JSC270 Assignment 4

Advaya Gupta, Chloe Nguyen

Notebook Link: Click here

Work Breakdown:
Both members contributed to every part of the project.
Below lists the member that was mostly responsible for each section.
Part 1 — Chloe: A-I; Advaya: J-K
Part 2 — Chloe: Motivation, Data, Exploratory Data Analysis;
Advaya: Model, Results
Presentation — Chloe & Advaya

1 Sentiment Analysis with a Common Twitter Dataset

A) The balance between the three classes of sentiments for the training dataset is:

Sentiment Label	Proportion
0	0.374159
1	0.187407
2	0.438434

(B) Tokenize the tweets. See colab notebook.

(C) Remove any URL tokens from each of the observations. See colab notebook.

(D) Remove all punctuation and special characters. Also, convert all tokens to lowercase only. See colab notebook.

Q: Can you think of a scenario when you might want to keep some forms of punctuation?

We may want to keep punctuation such as the ampersands (&) as it signifies that the next word is a username (when working with twitter tweets). Other punctuation such as $ or % can also contribute to the meaning of a sentence-.

Further, if we want to do Natural Language Generation, we would want to keep punctuation since they are integral to a sentence's structure.

(E) Stem tokens. See colab notebook.

(F) Remove stopwords. See colab notebook.

(G) Convert lists of words into vectors of word counts. See colab notebook.
The matrix of counts for the training dataset is 41151 by 1000- where 41151 is the number of tweets and 1000 is the word count. Hence, the length of our vocabulary is 1000.

(H) Fit a Naive Bayes model to data (using count vectors). See colab notebook.

Training accuracy: 0.667
Testing accuracy: 0.660

5 most probable words in each class and their counts:

Sentiment Label	Word	Count
0	[empty string]	255620
	coronaviru	18059
	covid19	12857
	food	7230
	price	9047
1	[empty string]	255620
	coronaviru	18059
	covid19	12857
	store	8192
	supermarket	7744
2	[empty string]	255620
	coronaviru	18059
	covid19	12857
	store	8192
	thi	7962

(I) Q: Would it be appropriate to fit an ROC curve in this scenario?

Yes. ROC curves are useful as they proide a graphical representation for evaluating the performance of our model. They can be used with imbalanced data- however, if the data is severely imbalanced then each data point may have a greater impact. From part 1a, our data is not severely imbalanced. Fitting an ROC curve is appropriate in this scenario.

(J) Fit a Naive Bayes model to data (using TF-IDF vectors). See colab notebook.

Training accuracy: 0.659
Testing accuracy: 0.653
We found that the training and testing accuracies for TF-IDF vectors were slightly lower than the accuracies using count vectors (training accuracy was 0.667 and testing accuracy was 0.660).

5 most probable words in each class and their counts:

Sentiment Label	Word	Count
0	[empty string]	255620
	coronaviru	18059
	covid19	12857
	food	7230
	price	9047
1	[empty string]	255620
	coronaviru	18059
	covid19	12857
	store	8192
	supermarket	7744
2	[empty string]	255620
	coronaviru	18059
	covid19	12857
	store	8192
	thi	7962

(K) Fit a Naive Bayes model to data (using TF-IDF vectors and using lemmatization instead of stemming). See colab notebook.

Training accuracy: 0.650
Testing accuracy: 0.598
We found that the training and testing accuracies using TF-IDF vectors and lemmatization were slightly lower than the accuracies using TF-IDF vectors and stemming (training accuracy at 0.659 and testing accuracy at 0.653) and were also lower than the accuracies using count vectors and stemming (training accuracy at 0.667 and testing accuracy at 0.660).

5 most probable words in each class and their counts:

Sentiment Label	Word	Count
0	[empty string]	255620
	contact	18054
	corpor	12857
	practic	8944
	farm	7230
1	[empty string]	255620
	contact	18054
	corpor	12857
	still	8177
	street	7744
2	[empty string]	255620
	contact	18054
	corpor	12857
	still	8177
	glove	6741

Bonus: Naive bayes is a Generative model.

