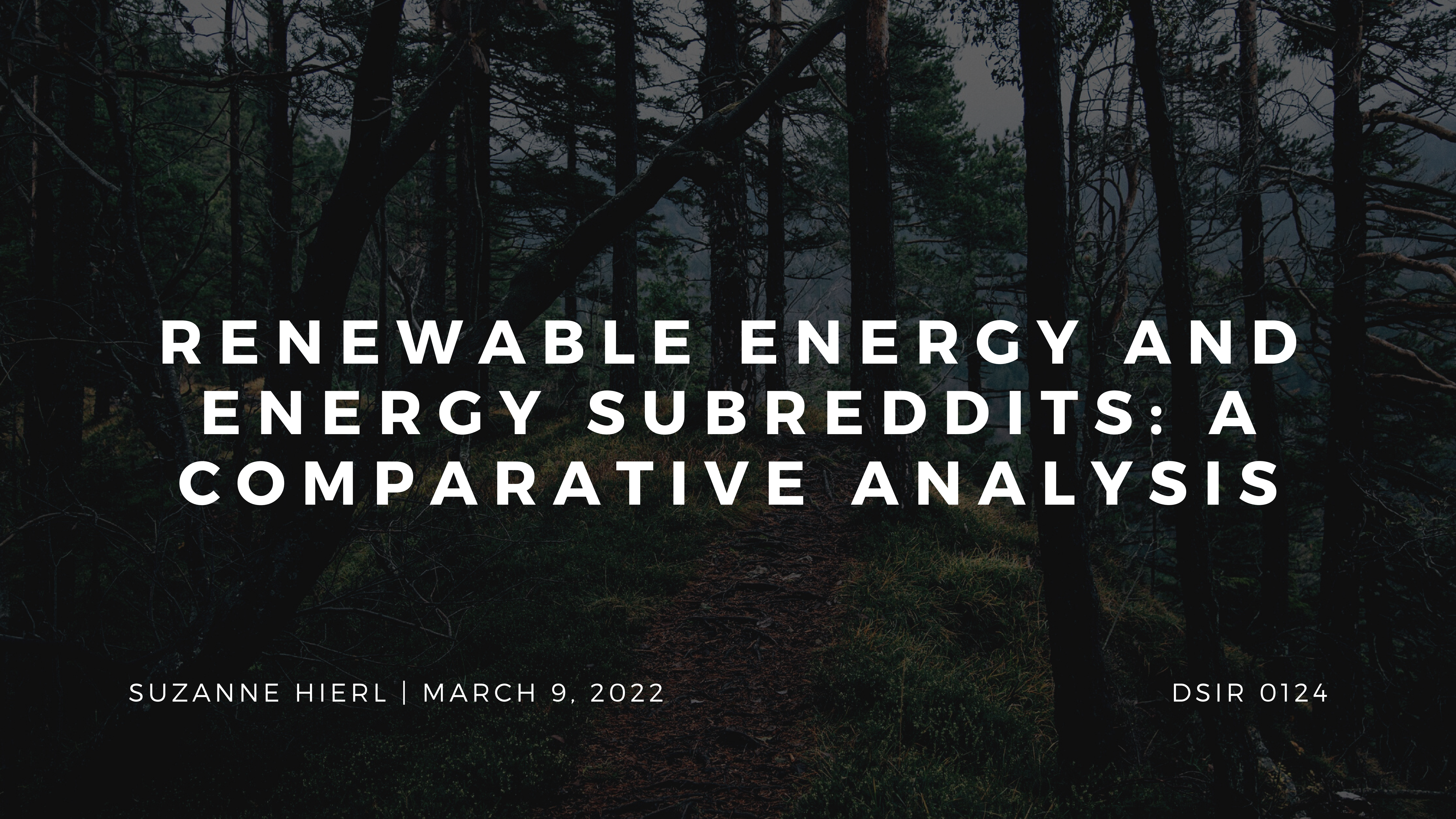


RENEWABLE ENERGY AND ENERGY SUBREDDITS: A COMPARATIVE ANALYSIS

SUZANNE HIERL | MARCH 9, 2022

DSIR 0124

INTRODUCTION

Faced with a rapidly changing climate and declining fossil fuel reserves, the shift from traditional fossil fuels to renewable sources of energy (including wind, solar, hydro, geothermal, and bioenergy) becomes increasingly important year to year. The language surrounding this shift, as observed in subreddits r/RenewableEnergy and r/Energy, is reflective of its complex nature.

