Subreddit Classifier Proof of Concept

Yiang Yuet Meng

Agenda

01 Background

02 Data Gathering/Cleaning

03 EDA

04 Modelling

05 Conclusion

06 Further Improvements

01 Background

Honda Customer Service main priority is to provide excellent service to our customers.

Recently, our customer service has been receiving a lot of irrelevant posts from Mazda. This is likely due to the year end promotion sales launched by Mazda. This has caused unnecessary stress to the team to keep up with the desired service level agreement.

Today my main objective is to develop a proof of concept using NLP and Machine Learning to analyze and segregate the posts and provide a solution to the customer service team.

02 Data Gathering and Cleaning

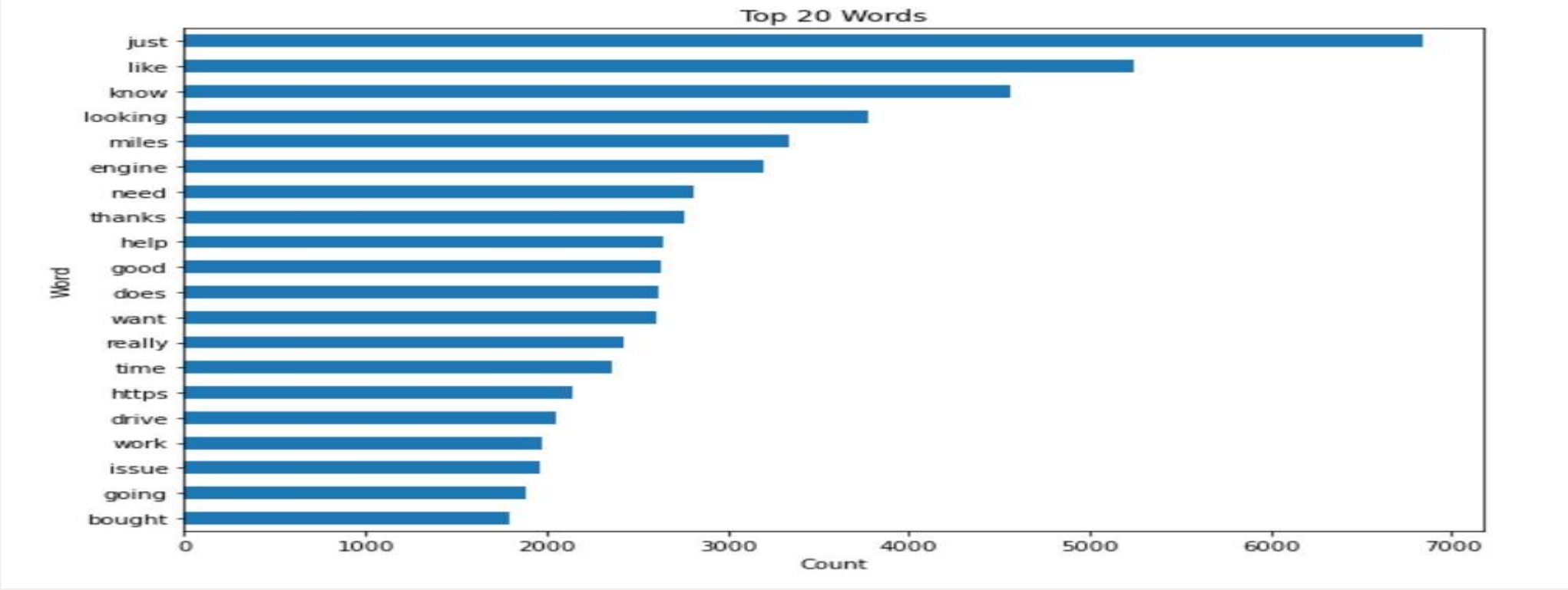
Data Gathering/Cleaning

Import from r/Honda and r/mazda.

Standard steps taken in data gathering and cleaning which includes removal of duplicate posts, deleted posts and lemmatizing of posts

Data Exploratory Analysis

Key steps are reviewing and clearing of unnecessary punctuations, numbers etc as well as experimenting of different stop words to be introduced. Result below show that the text is cleared of model name, punctuation, special characters.



03 Modelling

Modelling
Base Model

- Multinomial Naive Bayes classifier is chosen
- Train Score: 0.7804
- Test Score: 0.7032
- Accuracy: 0.7466
- Misclassification: 0.2534
- Precision: 0.668
- Recall: 0.5567
- Specificity: 0.8497

The model exhibits overfitting of train scores. I will use this model as the baseline model and explore other parameters and models

