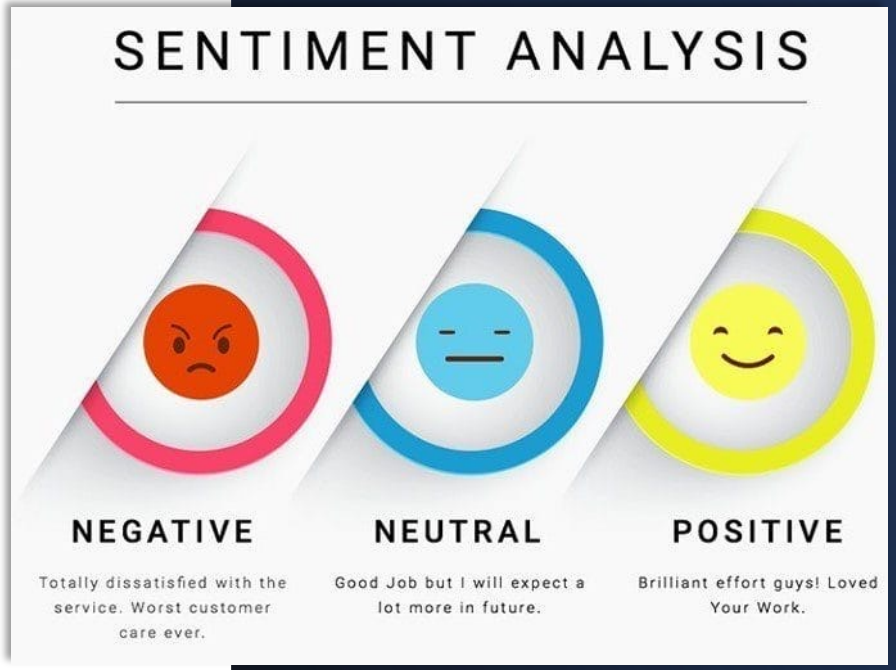
Tweeting Emotions: Decoding Sentiment in 140 Chars

Edward Wang

Fall 2024

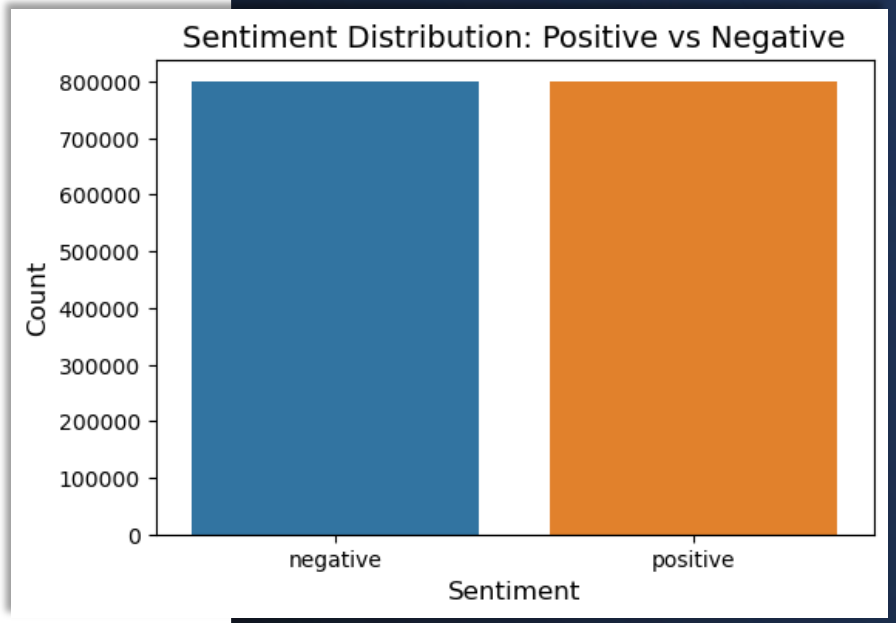
Introduction to Sentiment Analysis

- What is Sentiment Analysis?
  - Identifying the emotional tone of a text
  - Used to analyze opinions in social media, reviews, and feedback
- Importance of Sentiment Analysis
  - Customer feedback on new product release
  - Improving customer service quality
  - Business insights
    - Brand monitoring
    - Campaign Effectiveness