INTRODUCTION

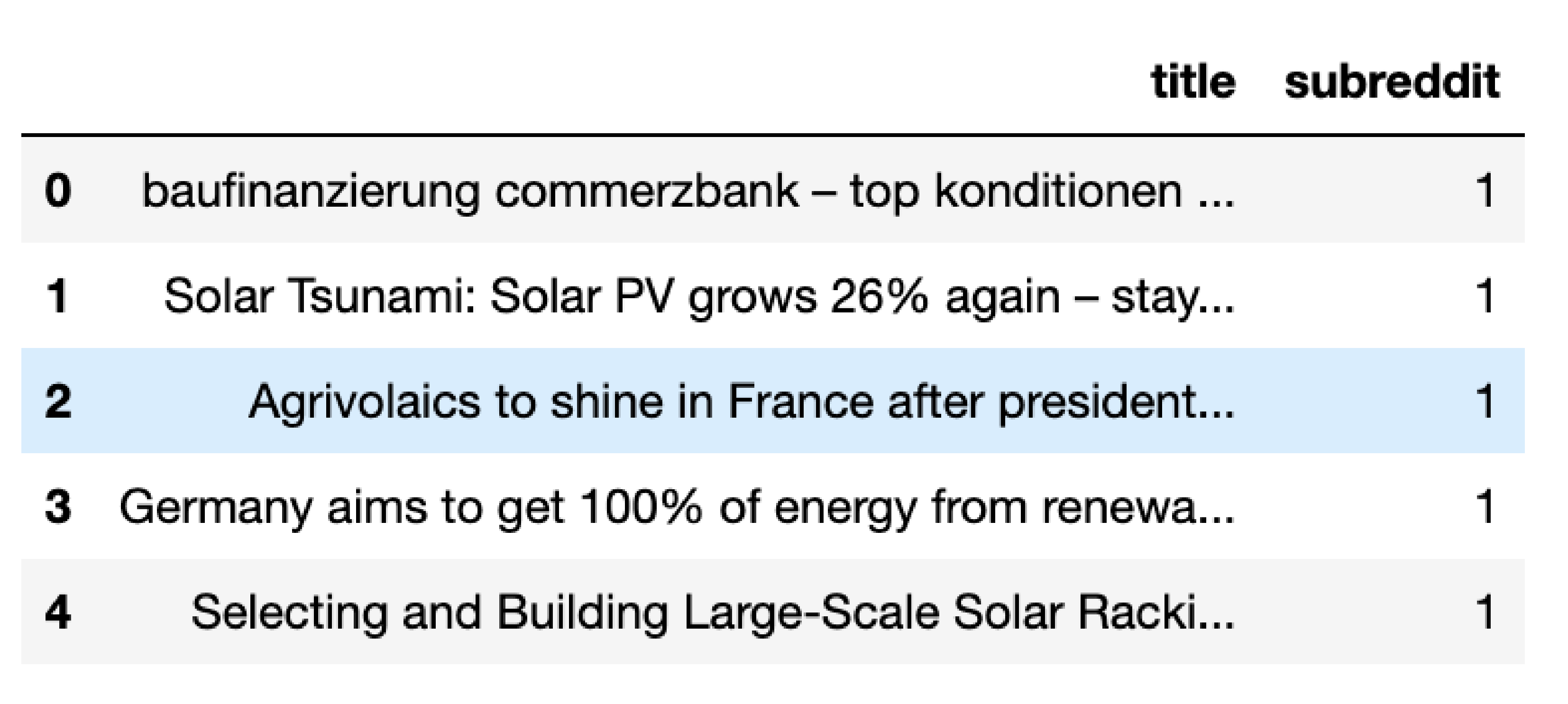
The analysis of the each subreddit and the comparison between the two has the following aims:

- Document the current language and public sentiment surrounding each topic.
- Identify linguistic overlap and differences, potentially to inform language or strategies to be used towards bridging the two subjects and ultimately strengthening support for Renewable Energy.

Data Collection and Processing

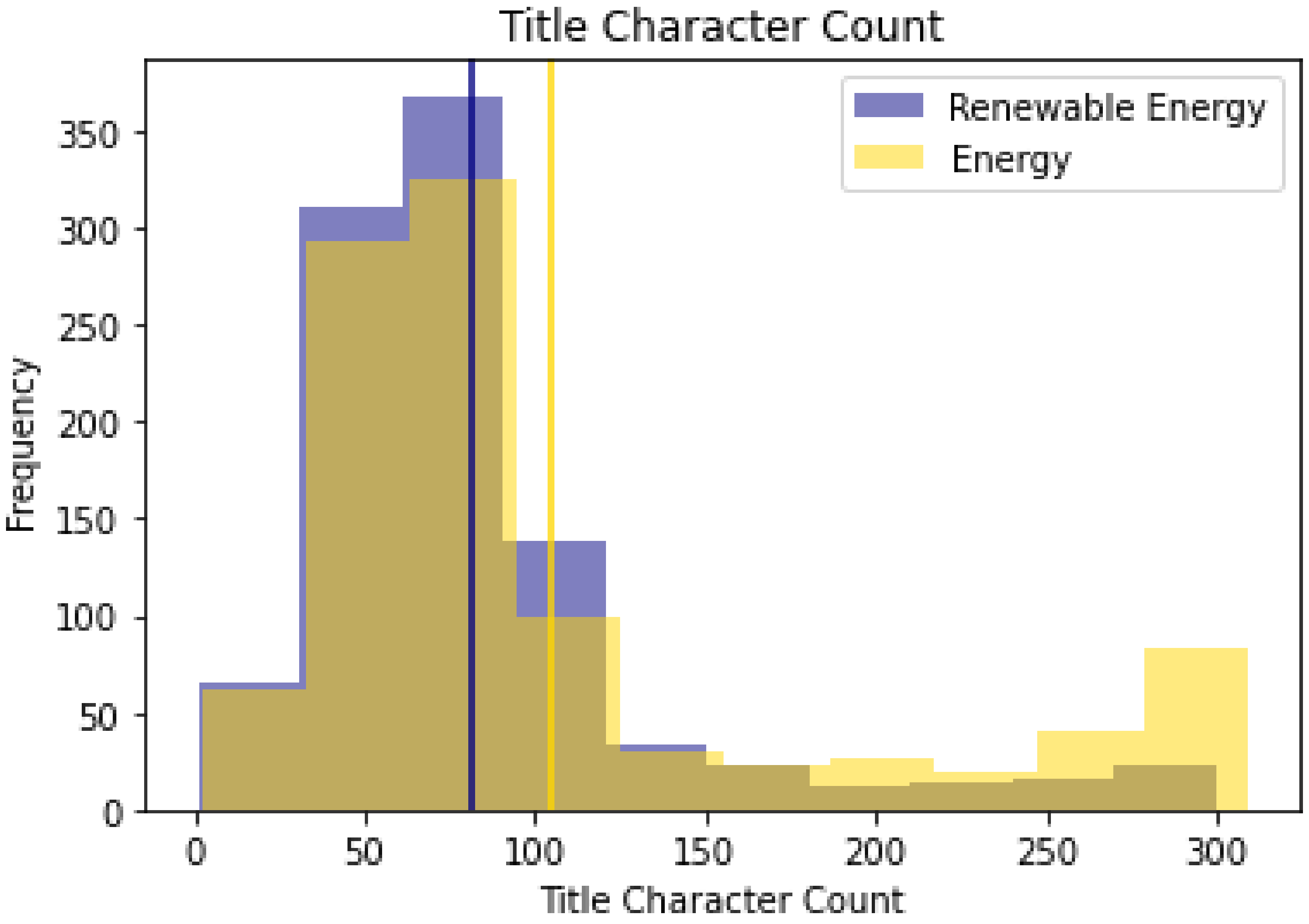
- Using Pushshift's Reddit API, gather data from 1000 most recent posts in each subreddit

