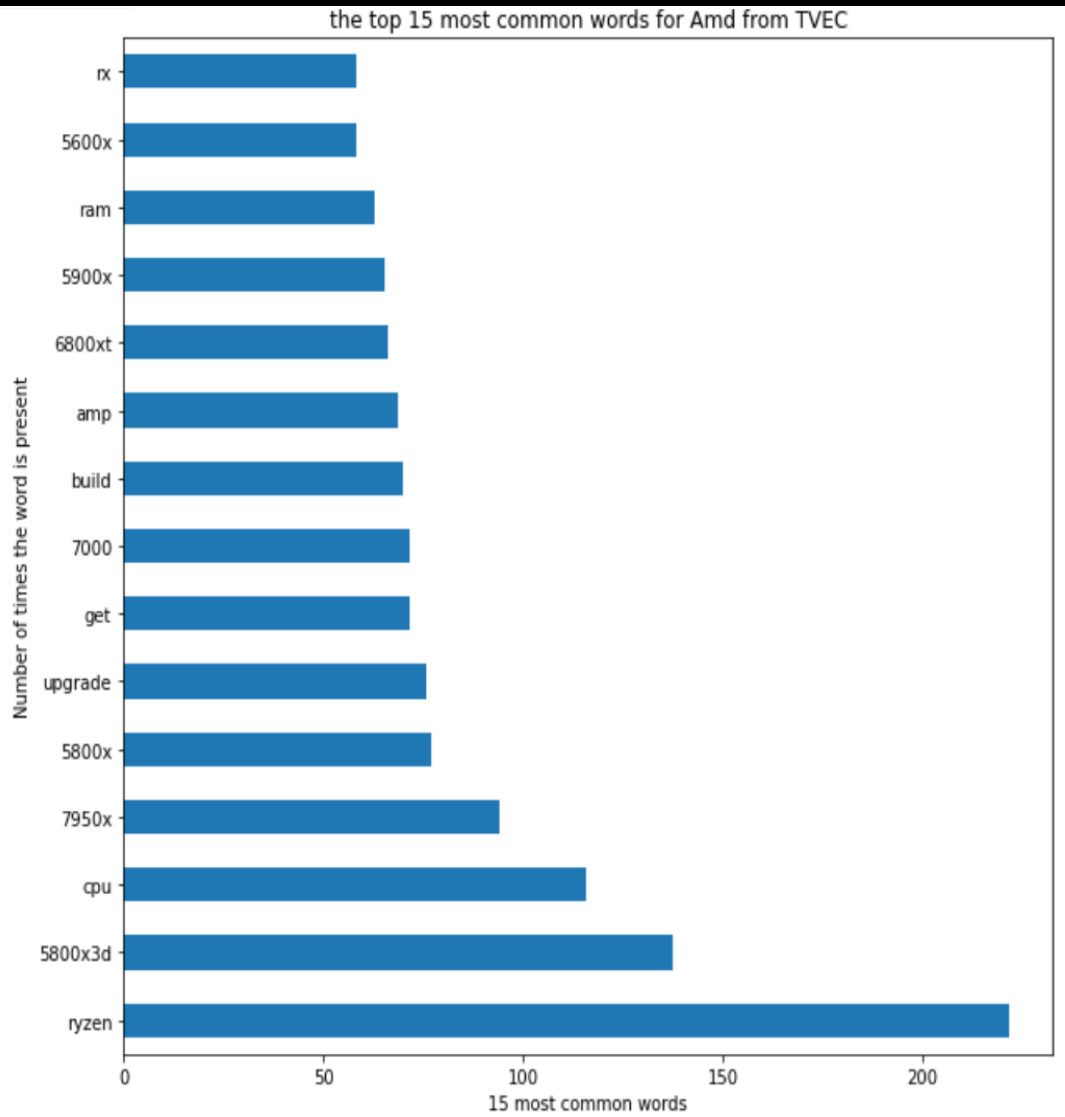
Preprocessing Analysis (countvectorizer)





