Reddit Classification
(Energy vs Renewable Energy)

By: Ramin Vafadary

Problem workflow

1. Problem statement
2. Data
3. EDA
4. Modeling
5. Conclusion

Problem statement and data

- The data is gathered by using pushshift(subreddit api)

text	score	num_comments	subreddit
submissions/comments	The submission score	The submission number of comments	Renewable energy=1 Energy=0

A Data frame with 4 columns and 40_000 rows

- Obtain the best model that can predict the subreddit category based on the submissions/comments text

preprocessing

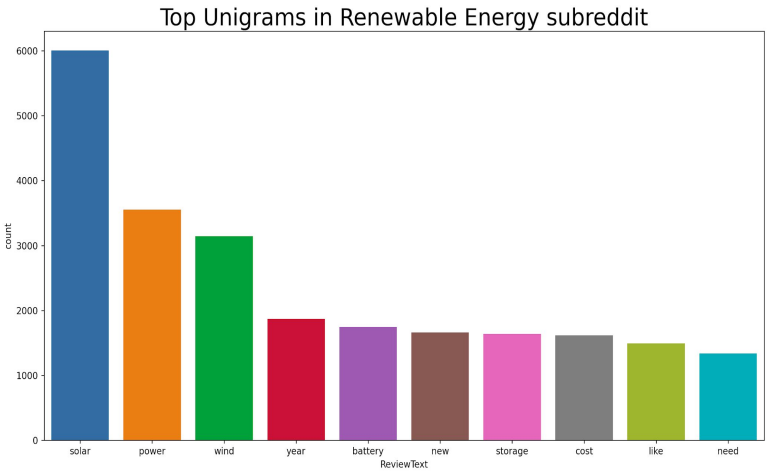
1. The text column is cleaned by :
   removing duplicates,HTML, Non-letter characters, Stop words And lemmatizing

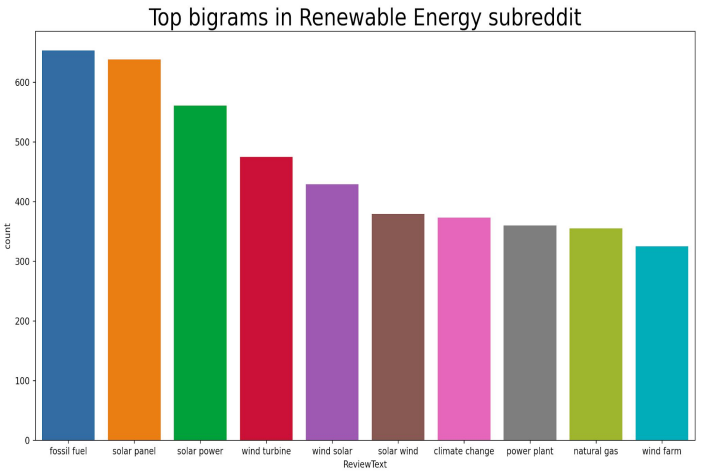
2. Some words are removed manually: {Energy, renewable, renew}

3. Some words are removed manually based on the EDA:

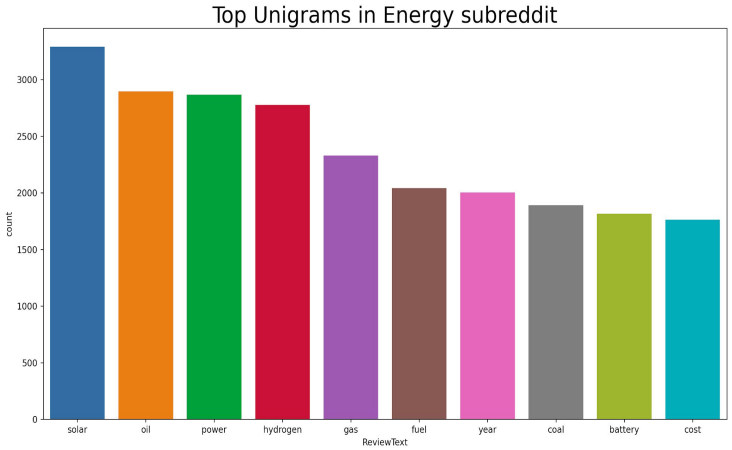
{www, https, np, reddit, en, wikipedia, com, org, wiki, youtube, watch, message}

