



# WEB APIS & CLASSIFICATION PROJECT

Prepared by Jean Brewster

# TABLE OF CONTENTS

- PROBLEM STATEMENT
- METHODOLOGY / APPROACH
- RESULTS
- CONCLUSIONS / NEXT STEPS







## PROBLEM STATEMENT

- Hired by a leading plant-based delivery service who wants to understand through Natural Language Processing (NLP) the posts of two subreddit communities to gain insights for a more data-driven marketing campaign.

# METHODOLOGY

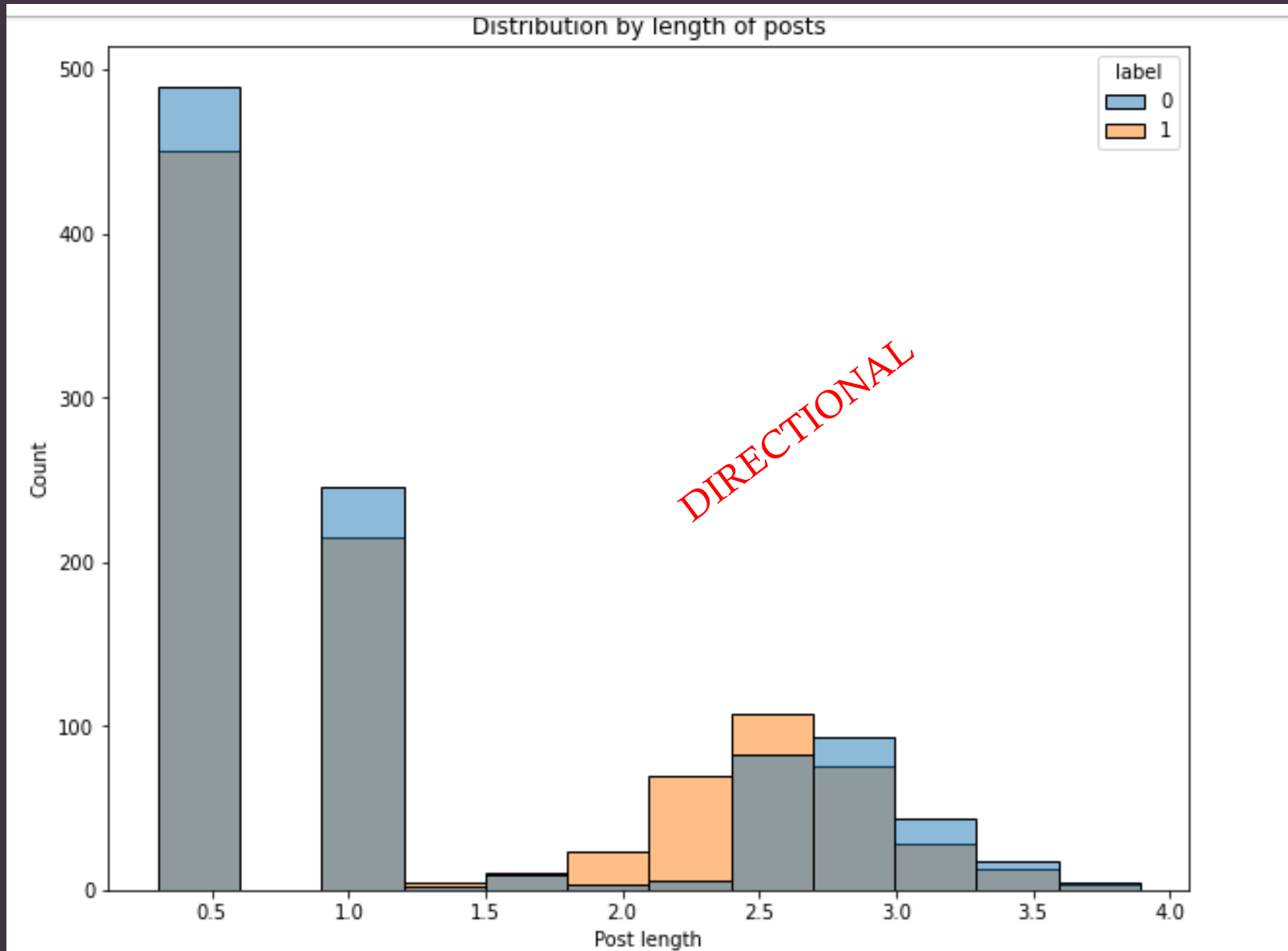
## DATA COLLECTION / CLEANING / EDA

- Used Pushshift's API to collect posts from two subreddits (r/vegan and r/plant-based diet)
- Collected 1,000 posts from each subreddit
- Removed unnecessary columns and kept relevant columns including, posts, number of comments, and author
- Preliminary EDA showed many similarities between the two subreddits