2 NLP with Twitter API

Problem Description & Motivation

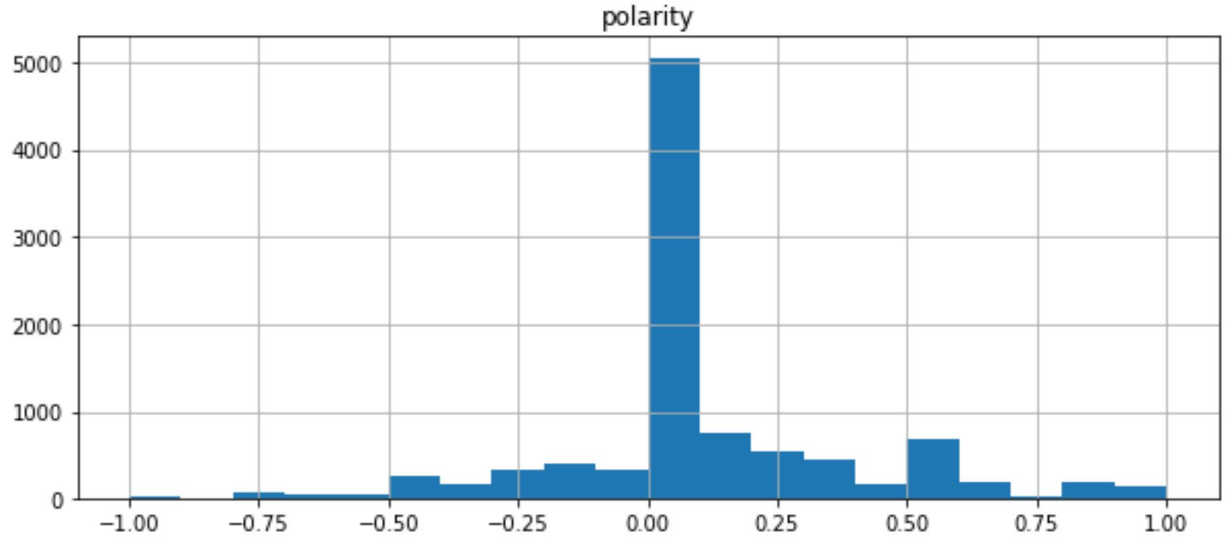
For this analysis, we were interested in taking a look at sentiments in Twitter tweets. Particularly, we wanted to explore the relationship between the polarity of tweets and its subjectivity. To do so, we decided to fit various models: some for the goal of inference and some for the goal of prediction.

