r/BMW & r/teslamotors

lynette ng

SG-DSI-20

Problem Statement & Project Goal:

- For the user to distinguish between two subreddits (r/BMW & r/teslamotors)

- To identify which subreddit to post in when met with similar subreddits

Classification Process

Part A

1. Data Collection by Scraping (using Reddit's API)

Part B

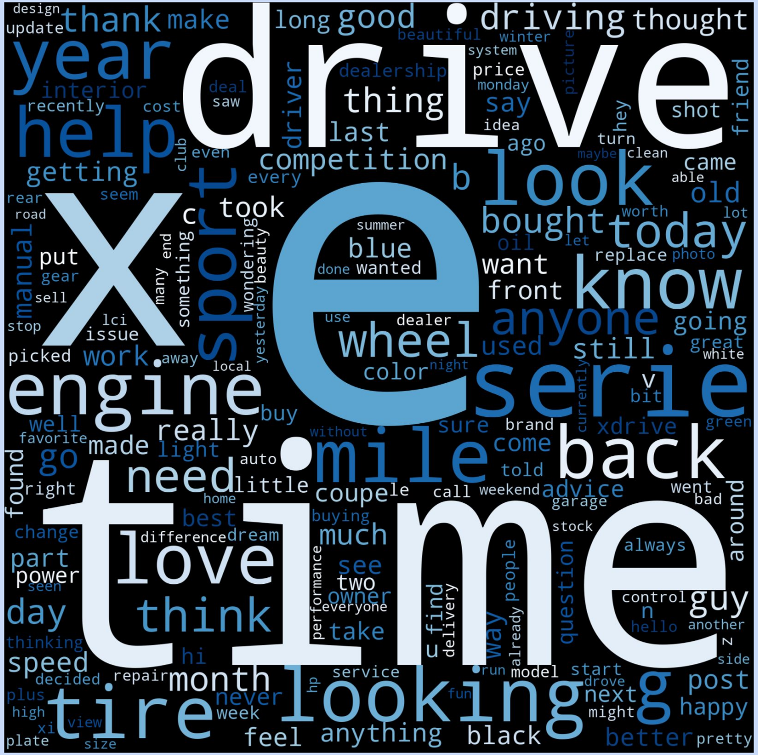
1. Data Cleaning
2. Pre-processing & EDA
3. Modelling
4. Evaluation & Conceptual Understanding
5. Conclusion/Recommendations

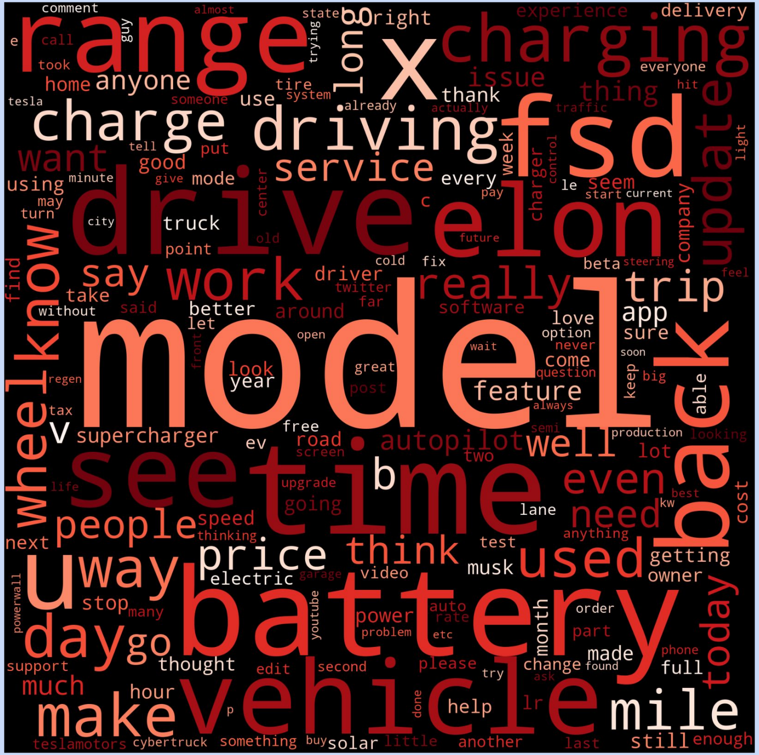
1. Data Collection - Scraping with Reddit's API

