Twitter Sentiment Analysis: A case study from Automotive Industry

Iron Hack Final Project(Sadaf Shahbaz)

Project Breif

Challange

Analysis and prediction of sentiments
using twitter data

Scenario

An automobile company wants to :

1. Maintain its Brand Reputation
2. What do users like/dislike about our products? Has the number of negative responses increased gradually?
3. To build machine learning models to predict sentiment.

Goal

1. Answer business questions
2. Build machine learning that can predict sentiment from twitter data.

PROCESS

Collect

- Scrapping data from twitter (oct -dec 22), 40K obs
- KW: Tesla, Mercedes, BMW, Audi

Extract & Preprocess

- Removing emojis, Urls, spaces etc
- Spelling Correction
- Lemmatization
- Removing Stopwords

Sentiment Extration

- Vador Polarity Score
- postive (score >1)
- neutral (score ==1)
- negative (score < 1)

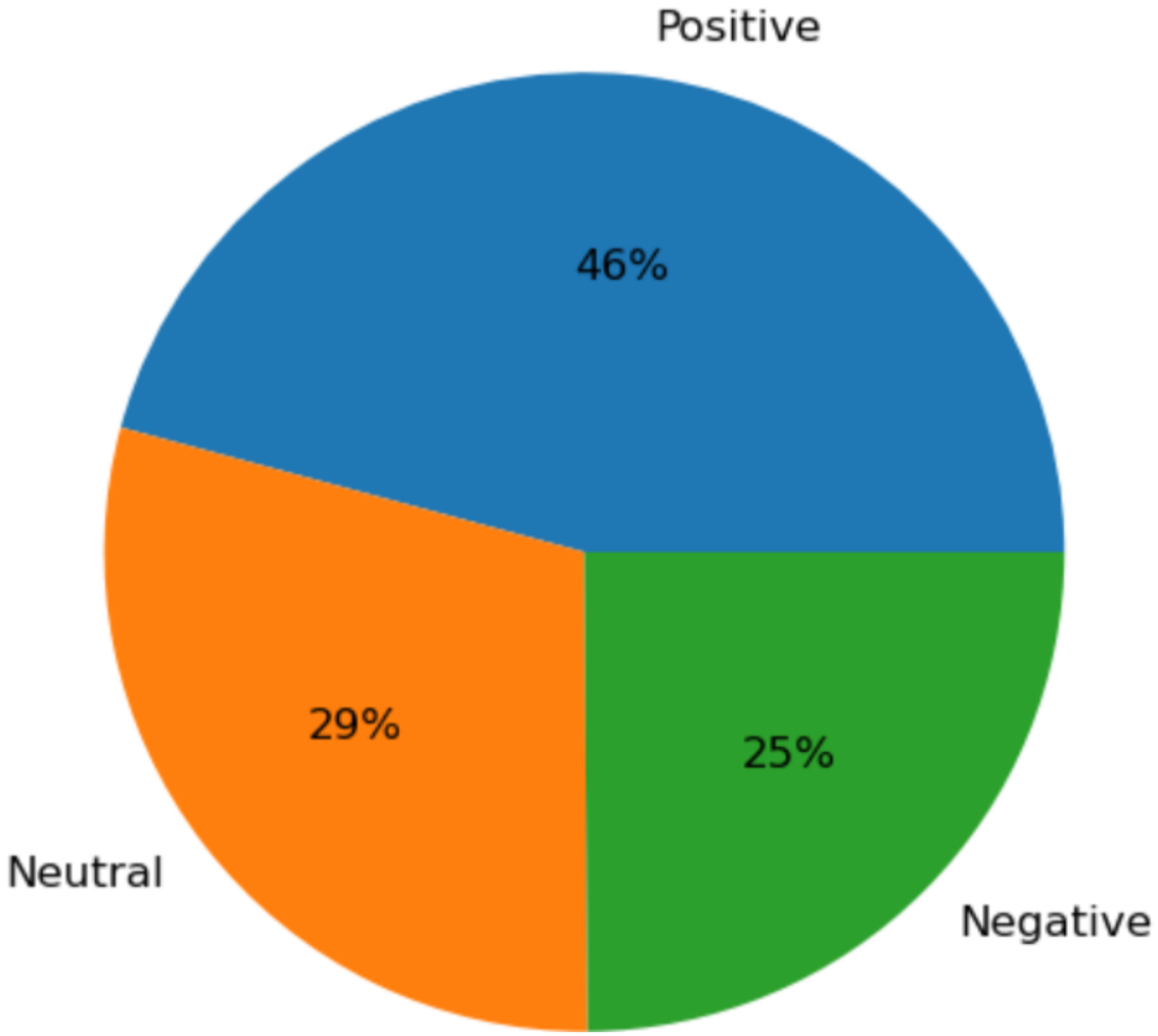
Analyze

- Sentiment distribution
- Trigram Analysis
- Topic Modelling

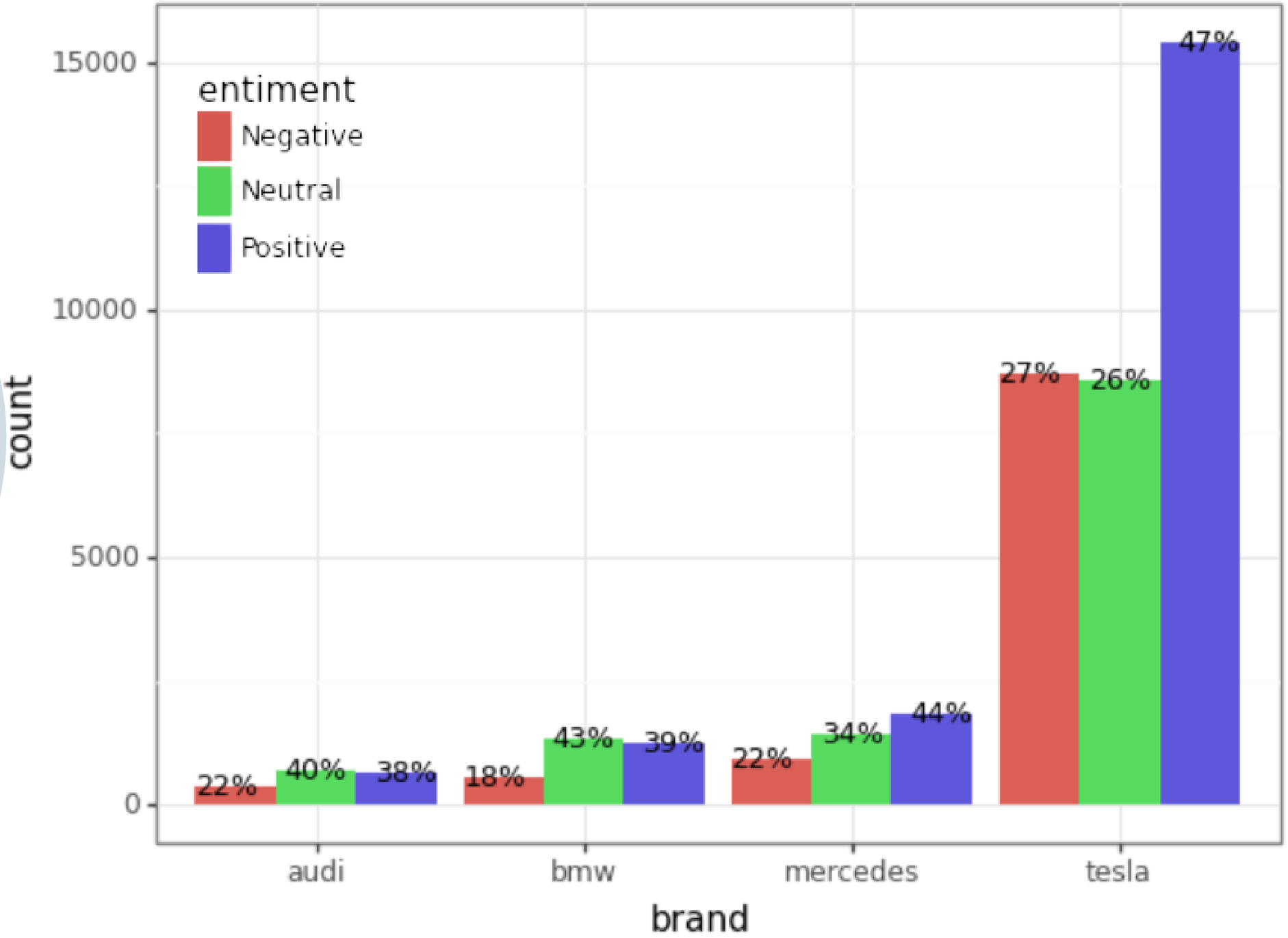
Modelling (Train, test, validation)

