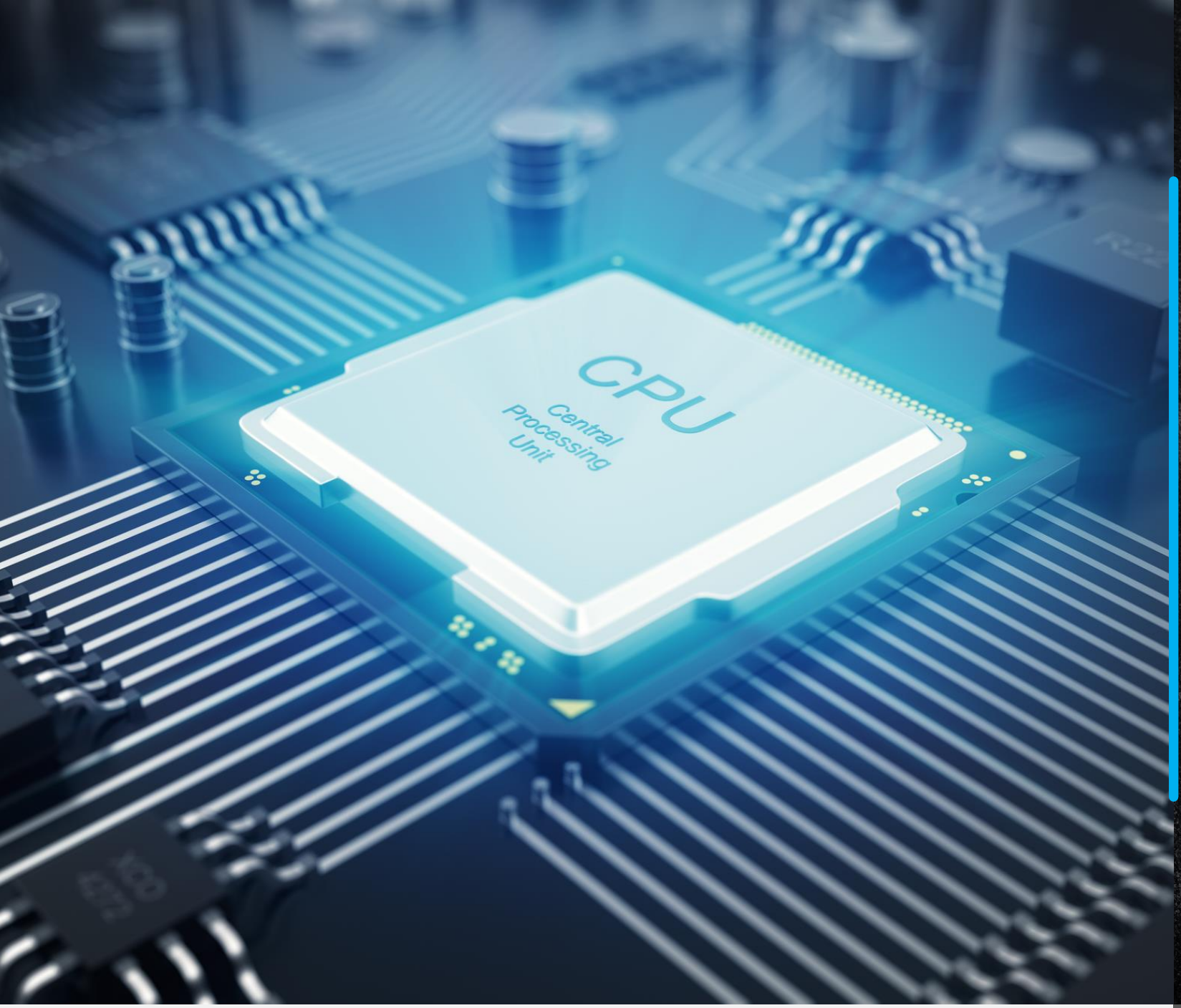


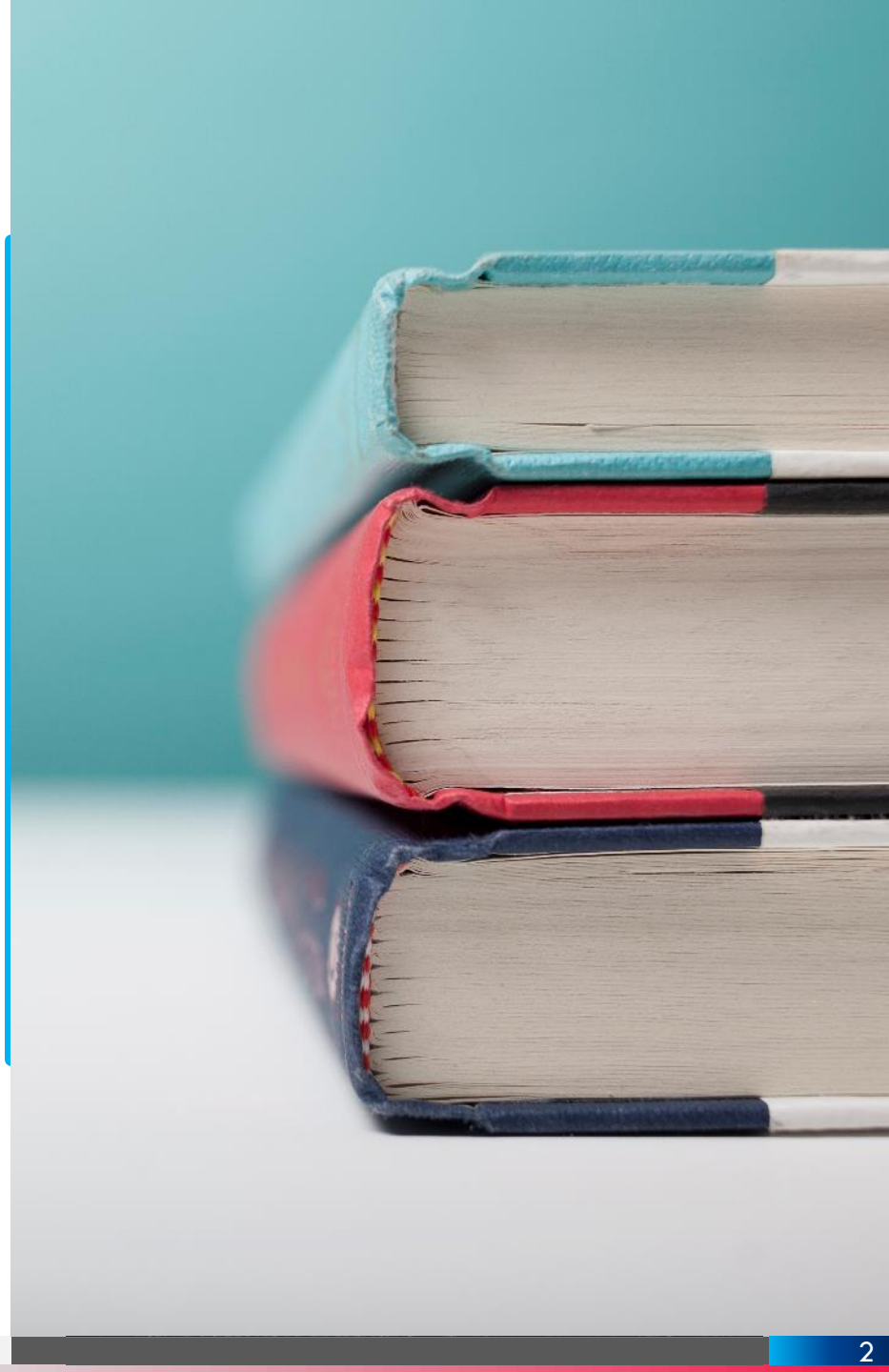

# Using Natural Language Processing (NLP) Modelling to Predict Desktop CPU Brand Popularity

Nor Rashidi Bin Norhashim

SG-DSI-27

# Content

- Background
- NLP Model Training
  - Overview
  - Data Cleaning
  - Exploratory Data Analysis
  - Base Model Benchmarking and Selection
  - Feature Engineering