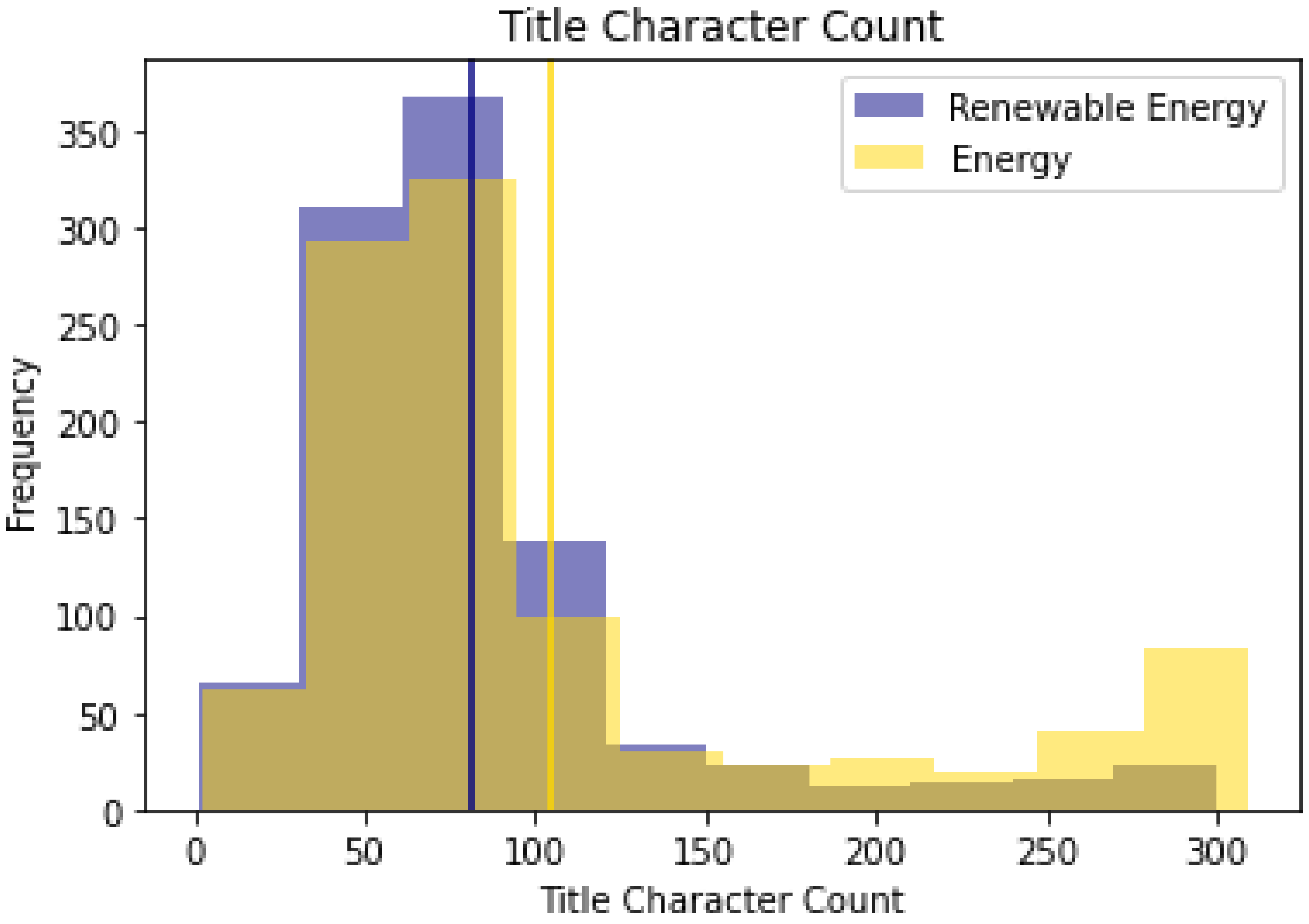
- Isolate title text for each post, with column to indicate which subreddit the text originated from



	title	subreddit
0	baufinanzierung commerzbank – top konditionen ...	1
1	Solar Tsunami: Solar PV grows 26% again – stay...	1
2	Agrivolaics to shine in France after president...	1
3	Germany aims to get 100% of energy from renewa...	1
4	Selecting and Building Large-Scale Solar Racki...	1

