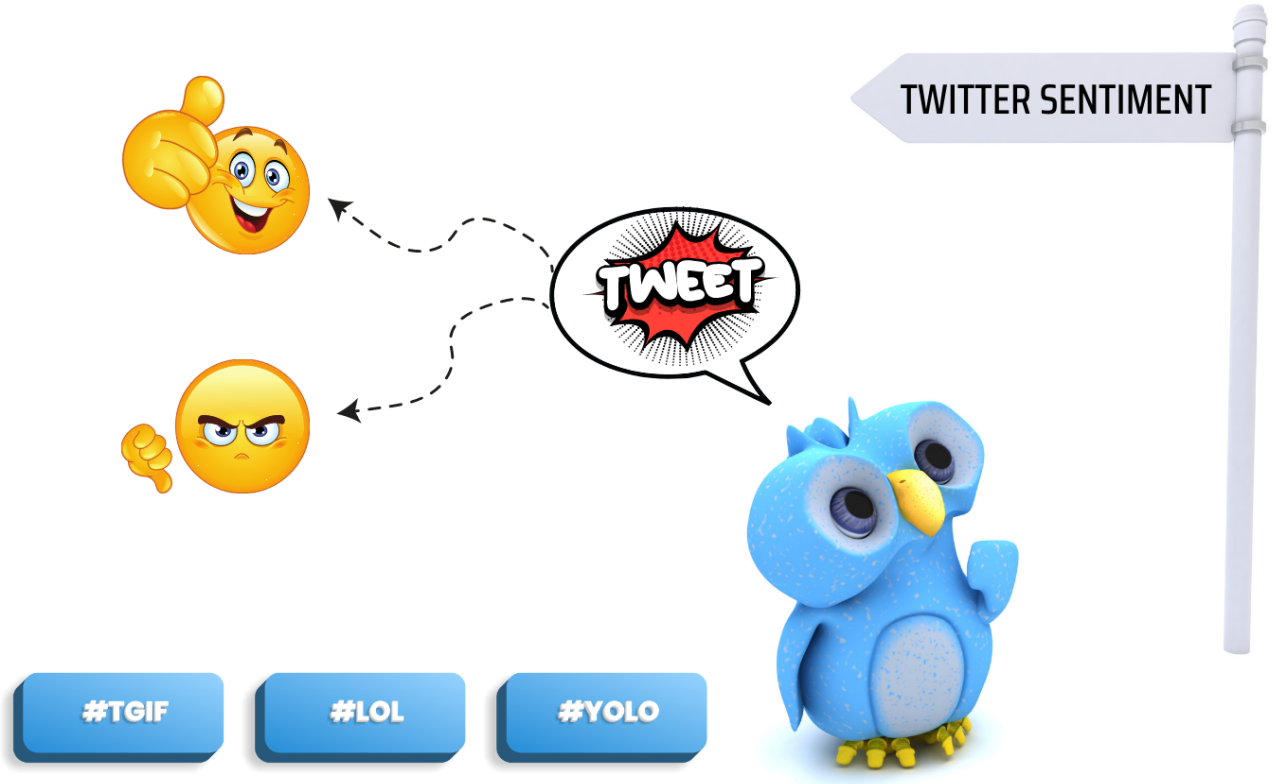
BrainStation Capstone:

Twitter Sentiment

By: Mohamed Emran

June 2023



INTRODUCTION

Every day we scroll through Twitter we might encounter negative Tweets. Twitter became a huge community in Social Media. Most of us use Twitter to look for news, sports, and business. Twitter currently has 353 Million Twitter users on its platform. At least 500 million Tweets are posted every day. Imagine how many fake accounts are being created every minute and how many offensive Tweets are posted daily.

How can we effectively identify and address negative tweets on Twitter in real time, thereby improving customer satisfaction and streamlining customer service efforts?

Capstone Objectives and Value Add

To build an effective Machine Learning model to identify any negative Tweets posted on Twitter. By analyzing user behavior to gain a deep understanding of specific topics, products, or brands. In addition, businesses can also monitor user in real-time, by responding to any issues before it escalates. This project aims to reduce negative feedback in real time in a concise manner. Addressing the negative feedback demonstrates their commitment to customer service and satisfaction.

BREAKDOWN OF DATA

Details on the Dataset:

http://help.sentiment140.com/for-students/

Citation: Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009), p.12*.

Data Collection and Classification

This is a real-world data set collected by the authors for their sentiment analysis. They have used a Twitter Search API to collect these tweets by using keyword searches. The dataset was classified as follows: 0 is a negative tweet, and 4 is a positive tweet.

DATA CLEANING AND PREPROCESSING

Starting Point

The original datasets consisted of 1.6 million rows, this is a great amount of data to work on modeling but due to computational I sample the data to 5%. In addition, to 6 features/columns.