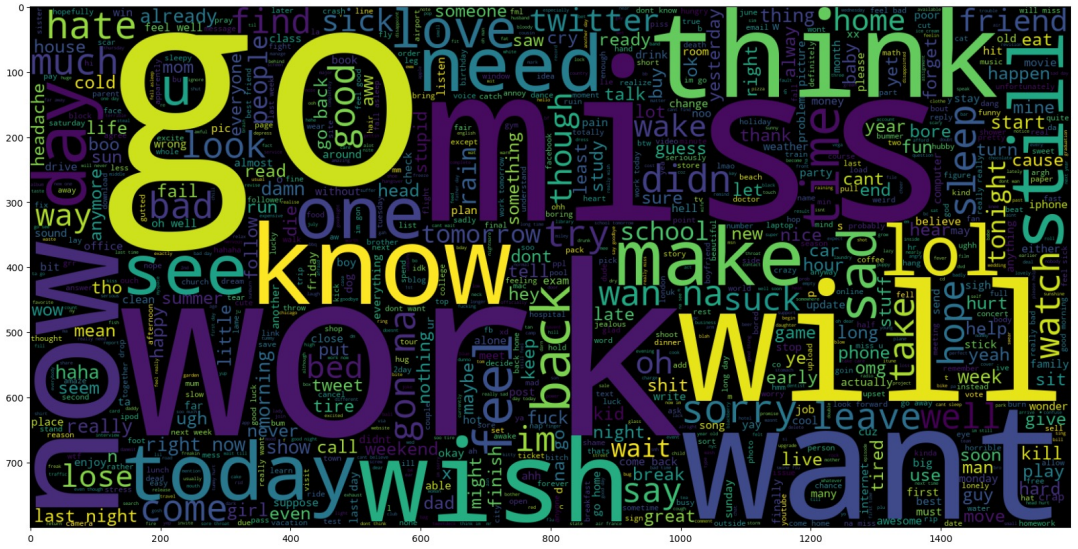


Dataset Overview

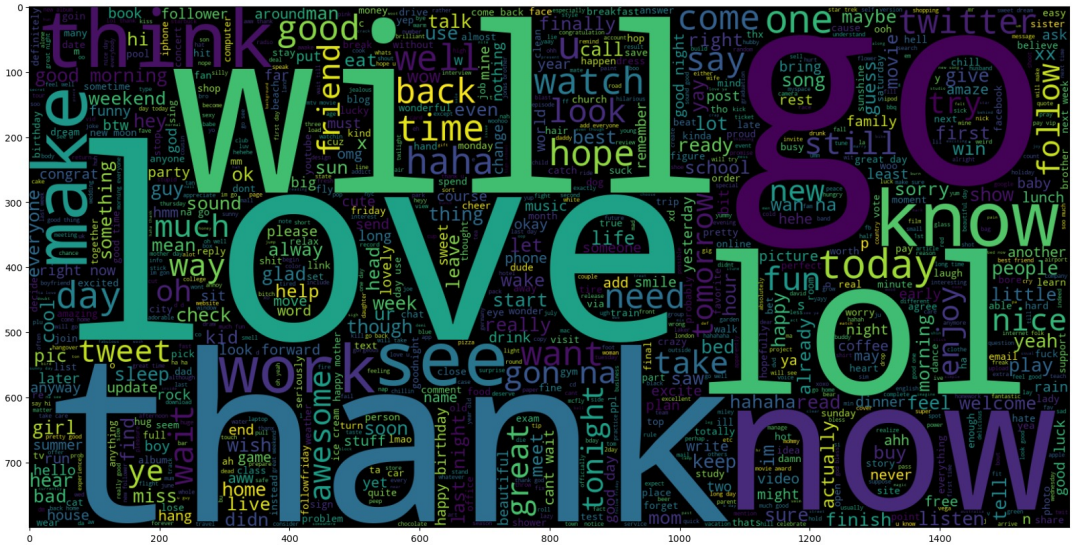
- 1.6 Million Labeled Tweets
  - Positive (4) or Negative (0) sentiment
  - No neutral Class
- 140 Characters
  - Restricted to original Twitter limit
- Balanced Classes
  - Equal number of positive and negative tweets
- Structure
  - Tweet ID
  - Sentiment Label
  - Tweet Text



