Text Classification Model for Subreddit

Leveraging NLP for Binary Text Classification

Dogs vs Finances

By: Damar Shipp

Contact: LinkedIn

TABLE OF CONTENTS

01
TARGET
Objective: Identify target audience

02
PROBLEM & SOLUTION
Objective: Classify Subreddit post based on their given text by a classification model.

03
DATA OVERVIEW
Objective: Show the data we used to create the model.

04
METHODOLOGY
Objective: Show what features were used to create the model.

05
EDA
Objective: Steps taken to ensure the model works properly.

06
MODEL PERFORMANCE
Objective: Show how the model performed.

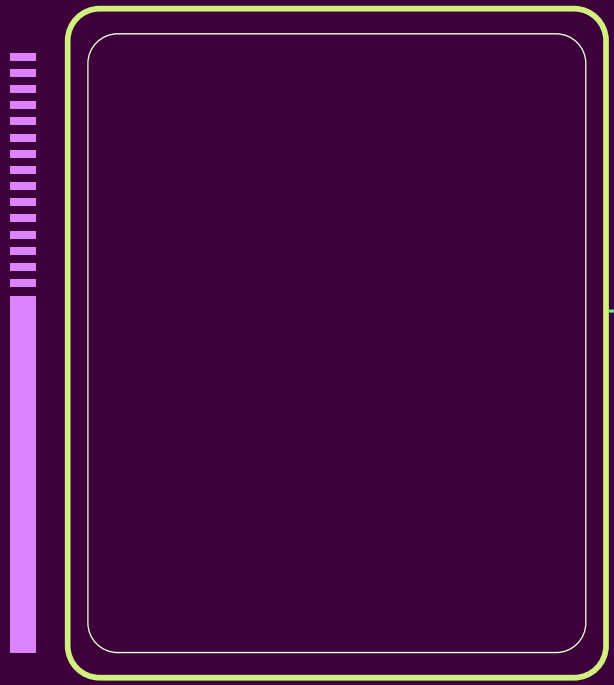
07
Conclusion
Objective: Show Findings and recommendations.

08
Next Steps
Objective: Show possible improvements for the model

Target Audience

Target:

Platform moderators and data analyst/ researchers



PROBLEM STATEMENT

Problem

- Reddit is comprised of numerous communities known as subreddits, where each subreddit is tailored to a specific interest, and misclassified posts can disrupt community discussions and reduce user engagement. To address this; I have tailored this project to develop and evaluate two supervised learning models that will determine which subreddit the post belongs to.

Solution

Develop and evaluate two supervised learning models that will determine which subreddit the post belongs to.

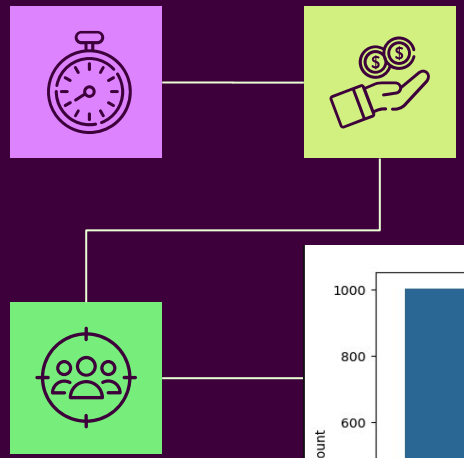
DATA OVERVIEW

SOURCE

COLLECTED posts from two subreddits (dogs & finances) using PRAW

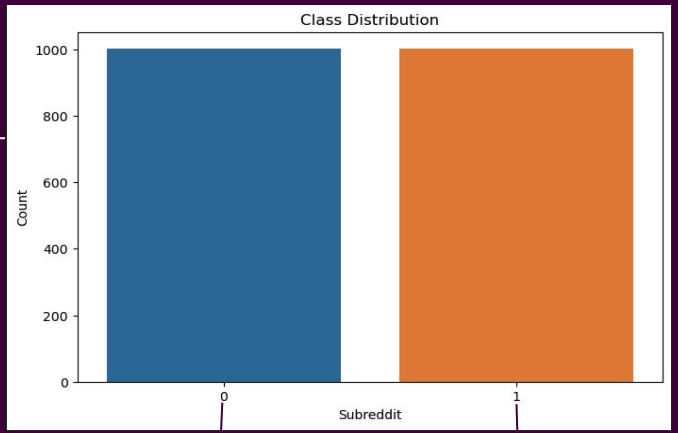
SIZE OF DATA

Dataset size: 2004
New Data Size: 1880



Challenge

Extra text



Finances Dogs

METHODOLOGY

Pipeline Overview:

1. Data preprocessing (stop words removal, lemmatization).
2. Feature extraction (vectorization).
3. Model selection: Logistic Regression and Random Forest.
4. Evaluation using accuracy and classification report

- Justification: Models chosen for interpretability and performance

Exploratory Data Analysis (EDA)

Visualizations:

- Term frequency distribution plots.

- Insights: Key words and patterns indicate distinct subreddit behaviors.

MODEL PERFORMANCE

METRICS:

- ACCURACY, PRECISION, RECALL, F1-SCORE.
- CONFUSION MATRICES FOR BOTH MODELS

Metric	Logistic Regression	Random Forest
Test Set Accuracy	99.83%	98.01%
Precision (dogs_subreddit)	1.00	1.00
Recall (dogs_subreddit)	1.00	0.96
F1-Score (dogs_subreddit)	1.00	0.98
Precision (personalfinance_subreddit)	1.00	0.96
Recall (personalfinance_subreddit)	1.00	1.00
F1-Score (personalfinance_subreddit)	1.00	0.98

HOW THE MODEL PERFORMED ON NEW DATA.

Metric	Logistic Regression	Random Forest
Test Accuracy	99.48%	96.99%

- BEST MODEL: LOGISTIC REGRESSION ACHIEVED THE HIGHEST SCORE ON ALL CATEGORIES. 1.00 = 100%

CONCLUSION & KEY FINDINGS

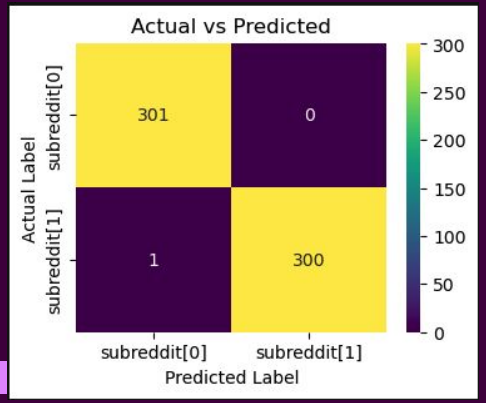
Key Findings:

• - Text classification is effective with proper preprocessing.

• Model Recommendation:

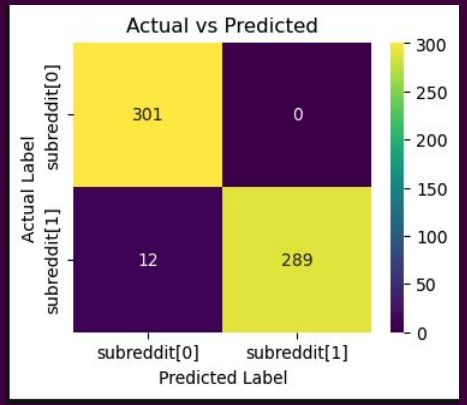
• - Logistic Regression model due to its high accuracy.

- Impact: Enhances subreddit content analysis for targeted actions.



Left side:
Logistics Regression Model
1 = Dogs
0 = Finances

Right side:
Random Forest Model
1 = Dogs
0 = Finances



Next Steps

Improvements:

- Incorporate more data.
- Extend the model to classify posts across more subreddits for testing.
- Experiment with other models and maybe more hyperparameters.

Applications:

- I would feel 100% positive using my model in live action as the accuracy is high enough to not make a drastic negative effect.

Do you have any QUESTIONS?

www.linkedin.com/in/damar-shipp-jr-614b71186

https://github.com/DamarTheMunginizer