Modelling

Model		Parameters	Train Score	Test Sore	Metrics	Remarks
Model 1	Random Forrest		0.7656	0.6567	Accuracy: 0.7447 Misclassification: 0.2553 Precision: 0.8092 Recall: 0.3593 Specificity: 0.954	exhibits overfitting of train scores. I will explore to tune the parameters. Recall is low.
Model 2	Multinomial Naive Bayes	cvec__min_df: 2, cvec__max_df=0.5, cvec__max_features: 5000, cvec__ngram_range: (1,1)	0.7226	0.722	Accuracy: 0.722 Misclassification: 0.278 Precision:0.6027 Recall: 0.6159 Specificity: 0.7796	Tuning of CountVectorizer It is overall better than the baseline model. It is slightly more accurate. There is no sign of overfitting of training set in this case. Recall improves
Model 3	Random Forrest	cvec__min_df: 3, cvec__max_df=0.75 cvec__max_features: 5000, cvec__ngram_range: (1,1)	0.746	0.7443	Accuracy: 0.7443 Misclassification: 0.257 Precision:0.7411 Recall: 0.4203 Specificity: 0.9203	Tuning of CountVectorizer It is better than the Model 2. There is no sign of overfitting of training set in this case

Modelling

Model		Parameters	Train Score	Test Sore	Metrics	Remarks
Model 4	Random Forrest	cvec__min_df: 2, cvec__max_df= 1.0, cvec__max_features: 4000 cvec__ngram_range: (1,1) rf__n_estimators = 150 rf__max_depth = None	0.7458	0.7497	Accuracy: 0.7497 Misclassification: 0.2503 Precision:0.7488 Recall: 0.4343 Specificity: 0.9209	Introduce hyperparameter tuning of Random Forrest It is overall better than the Model 3. There is no sign of overfitting of training set in this case. Metrics having True Positive is critical for the model selection I have chosen this model due to high test score, high precision

Evaluation of Model 4

True Negative : 4725, False Positive: 406, False Negative: 1576, True Positive: 1210

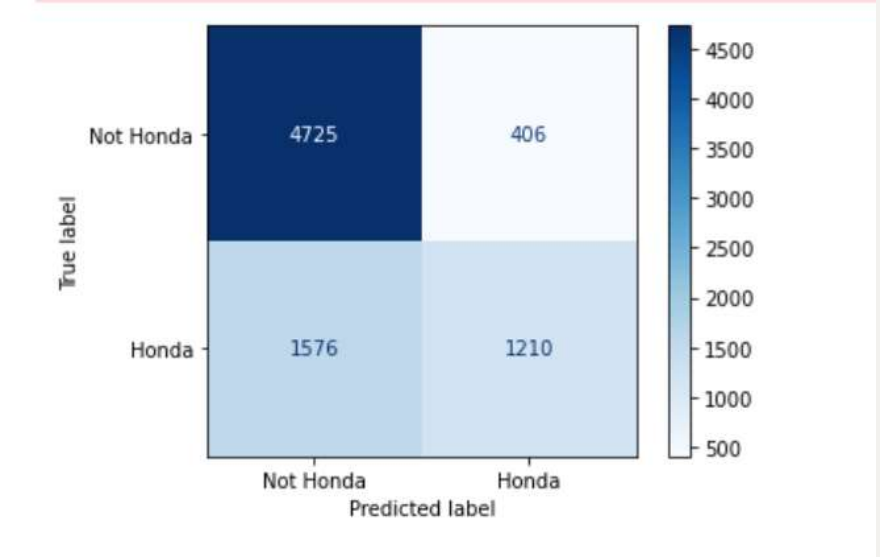
Accuracy: 0.7497

Misclassification:0.2503

Precision: 0.7488

Recall: 0.4343

Specificity: 0.9209



Our model correctly predicts 74.97% of observations.

Among posts that our model predicted to be in /r/honda, we have 74.88% of them correctly classified.

Among posts that are in /r/honda, our model has 43.43% of them correctly classified.

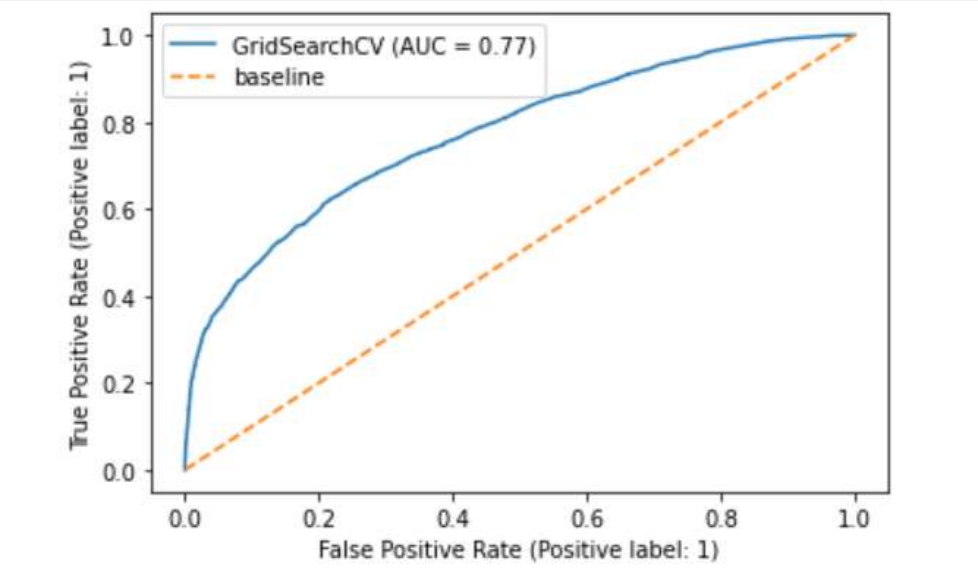
Among posts that are in /r/mazda, our model has 92.09% of them correctly classified

04 Evaluation

Evaluation of Model 4

ROC AUC is 0.77. The positive and negative populations are clearly separated.