Developed 10 ML models, and applied hyper parameter tuning on 5

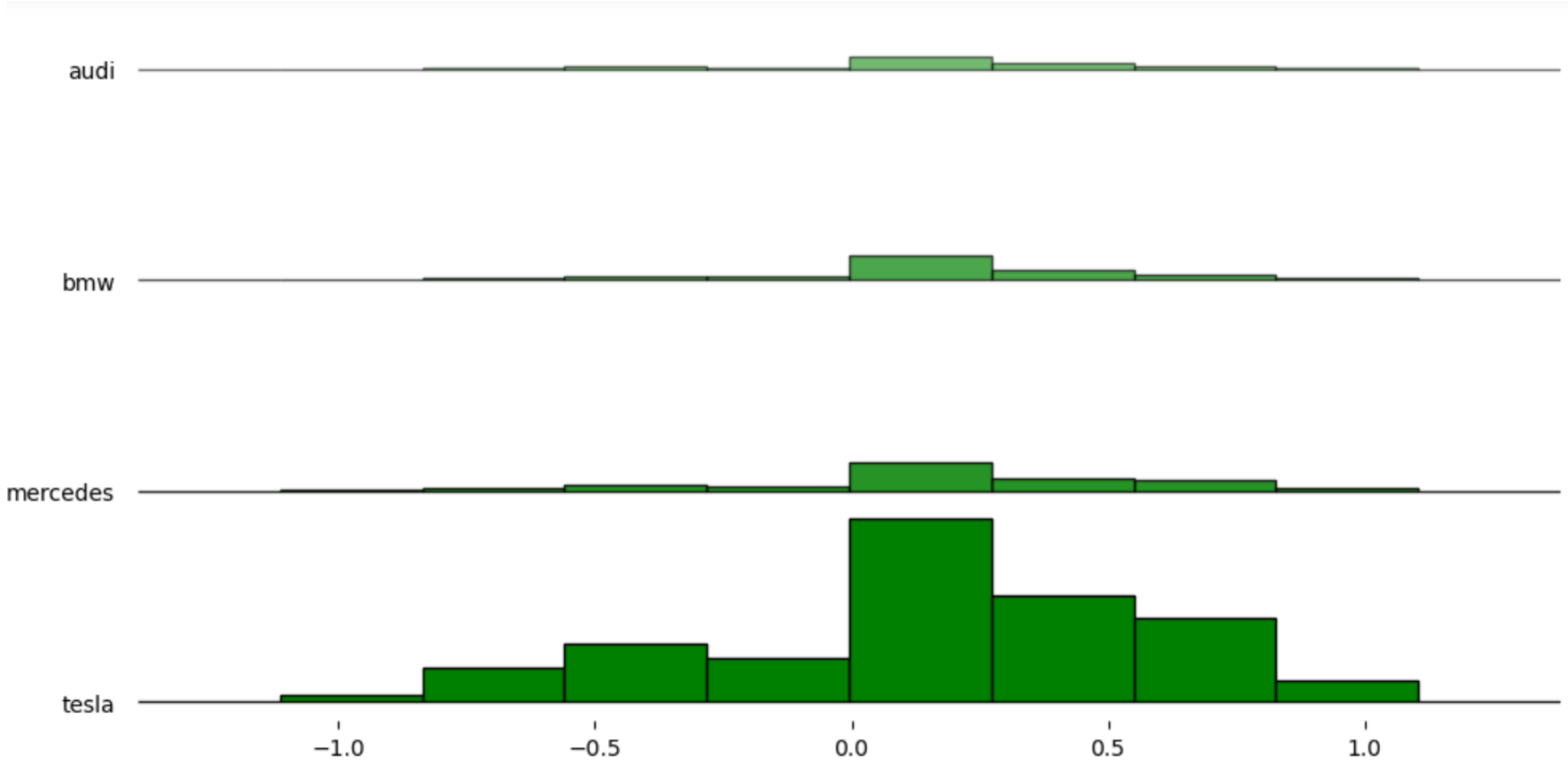
Overall Sentiment Analysis



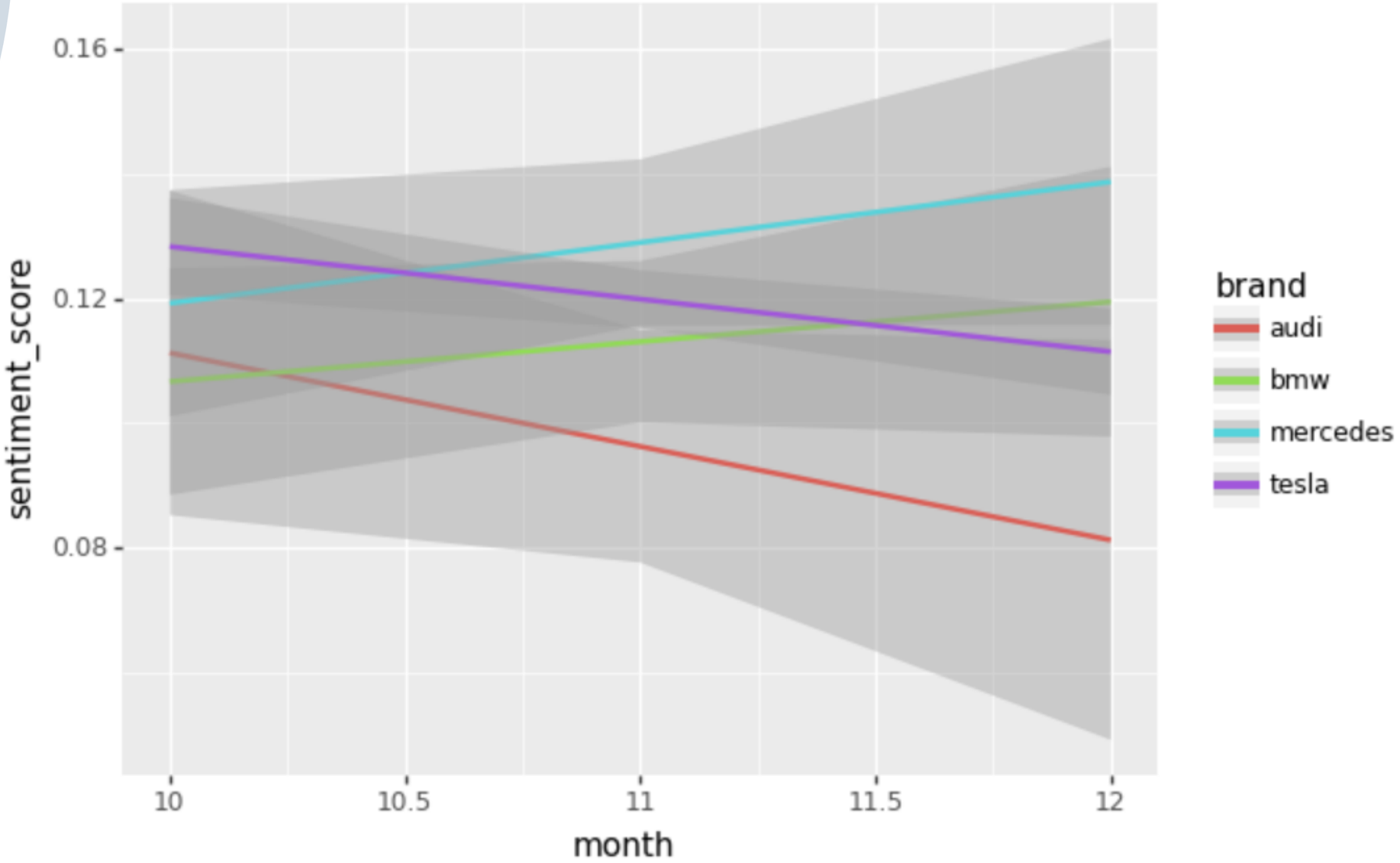
Distribution of Sentiment for different Automobile Brands