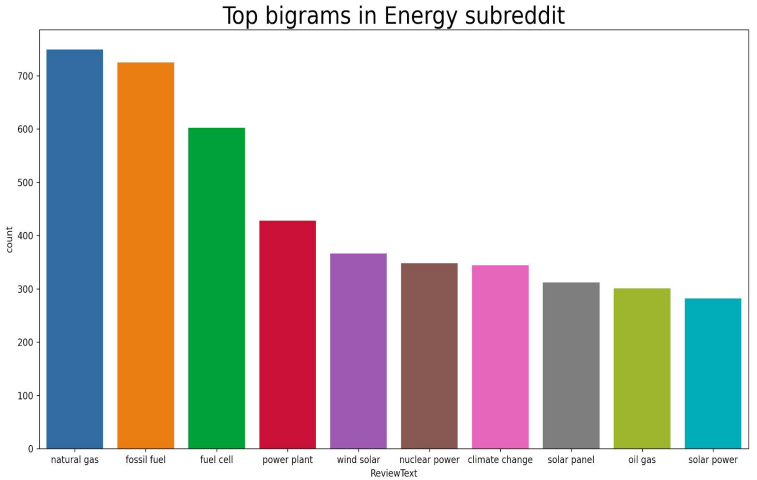
EDA (the top words in **Renewable Energy** subreddit)



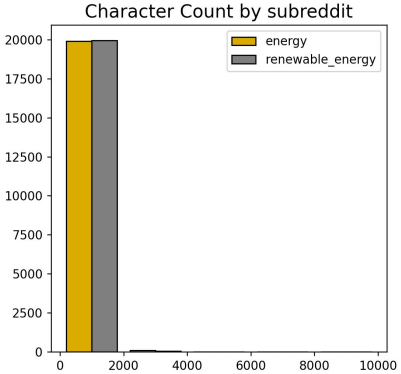


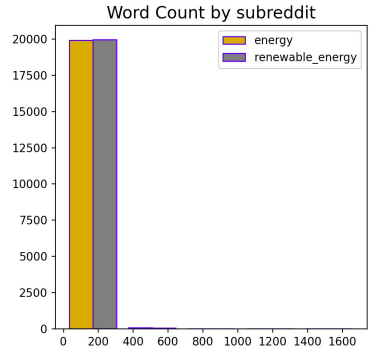
EDA (the top words in **Energy** subreddit)