Exploratory Data Analysis: Length of Title



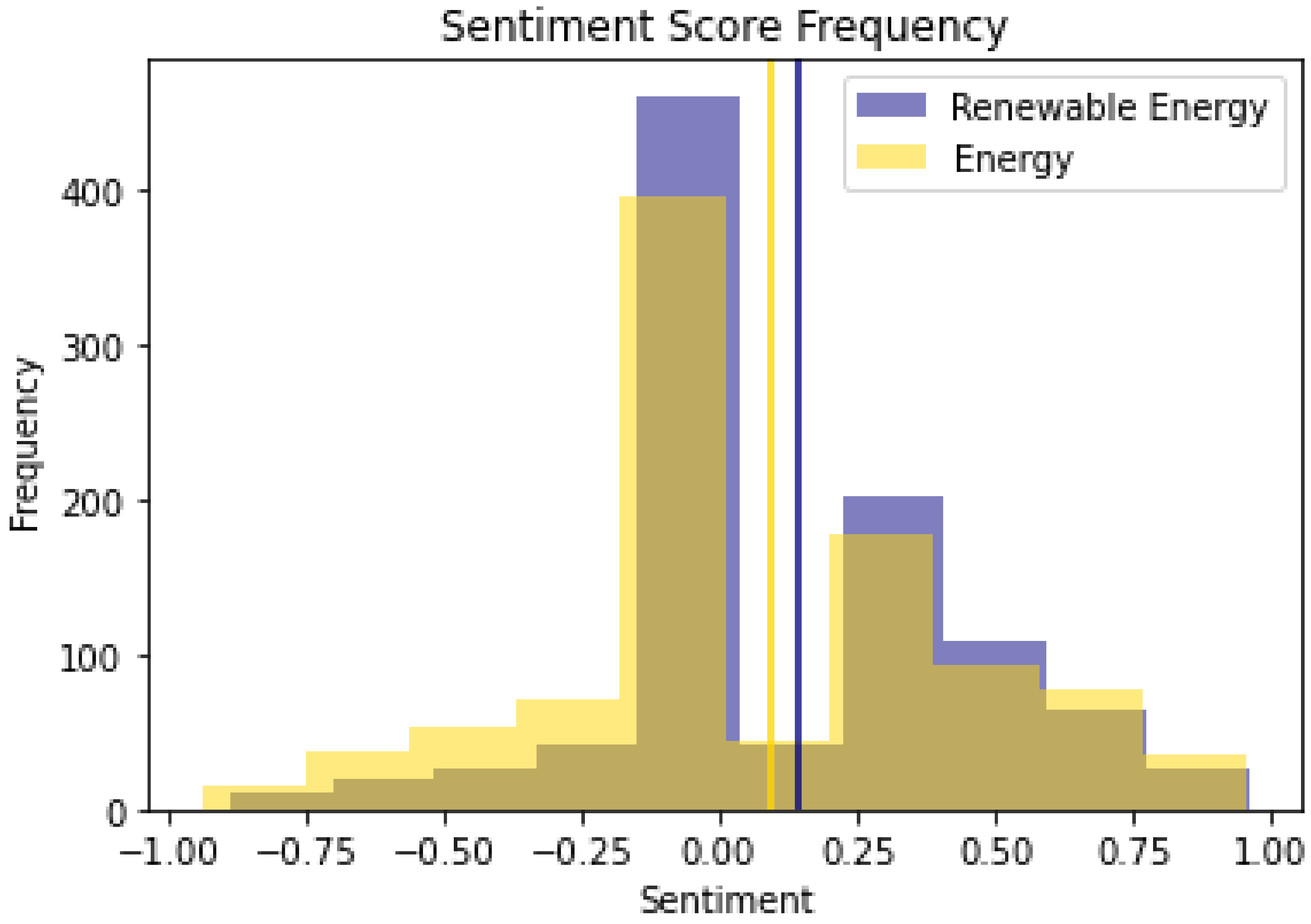


Exploratory Data Analysis: Sentiment Score

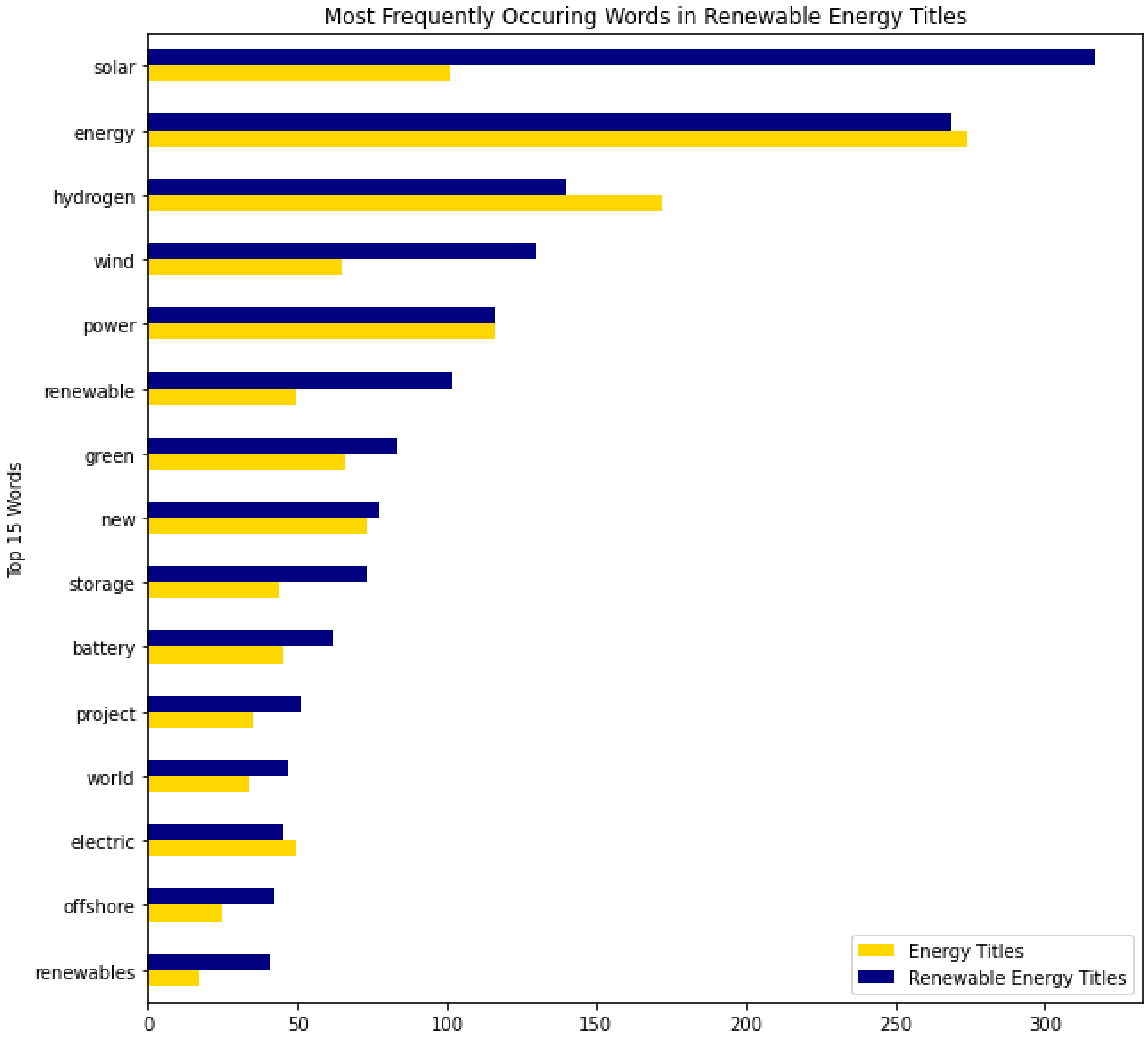
Mean sentiment score

