Social Media Sentiment Analysis

Quinn Meier

Why?

- Social media is essential to any marketing or analytics team
- People spend an average of 147 minutes per day on social media
- Key insights into consumer trends, sentiment, behavior

Data

- Kaggle dataset

- 732 social media posts from Facebook, Instagram, and Twitter

- Text, Hashtag, Timestamp, Platform, Likes, Shares, Country, Username,

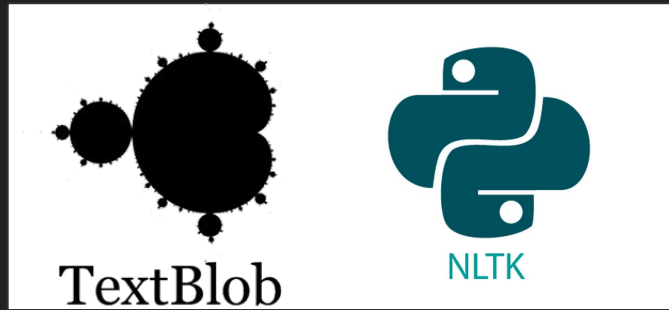
Method

- Multiclass classification problem

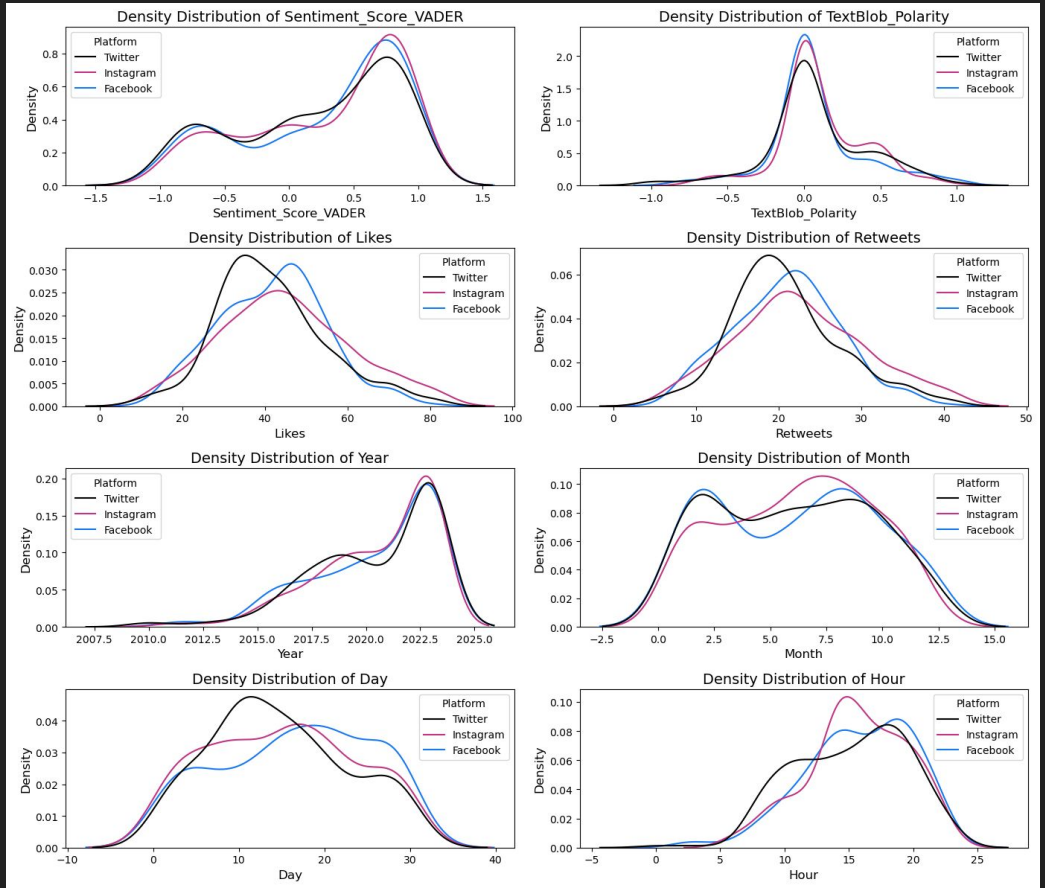
- Sentiment analysis from two different models NLTK VADER and Text Blob