The Data

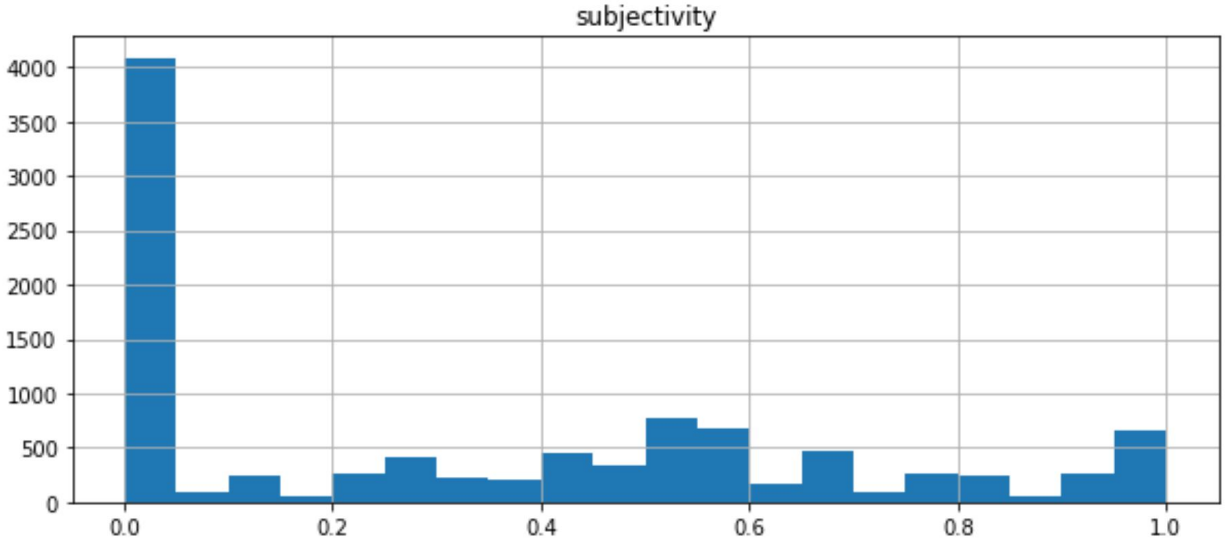
Our dataset contains 9999 observations and 4 features: text, favorite_count (count of likes on a specific tweet), polarity (ranges from -1, negative, to 1, positive) and subjectivity (ranges from 0, objective, to 1, subjective). The features text and favorite_count were taken using the Twitter API. The features polarity and subjectivity were calculated using TextBlob. The data was extracted using the Twitter API with the parameters: q="#", lang="en", since="2021-01-01", result_type="mixed", tweet_mode="extended". This dataset contains a mix of both popular and real time tweets (in extended mode) that include at least one hashtag symbol and is in English. We chose to query only English tweets since we wanted to avoid translation misinterpretations. Note that since we are using the free version of the Twitter API, we are limited to extracting tweets within 7 days. Hence, the since parameter did not matter as long as we chose a date that is over a week before the current date.

Exploratory Data Analysis

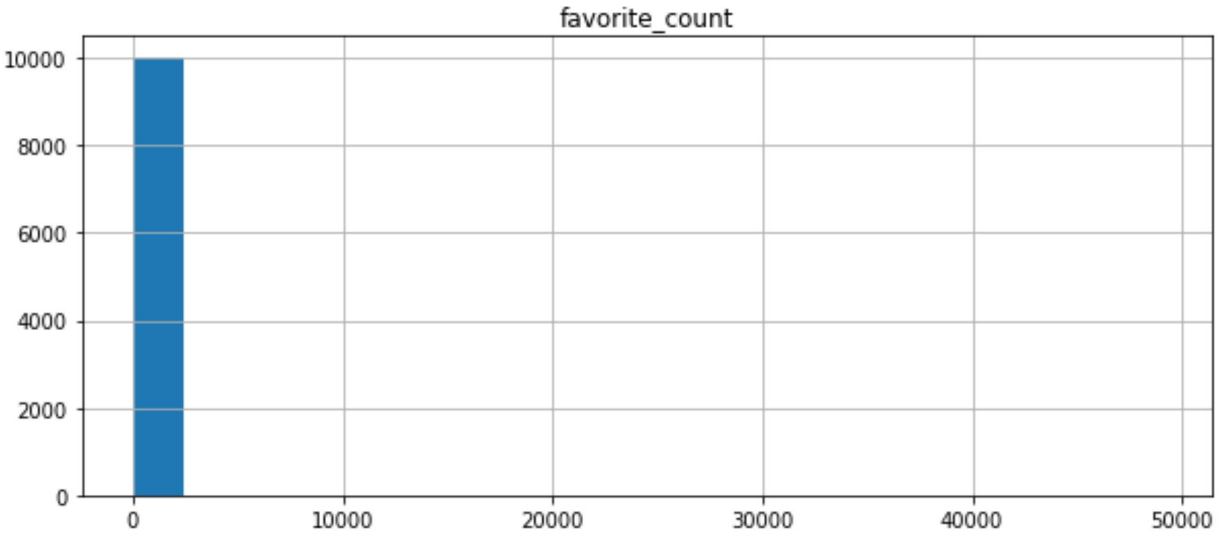
The domain of polarity ranges from -1, representing negative, to 1, representing positive. The distribution of polarity is roughly normal with a spike at 0. The mean polarity is 0.087396 and the median polarity is 0.