I will not proceed to adjust my threshold as the ROC AUC is reasonable ok. Besides, the objective is mainly to segregate the posts to reduce the unnecessary workload of Customer Service Desk.



Top Important Features and Probability

Top important Features and Probability

	0
coupe	0.023005
prelude	0.009517
acura	0.009140
type	0.008629
http	0.008351
hybrid	0.007649
pilot	0.005989
touring	0.005892
swap	0.005755
grand	0.004394
just	0.003803
know	0.003593
infotainment	0.003505
sport	0.003232
like	0.003221
looking	0.003209

miles	0.002938
engine	0.002933
thanks	0.002731
good	0.002661
wanna	0.002659
speed	0.002601
need	0.002595
does	0.002585
help	0.002575
dealer	0.002414
passport	0.002394
https	0.002393
want	0.002381
insight	0.002368
imgur	0.002200
wondering	0.002143
guys	0.002135
bose	0.002130

wrongly classified to Mazda post

Honda post wrongly classified to Mazda post	
('just', 512),	('does', 183),
('like', 373),	('thanks', 181),
('know', 321),	('really', 174),
('looking', 283),	('time', 171),
('miles', 228),	('help', 168),
('want', 216),	('need', 162),
('good', 211),	('https', 159),
('sport', 186),	('going', 155),
('engine', 145),	('dealership', 154),
('used', 135),	('work', 150),
('bought', 135),	('issue', 148),
('driving', 134),	('drive', 148),
('price', 130),	('cars', 148),
('touring', 130),	('think', 128),
('wondering', 129),	('door', 124

Those in red belong to the top most important words. Top common words 'just','like','know','looking','miles','want','good','sport', 'engine','touring'. These words are relatively common among all car brands. So, model is not able to predict it properly

wrongly classified to Honda post

Mazda post wrongly classified to Honda		
('just', 390),	('replaced', 125),	('getting', 155),
('engine', 330),	('time', 125),	('start', 152),
('know', 240),	('does', 119),	('driving', 145),
('like', 237),	('looking', 118),	('problem', 144),
('miles', 237),	('transmission', 117),	('check', 142),
('light', 175),	('good', 114),	('cars', 142),
('help', 172),	('check', 113),	('auto', 139),
('start', 170),	('going', 110),	('guys', 139),
('need', 144),	('work', 109),	('sure', 138),
('problem', 137),	('thanks', 107),	('hybrid', 138),
('issue', 135),	('fine', 105),	
('want', 133),	('code', 102),	
('driving', 131),	('change', 102),	
('battery', 127),	('guys', 102),	
('https', 126),	('power', 101)]	

Above mazda posts are wrongly classified as r/honda post. Top common words 'just','like','engine','know','miles','hybrid'. The word 'hybrid' is likely the word that the model picks up and identify as honda

Insights

The top important features pertaining to Honda:
Coupe
Sports
Hybrid
Infotainment
Miles

Based on above top important features, Honda is synonymous with economical, fuel-efficient automobiles. it's the classic models, coupe, remains the best of the bunch. These two characteristics can deceive one into thinking they're simple economy cars or another mildly sporty coupe.
There is also frequent mention of infotainment system. Honda issues a recall for nearly 608,000 vehicles because of instrument panel software problems. The recall affected the 2018-2020 Odyssey, 2019-2020 Passport and 2019-2021 Pilot vehicles with defective instrument panel control modules. So, that explains frequent mention of infotainment in Honda posts.

05 Conclusion

Conclusion

My Random Forrest classifier performed well with a test accuracy score of 74.97%. This is within expectations because the topics of our two chosen subreddits is closely related.

On top of it, the mentioned brand name are removed from the train data so that we can subject the model with more stringent conditions.

Subreddit Classifier - A proof of concept web application was developed to demonstrate potential use case.

My application can segregate the posts well. This has been further verified with some mock test messages.

With this, I will seek to further develop the application in our production system.

Further Improvements

Optimize the Stop Words

Feature Importance is key to the accuracy of model. Thus, stop words need to be reviewed and added in as we drill deeper in the analysis

Train with more Reddits

With more Reddits as training set, we can obtain more features that are significant to the predicting.

Try Other Ensemble Models

Examples are XGBoost Classifier

Expand to other irrelevant posts

Expand to include segregation of irrelevant posts from other car competitors