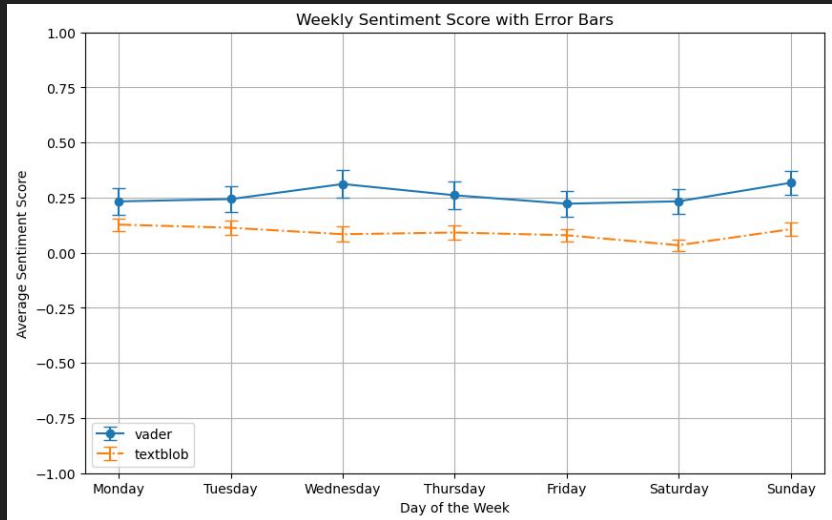
Data Wrangling

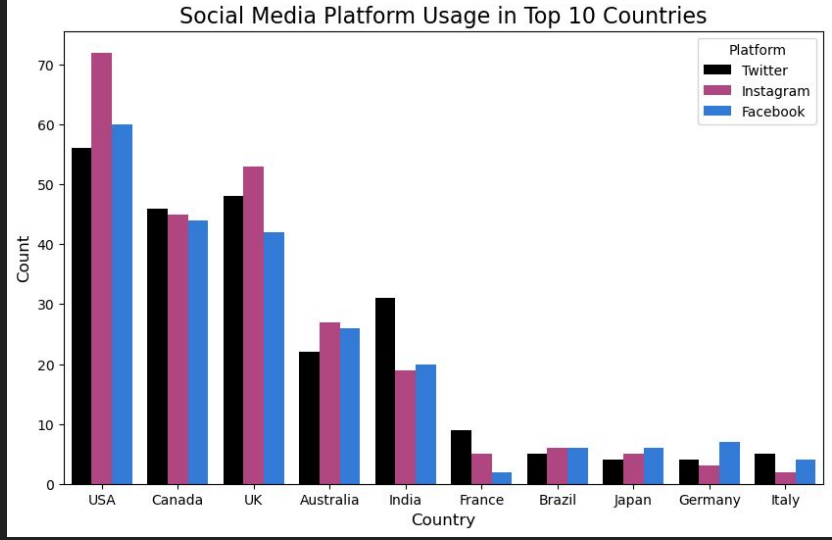
- Index columns (2)
- Duplicate values (26 posts, 1 category - twitter)
- Remove excess whitespace, punctuation, numbers, emojis
- Tokenization, Lemmatization, Removing Stop words
- Text blob and NLTK Vader to assign numeric scores to text data which will be used to train models
- -1 to 1



EDA


Density Distribution of Sentiment_Score_VADER
Density Distribution of TextBlob_Polarity
Density Distribution of Likes
Density Distribution of Retweets
Density Distribution of Year
Density Distribution of Month
Density Distribution of Day
Density Distribution of Hour
Platform
Twitter
Instagram
Facebook
Density
Sentiment_Score_VADER
TextBlob_Polarity
Likes
Retweets
Year
Month
Day
Hour


Weekly Sentiment Score with Error Bars
Average Sentiment Score
vader
textblob
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week


Social Media Platform Usage in Top 10 Countries
Platform
Twitter
Instagram
Facebook
Count
USA
Canada
UK
Australia
India
France
Brazil
Japan
Germany
Italy
Country

