Mildly Interesting and Interesting af

By: Christopher Sycamore

Introduction

Me: I’m a data science student currently enrolled in the General Assembly Data science immersive program.

Objective: Looking to build the best classification model to determine which subreddit posts are mildly interesting vs interesting af.

Hypothesis: Based on the user determining which post is interesting af vs mildly interesting, I have low confidence my model will be able to reach an accuracy score over 70%. I predict ExtraTrees will be the best model.

How to access Reddit API

PRAW(Python Reddit API Wrapper) - is a special wrapper used to access Reddit’s API with python.

Creating an account

The subreddits: Interesting af Top 5 posts

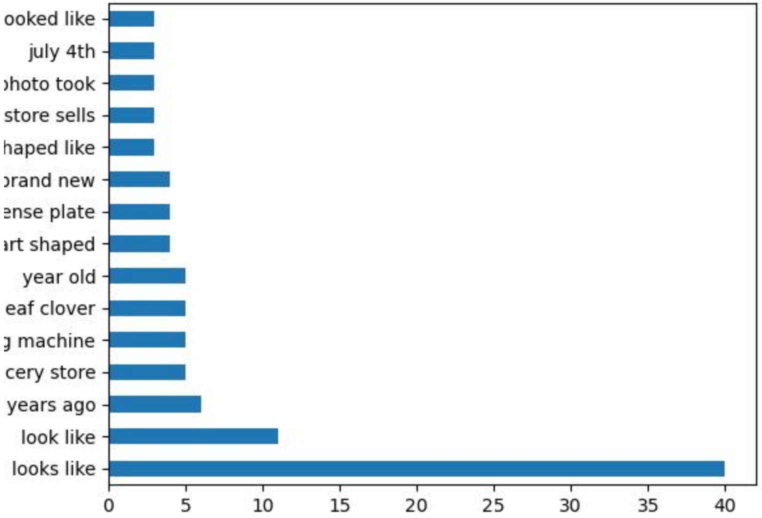
Date: 2022-03-01	In 1996 Ukraine handed over nuclear weapons to Russia "in exchange for a guarantee never to be threatened or invaded".
Date: 2022-02-28	Ukrainian ambassador to the UN pretty much tells Putin to kill himself: "If he wants to kill himself, he doesn't need to use nuclear arsenal. He has to do what the guy in Berlin did in a bunker in May 1945"
Date: 2021-06-30	“The dog on the Left is award winning showdog named Arnie an AKC French Bulldog..The dog on the right is Flint, bred in the Netherlands by Hawbucks French Bulldogs - a breeder trying to establish a new, healthier template for French Bulldogs.
Date: 2022-02-06	My turtle follows me and seeks out affection. Biologist have reached out to me because this is not even close to normal behavior. He just started one day and has never stopped. I don’t know why.
Date: 2021-05-02	I created a photorealistic image of George Washington if he lived in the present day.

The subreddits: Mildly Interesting Top 5 posts

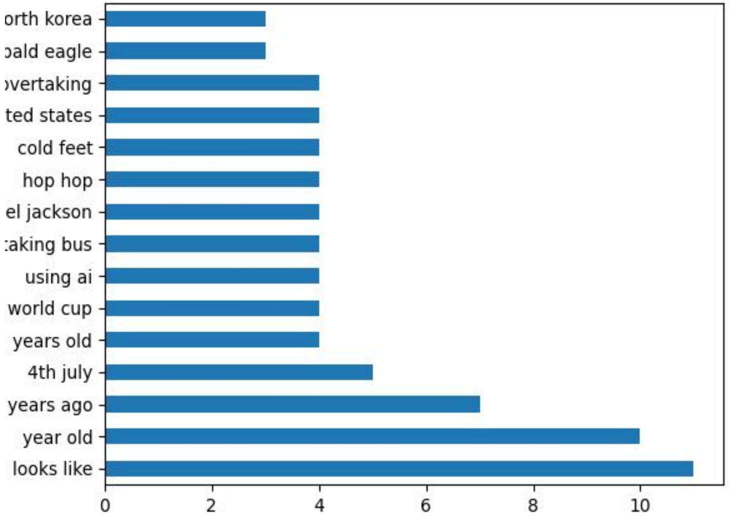
Date: 2022-03-01	Local Mexican restaurant used to be a Chinese restaurant. Instead of painting over a mural, they just put sombreros on the pandas.
Date: 2022-02-28	Found the cliff this Clif bar came from.
Date: 2021-06-30	I bought some suspiciously perfect bananas yesterday.
Date: 2022-02-06	My hotel phone in Iceland has a special button that will wake you up if there are northern lights in the sky.
Date: 2020-06-03	I have a hereditary gap in my eyebrow