  - Hyperparameter Tuning
- Production Model
  - Performance
  - Error Analysis
  - Deployment and Recommendations
- Conclusion



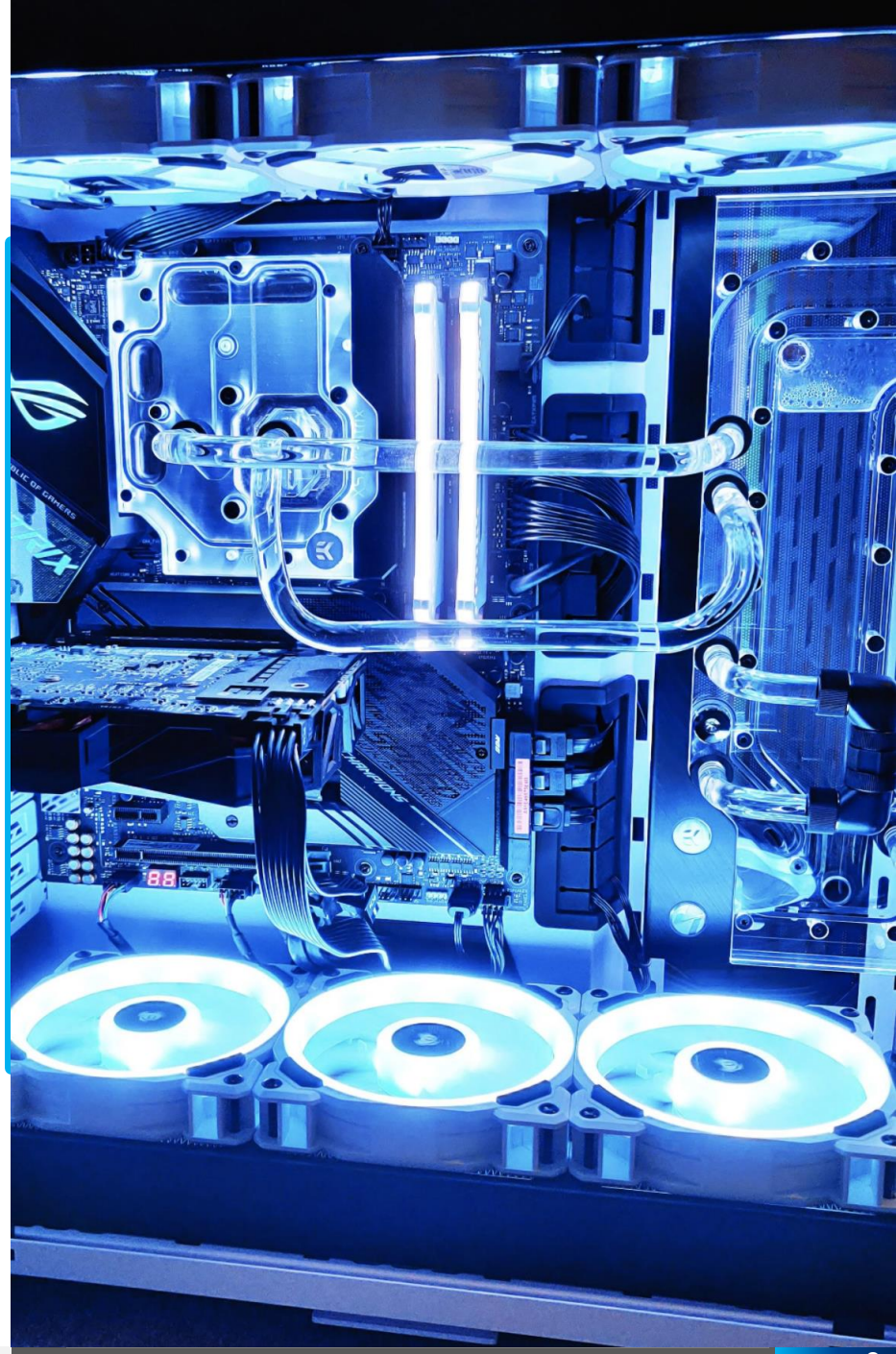
# Background

## Problem Statement

As an aspiring entrepreneur embarking on a new custom Gaming/Enthusiast Desktop PC startup, which desktop CPU brand should I carry to minimize dead stock/ slow moving inventory?