The feature subjectivity ranges from 0, representing objective, to 1, representing subjective. The distribution of subjectivity is trimodal- with modes at roughly 0, 0.5 and 1. The mean subjectivity is 0.332452 and the median subjectivity is 0.3.



The feature favorite_count represents the number of likes for a specific tweet. The minimum favorite_count is 0 and the maximum favorite_count is 48940. The mean of the distribution is 8.59666 and the median is 0. Further, we noted that there were only 3 tweets with over 1000 likes. Originally, we wanted to explore favorite_count as well as polarity and the association with subjectivity. However, after retrieving the data, we found that favorite_count was extremely imbalanced- as per the plot. Hence, we decided to exclude that feature from our analysis.



We used TextBlob to generate features polarity and subjectivity from tweet texts. By default, TextBlob processes raw text (some steps include: splitting punctuation from words

and lemmatizing tokens) and calculates sentiments using the PatternAnalyzer method (based from the pattern library). Because TextBlob's sentiment calculator typically handles raw text as input, we did not remove punctuation, character cases, stopwords or lemmatized.

Models

We fit multiple regression models for prediction of subjectivity as a function of sentiment polarity. We then compared the models based on Root Mean Squared Error (RMSE). Note that we did not perform simple linear regression since the response variable (subjectivity) is not normally distributed. The following models were fit:

1. Linear regression with the formula:

$$subjectivity_i = \beta_0 + \beta_1 polarity_i^2 + \beta_2 polarity_i + \epsilon_i$$

   We refer to this model as the Quadratic Regression model (or "Quad") in the rest of this report.
   Quad is a supervised model which estimates coefficients for the best fit quadratic equation for the response with respect to the explanatory variable. If the data follows a quadratic pattern, then this model should give good results. However, this is also its limitation, since we are making an assumption about the relationship between these variable that may not be true. We can use R^2 and RMSE values for evaluation of this algorithm.

2. Linear regression with the formula:

$$subjectivity_i = \beta_0 + \beta_1 |polarity_i| + \epsilon_i$$

   We refer to this model as the Linear Regression Model (or "Abs") in the rest of this report.
   Abs is a supervised model which estimates coefficients for the best fit linear equation for the response with respect to the explanatory variable. If the data follows a linear pattern with respect to the absolute of the explanatory variable, then this model should give good results. However, this is also its limitation, since we are making an assumption about the relationship between these variable that may not be true. We can use R^2 and RMSE values for evaluation of this algorithm.

3. A Generalized Additive Model (GAM) with a basis spline smoother with 10 degrees of freedom. A generalized additive model is a model which estimates coefficients for a linear combination of splines that are functions of the explanatory variable. Our model is of the form

$$\hat{subjectivity_i} = \beta_0 + \beta_1 polarity_i + f(polarity_i)$$

where $f(polarity_i)$ represents the smoothing function.
GAMs can model very complex relationships since we are not limited to polynomial relationships. However, it is possible that GAM overfits the training data. Another limitation is that GAMs are solely used for prediction and are not very useful for inference. We can use RMSE for evaluation of this algorithm.

4. A KNN Regression Model (KNR) with $k = 17$. This model finds the 17 nearest polarity values (based on euclidean distance) from the training set and averages their corresponding subjectivity values to generate a prediction. We use a high k in order to reduce the chances of overfitting. As with GAMs, KNR can only be used for prediction. We can use RMSE for evaluation of this algorithm.

Results

Regression results suggest that subjectivity values are positively associated with absolute value of polarity. The quadratic regression model suggests that subjectivity grows faster for higher strengths of polarity, while Abs suggests that subjectivity grows linearly with magnitude of polarity.