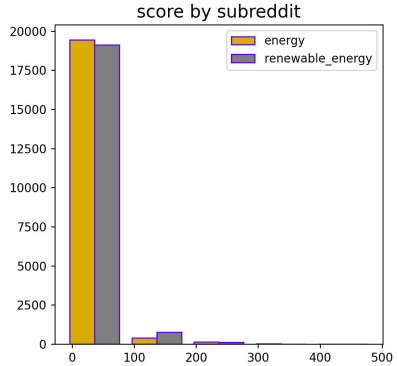


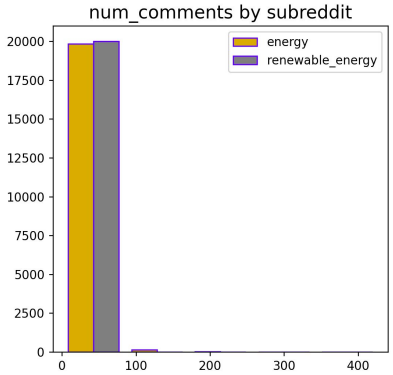


Feature Engineering and feature selection

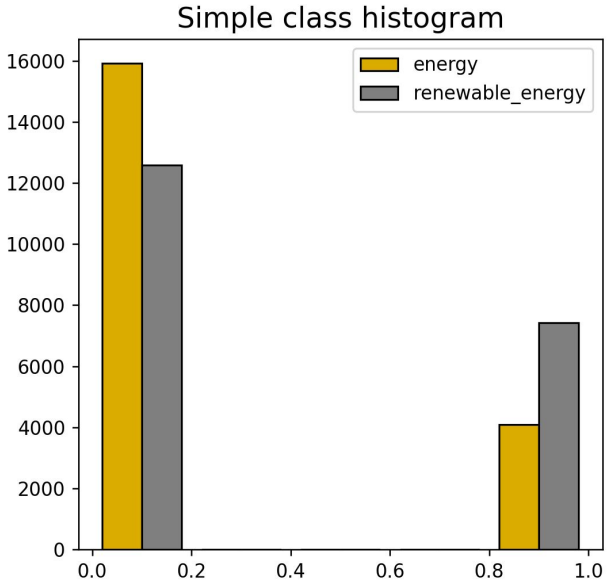


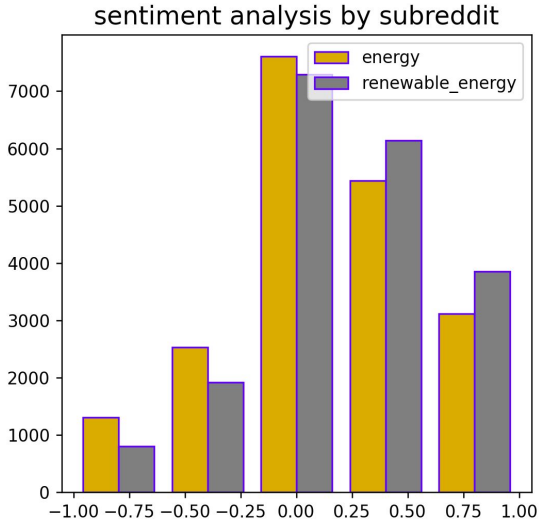






Feature Engineering and feature selection





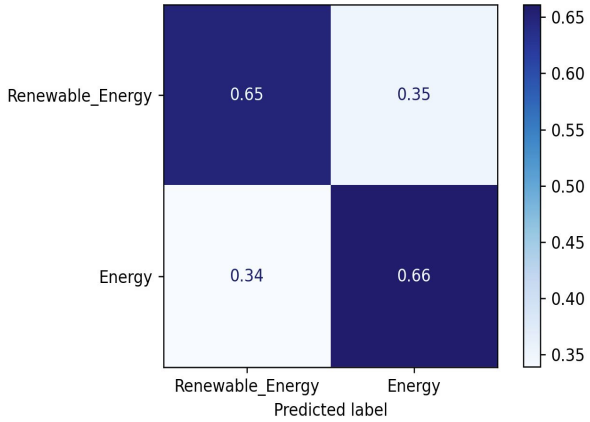
List of popular Energy words={oil, gas, hydrogen, plant, fuel, nuclear}

List of popular Renewable Energy words={green, wind, geothermal, tidal, solar, biomass}

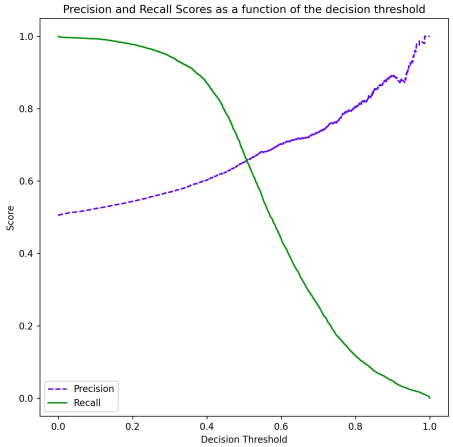
Modeling(Logistic Regression)

