To compare the performance of our LSTMs and DistilBERT models, we analyzed two different text representations: bag-of-words and sentiment analysis.

For the Kaggle dataset, the model was not that accurate for any time period. For instance, “january” and “private” are strongly negatively correlated to stock price in the monthly case. logistic regression performed surprisingly well on the API dataset with F1 scores of both labels above 0.65. We believe that the main difference is the shorter timespan. Since the dataset is restricted to tweets posted in a weeklong interval, current events will heavily affect both stock price and text information. It is important to note that in our analysis below, we do not train our model on 2024.

For sentiment analysis, we used a pre-trained model from Hugging Face, twitter-roberta-base-sentiment-latest. Unlike the bag-of-words case, this method was not successful on any dataset or time period. We believe this is a result of the assignment of one label to an entire tweet.

The LSTM model demonstrated accuracies of around 50% across all labeling schemes. A notable aspect of the model's behavior was its significantly high recall, particularly in the weekly dataset. The performance on the curr_tweets dataset mirrored these trends, with the weekly model showing a high recall of 80.81%. This high recall indicates the model's strong ability to identify tweets that correlate with downward movements in Tesla's stock price. While this sensitivity to negative price movements is advantageous in risk management, it also led to a high rate of false positives.

Overall, the results that we received from the DistilBERT were quite interesting, but still subpar to what we had hoped to achieve. With the quarterly model, our results were consistently the lowest across several different initialization points, with the validation and testing accuracy hovering right around ~50%. While these results were not exactly as we had hoped prior to our experimental analysis, it is interesting to see the difference between the confusion matrices of each of the different models. For example, within the weekly DistilBERT model, we see that despite having uniformly distributed labels, weekly labeling might capture the immediate impact of such events and subsequently output “0” labels. quarterly labeling might smooth out these effects, which is why it would then show a much more evenly distributed confusion matrix with close precision and recall measures, and have different impacts depending on the time frame.