Final Report
Social Media Capstone

Problem Identification

This project uses Natural Language Processing (NLP) techniques to analyze sentiments and trends in the Social Media Sentiments Analysis Dataset from Kaggle. By applying NLP methods, we seek to extract meaningful insights from user-generated content across various social media platforms. This is important to study as social media is an integral tool for anyone seeking to understand their potential customer base and market to them appropriately. Understanding the sentiments and trends prevalent in social media content can provide valuable insights for businesses seeking to capitalize on this easy access to consumers.

The data utilized in training these models comes from Kashish Parmar's social media sentiment analysis dataset from Kaggle. The dataset consists of 14 attributes of 732 social media posts categorized as positive, neutral, or negative in sentiment, with no missing values. Several supervised machine-learning classification models were used to address this problem, which will be explored more deeply in this report.

Data Wrangling

For the data-wrangling portion of this project, the first step was to remove unnecessary columns and duplicate posts from the analysis. This led to two redundant index columns and 26 posts being removed from the dataset. The next step was removing unnecessary whitespace, punctuation, and emojis from the text data for easier processing. Additionally, Twitter had been separated into two different categories due to differing whitespace, which was also addressed. After the text data has been cleaned up, the process of preparing it for processing can be performed by performing tokenization - breaking up the sentence into words, lemmatization - the process of taking those tokens and getting their roots, and the removal of stop words such as the, of, a, etc.

After completing these steps, the processed text was fed into two sentiment analysis models: NLTK Vader and TextBlob. These models assign numeric scores to the processed text, ranging from -1 to 1, where -1 to -0.05 is assigned a Negative sentiment, -0.05 to 0.05 is assigned a Neutral sentiment, and the rest are assigned to positive. The final shape of the dataset was 706 unique social media posts across 19 columns.

Exploratory Data Analysis

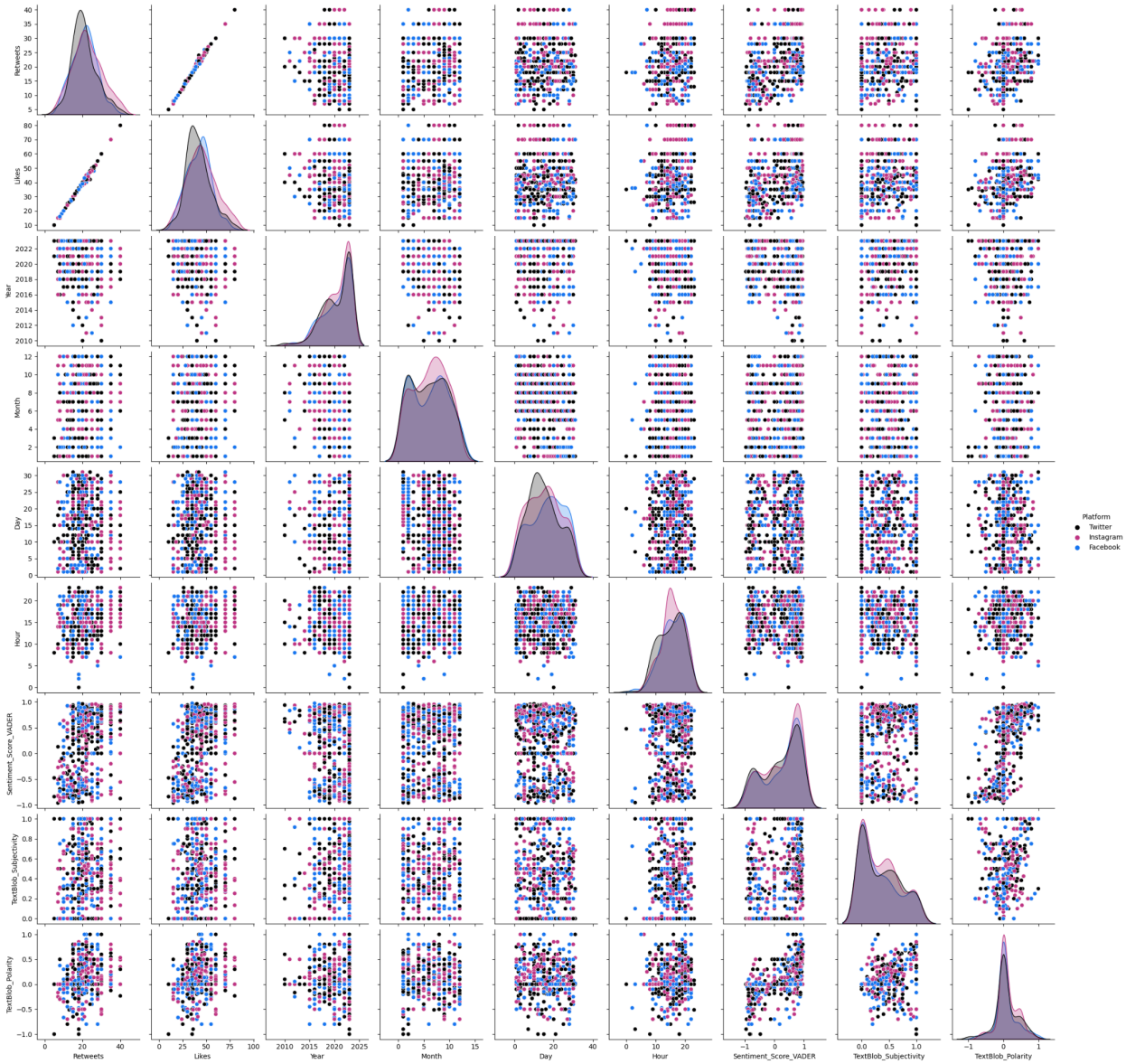
Once the data had been processed, exploratory data analysis could begin. This included examining the structure and searching for trends, patterns, and statistical relationships between the dataset's features, both visually and numerically.