Renewable Energy: 0.1451

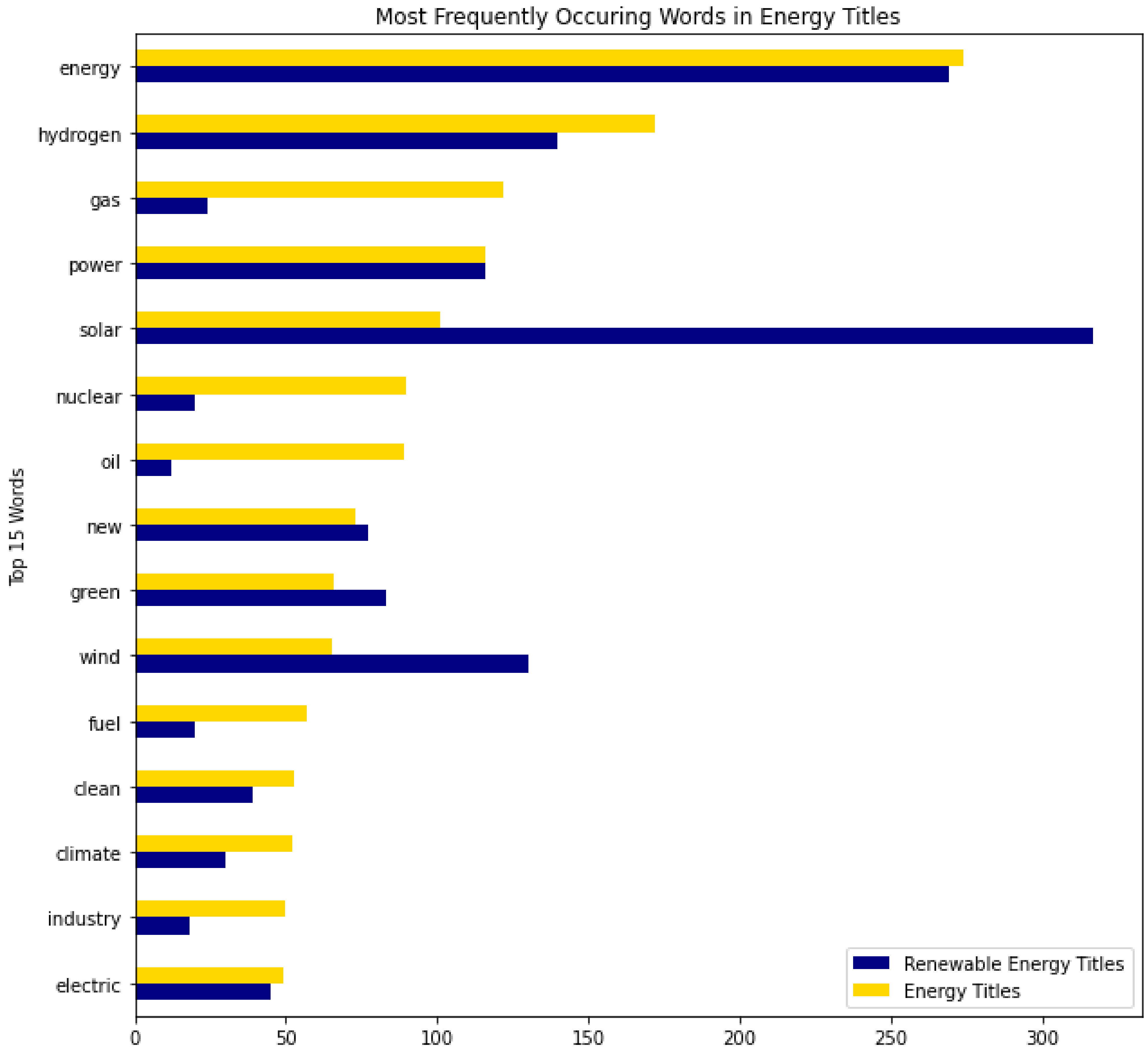
Energy: 0.0934



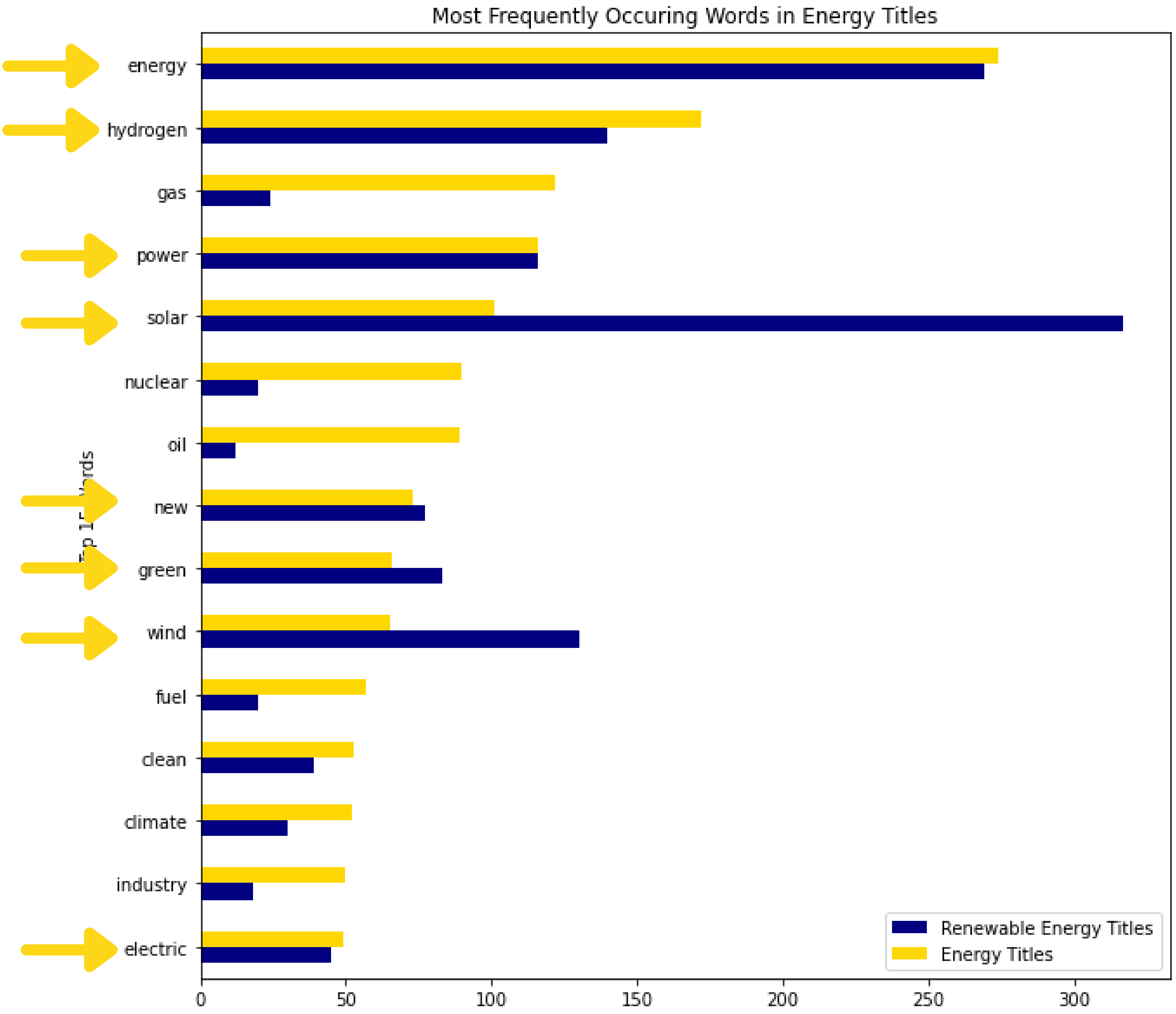
Exploratory Data Analysis: Most Frequently Used Words