## PREPROCESSING / MODEL DEVELOPMENT

- Used Count Vectorizer and TFIDF Vectorizer to transform text to numerical features
- Removed stop words to focus on important words
- Developed/Evaluated Models
  - Split the data into training and testing for validation
  - Baseline Accuracy Score
  - Multinomial Naïve Bayes
  - Random Forest

# DATA OBSERVATIONS

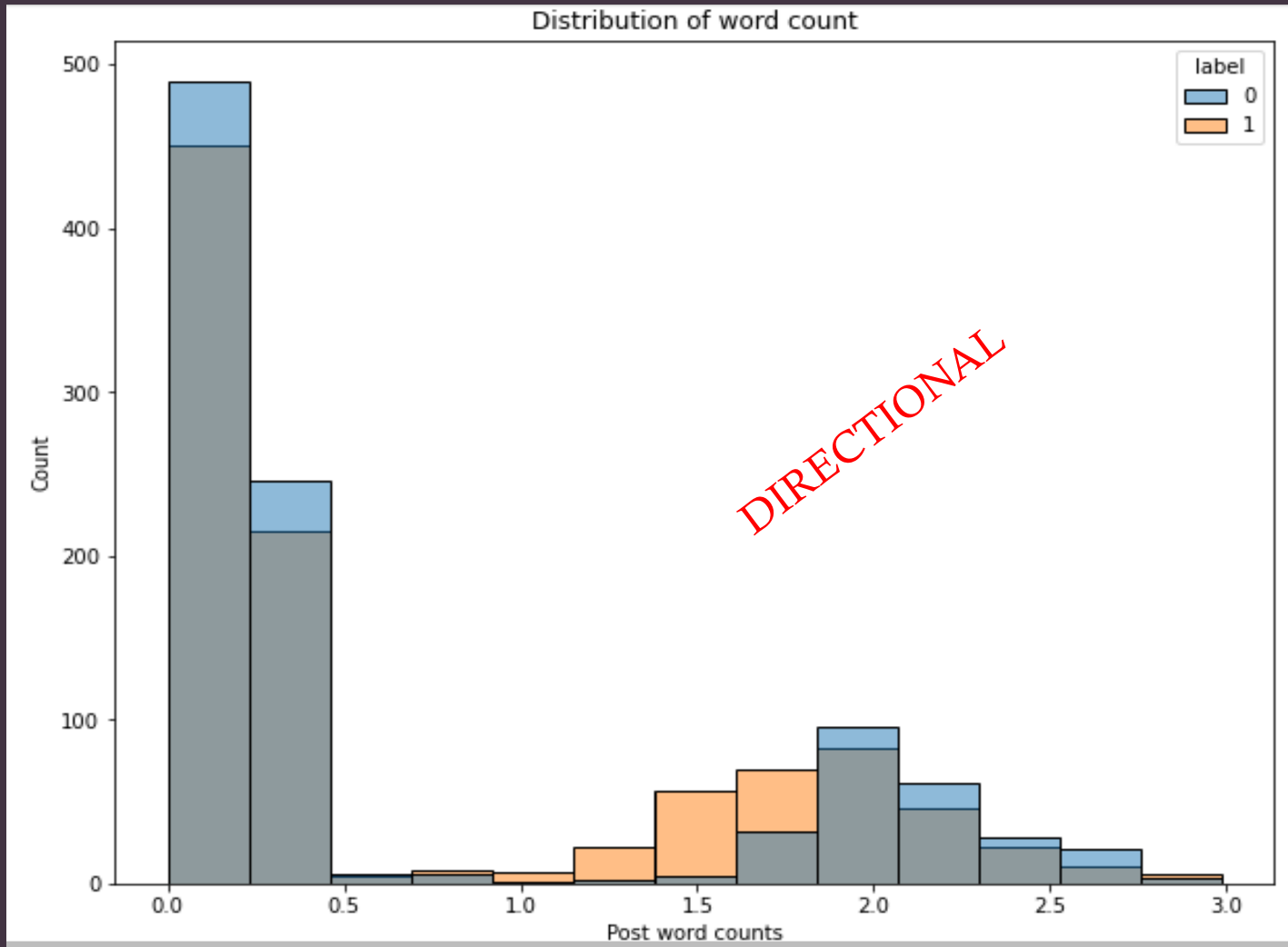


0: Vegan

1: Plant-based diet

## DISTRIBUTION OF POSTS LENGTH

- Converted the data to a log scale because distribution was very skewed
- Majority of posts in both vegan and plant-based are not long posts.
- The posts that are longer tend to come from plant-based subreddit