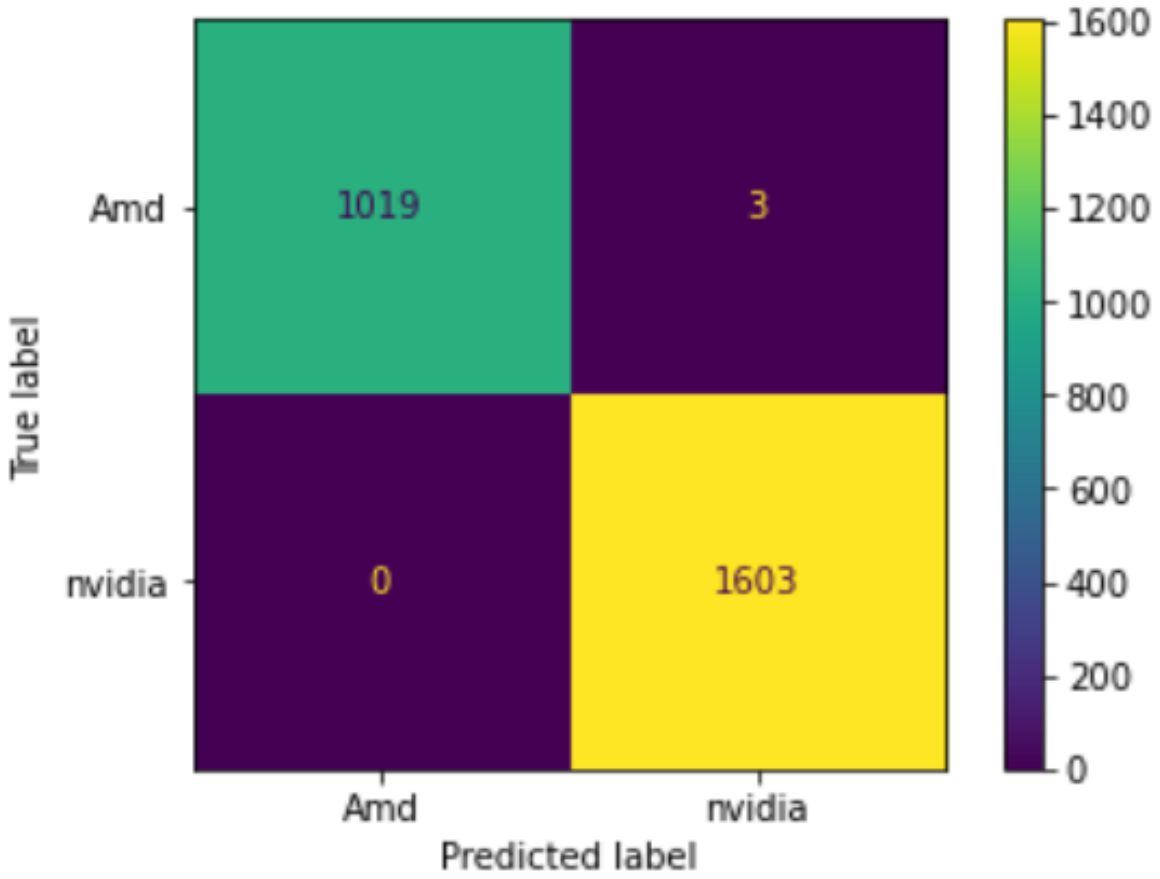
Preprocessing Analysis (TFDIFVectorizer)