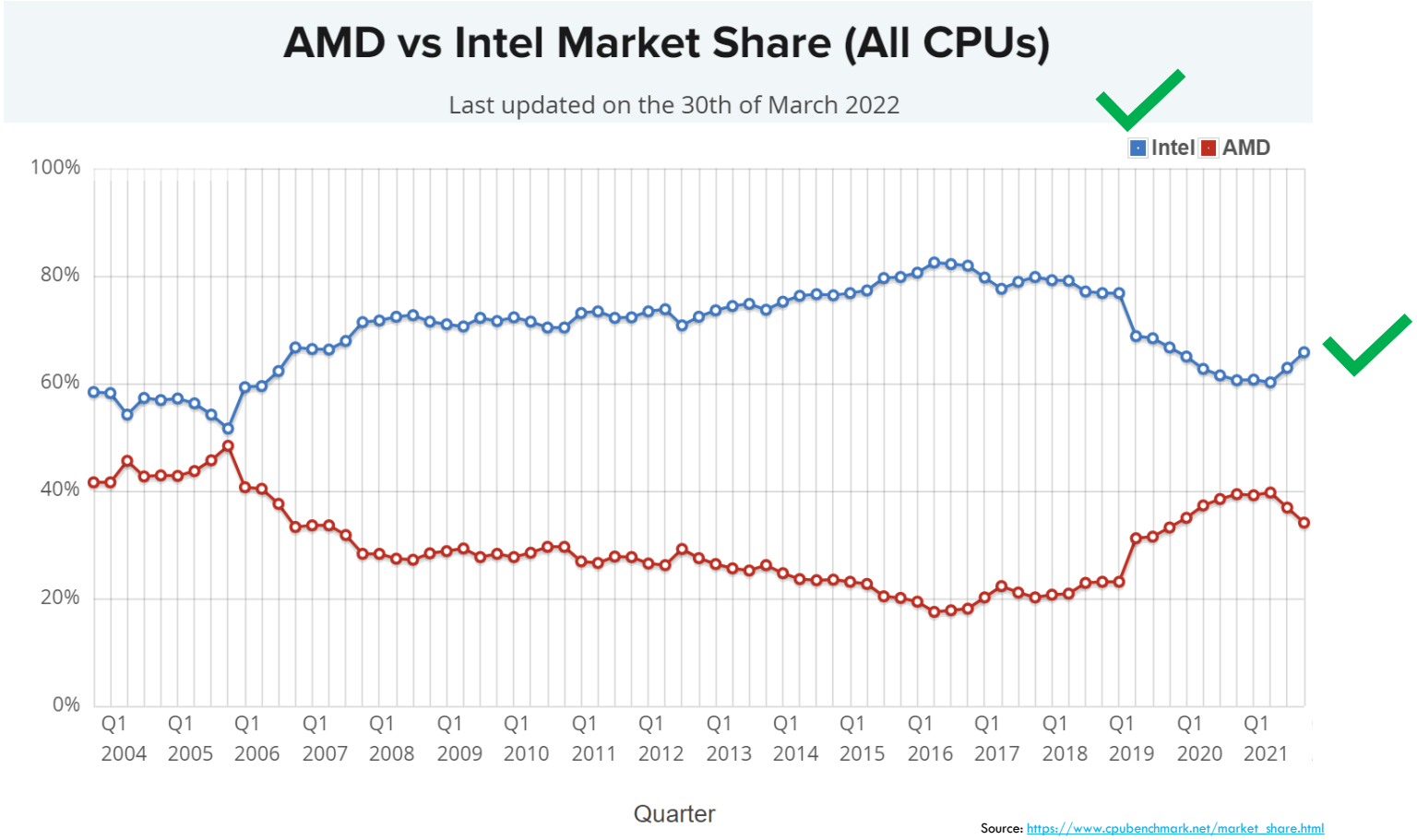
## Key Considerations

- Low initial capital
  - Parts brought in need to be quickly sold to obtain more funds to grow the business.
- AMD vs. Intel in the PC Hardware Ecosystem
  - Choice of CPU brand will dictate other parts like motherboard, DRAM etc. and are not interchangeable.
- Target market group
  - PC Gamers and Enthusiasts (PC Master Race)



# Background

## AMD vs. Intel (All CPUs)




**BUT!** Include Server, Laptop, Desktop CPUs for both Consumer and Enterprise customers

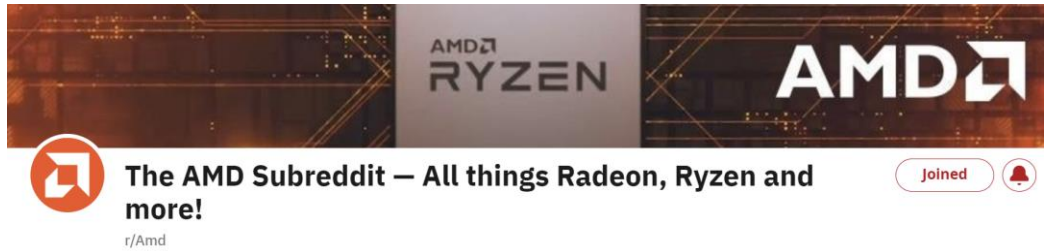
How then can I know which CPU brand to choose?