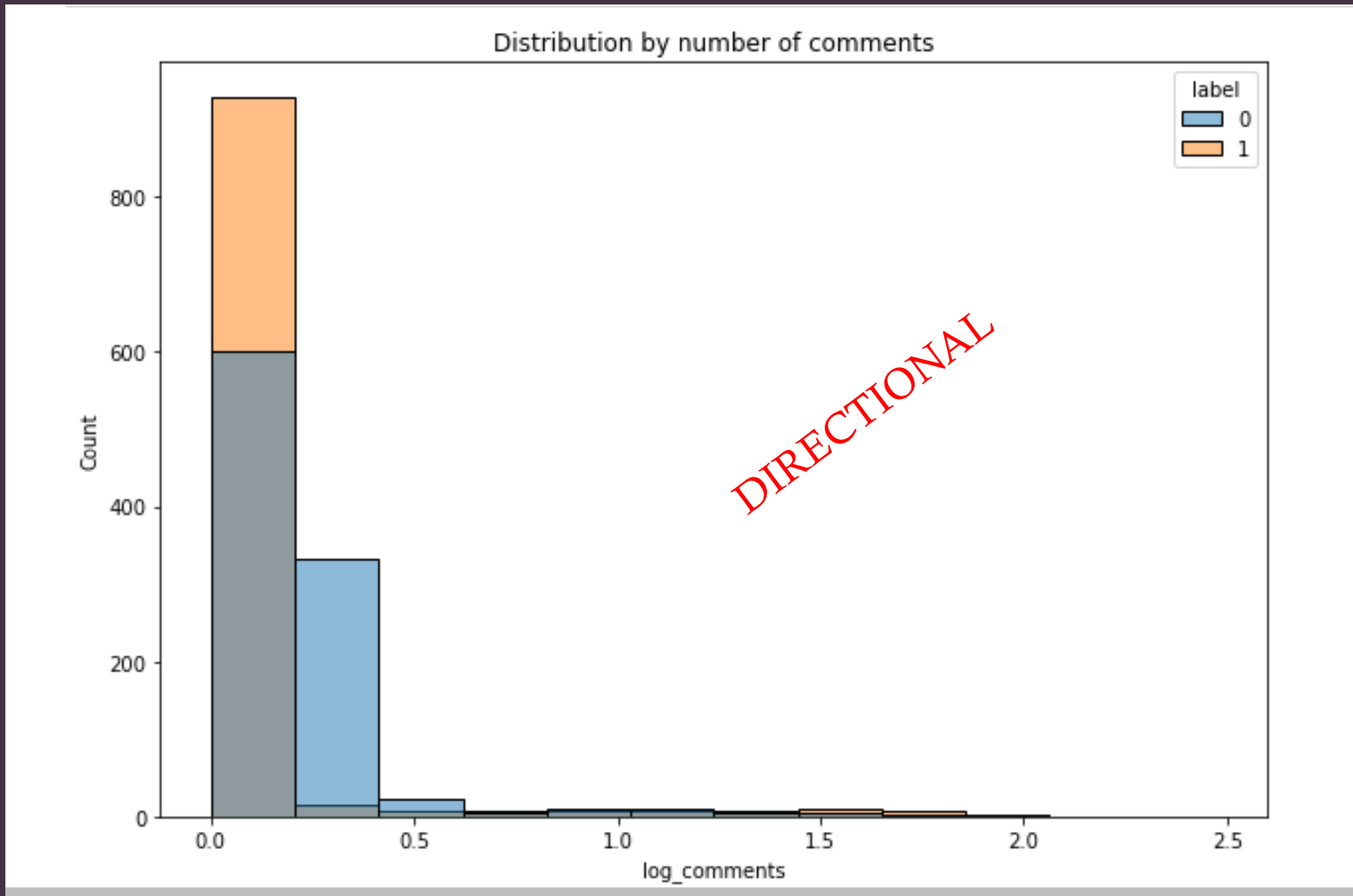


0: Vegan

1: Plant-based diet

## DISTRIBUTION OF WORD COUNT

- Note: Data was converted to a log scale because distribution was very skewed.
- Posts on Reddit tend to be short so it's not surprising that a majority of both vegan and plant-based tend to have fewer words.
- Where there are higher word count is coming mostly from the plant-based subreddit.