Exploratory Data Analysis: Most Frequently Used Words



Exploratory Data Analysis: Most Frequently Used Words



Production Models

Logistic Regression

Best parameters:

- Max iteration of 1000 (LR)
- 'lbfgs' solver (LR)
- Binary as true (CV)
- Max df of 1 (CV)
- Min df of 0 (CV)
- ngram range of (1,2) (CV)
- Stemming tokenizer (CV).

Cross val score: 0.6897

Accuracy: .67

Random Forest

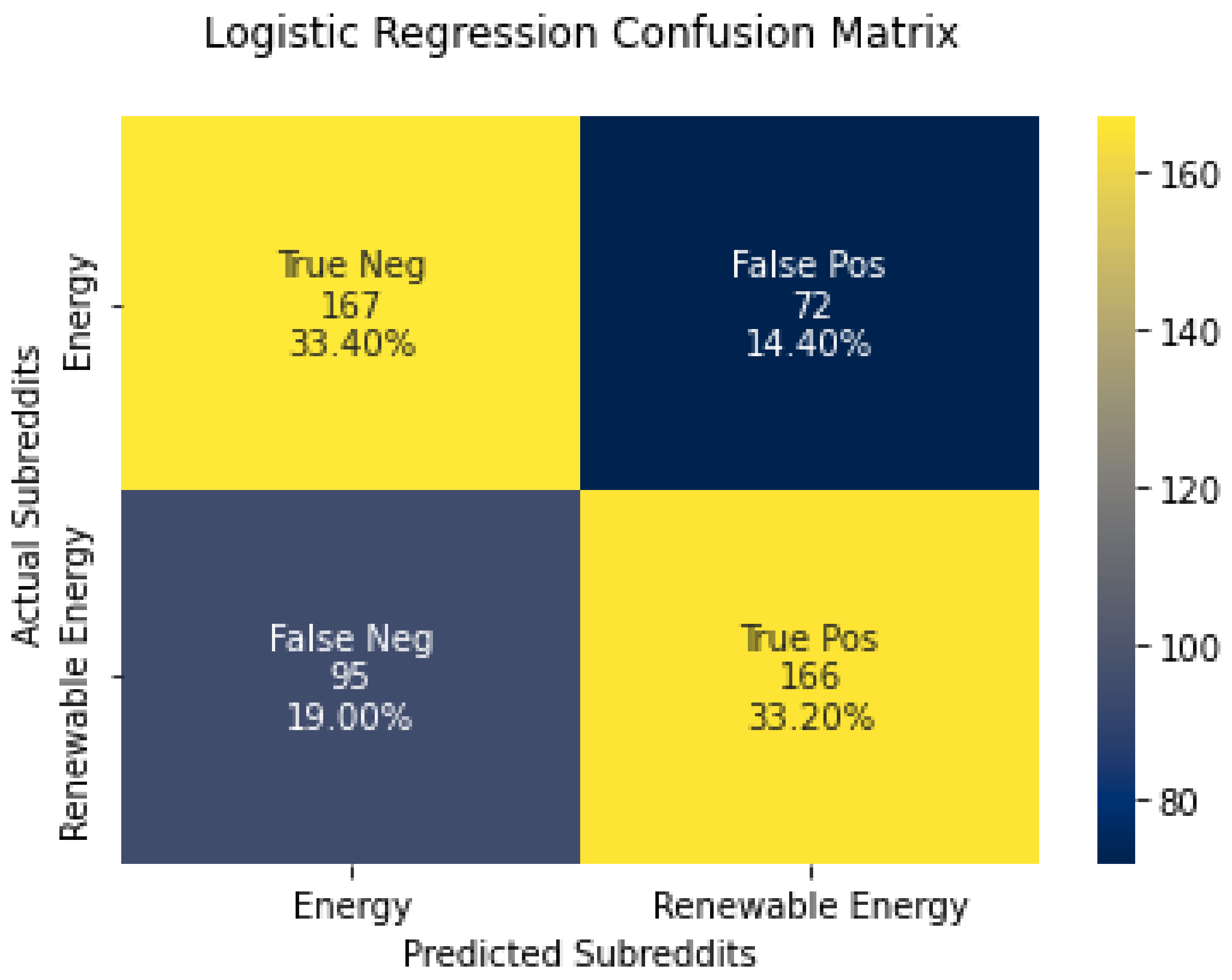
Best parameters:

- No max depth (RF)
- Automatically determined (sqrt(n_features)) max features (RF)
- Binary as false (CV)
- Max df of 1 (CV)
- Min df of 2 (CV)
- ngram range of (1,2) (CV)
- Stemming tokenizer (CV).