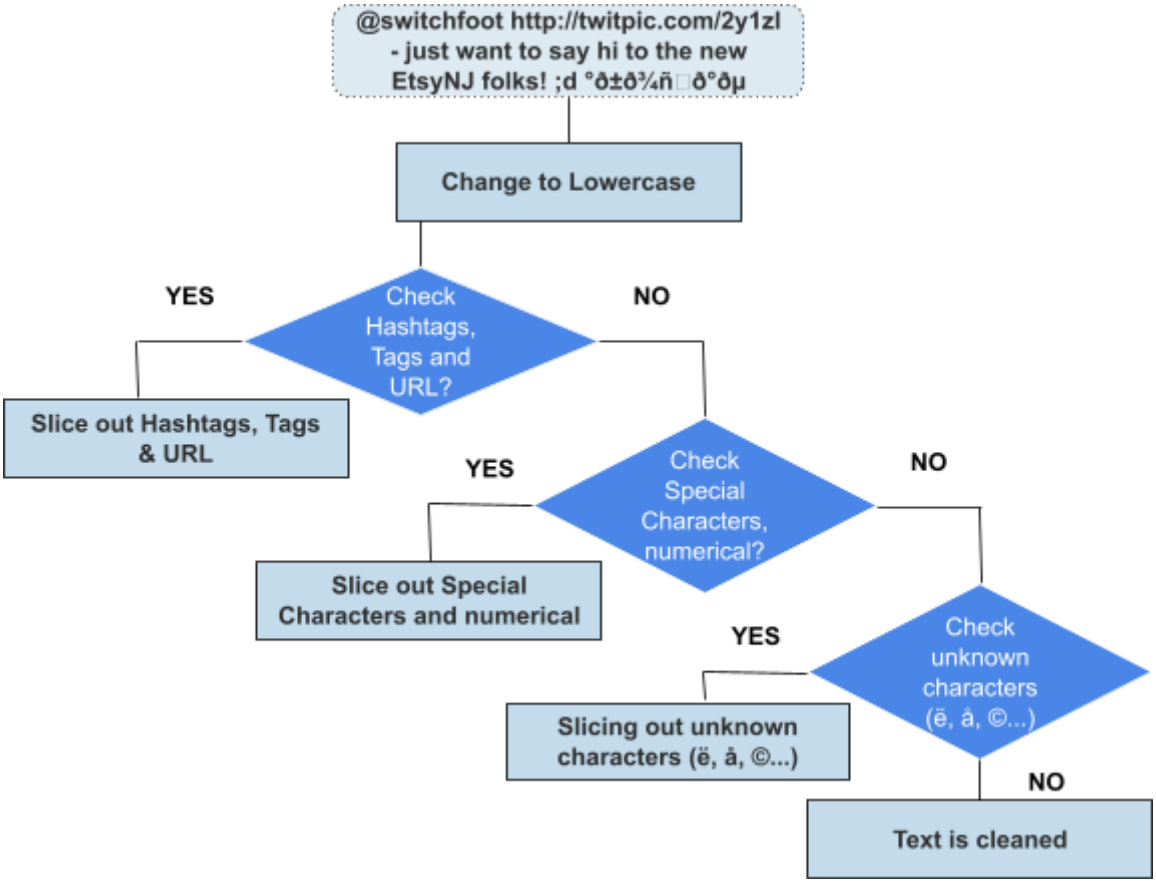
Cleaning

The dataset was checked for any duplicate and missing data. Before sampling data to 5% there were several processes to take to have the text in a plain textual format.

For example, having a text from this format:

“@switchfoot http://twitpic.com/2y1zl - just want to say hi to the new EtsyNJ folks! ;d °ð±ð¾ñ¸ð°ðµ” to an actual plain textual format

Output: “just want to say hi to the new etsynj folks”



Data Exploration

After the data was cleaned the distribution of the data was checked, and it turned out we have a normal distribution with 50% of the data as class 0 and 49.9% as class 1 as shown in Figure 0. In addition, to the distribution of the week, it shows that weekends tend to be more compared to weekdays. We can tell that people tend to spend quality time on the weekends tweeting, though negative tweets are mostly done during the weekend compared to the weekdays.