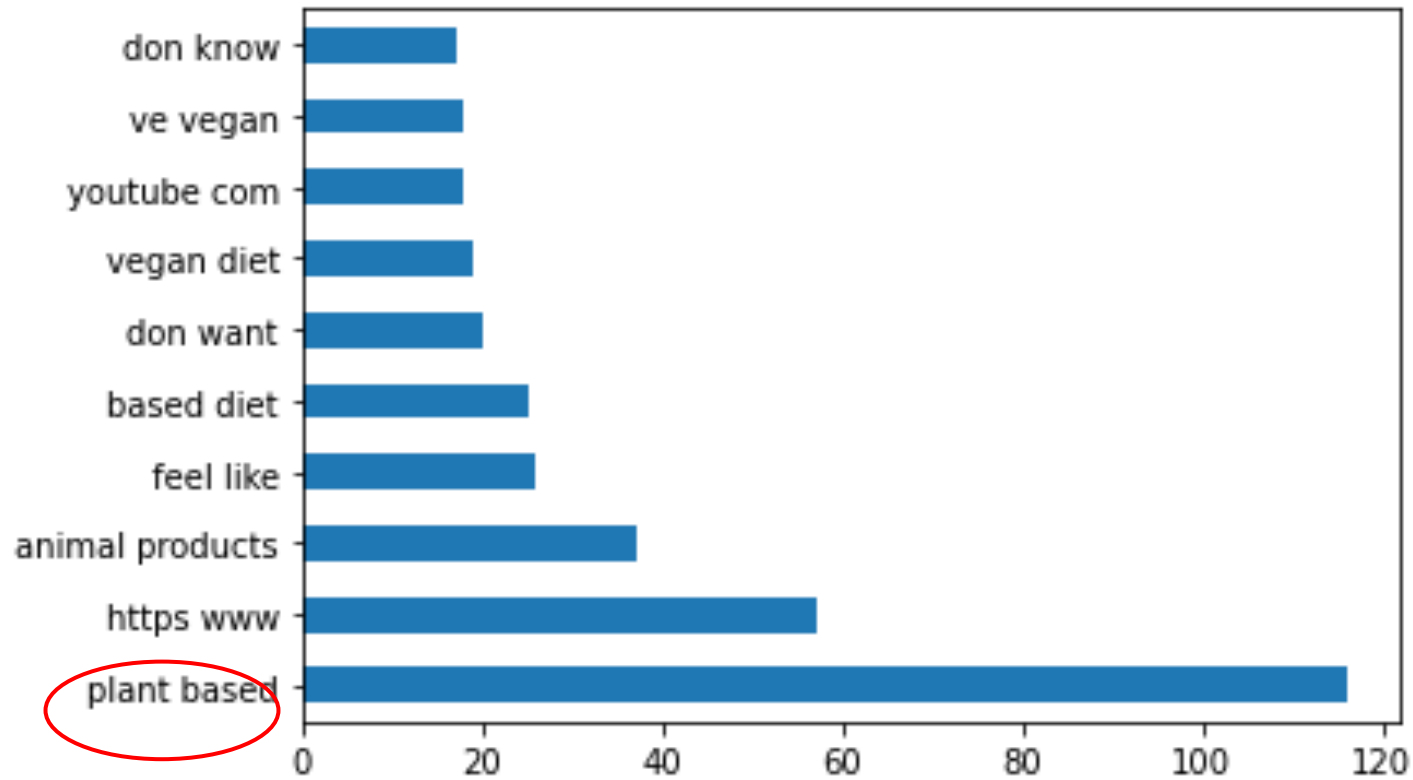
0: Vegan

1: Plant-based diet

## DISTRIBUTION OF NUMBER OF COMMENTS

- Note: Data was converted to a log scale because distribution was very skewed.
- A key measure of engagement is number of comments, and you can see from the chart, both subreddits have relatively low comments and engagement





## TOP 10 WORDS

Even after using Bigram and StopWords features, the top 10 words are not particularly insightful

# MODEL EVALUATION

## KEY METRICS

VECTORIZER	MODEL	TRAIN SCORE	TEST SCORE	BEST PARAMETERS
NA	Baseline	0.50	NA	NA
CountVectorizer	Multinomial Naive Bayes	0.611	0.55	{'cvec__max_df': 0.9, 'cvec__max_features': 2500, 'cvec__min_df': 3, 'cvec__ngram_range': (1, 1)}
TfidfVectorizer	Multinomial Naive Bayes	0.632	0.584	{'tvec__max_features': 3000, 'tvec__ngram_range': (1, 1), 'tvec__stop_words': 'english'}
TfidfVectorizer	Random Forest	0.617	0.624	{'tvec__max_features': 1000, 'tvec__ngram_range': (1, 2), 'tvec__stop_words': 'english'}

- While all models, performed better than the baseline accuracy score of 0.50, the Random Forest model with TfidfVectorizer has the best predictive performance on this classification problem.
- Max\_features as 1000 which is the smallest number among the max\_features parameters tested for, and n\_gram range as 1-2, which prefers up to two-words and the removal of stop words.



## CONCLUSION / NEXT STEPS

- Measurable increase in the predictive performance of all models compared to the baseline accuracy.
- Similarities between vegan and plant-based subreddits were challenging for the models and resulted in relatively low accuracy scores.
- Next steps include, exploring other relevant subreddits; collecting more training data; doing more pre-processing (e.g., removing more stop words); conducting sentiment analysis; and trying additional models like boosting and K-Nearest Neighbors.