Sentiment Score Distribution

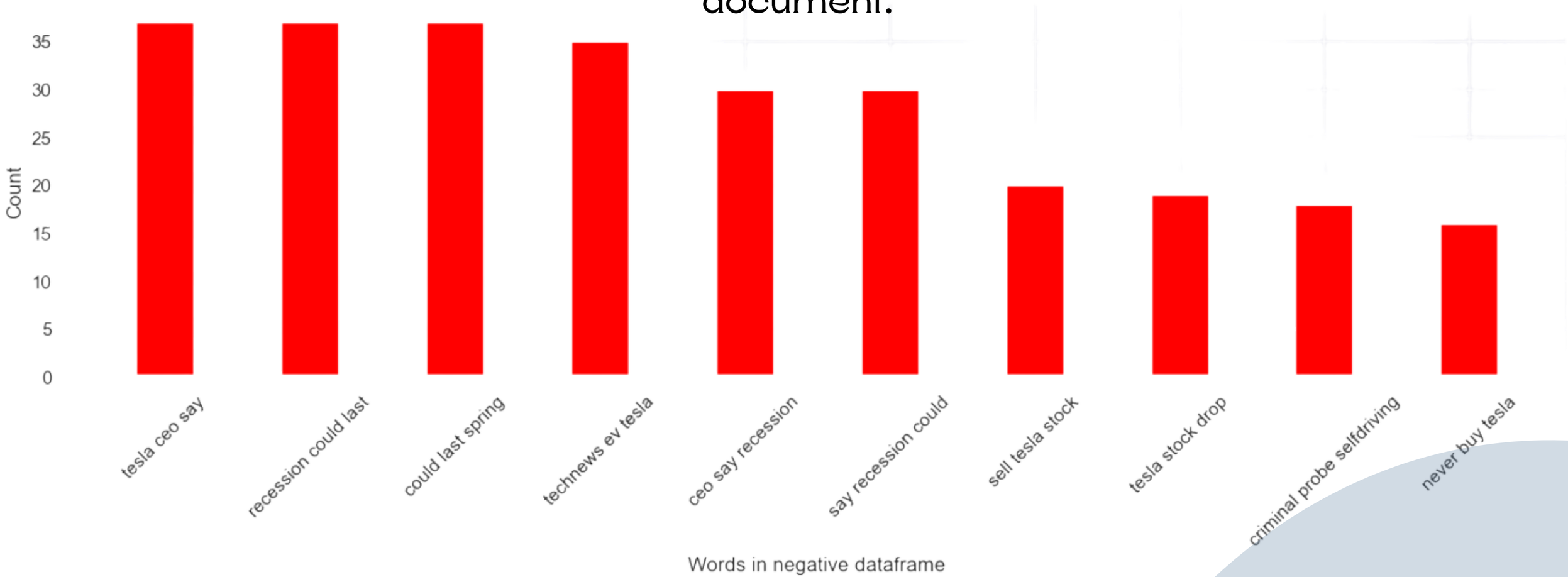


Brand Sentiment Over Time

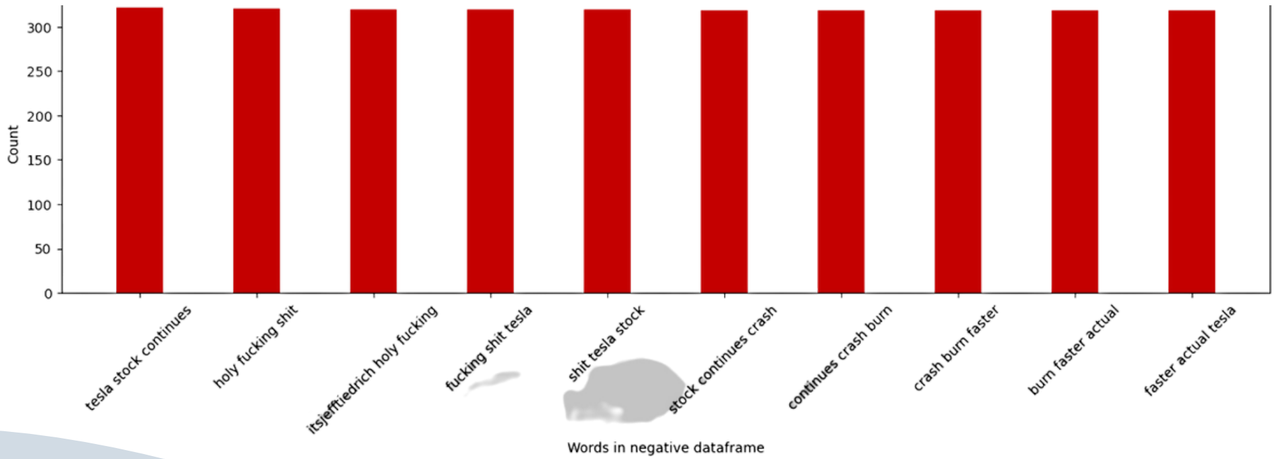


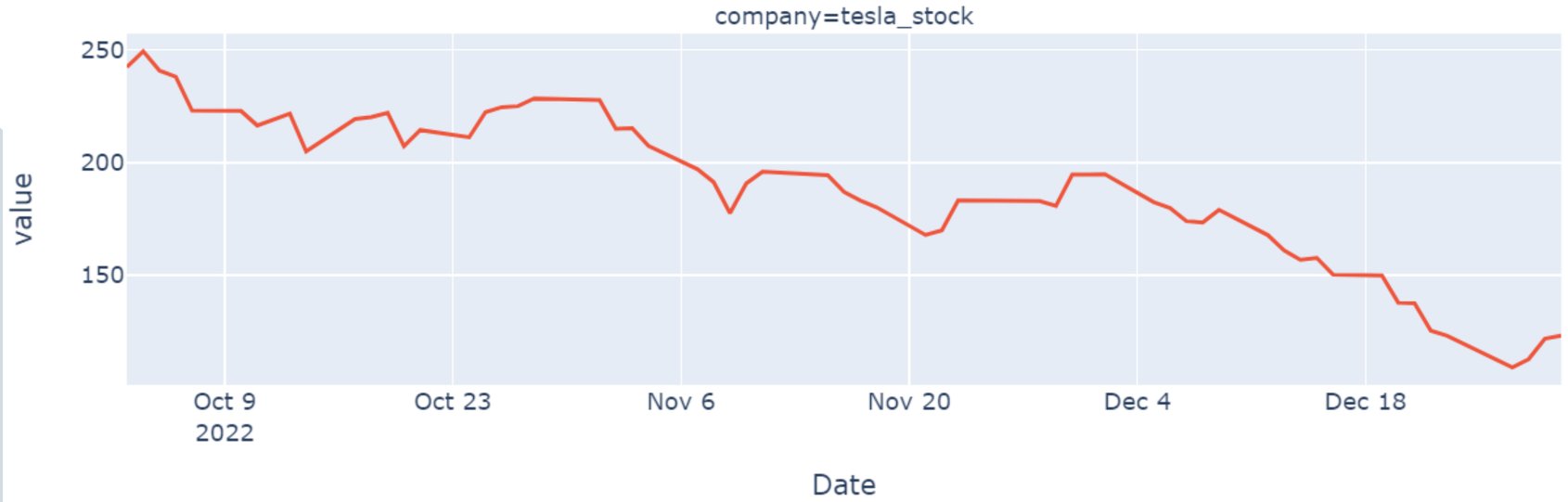
Top 10 words found in NEGATIVE Dataframe(Trigram Analysis)

N-grams are continuous sequences of words or symbols or tokens in a document.



Trigram Analysis





Hypothesis Testing

The **Granger causality test** is a statistical hypothesis test for