# NLP Model Training

# Overview

10,000 posts each from r/AMD and r/Intel were scraped to train the model.

## r/AMD




## r/Intel


10,000 posts from r/BuildaPC will be used to quantify which brand is more popular.


Example Post:

r/buildapc · Posted by u/thecerealkidd 32 minutes ago

**Is a 550w Corsair RM550x enough to run a Zotac RTX 3060Ti Twin Edge OC?**

Build Upgrade

I'm looking to upgrade my GPU to a 3060ti but I'm not sure if my current PSU can run it without any hiccups. My full system specs are:

Processor: Ryzen 5 5600x
CPU Cooler: Arctic Freezer 34 Esports Duo
Motherboard: MSI B450m Bazooka Plus
RAM: 16gb 3200mhz ddr4 (Dual Channel)
Storage: 240GB NVMe SSD and 2TB HDD
GPU: Palit GTX 1050 2GB
PSU: Corsair RM550x

Thanks for the feedback!!

# Data Cleaning

Subreddits r/AMD vs. r/Intel

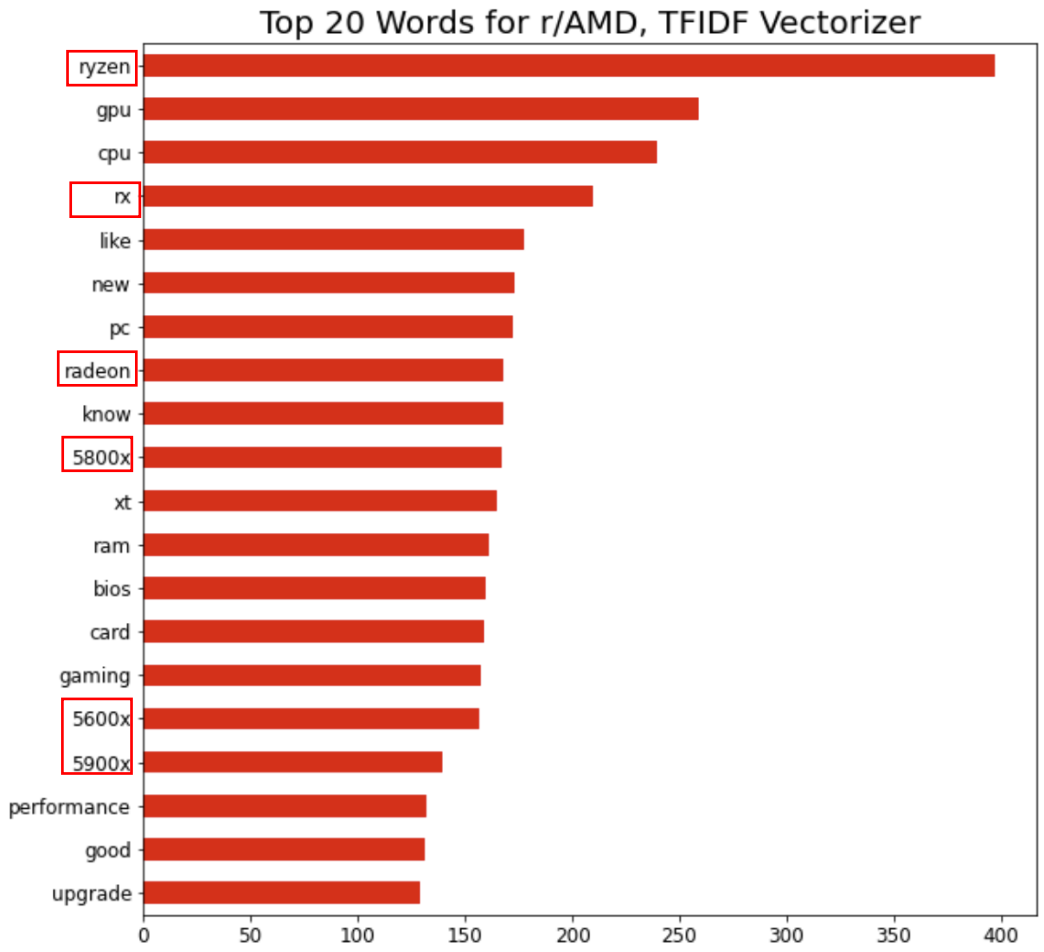
Following was done to clean the reddit title and posts:

1. Links were removed from the posts. Regex selector = r"http.:+[^\s]+[\w]“.
2. Markdown codes such as ‘’, ‘<’, ‘>’, /n were removed.
3. Common words were excluded from model using the spaCy (https://spacy.io/) list of English stop-words.
4. Dead giveaway words like AMD and Intel were excluded as well.
5. Reddit posts and titles were merged into one ‘Text’ column.

