Twitter Sentiments Analysis

How did I approach this problem(process)?:
Most of the process which I followed can be seen from my notebook. I will briefly talk about it here. I started off by cleaning the data, carrying out data visualizations, and data preprocessing before feeding the data to the Logistic regression. I chose the logistic regression model for twitter sentiment analysis since it's a binary classification-based problem and I also wanted to try a simple algorithm first. There are different ways to approach this problem. I started from data text processing, cleaning the raw data like removing special characters, stop words, hyperlinks, white spaces and other irrelevant information. After cleaning the raw data, I tokenized the data(splitting text into smaller chunks) and then applied a stemming method(removing and replacing suffixes from a token to obtain the base form of the word).

The second phase is converting the token to numbers and it can be done in different ways.Bag of words, TF IDF and word2vect are the techniques to convert tokens into machine readable format(e.g vectors).I used a bag of words for twitter sentiment analysis. Once whole data converts into vector form,then we can feed into different Machine learning classifiers. Here I used Logistics regression.

By looking at these values[Training Accuracy : 0.8240, Test Accuracy :0.7575, f1 score : 0.7624], the model is performing reasonably well. It's not overfitting. Though, the overall accuracy value can be improved by further optimizations which I discuss in the future work directions.

What can be done to improve the results?

- Different classifiers for Twitter Sentiments analysis(like SVMs, Decision Trees, Neural Network)
- Investigate ways to clean and improve the data quality
- Use different vectorizing techniques(e.g BERT word embedding, Word2vec)