Most used Bi-grams

Mildly interesting



Interesting af



Natural Language Preprocessing

NLP(Natural Language Processors) lets machines conduct text analysis, speech analysis and understand emotions, expressions and intent.

https://www.digitalaptech.com/natural-language-processing-definition-techniques-components-and-more/#:~:text=Natural%20Language%20Processing%20combines%20machine,understand%20emotions%2C%20expressions%20and%20intent.

Two NLP Preprocessors are:

Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer

Count Vectorizer

Types of Classification models

Logistic Regression

K-Nearest Neighbours

Kernel SVM

Naive Bayes

Decision Tree Classification

Random Forests

ExtraTrees Classifier

Random Forest Confusion Matrix

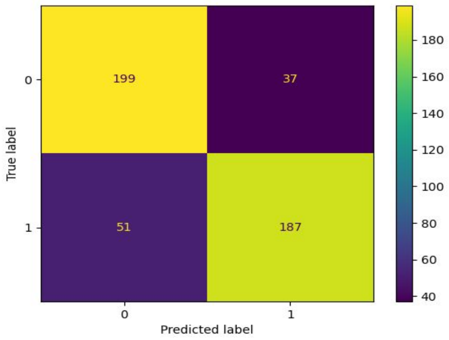
Using a Tf-IDF preprocessor:

- Accuracy = 81%
- R2 score = 78.9%

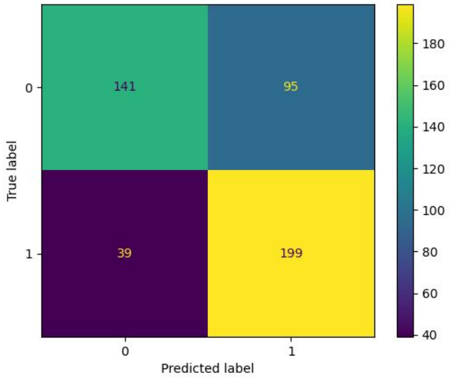
Using CountVectorizer preprocessor:

- Accuarcy = 71.7%
- R2 score = 79.7%

Tf-IDF CM



CountVecotorizer CM



Logistic Regression

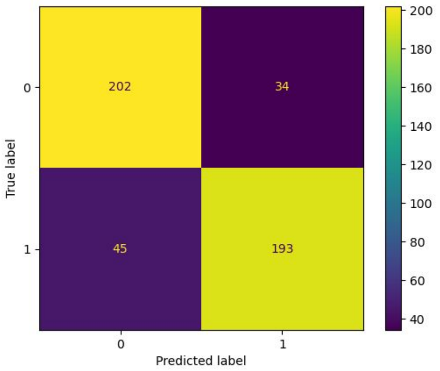
Using a Tf-IDF preprocessor:

- Accuracy = 83.3%
- R2 score = 83.3%

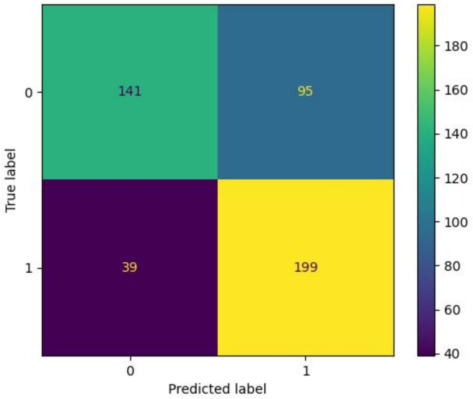
Using CountVectorizer preprocessor:

- Accuracy = 71.7%
- R2 score = 71.7%

Tf-IDF CM



CountVecotrizer CM



Extra Trees

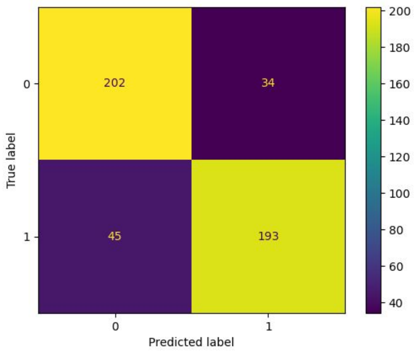
Using a Tf-IDF preprocessor:

- Accuracy = 82.2%
- R2 score = 82.2%

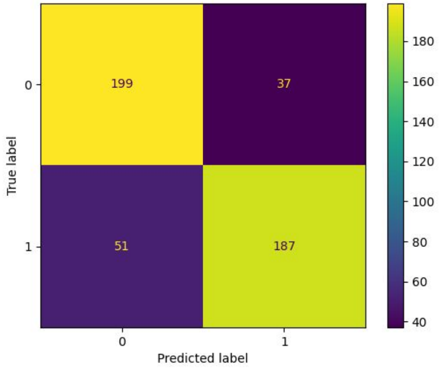
Using CountVectorizer preprocessor:

- Accuracy = 80.5%
- R2 score = 80.5%

Tf-IDF CM



CountVectorizer CM



Winner

Logistic Regression - Tf IDF vectorizer

Using a Tf-IDF preprocessor:

- Accuracy = 83.3%
- R2 score = 83.3%

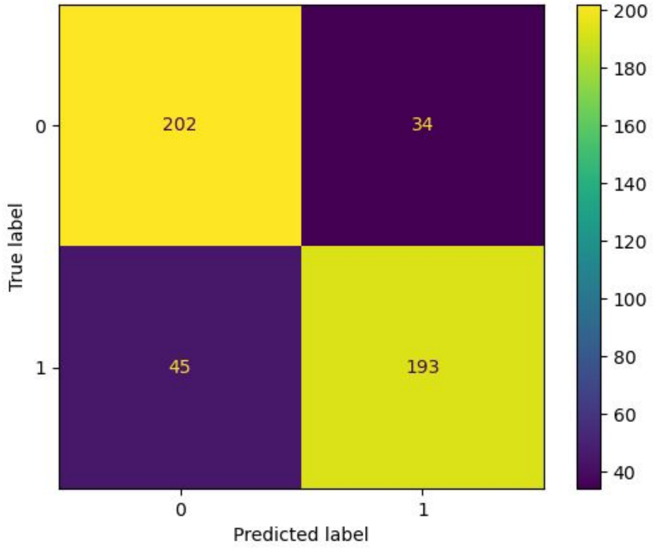
Best Params:

'tvec__max_df': 0.9,

'tvec__max_features': 2000,

'tvec__min_df': 2,

'tvec__ngram_range': (1, 2)



Conclusion

Hypothesis: Based on the user determining which post is interesting af vs mildly interesting, I have low confidence my model will be able to reach an accuracy score over 70%. I predict ExtraTrees will be the best model.

Conclusion: My models worked better than I anticipated with all three models having an accuracy of over 70%. ExtraTrees was not the best model, Logistic Regression was the best model.

Takeaway: This was a fun challenge. I wish I had managed my time better.