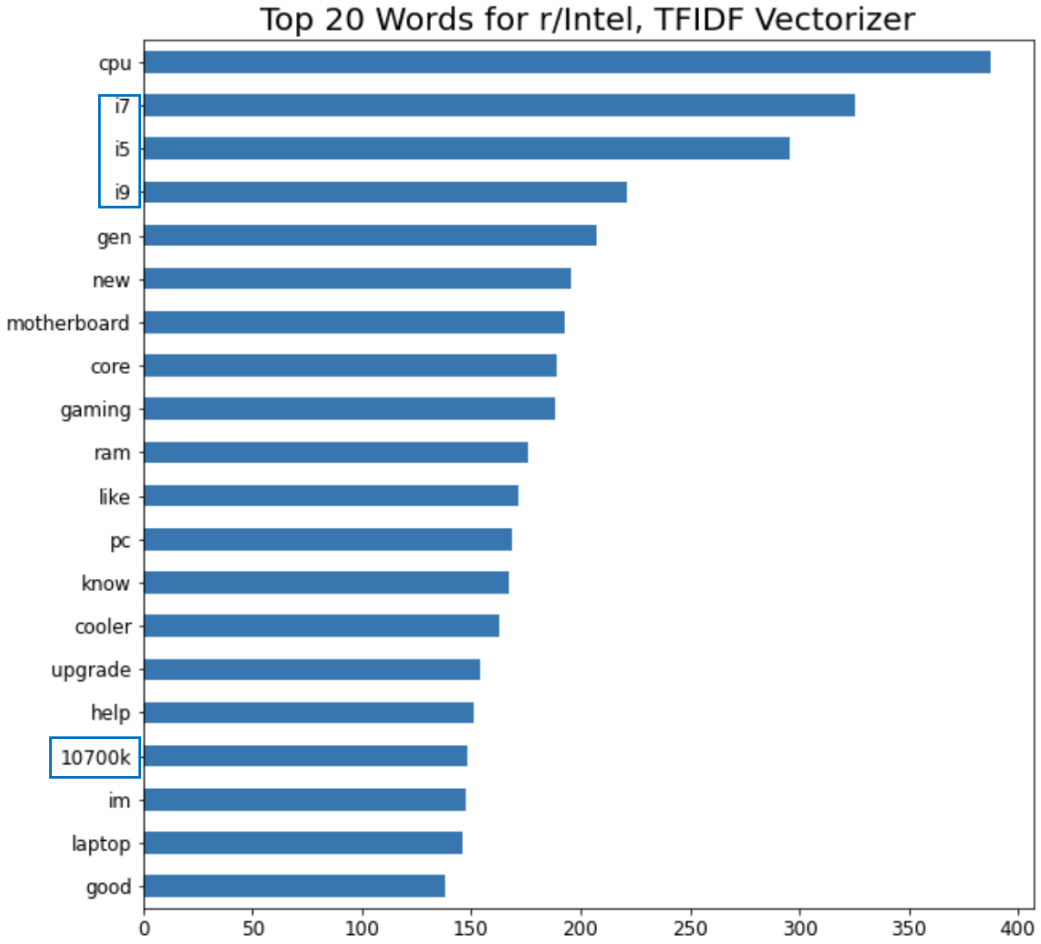
# Exploratory Data Analysis

Comparing Top Words

**r/AMD**



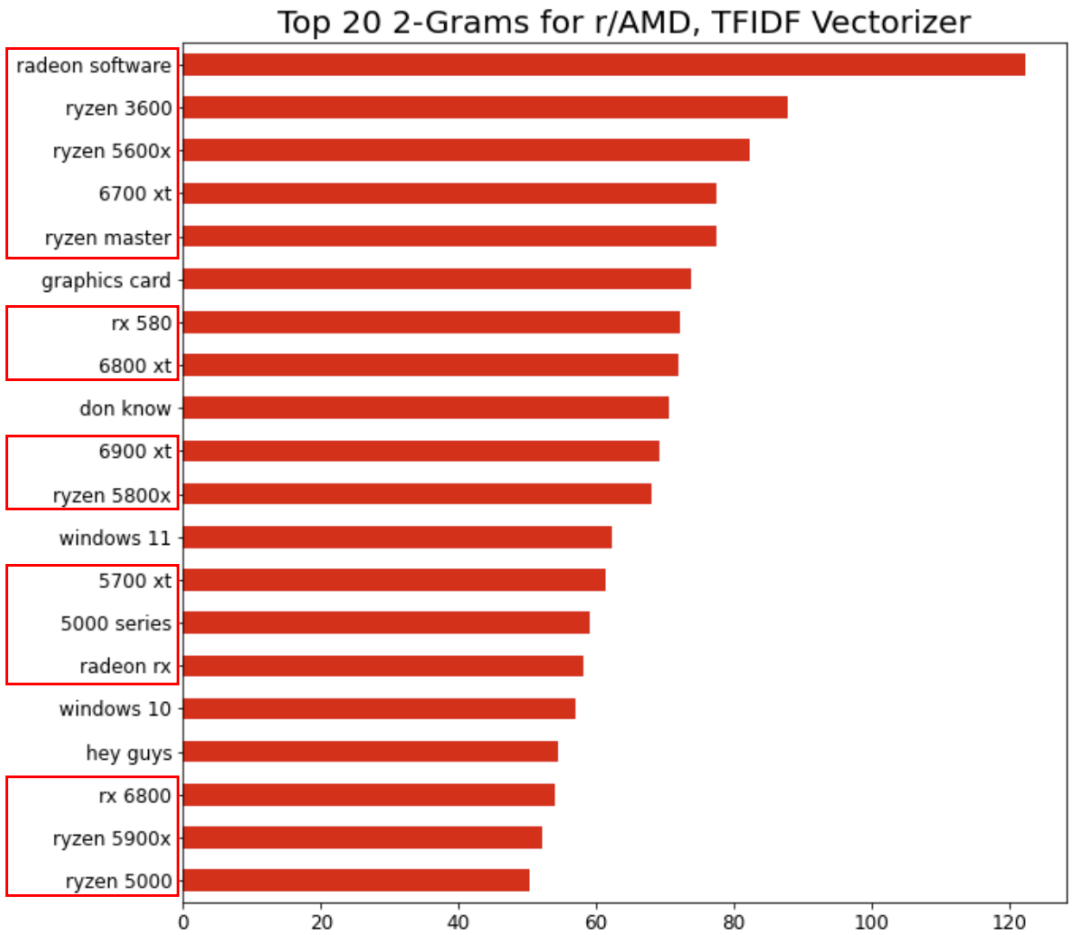

**r/Intel**



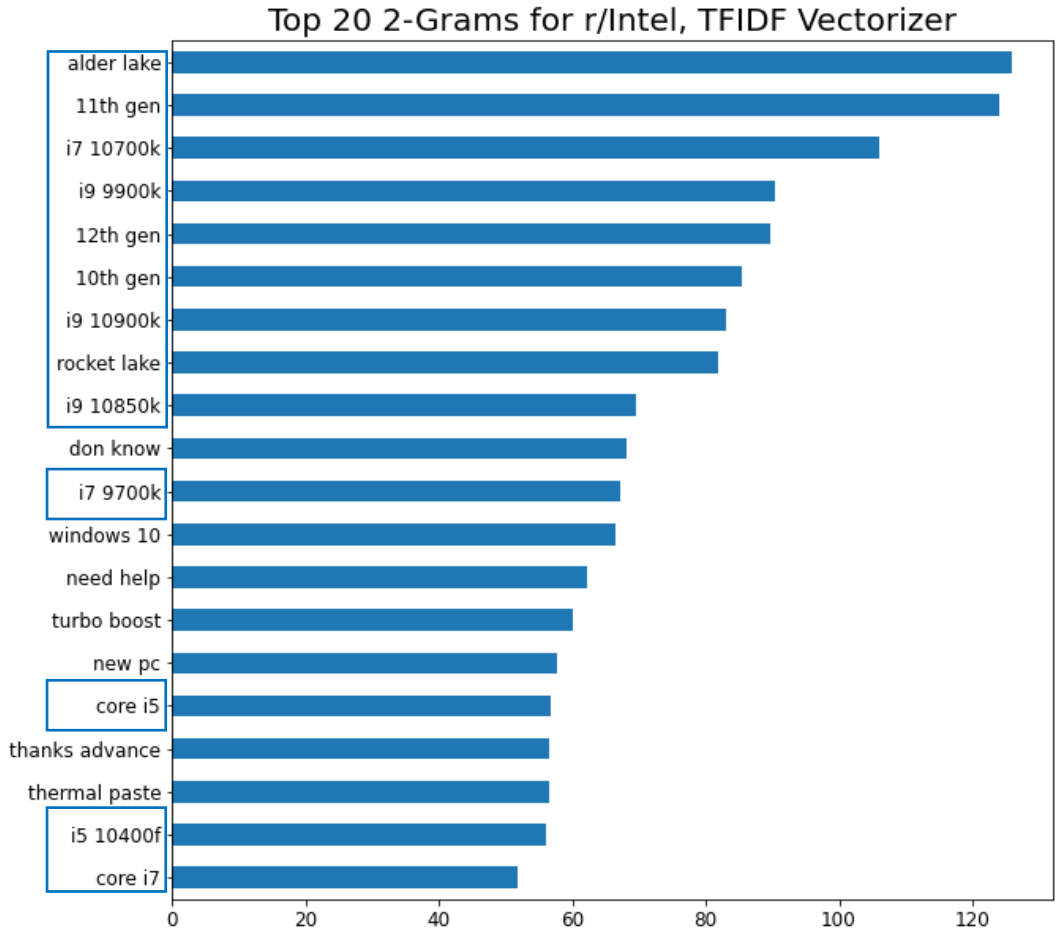
# Exploratory Data Analysis

Comparing Top Bigrams

**r/AMD**




**r/Intel**



# Base Model Benchmarking and Selection