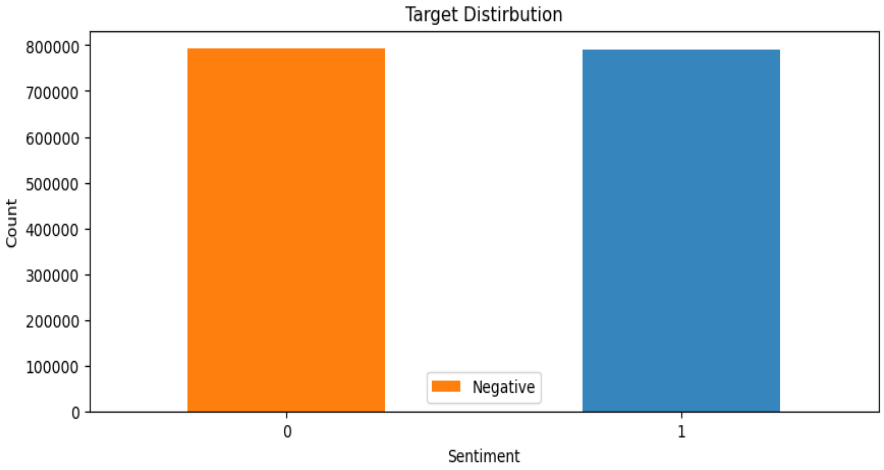


Figure 0

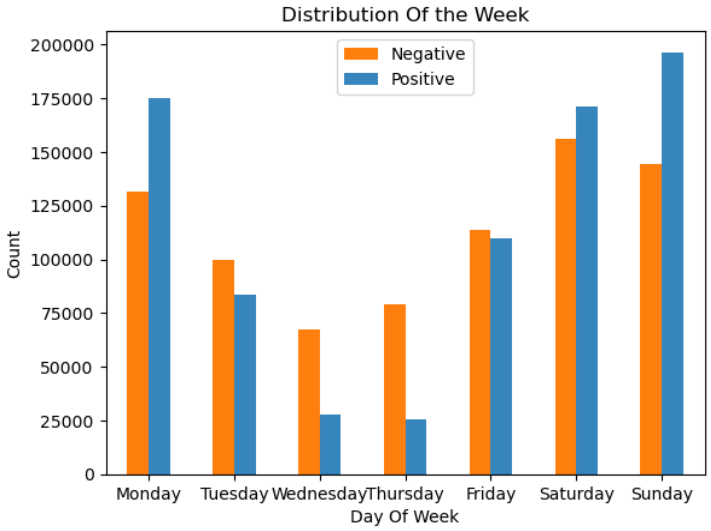


Figure 1

MODELING AND EVALUATION

Modeling Procedure:

- Data Transformation
- Vectorization (converting Tweet text data using TF-IDF into individual features)
- Scale data
- Run model
- Check model
- Manual model hyperparameter optimization
- Advanced modeling (ML pipeline and GridSearch)
- Best model evaluation

Model Used:

- Logistic Regression
- SVM (Support Vector Machine)
- Random Forest
- XGBoost

The main reason for selecting these types of models is they work well in classification

by using a supervised learning model.

Insights from Modeling:

After completing the vectorization and modeling process. The model effectively classified words, revealing that tweets with the highest indications were very positive, while those with the lowest indications tend to be negative (Figure 4). Such insights are valuable in showing the overall sentiment.