	Value
Intercept	0.2324
$sentiment^2$	1.0072
$sentiment$	0.0064
R^2	0.329

Figure 1: Results for quadratic regression model

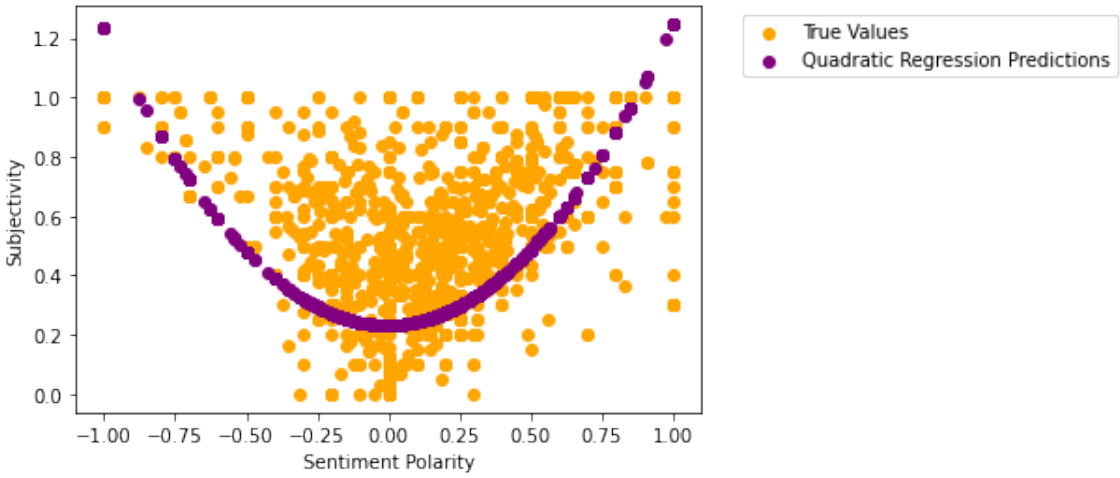


Figure 2: Quadratic regression predictions on test set

Since Abs has a higher R^2, it provides a better explanation of the trends in the training data.

	Value
Intercept	0.1471
$\lvert sentiment \rvert$	0.985
R^2	0.528

Figure 3: Results for linear regression model "Abs"

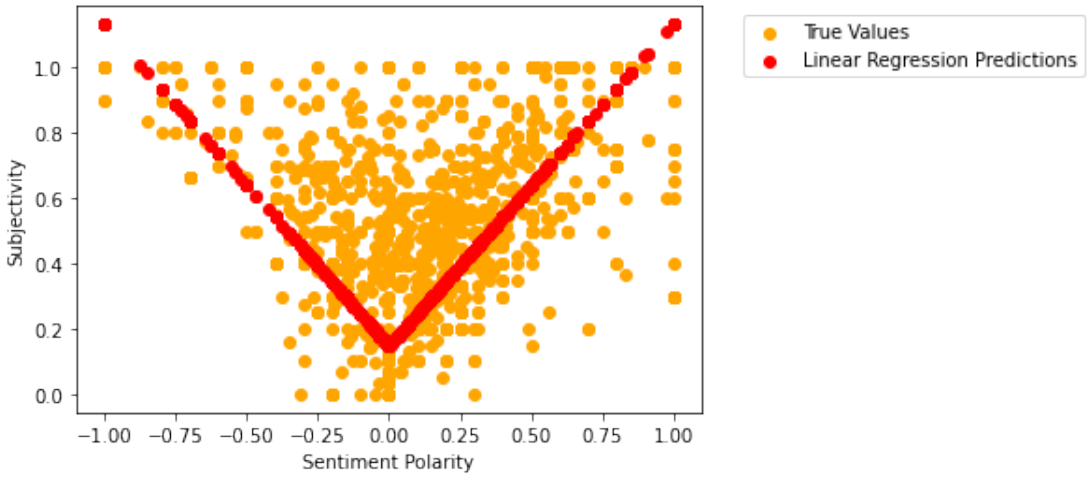


Figure 4: Linear regression predictions on test set

For prediction, we used all four models, and compared on the basis of RMSE values on the test set.