## Parameters

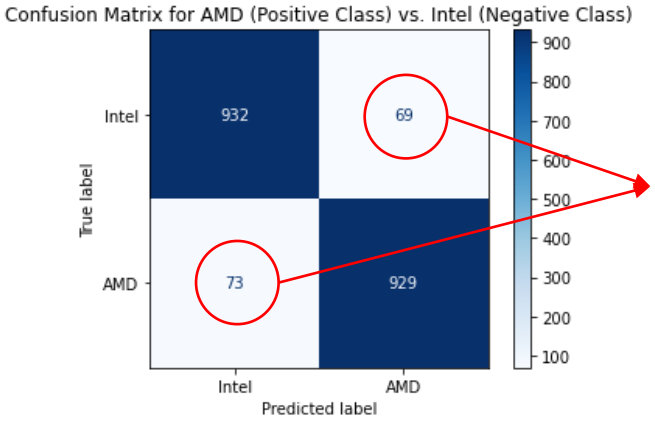
| | | |
|---|---|---|
| Vectorizers | CountVectorizer() | 'max_features': [5000], |
| | | 'n_gram_range': [(1,1)] |
| | TfidfVectorizer() | 'stop_words' : [new_spacy], |
| Classifiers | LogisticRegression() | Default |
| | MultinomialNB() | |
| | RandomForestClassifier() | |