An explanation of the variables in the dataset is helpful for context.

- Text: the text of the social media post
- Sentiment: the sentiment assigned in the dataset - some are simply positive, neutral, and negative, while others are more specific, i.e., Calm, Bitter, Affectionate, etc.
- Timestamp: date time of when the post was created
- User: username of the poster
- Platform: where the post was created: Facebook, Twitter, or Instagram
- Shares: the number of times the post was shared, retweeted, reposted, etc.
- Likes: the number of likes the post received
- Country: country where the post was created

The timestamp is also broken up into year, month, day, and hour components and saved in the dataset.

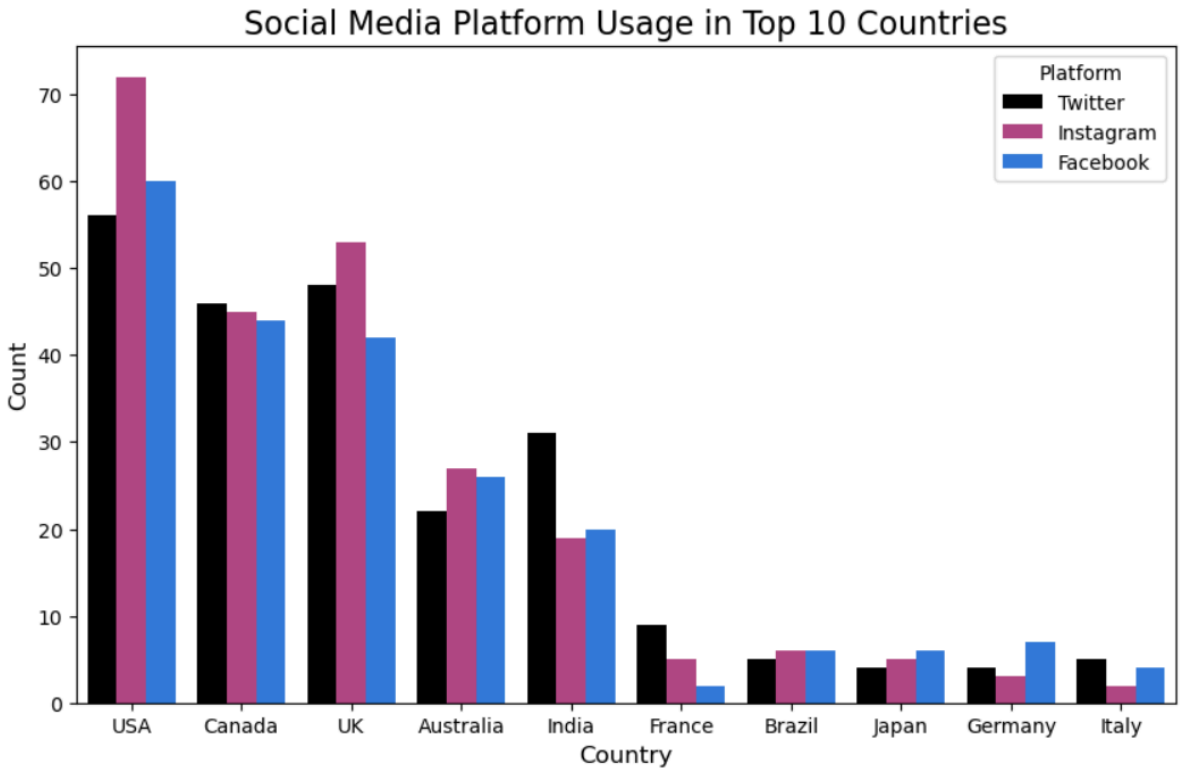
Below is a combined visualization plotting the above variables against each other. This is a portion of the Seaborn library known as a pairplot. The light blue data points represent posts created on Facebook, the pink points represent posts on Instagram, and the grey points represent posts on Twitter. Several trends become apparent, such as the similarity of overall sentiment across the platforms with the two different models and the difference in the distribution of the two models, where TextBlob tends to skew much more neutral. At the same time, VADER reads the dataset more positively. Another trend observed is the strong correlation between likes and shares across all platforms. This falls in line with what is expected of social media posts.