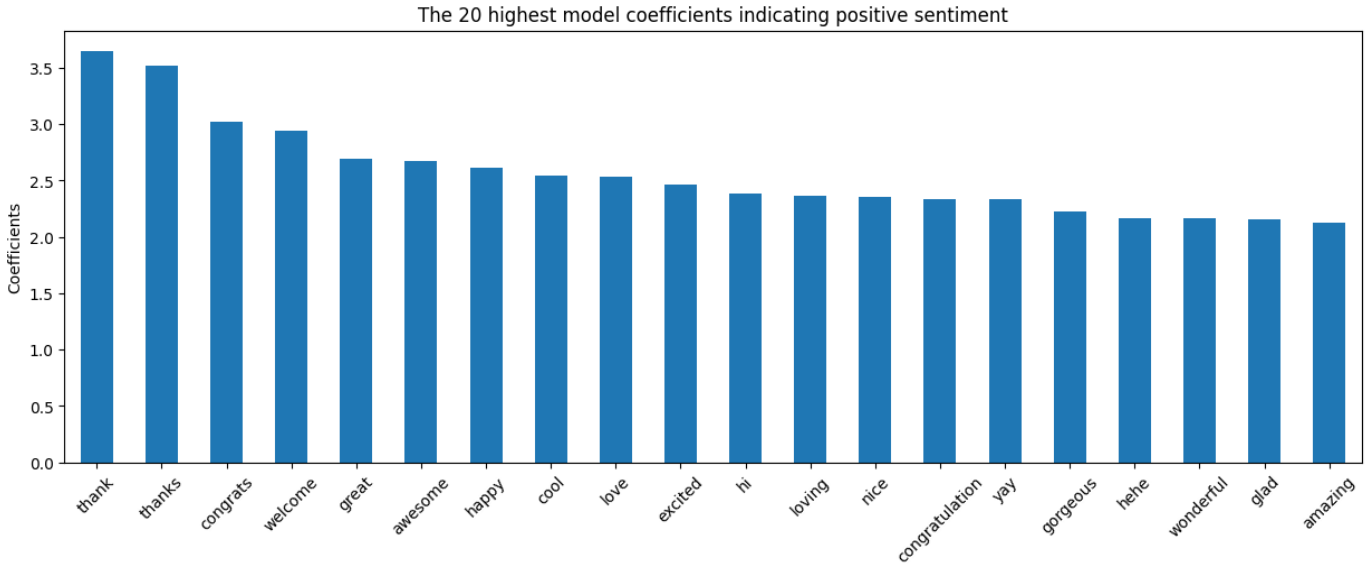


Figure 3

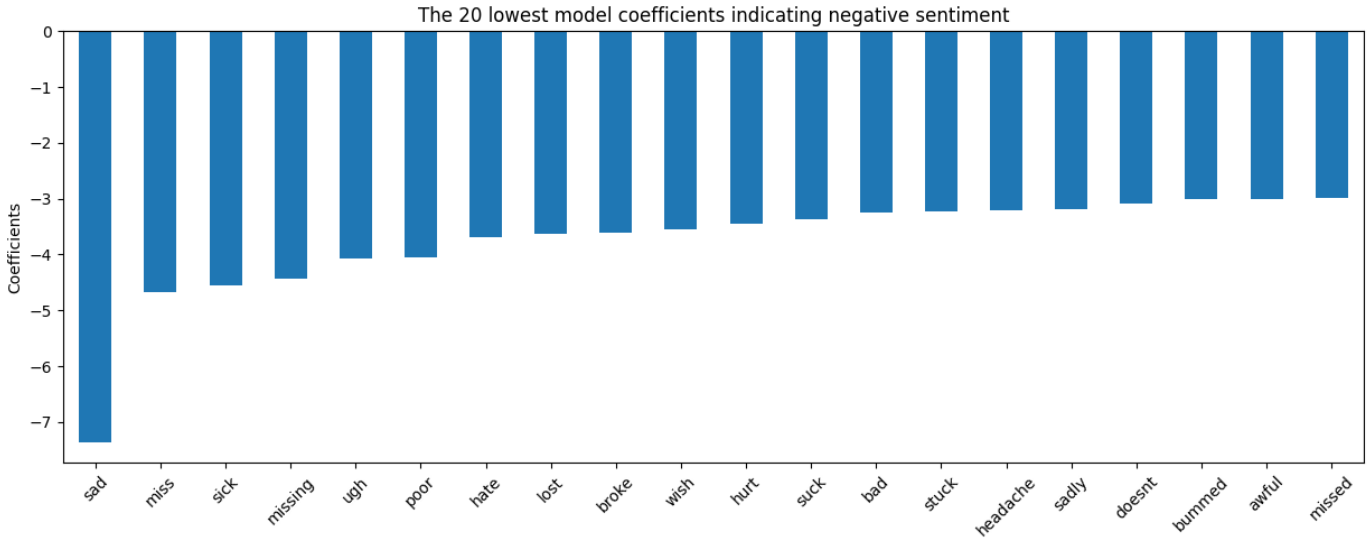


Figure 4

Final Model and Results:

After running multiple models using a machine learning pipeline with GridSearch, for optimal hyperparameters, it is evident that most of the models perform well. Surprisingly, the Baseline Logistic Regression stands out as the top performer in terms of both speed and computational efficiency.

GridSearch Model	Optimal Hyperparameters	Train Score	Validate Score	Test Score
Logistic Regression	C=1, max_iter=10000, penalty='l1', random_state=1,,solver='liblinear'	76.5%	76%	74.9%
LinearSVC (SVM)	C=0.1, max_iter=10000, random_state=1, penalty='l2'	76.6%	76.4%	74.5%
Random Forest	max_depth=11, n_estimators=200, random_state=1	70.6%	70.2%	69.2%
XGBoost	learning_rate=0.3, max_depth=9 n_estimators=200, n_jobs=None, random_state=1	81.8%	73.2%	72.2%

Baseline Model	Optimal Hyperparameters	Train Score	Validate Score	Test Score
Logistic Regression	C = 1, penalty = 'l2', solver = 'lbfgs'	77%	74.3%	74.1%

CONCLUSION

Based on all the models conducted, the Baseline Logistic Regression and the pipeline conducted on Logistic Regression and SVM, indicate that these models are the most effective classification in tweets, with test scores of 74%. Both Logistic Regression and SVM are promising models for classifying tweet sentiments. However, the Baseline Logistic Regression model is the one selected in terms of computational efficiency and business requirement.

In the next phase, we aim to enhance our model by providing a new dataset. This additional dataset will allow us to further train and fine-tune our models making them more robust and accurate in classifying positive and negative tweets.