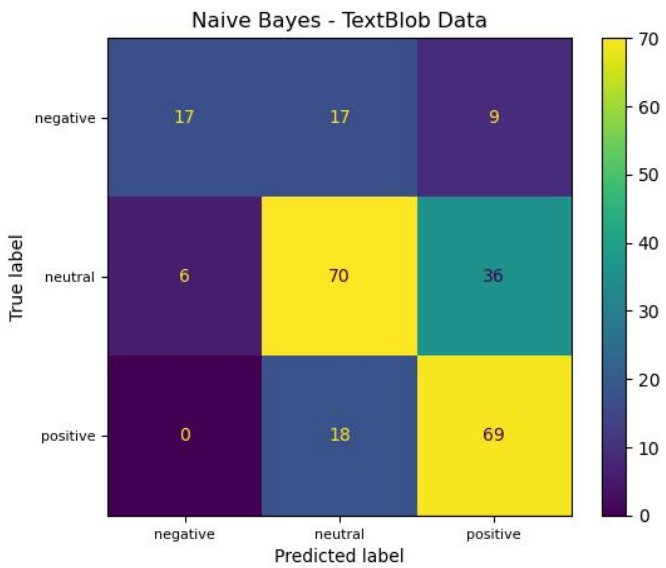
Modeling

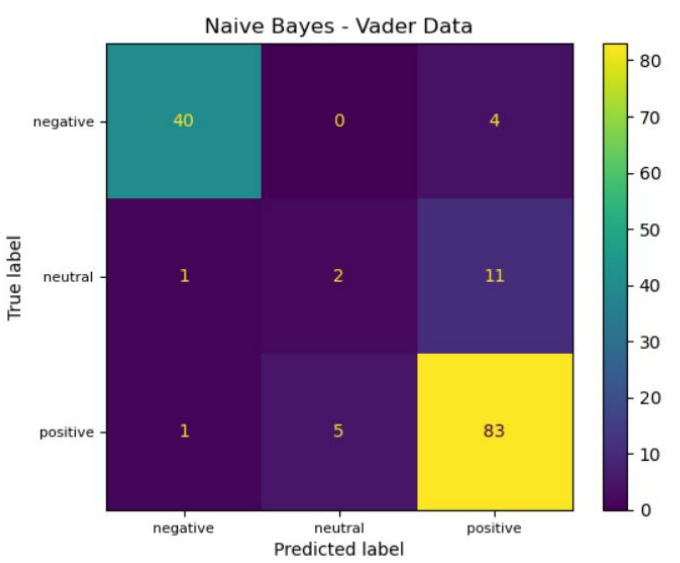
- 80/20 train-test split

- Logistic Regression
- Support Vector Machines
- Random Forest
- Naive Bayes
- Gradient Boosting

Performance

- Naive Bayes
- Vader





1	***features***									
2										
3	***dataframe shape***									
4	***models***	Logistic Regression	Logistic Regression	SVM	SVM	Random Forest	Random Forest	Naive Bayes	Naive Bayes	Gradient Boosting	Gradient Boosting
5		VADER	TextBlob	VADER	TextBlob	VADER	TextBlob	VADER	TextBlob	VADER	TextBlob
6	***hyperparameters***										
7	random_state	123	123	123	123	123	123	123	123	123	123
8	training sample size	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
9	solver / kernel / criterion	liblinear	liblinear	n/a	n/a	n/a	n/a	n/a	n/a	mlogloss	mlogloss
10	C	10	100	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
11	penalty	l2	l1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
12	max_depth	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	5	3
13	learning_rate	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.1	0.1
14	n_estimators	n/a	n/a	n/a	n/a	200	200	n/a	n/a	200	200
15	***performance***										
16	accuracy	0.844	0.740	0.844	0.657	0.748	0.690	0.850	0.645	0.803	0.669
17	F1	0.831	0.728	0.833	0.645	0.706	0.659	0.834	0.637	0.781	0.651

Future Work

- Could apply analysis to a particular company, brand, or product
- Day of week
- Country
- Tracking world events, product launches etc.

Questions?