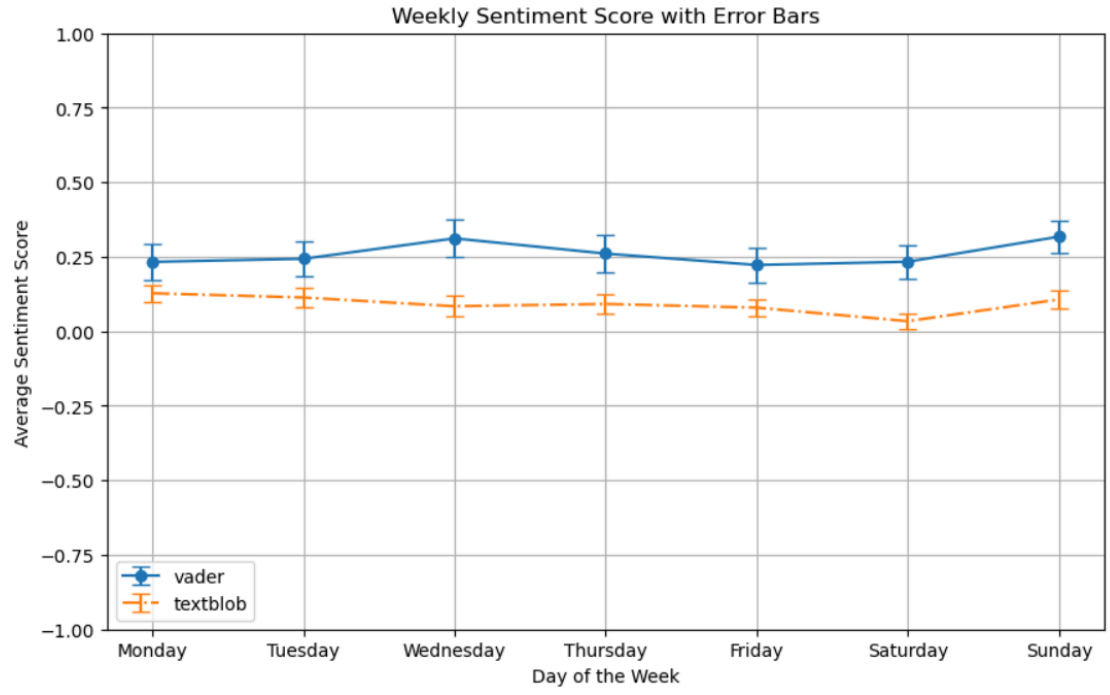


Other examples of trends investigated include the share of social media platform usage by country for the top 10 most common countries in the dataset. From that, it can be seen that although instagram was the most popular in the United States, in other countries twitter was the most popular, in some cases by a large margin. The plot showing these popularities can be seen below.