Word Cloud – Negative vs Positive



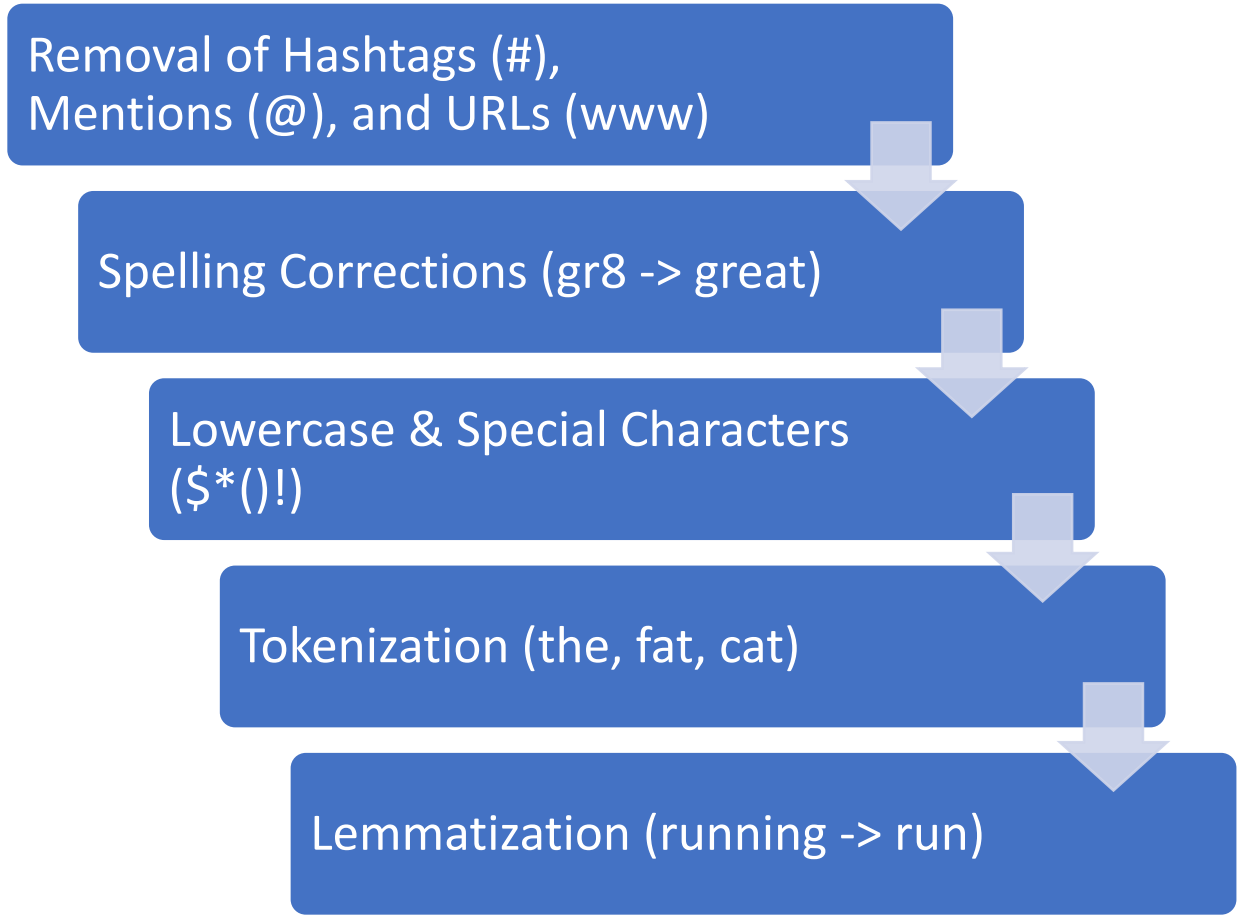
Negative Sentiment



Positive Sentiment

Text Preprocessing

- Why Preprocessing?
  - Reduce noise in data
  - Improve model accuracy
- Why not remove StopWords?
  - Preserves meaning:
    - “The product is not good” (Negative)
    - “Product good” (Positive)