determining whether one time series is useful in forecasting another

Null Hypothesis: Past twitter sentiment values are not statistically significant for stock price prediction.

Alternative Hypothesis: Past twitter sentiment values are statistically significant for stock price prediction

ADF Test for Stationarity

Sentiment score = stationary

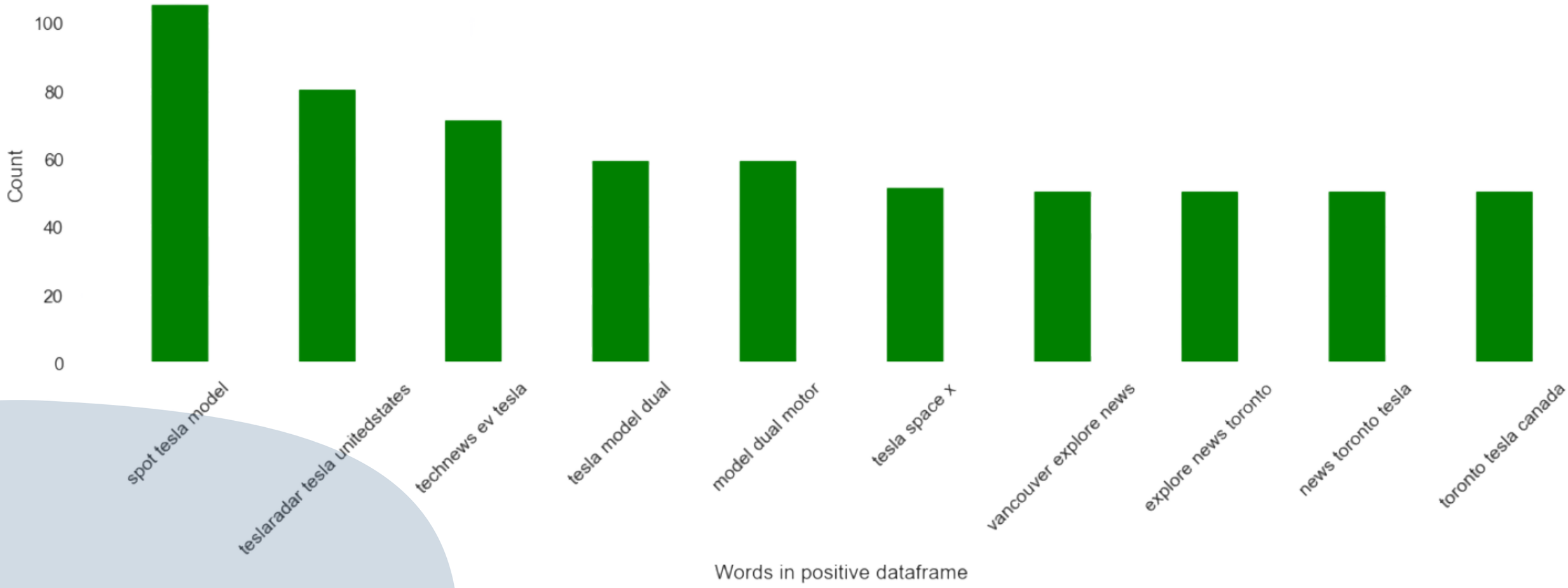
Tesla Stock = stationary after first difference