## Model Performance

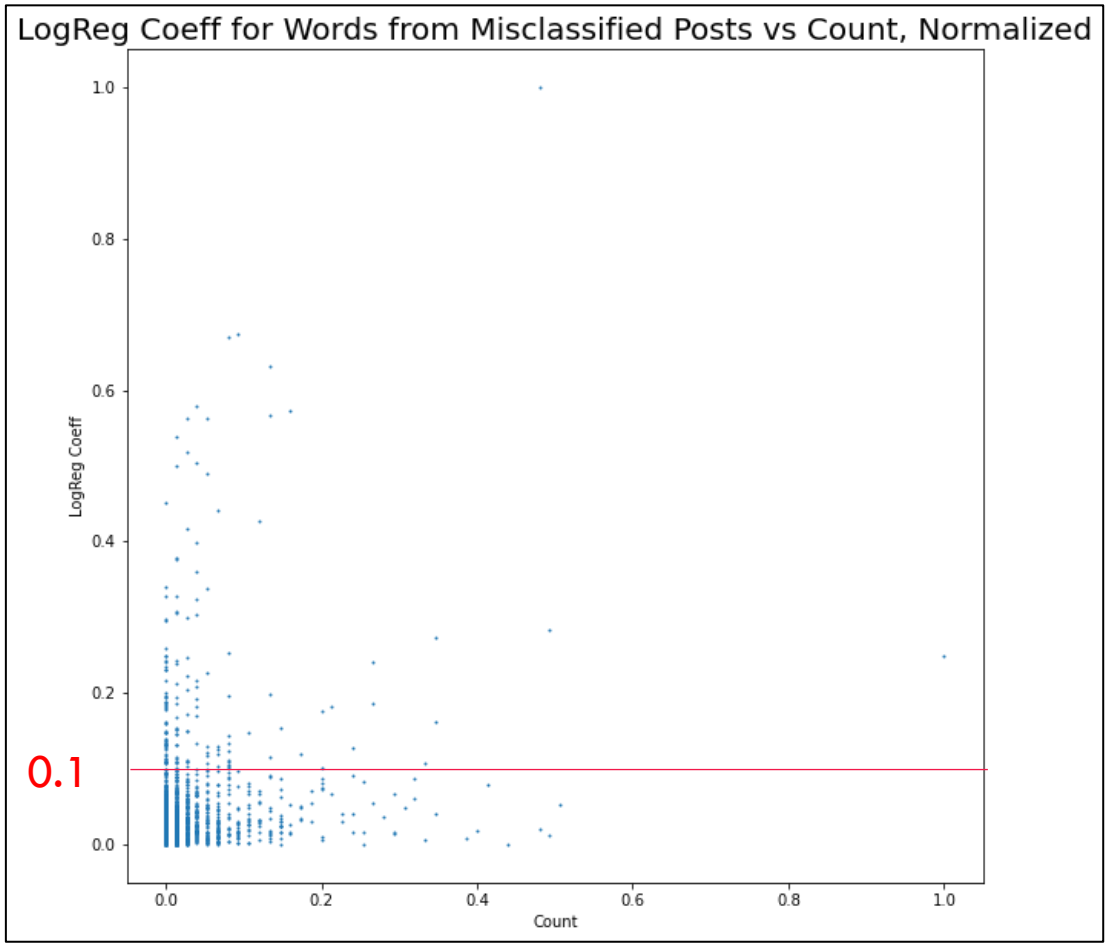
| | | Accuracy Score | | |
|---|---|---|---|---|
| **Vectorizer** | **Classifier** | **Cross-Val Score** | **Train Score** | **Test Score** |
| **CountVectorizer(stop_words=['amd', 'intel'])** | **LogisticRegression()** | **91.58%** | **98.62%** | **91.66%** |
| CountVectorizer(stop_words=['amd', 'intel'] + spaCy)) | LogisticRegression() | 91.95% | 98.51% | 92.21% |
| CountVectorizer(stop_words=['amd', 'intel'] + spaCy])) | MultinomialNB() | 91.42% | 92.86% | 91.06% |
| CountVectorizer(stop_words=['amd', 'intel'] + spaCy)) | RandomForestClassifier() | 92.11% | 99.87% | 91.96% |
| **TfidfVectorizer(stop_words=['amd', 'intel'] + spaCy))** | **LogisticRegression()** | **92.74%** | **95.71%** | **92.91%** |
| TfidfVectorizer(stop_words=['amd', 'intel'] + spaCy)) | MultinomialNB() | 91.00% | 92.58% | 91.51% |
| TfidfVectorizer(stop_words=['amd', 'intel'] + spaCy)) | RandomForestClassifier() | 91.94% | 99.87% | 92.46% |

+1.25%

# Feature Engineering




Tokenized




| E.g. Words Dropped from Model | |
|---|---|
| Numbers | 12 |
| | 120 |
| | 1200 |
| Non-Descriptive | add |
| | experience |
| | feel |
| | hard |
| Other Brands | youtube |
| | creative |
| | crucial |
| | evga |

| | | Accuracy Score | | |
|---|---|---|---|---|
| **Vectorizer** | **Classifier** | **Cross-Val Score** | **Train Score** | **Test Score** |
| TfidfVectorizer(stop_words=['amd', 'intel'] + spaCy) | LogisticRegression() | 92.74% | 95.71% | 92.91% |
| TfidfVectorizer() w/ added Stop-Words | LogisticRegression() | 93.41% | 96.20% | 93.11% |

+0.20%

# Hyperparameter Tuning

## Hyperparameters Tuned

| | | |
|---|---|---|
| Vectorizers | TfidfVectorizer() | **'max_features':** [5000, 10_000, 20_000, 30_000, 40_000, 50_000] |
| | | **'n_gram_range':** [(1,1), (1,2), (1,3)], |
| | | **'tvec__max_df':** np.linspace(0.9,1.0,6) |
| | | **'tvec__min_df':** np.linspace(1,10,10) |
| Classifiers | LogisticRegression() | **'lr__C':** np.linspace(0.001,1,11) |
| | | 'lr__solver': ['saga'], 'lr__penalty': ['elasticnet'], **'lr__l1_ratio':** np.linspace(0, 1, 11) |

| | | Accuracy Score | | | |
|---|---|---|---|---|---|
| **Vectorizer** | **Classifier** | **Cross-Val Score** | **Train Score** | **Test Score** | **Hyperparameters** |
| TfidfVectorizer() w/ added Stop-Words | LogisticRegression() | 93.41% | 96.20% | 93.11% | 'tvec__max_features' = 5_000, 'tvec__n_gram_range' = (1,1) |
| TfidfVectorizer() w/ added Stop-Words | LogisticRegression() | 93.56% | 96.83% | 93.71% | 'tvec__max_features' = 30_000, 'tvec__n_gram_range' = (1,1), 'lr__C': 1.0 |