The trends of how sentiment changes over time were also investigated with line plots across all of the DateTime components of Year, Month, Day of the Month, Day of the Week, and Hour. An example of one of these plots is seen below, across the days of the week. It is clear that the two days that have the highest average VADER sentiment

score are Wednesday and Sunday. It would be interesting to see if this trend was still evident with a larger sample of data. Additionally, the wednesday positive trend seems to vanish with the TextBlob sentiment analysis, indicating that the trend may indeed be illusory with Wednesday, as one might expect from intuition.





Preprocessing and Training

For the preprocessing and training portion of this analysis several factors are considered. The primary consideration is splitting the dataset into training and test sets to prevent overfitting and to ensure the models are more generalizable. In this case, the datasets were split into 80% training and 20% test. This will be used in addition to grid search cross validation. Several other typical considerations for this portion of the analysis would be transforming categorical variables into dummy variables and scaling the data. In this case, neither of those apply as we are not using any categorical variables in the training of the models and scaling doesn't make sense to apply to text data.

Modeling

Several different machine-learning classification algorithms were utilized for the modeling phase of this analysis. Each will be explored individually to discuss its respective strengths and weaknesses. Each model underwent grid search cross-validation to optimize its hyperparameters and was evaluated using accuracy, F1 scores, and confusion matrices to determine which performed the best. A table of those hyperparameters and the evaluation scores for each model is provided below.

The first model considered was logistic regression. Logistic regression is the standard baseline for almost all classification problems and should be the starting point for any modeling of this kind. It is straightforward to implement, interpret, and explain. The model performed reasonably well, as expected, with an accuracy for VADER of 84.4% and an F1 of 0.831. Logistic regression also performed the best of all the models on the TextBlob sentiments, as seen in the metrics table below.

The second model considered was support vector machines (SVM). SVM performs well for datasets with many features. It measures some degree of "closeness" in the hyperspace of the parameters of the dataset. These features of SVM allow it to capture complex relationships between parameters that would otherwise be virtually impossible to tease out. On the other hand, these same factors cause SVM to be less interpretable than other models like logistic regression or decision trees and become computationally prohibitive with large datasets. In this case, the latter was not an issue. SVM also requires careful tuning of its hyperparameters to achieve the best results. In this case, VADER SVM scored an accuracy of 84.4% and an F1 of 0.833.

The third model considered was random forest. Random forest is an ensemble of decision trees that improves prediction accuracy and controls overfitting by averaging multiple deep decision trees trained on different parts of the training set. RF generally represents a significant improvement in accuracy over a single decision tree at the cost of reduced interpretability. In this instance, RF underperformed relative to the other models with a VADER accuracy of 74.8% and an F1 score of 0.706.

The fourth model considered was a Naive Bayes Classifier. The “Naive” assumption is that the features are independent of one another, providing a simple and powerful framework for classifcation. Naive Bayes is particularly well suited for text processing, as seen here, as Naive Bayes was the most performant and accurate of all the models. With a VADER accuracy and F1 of 85% and 0.834, respectively. That being said, Naive Bayes performed the worst of all of the models with the TextBlob sentiments.

The fifth and final model considered was gradient boosting (GB). GBMs are ensemble methods that combine multiple weak learners to reduce a loss function. GB updates weights by computing the negative gradient of the loss function with respect to the predicted output. The loss function is the difference between the actual and the predicted variables. Gradient boosting can achieve some of the best accuracies possible with machine-learning models; however, it tends to overfit if not tuned properly and is less interpretable than other models. In this case, gradient boosting was a middle-of-the-road model with a VADER accuracy score of 80.3% and an F1 of 0.781.

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

Conclusions and Future Research

As seen above, Naive Bayes and Vader were the strongest combination in this case. Vader outperformed TextBlob across the board as TextBlob was far less sensitive to sentiment both positive and negative, as seen in the confusion matrices above.

With more extensive data covering a particular product or company, those companies could evaluate consumer mindsets at a specific moment about their company, product, and the state of the world. This would be invaluable data for these companies seeking to understand how consumers respond to their marketing, new product launches, app updates, etc. With more detailed location data from underrepresented countries, a more detailed analysis of sentiment trends by location and time could be performed to determine if certain regions or nations have different sentiment trends than others and if certain prominent world events, both long and short-term, had significant impacts on sentiment in the affected areas and globally, i.e., measuring the effects of the local sports team winning a championship vs. a global economic downturn. Another example of research that could be done is examining if certain days of the week have apparent differences in sentiment (more negative on Mondays, more positive near the weekend, for example), which could be leveraged by a marketing team to plan their media campaigns.