Granger Causality Test

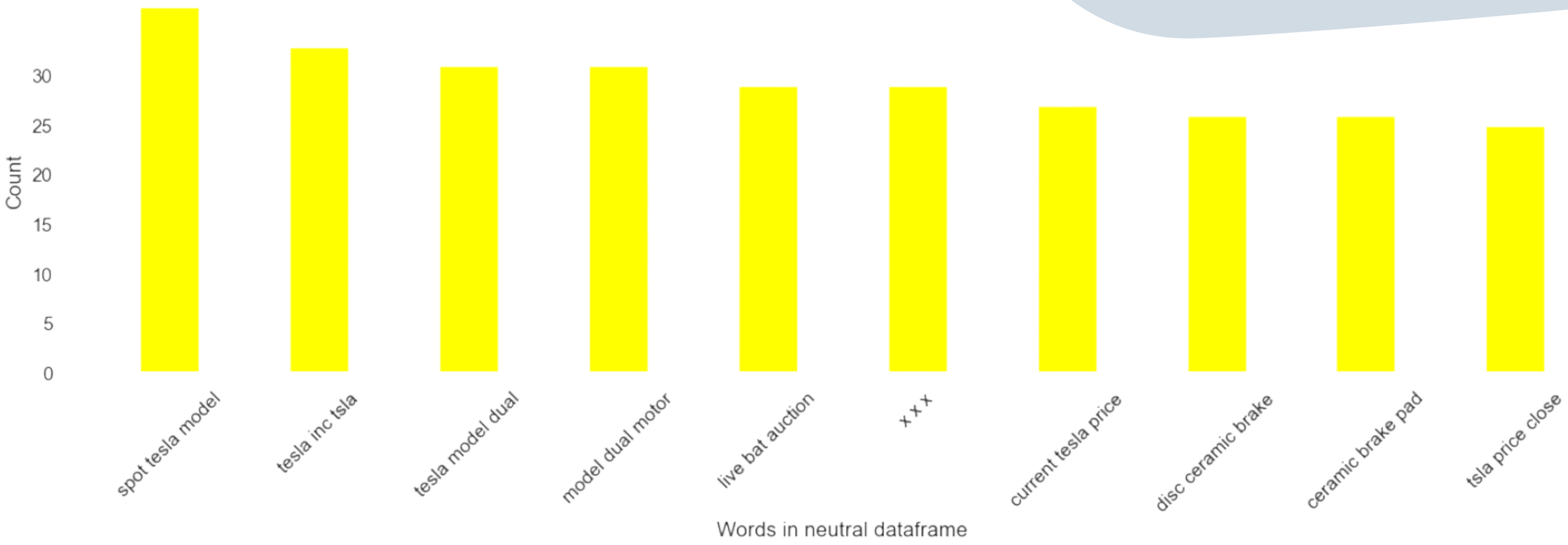
Reject Null hypothesis as p value < 0.05.

	sentiment_score_x	tesla_stock_x
sentiment_score_y	1.0000	0.0
tesla_stock_y	0.0002	1.0

Top 10 words found in POSITIVE Dataframe(Trigram Analysis)



Top 10 words found in NEUTRAL Dataframe(Trigram Analysis)



Topic Modelling (LDA)

Unsupervised Machine Learning Method

Topic 0: tesla stock year tsla ceo investor space via value say
Topic 1: bmw ev mercedes lose driver last guy front amp own
Topic 2: tesla go like would buy share much get many charge
Topic 3: mercedes model thats apple ever right china mercedesbenz everyone always
Topic 4: tesla drive want mercedes dont car im doesnt say bad
Topic 5:tesla wait first stop news love turn state yeah spot
Topic 6: twitter tesla get phone im work keep owner change spacex
Topic 7: tesla money he new market start amp every get free
Topic 8: tesla car audi sell vehicle electric price battery year best
Topic 9: tesla one think make even still company need could get

Topic Modelling (LDA)

1. Tesla stock price and investors
2. BMW vs Mercedes
3. Tesla Share
4. Popularity & love for Mercedes in China
5. Tesla vs Mercedes
6. News about first spot of Tesla
7. Tesla or Twitter
8. Tesla money being invested somewhere else
9. Best selling electric vehicle
10. Company Need

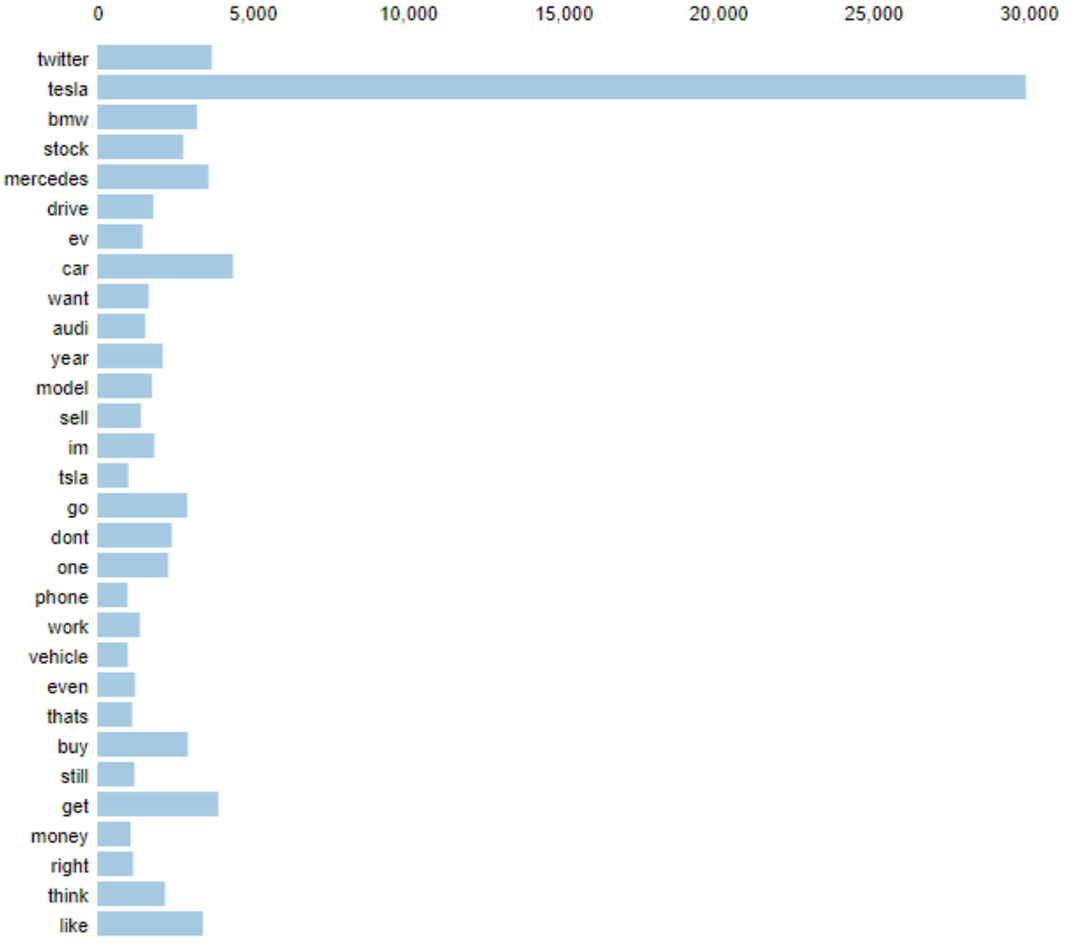
Selected Topic: 0 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)
λ = 1

0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms[1]