+0.60%

# Production Model

# Production Model Performance

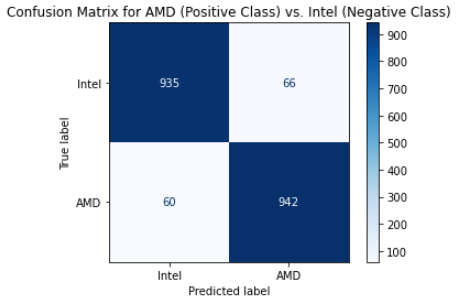
| | | Accuracy Score | | | |
|---|---|---|---|---|---|
| **Vectorizer** | **Classifier** | **Cross-Val Score** | **Train Score** | **Test Score** | **Hyperparameters** |
| CountVectorizer(stop_words=['amd', 'intel']) | LogisticRegression() | 91.58% | 98.62% | 91.66% | Default |
| TfidfVectorizer() w/ added Stop-Words | LogisticRegression() | 93.56% | 96.83% | 93.71% | 'tvec__max_features' = 30_000,<br>'tvec__n_gram_range' = (1,1),<br>'lr__C': 1.0 |

+2.05%

| Word | LogReg Coeff |
|---|---|
| ryzen | 10.0 |
| 5800x | 6.4 |
| 5900x | 6.1 |
| x570 | 5.8 |
| radeon | 5.6 |
| rx | 5.3 |
| 5600x | 5.3 |
| xt | 5.0 |
| 5950x | 5.0 |
| 6900xt | 4.7 |
| b550 | 4.6 |
| 6800xt | 4.4 |
| pbo | 4.2 |
| 6700xt | 4.1 |
| fsr | 3.7 |
| 21 | 3.5 |
| r5 | 3.3 |
| vega | 3.3 |
| 5600g | 3.3 |

| Word | LogReg Coeff |
|---|---|
| i7 | -9.3 |
| i5 | -8.1 |
| 10900k | -6.8 |
| i9 | -6.5 |
| 10700k | -6.2 |
| z490 | -6.2 |
| 12900k | -5.4 |
| i3 | -5.0 |
| z690 | -5.0 |
| lake | -4.5 |
| 9900k | -4.4 |
| 10850k | -4.3 |
| 12700k | -4.0 |
| 10600k | -3.9 |
| 12600k | -3.9 |
| optane | -3.5 |
| xeon | -3.4 |
| xe | -3.3 |
| uhd | -3.2 |

# Production Model Error Analysis



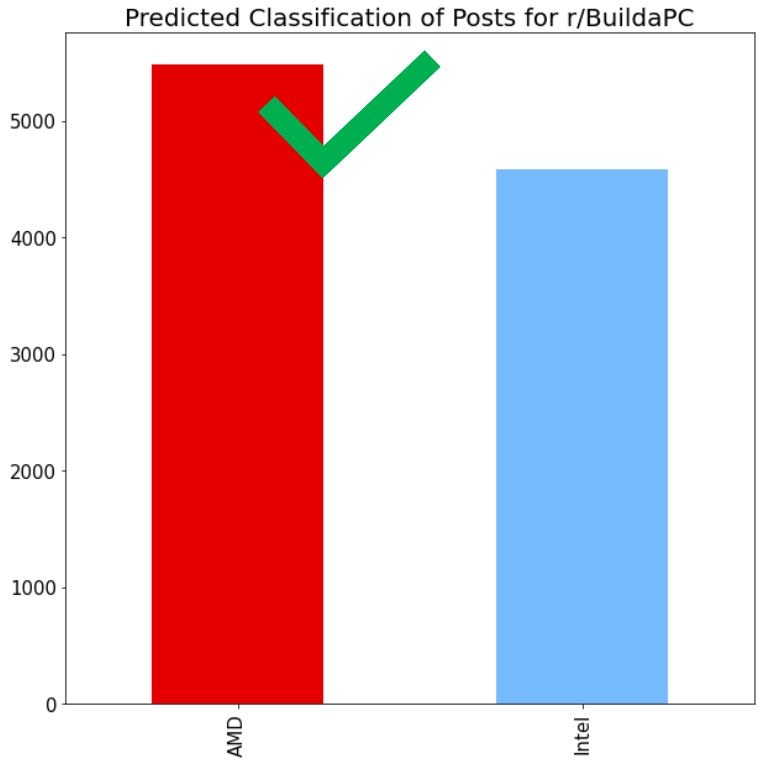

Error Analysis on Misclassified Posts

| Misclassification Category | Example of Post |
|---|---|
| Comparison post between AMD products vs Intel Products | "which would you choose between a ryzen 5800x and a 10850k? looking for the best gaming performance and futureproof (not looking to upgrade CPU in the next 3-4years), which option would be the best for me?" |
| General post that can be applicable to both AMD/Intel | "What do you call Intel's way of advertising? Its for my essay I forgot what its called when they showcase their new products on the stage and a lot of people are watching? is there a term to call this?" |
| Posts that don't talk about the products at all | “Honestly, this is a way better sub than /r/AMD The difference between this sub and /r/AMD , is that you can criticize Intel here and you won't immediately be called shill, gaslighted or told that you're spreading FUD. Whereas any criticism of AMD anywhere elicits an immediate attacking and scathing response from AMD fanboys, who seem to be extremely insecure.” |

# Production Model Deployment and Recommendations

Deploying on r/BuildaPC

| | | Accuracy Score | | |
|---|---|---|---|---|
| **Vectorizer** | **Classifier** | **Cross-Val Score** | **Train Score** | **Hyperparameters** |
| TfidfVectorizer() w/ added Stop-Words | LogisticRegression() | 93.04% | 96.82% | 'tvec__max_features' = 20_000, 'tvec__n_gram_range' = (1,1), 'lr__C': 1.0 |




Therefore, I will be choosing to carry **AMD** CPUs for my custom PC building startup.

# Conclusion

Possible Future Works

- Run sentiment analysis to see whether posts are positive or negative against AMD and Intel.
- Do multiclass classification for the GPU market (Nvidia RTX, AMD Radeon, Intel Arc) to see which GPU brand I should carry.
- Explore black-box models/deep learning models to improve accuracy of the model.

# Thank You

Nor Rashidi Bin Norhashim

Rashidi.norhashim@gmail.com