- Scraped using a while (post_count<1000) loop
- Scraped r/BMW & r/teslamotors from:
  - Top of all time
  - New
  - Hot (default)
- Issues encountered: Not enough posts
- After cleaning & removing duplicates:
  - r/BMW: 1198 rows, only 200+ with selftext
  - r/tslamotors: 1149 rows, only 200+ with selftext

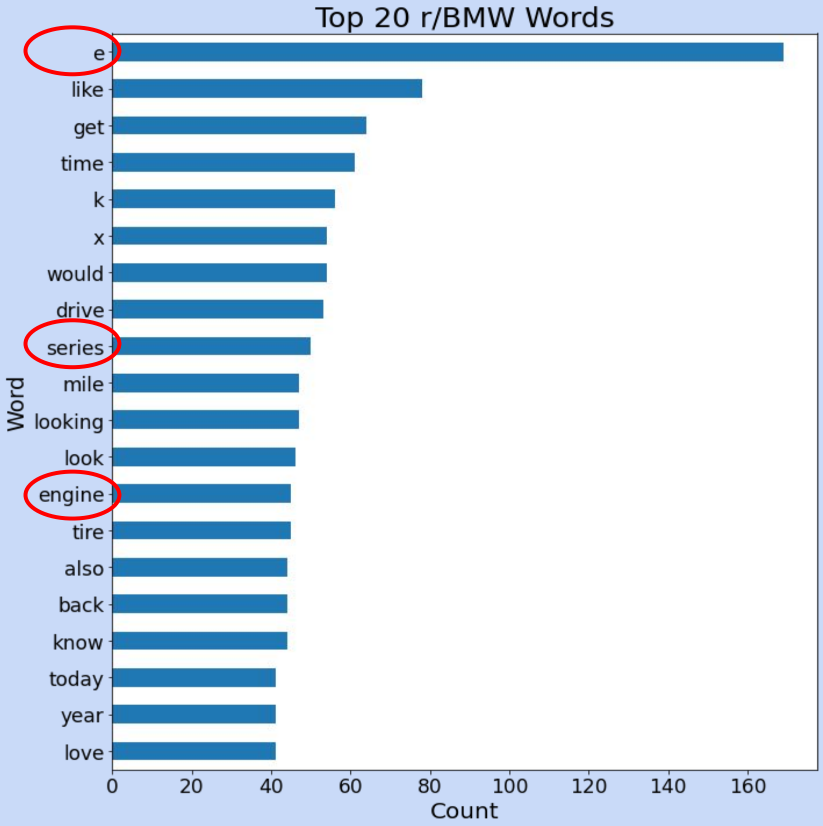
2. Text Pre-processing & EDA

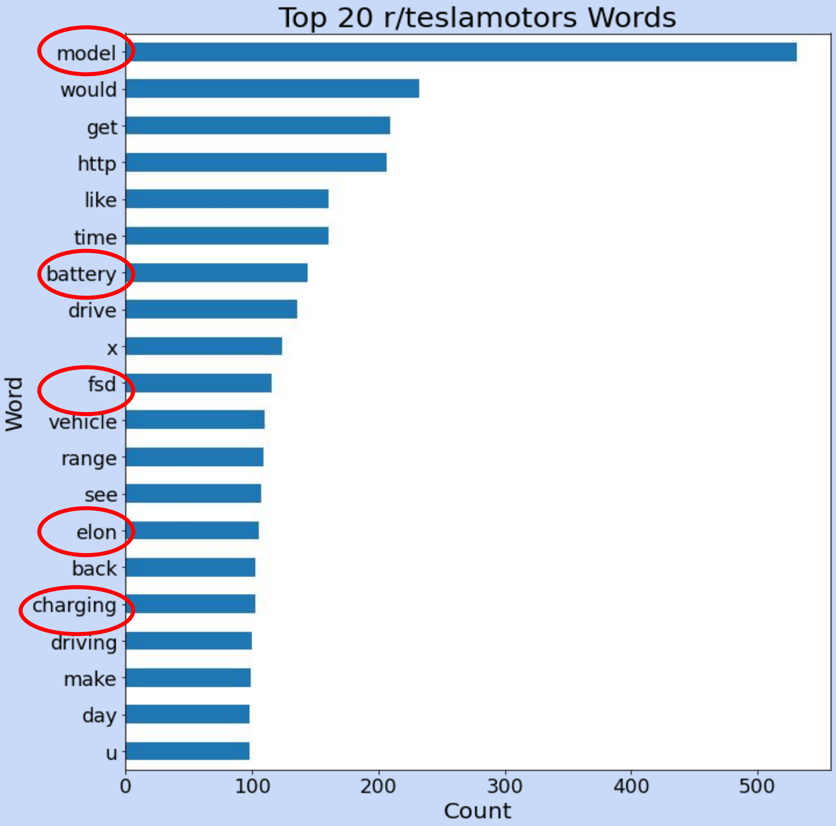
- Added some stopwords manually & Lemmatized