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

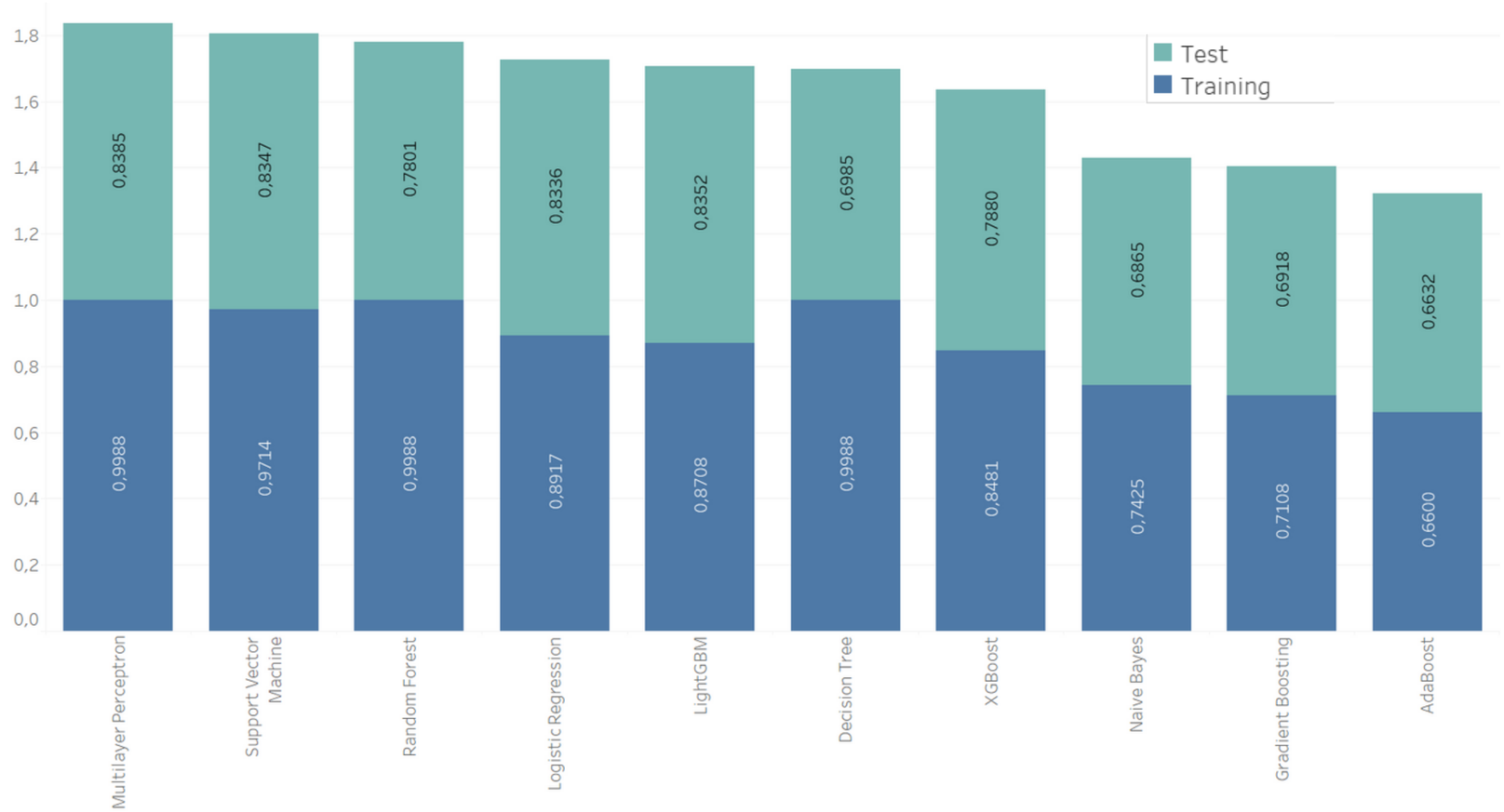
Conclusion

1. what are ppl liking about a brand
2. Tesla is not luxurious
3. Tesla Stock price
4. Tesla needs new CEO

Modelling

Accuracy score of 10 models

- Train-Test-validation: 50%, 33%, 17%
- Best models: MLP, SVC, LGBM, logistic Regression, XGBoosting



Scores...

	Models	Accuracy Training	Accuracy test	F1 Macro	Precision	Recall	Fbeta_half	Cross Validation
8	Multilayer Perceptron	0.998807	0.838515	0.829478	0.829374	0.830305	0.829329	0.828157
6	Support Vector Machine	0.971375	0.834738	0.820620	0.834795	0.818516	0.827442	0.833023
9	Logistic Regression	0.891704	0.833575	0.820335	0.830631	0.818522	0.825366	0.828586
3	LightGBM	0.870808	0.835246	0.819968	0.832562	0.823994	0.825401	0.826248
4	XGBoost	0.848051	0.788028	0.768134	0.795108	0.770087	0.780034	0.787605
0	Random Forest	0.998807	0.780110	0.752642	0.788898	0.753499	0.768215	0.779018
5	Decision Tree	0.998807	0.698460	0.680597	0.679717	0.684147	0.679755	0.685892
1	Gradient Boosting	0.710844	0.691777	0.654766	0.726693	0.652781	0.683434	0.691617
7	Naive Bayes	0.742474	0.686547	0.653633	0.754386	0.631738	0.701543	0.681408
2	AdaBoost	0.659988	0.663228	0.629671	0.691253	0.642527	0.653252	0.650922

Results after Hyper Parameter Tunning

Models	Accuracy Training	Accuracy Validation	F1 Macro	Precision	Recall	Fbeta_half	Cross Validation
Support Vector Machine	0.926769	0.862173	0.850454	0.853587	0.852009	0.851766	0.85535(
Multilayer Perceptron	0.996565	0.860598	0.850036	0.851543	0.850558	0.850695	0.851534
Logistic Regression	0.921092	0.858738	0.847487	0.851173	0.847326	0.849261	0.850341
XGBoost	0.863031	0.800343	0.778622	0.805509	0.779420	0.790693	0.795716
LightGBM	0.632174	0.621297	0.608123	0.631136	0.637004	0.615244	0.624969

SVC

	precision	recall	f1-score	support
negative	0.82	0.71	0.76	1710
neutral	0.85	0.97	0.91	2023
positive	0.89	0.87	0.88	3254
accuracy			0.86	6987
macro avg	0.85	0.85	0.85	6987
weighted avg	0.86	0.86	0.86	6987

MLP

	precision	recall	f1-score	support
negative	0.77	0.74	0.75	1710
neutral	0.86	0.93	0.89	2023
positive	0.89	0.86	0.87	3254
accuracy			0.85	6987
macro avg	0.84	0.84	0.84	6987
weighted avg	0.85	0.85	0.85	6987

Logistic Regression

	precision	recall	f1-score	support
negative	0.82	0.72	0.77	1710
neutral	0.85	0.94	0.90	2023
positive	0.88	0.88	0.88	3254
accuracy			0.86	6987
macro avg	0.85	0.85	0.85	6987
weighted avg	0.86	0.86	0.86	6987