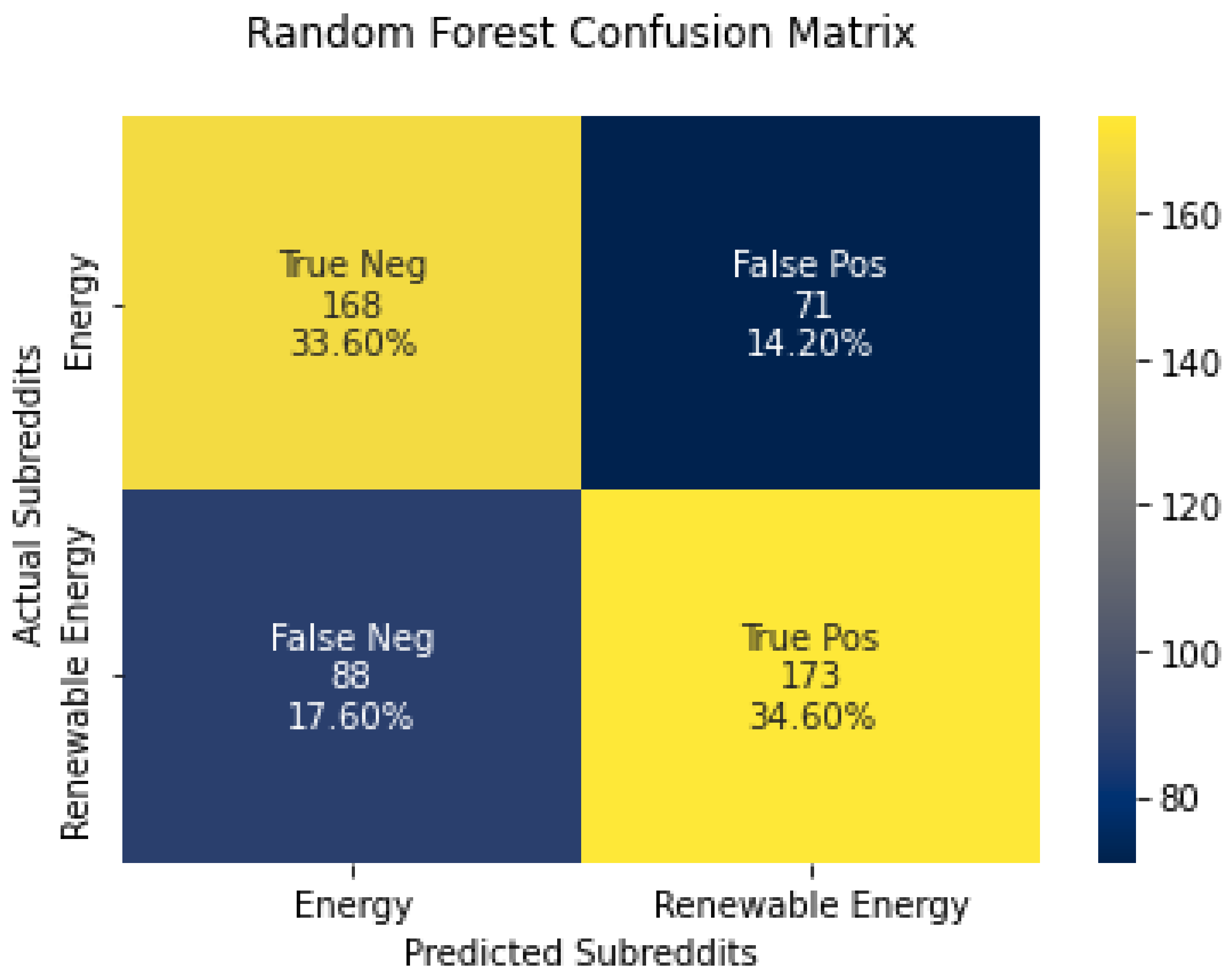
Cross val score: 0.7011

Accuracy: .68

Confusion Matrices

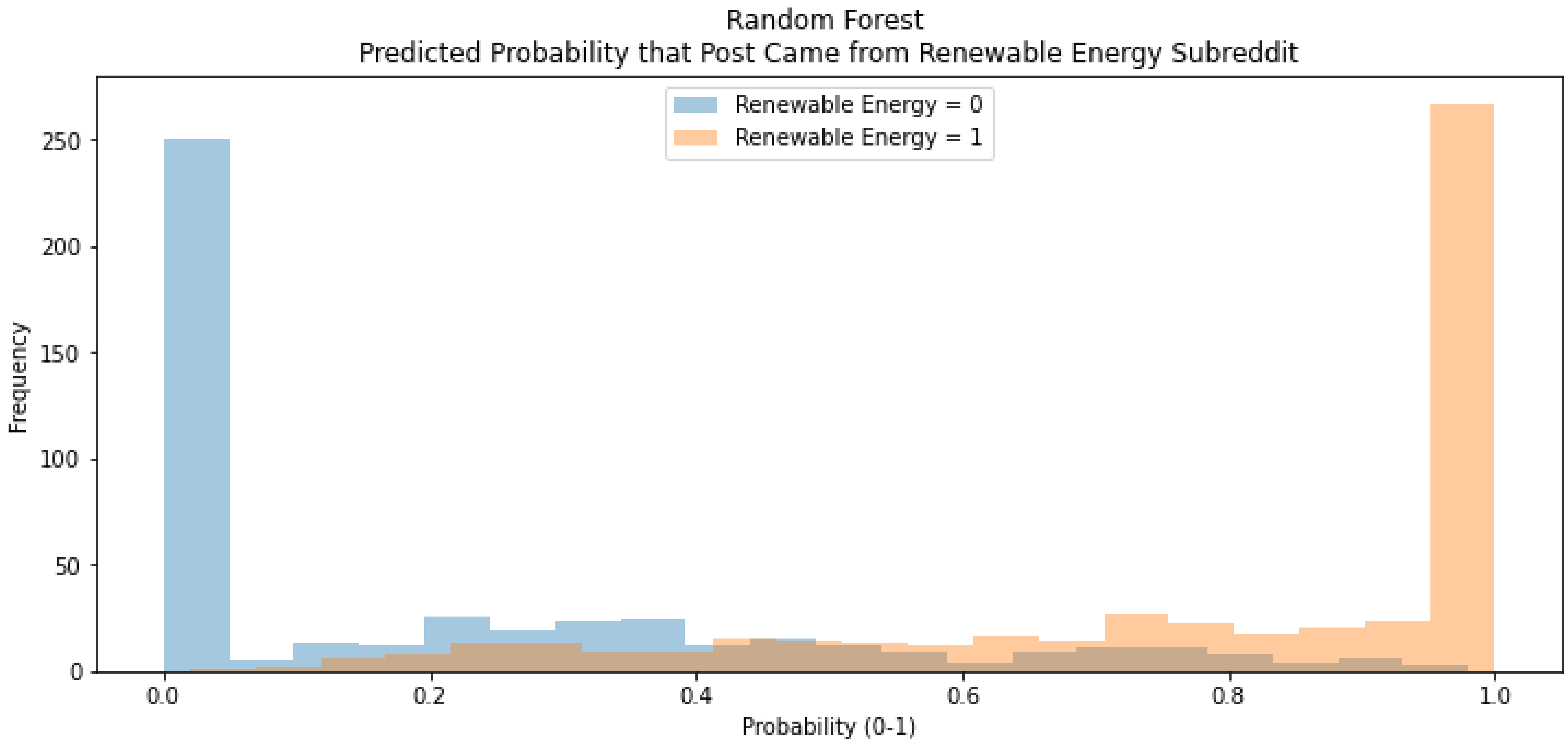


Logistic Regression Confusion Matrix

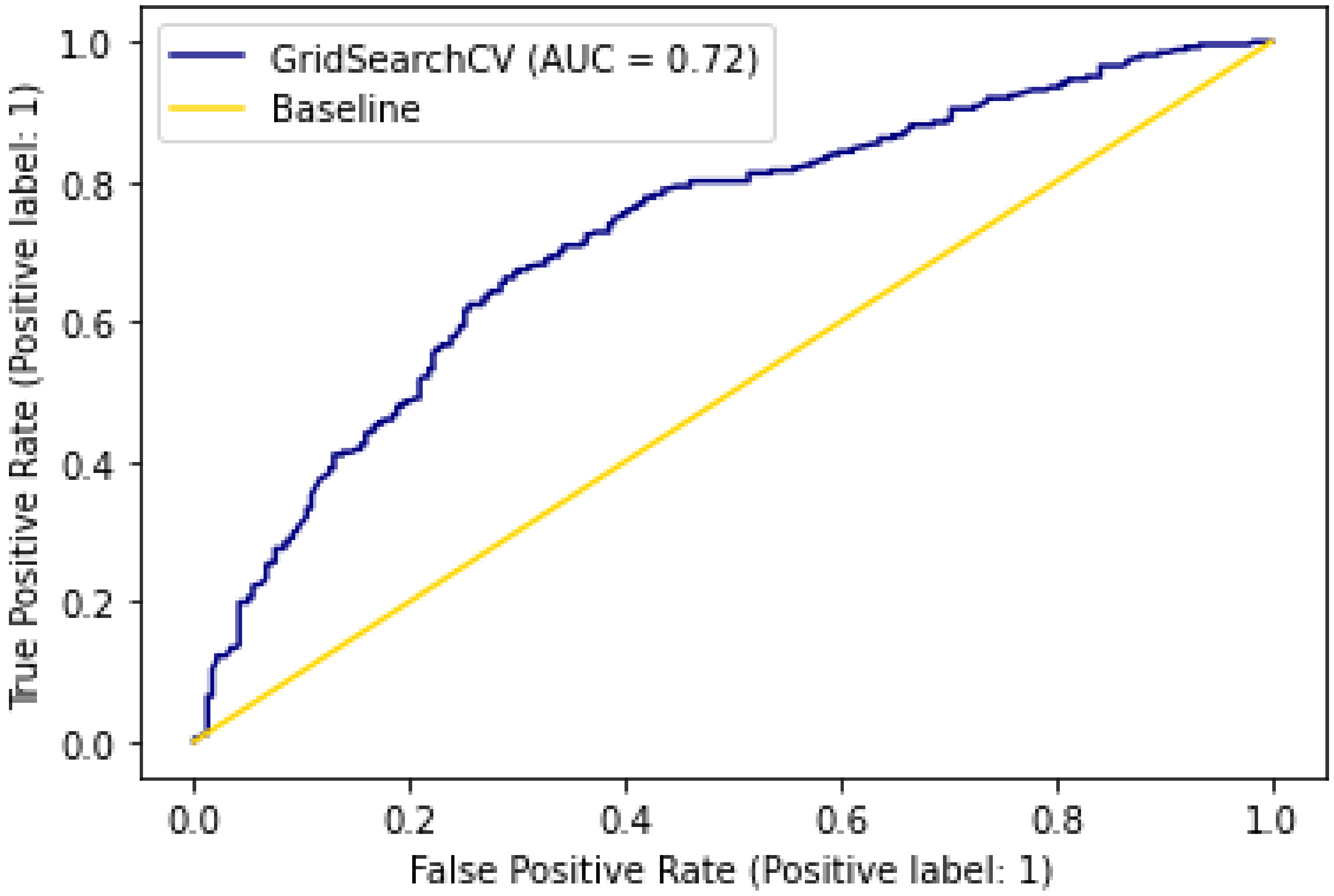


Random Forest Confusion Matrix

Specificity vs. Sensitivity



ROC Curve


GridSearchCV (AUC = 0.72)
Baseline
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

Conclusions and Next Steps

- Overlap in the language used between subreddits.
- The slightly higher sentiment score for Renewable Energy (0.1451 vs 0.0934).
- A more successful model could be developed through using a greater volume or more robust sampling of text (e.g. self text) from both subreddits, as well as testing additional models and parameters.
- If this analysis was conducted at future dates, as a time series analysis, it could be used to document and potentially forecast upcoming trends or shifts in either sector.
-