Base_accuracy=0.5

Train_Accuracy	0.87
Test_Accuracy	0.66
Sensitivity	0.67
Specificity	0.64

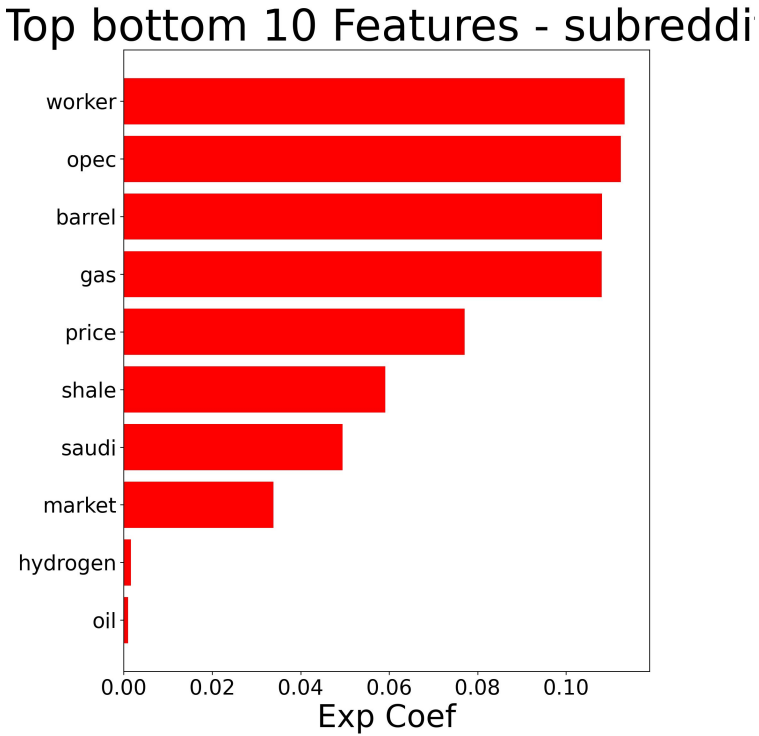


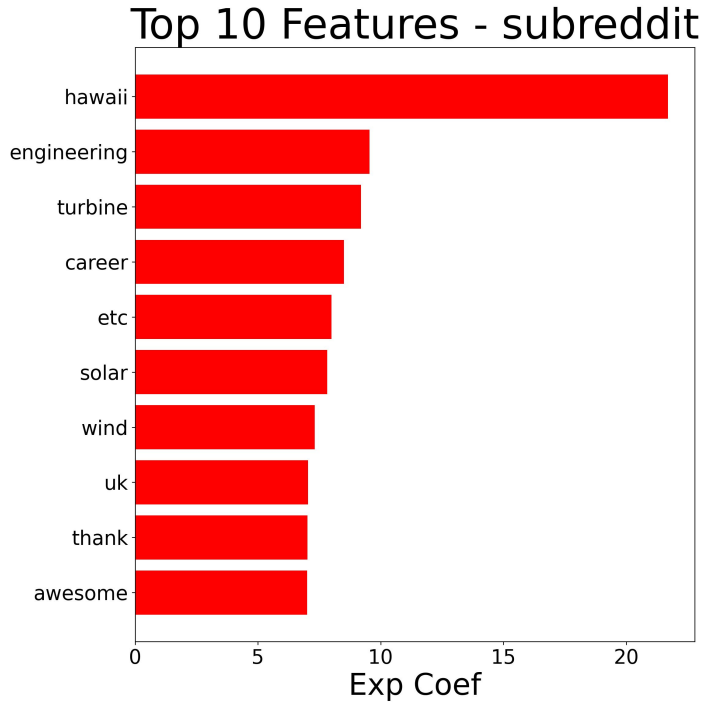
Confusion Matrix



Precision-Recall

Model interpretation

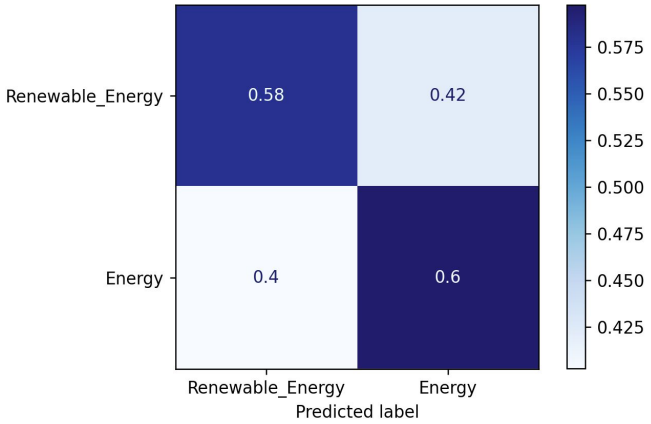




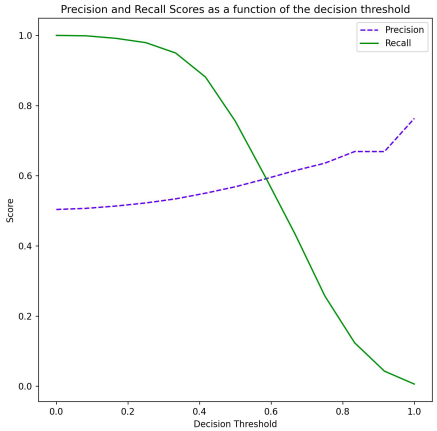
Why Hawaii:
-state's isolated location
-lack of fossil fuel resources
-the electricity price is the highest in the U.S

Modeling(KNN)

Train_Accuracy	0.67
Test_Accuracy	0.59
Sensitivity	0.57
Specificity	0.6

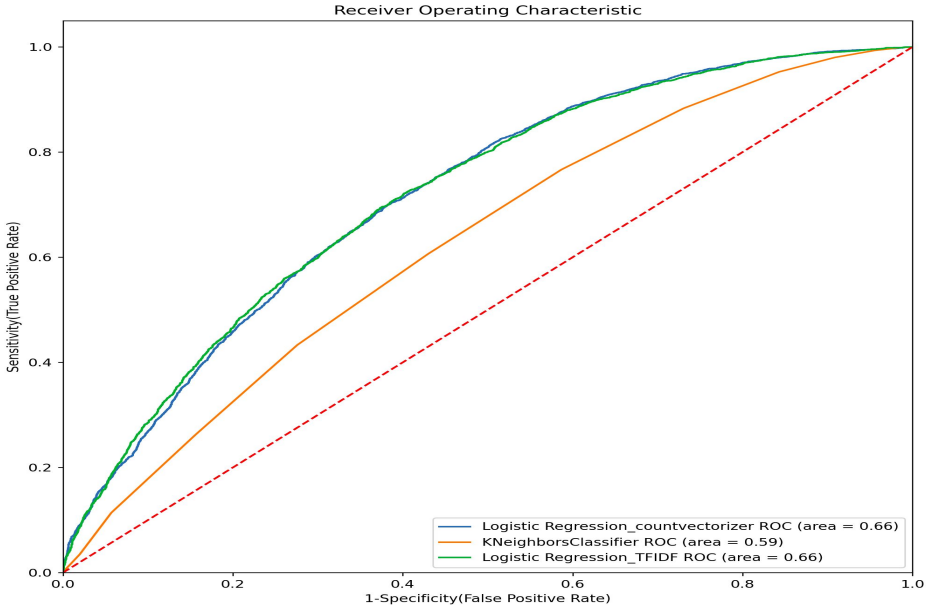


Confusion Matrix



Precision-Recall

Evaluating models



Conclusion

- Since the two subreddits are very similar in content, so the accuracy score is small

  Suggestion ⟶ Instead of having Energy subreddit we should create Non-Renewable Energy subreddit

- TFIDF is better in identifying the list of important words
- Logistic Regression is better in modeling when inference of the coefficients is important

- Future work: try other supervised models, do some more feature engineering, put some weights on most important words when modeling, make our own list of stop words