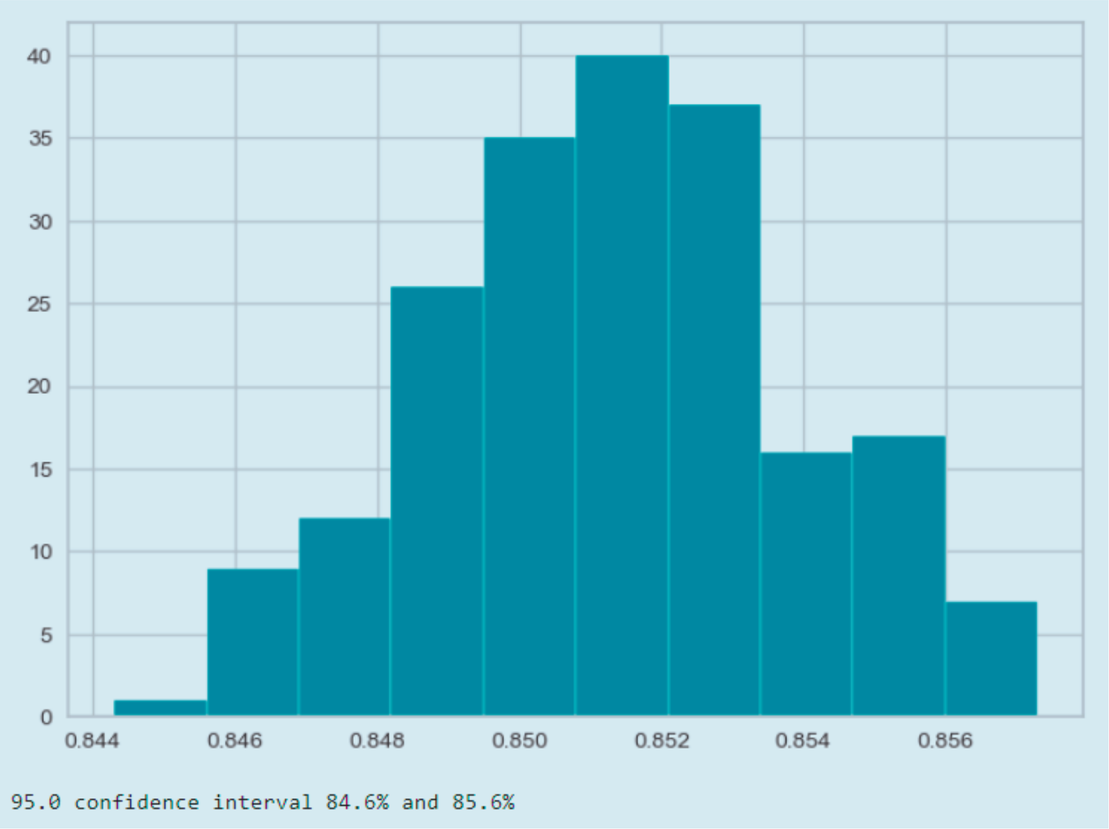
XGBoosting

	precision	recall	f1-score	support
negative	0.78	0.64	0.71	1710
neutral	0.72	0.98	0.83	2023
positive	0.89	0.77	0.82	3254
accuracy			0.80	6987
macro avg	0.80	0.80	0.79	6987
weighted avg	0.81	0.80	0.80	6987

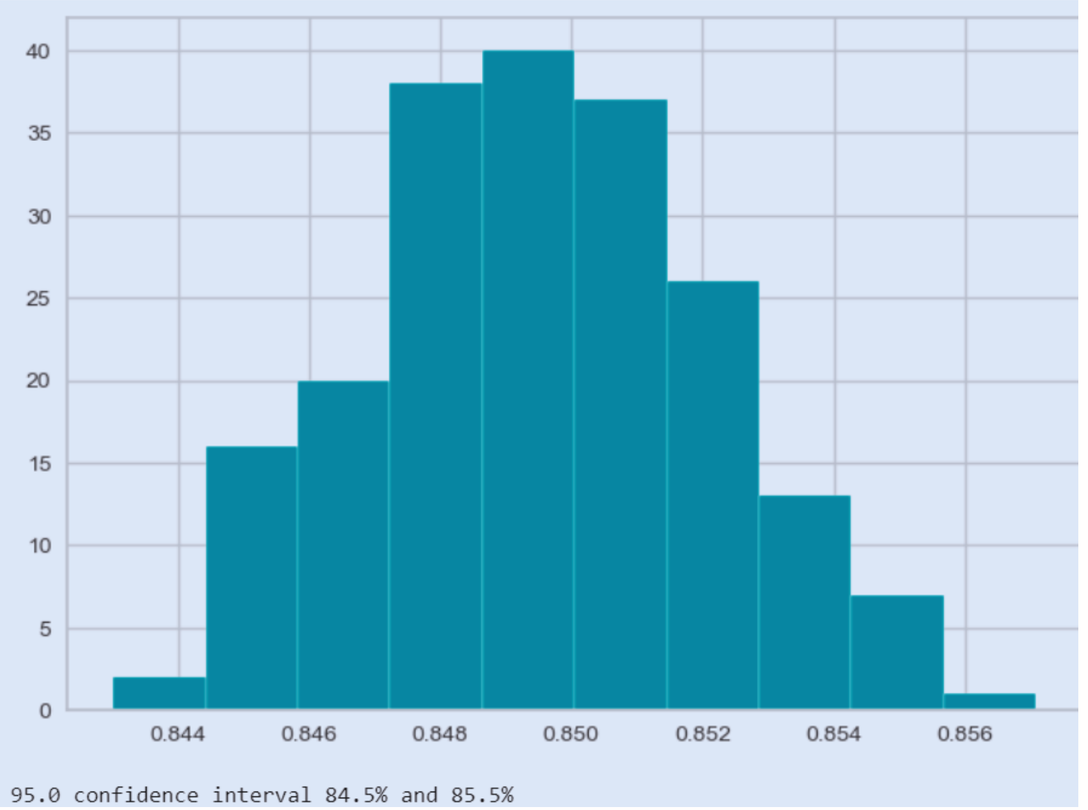
LightGBMC

	precision	recall	f1-score	support
negative	0.84	0.63	0.72	1710
neutral	0.79	0.98	0.88	2023
positive	0.88	0.86	0.87	3254
accuracy			0.84	6987
macro avg	0.84	0.82	0.82	6987
weighted avg	0.84	0.84	0.83	6987