2. Text Pre-processing & EDA

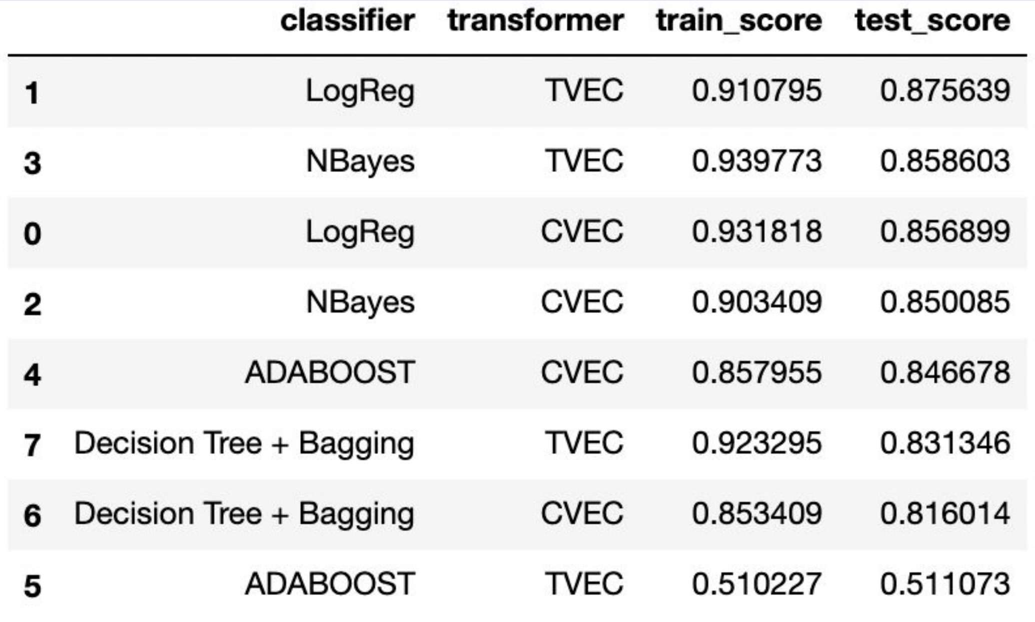




3. Modelling

- Vectorizers used: CountVectorizer & TfidfVectorizer for all models
- Classifiers used:
    - Logistic Regression
    - Naive Bayes
    - AdaBoost (base: DecisionTreeClassifier)
    - DecisionTreeClassifier + Bootstrap Aggregation

3. Modelling - Performance



	classifier	transformer	train_score	test_score
1	LogReg	TVEC	0.910795	0.875639
3	NBayes	TVEC	0.939773	0.858603
0	LogReg	CVEC	0.931818	0.856899
2	NBayes	CVEC	0.903409	0.850085
4	ADABOOST	CVEC	0.857955	0.846678
7	Decision Tree + Bagging	TVEC	0.923295	0.831346
6	Decision Tree + Bagging	CVEC	0.853409	0.816014
5	ADABOOST	TVEC	0.510227	0.511073