Algorithm	RMSE
Quad	0.2808
Abs	0.2343
GAM	0.1957
KNR	0.0370

Figure 5: RMSE values for prediction using our models

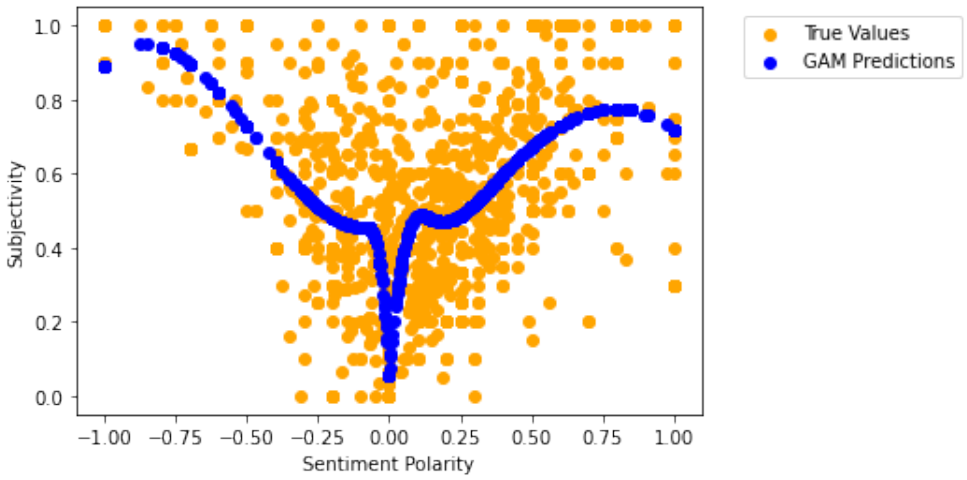


Figure 6: Generalized Additive Model predictions on test set

While the GAM performed better in prediction than the linear regression models, it was outperformed by the KNR model. This can be explained by the high variance in the data, which is better explained by a spread as is produced by KNR than a single (smooth) curve as produced by the GAM.

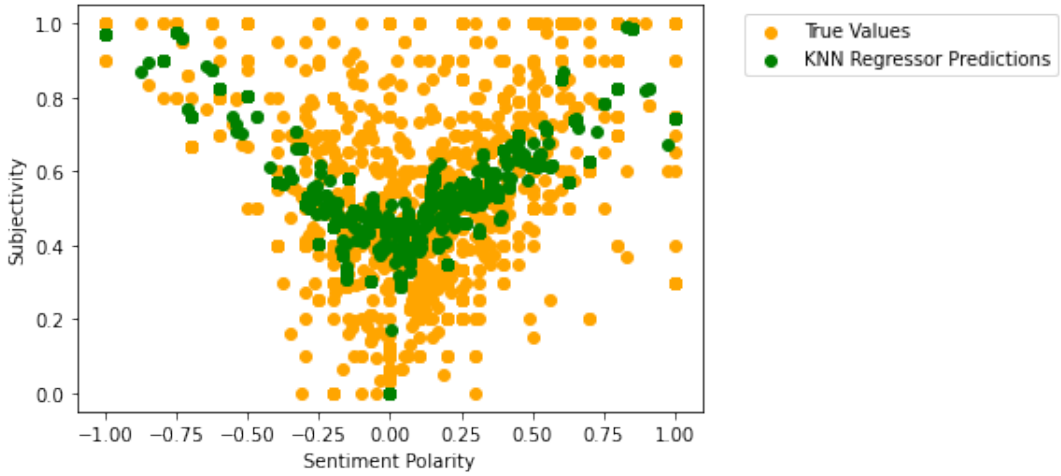


Figure 7: KNN regression predictions on test set

Conclusions

From all our models, we can see that there is a positive association between magnitude of polarity and subjectivity. Since spread in the data is high, models that estimate using smooth curves are outperformed by models like KNN since they model this spread better.