Bootstrapped confidence Interval for Fbeta_half score

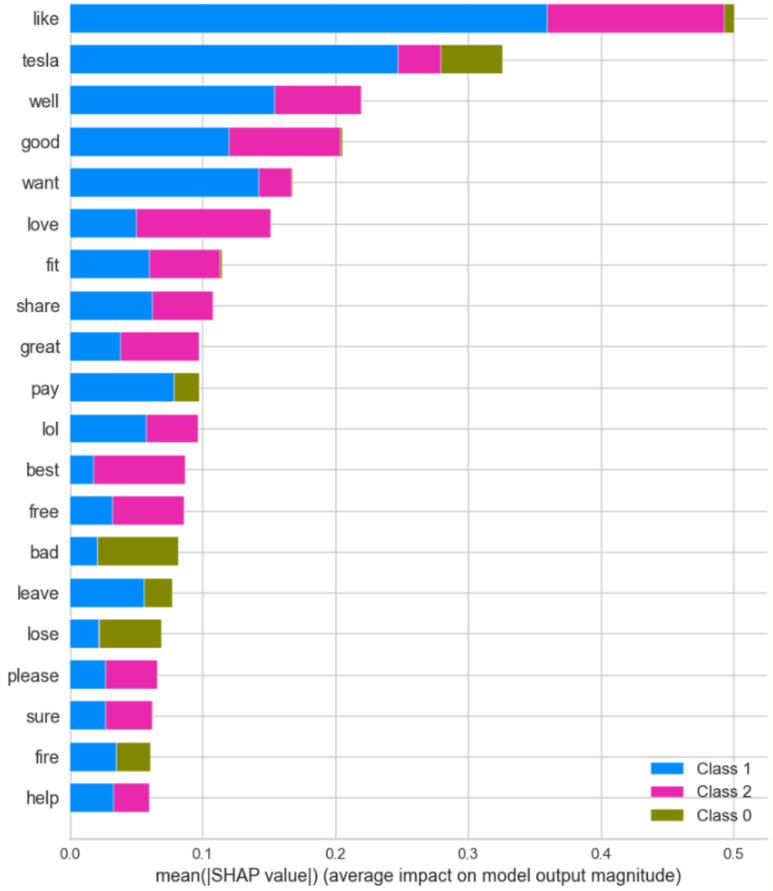


SVC

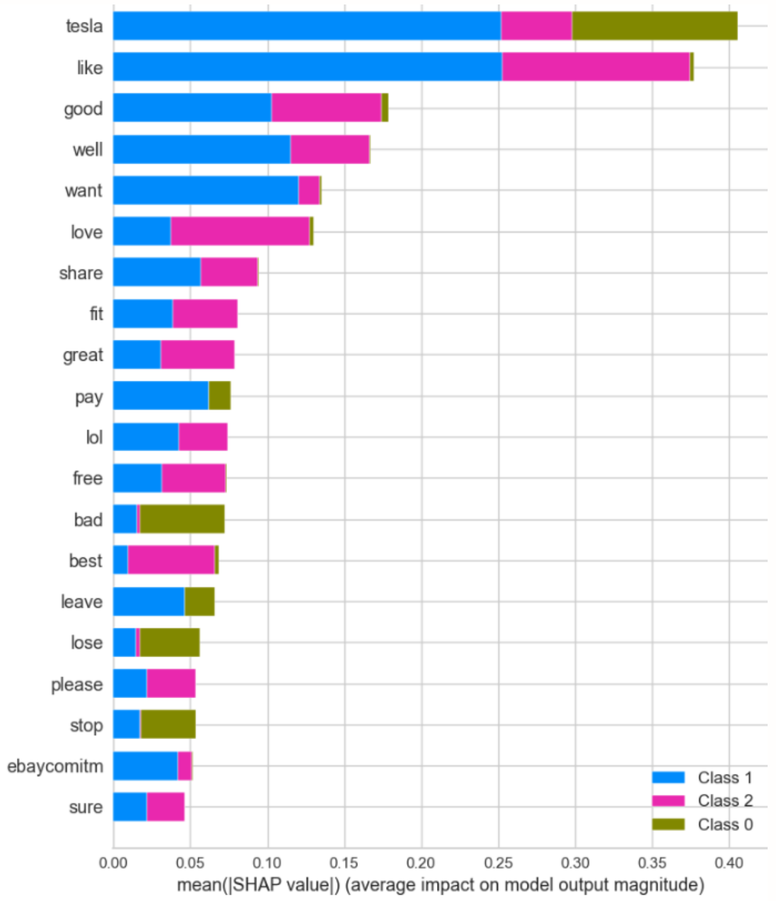


Logistic Regression

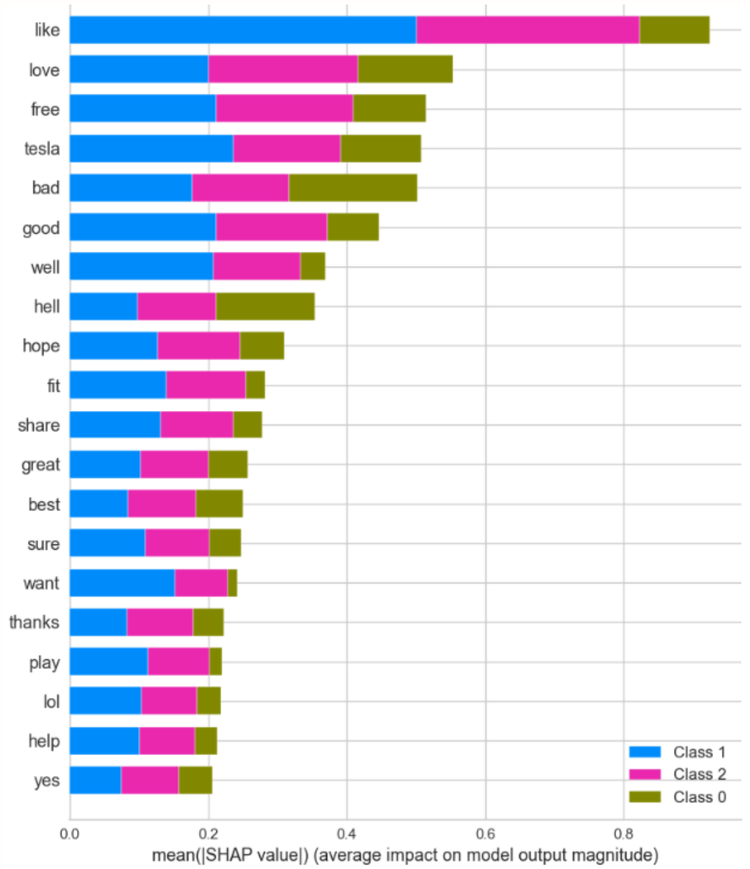
Feature Importance for sentiment classification(Shap Values)



LightGMB



XGBoosting



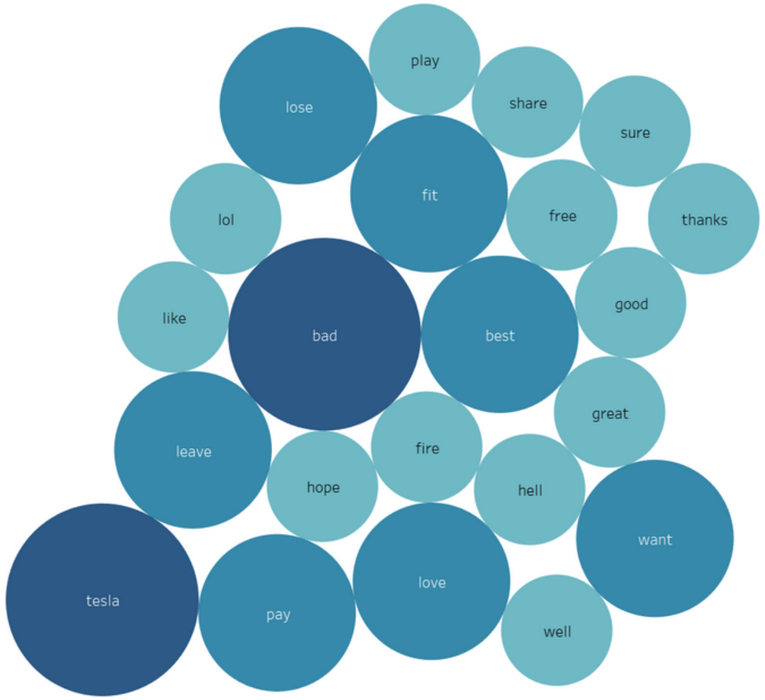
Logistic Regression

class 0: Negative
class 1: Neutral
class 2: positive

Words Frequencies

Postive Sentiment

Negative Sentiment



Neutral Sentiment

Next Steps

1. Preliminary Analysis
2. Add more languages, location for twitter data collection etc and do in-depth analysis of other brands.
3. Sentiment Analysis and stock price prediction

Thank You