- LogReg performed the best with TVEC - params:

  {'cvec__max_features': 1800,

  'cvec__min_df': 3,

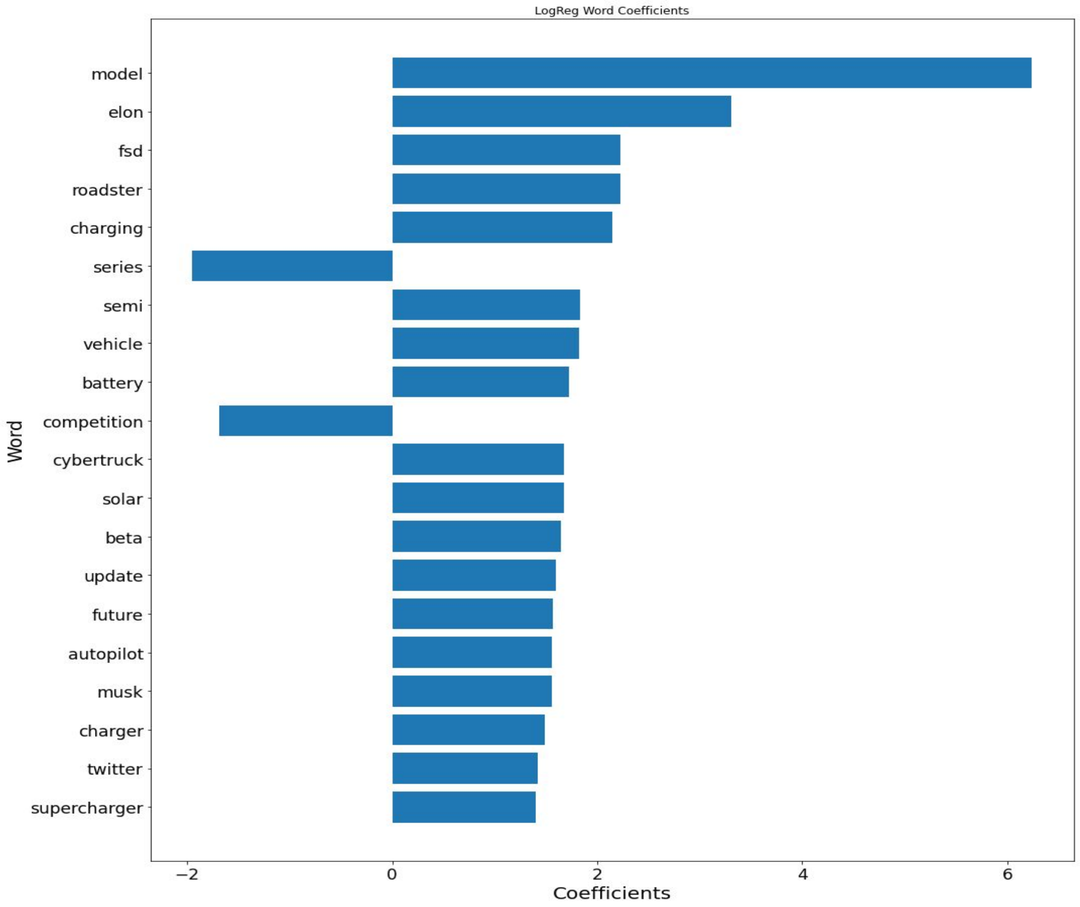
  'cvec__ngram_range': (1, 1),

  'lr__C': 0.8}

- AdaBoost performed the worst with TVEC - AdaBoost tends to not perform well with noisy datasets

4. Evaluation

- Based on our best model: Most words were indicators of r/tslamotors rather than r/BMW
- BMW were mostly model-specific terms, while Tesla mostly consisted of EV-specific terms



5. Conclusion

- Best classification model: Logistic Regression with TfidfVectorizer
- Areas of improvement:
  - If posts have been completely scraped, might improve model if comments are scraped as well due to the lack of selftext in posts
  - Remove more stop words
  - Try more models (i.e. SVM, k-NN)

Thanks & Regards

QUESTIONS?