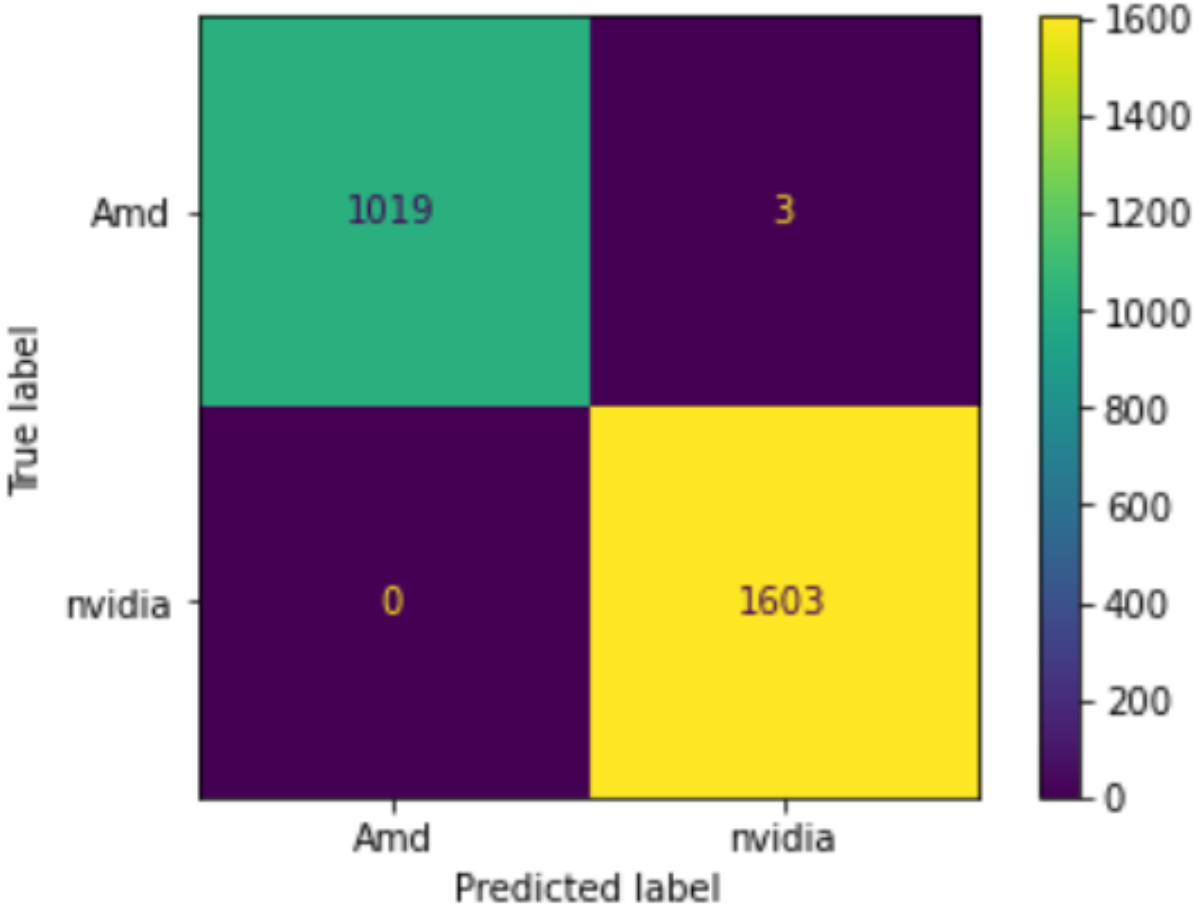
Modeling Analysis

- I did Count Vectorizer and Random Forest
- Score for Training: 1.0
- Score for Testing: 0.998



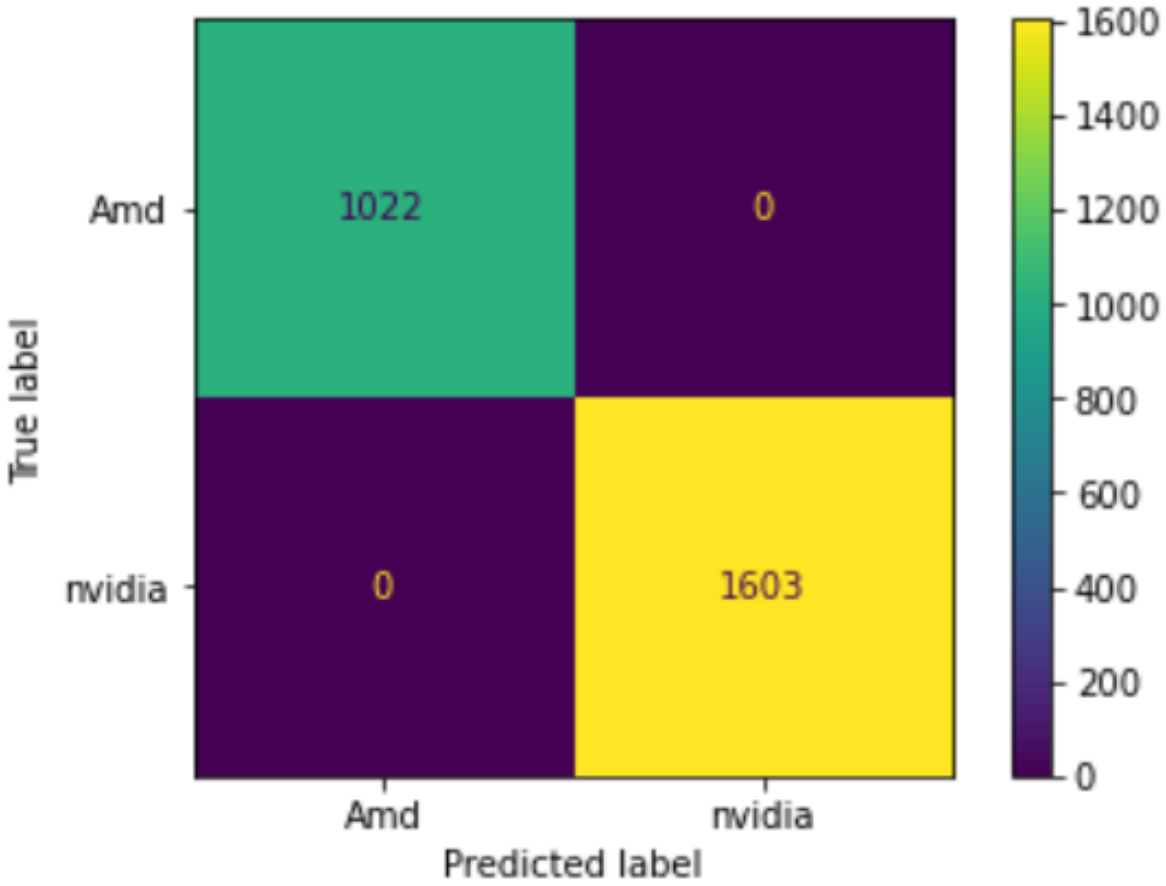
Modeling Analysis

- I did Count Vectorizer and Decision Tree
- Score for Training: 1.0
- Score for Testing: 0.998



Modeling Analysis

- I did Count Vectorizer and Logistic Regression
- Score for Training: 1.0
- Score for Testing: 1.0



Conclusion

- Logistic Regression was the best model because we saw no False Positive or False Negatives.