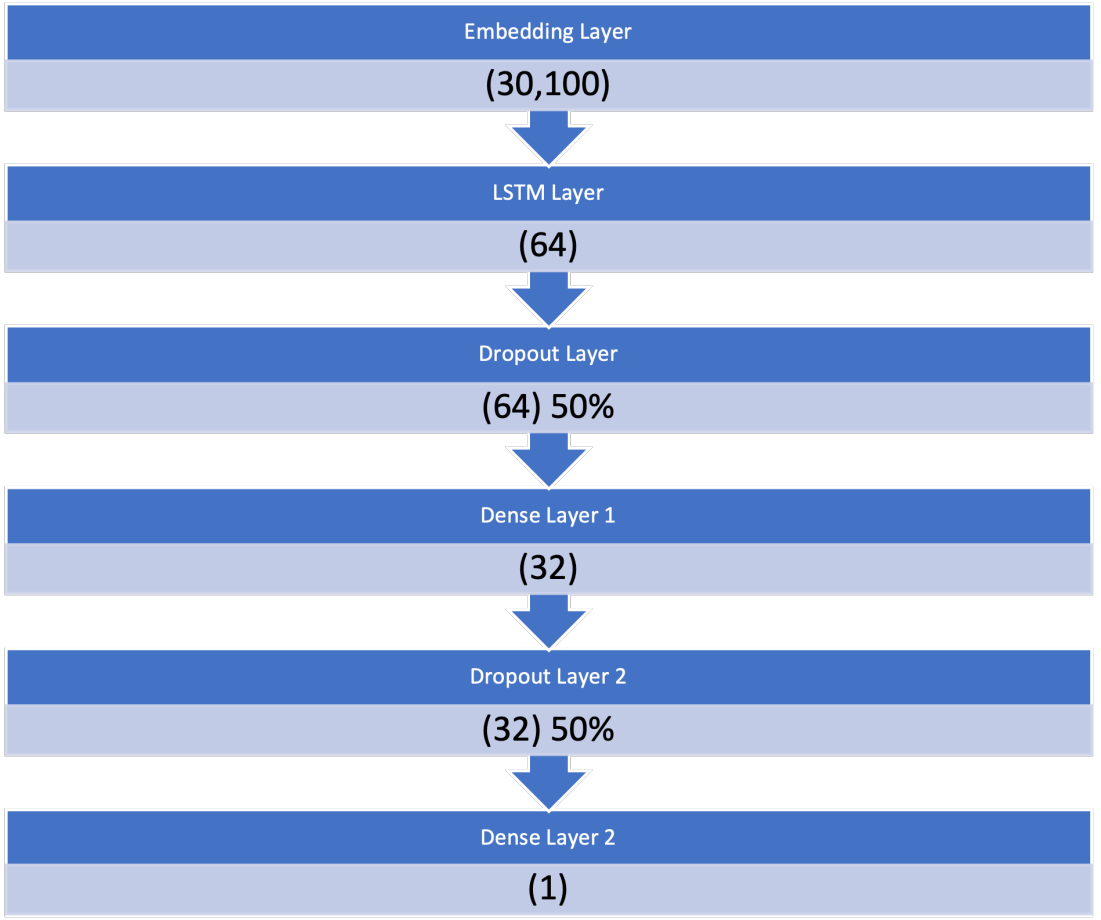


Traditional Models: Naïve Bayes & Logistic Regression

Naïve Bayes	Probabilistic model assuming independent features (words) Calculates class likelihood based on words in a tweet Fast & easy to implement
Logistic Regression	Linear model predicting class probability based on weighted features Uses TF-IDF (Term Frequency - Inverse Document Frequency) to evaluate word importance

Advanced Model: RNN with LSTM

- Recurrent Neural Network (RNN)
  - Designed for sequential data
    - e.g., text, time series
  - Remembers previous words to improve predictions
- Long Short-Term Memory (LSTM)
  - A special type of RNN
  - Solves the problem of long-term dependencies
    - Retains information over longer sequences

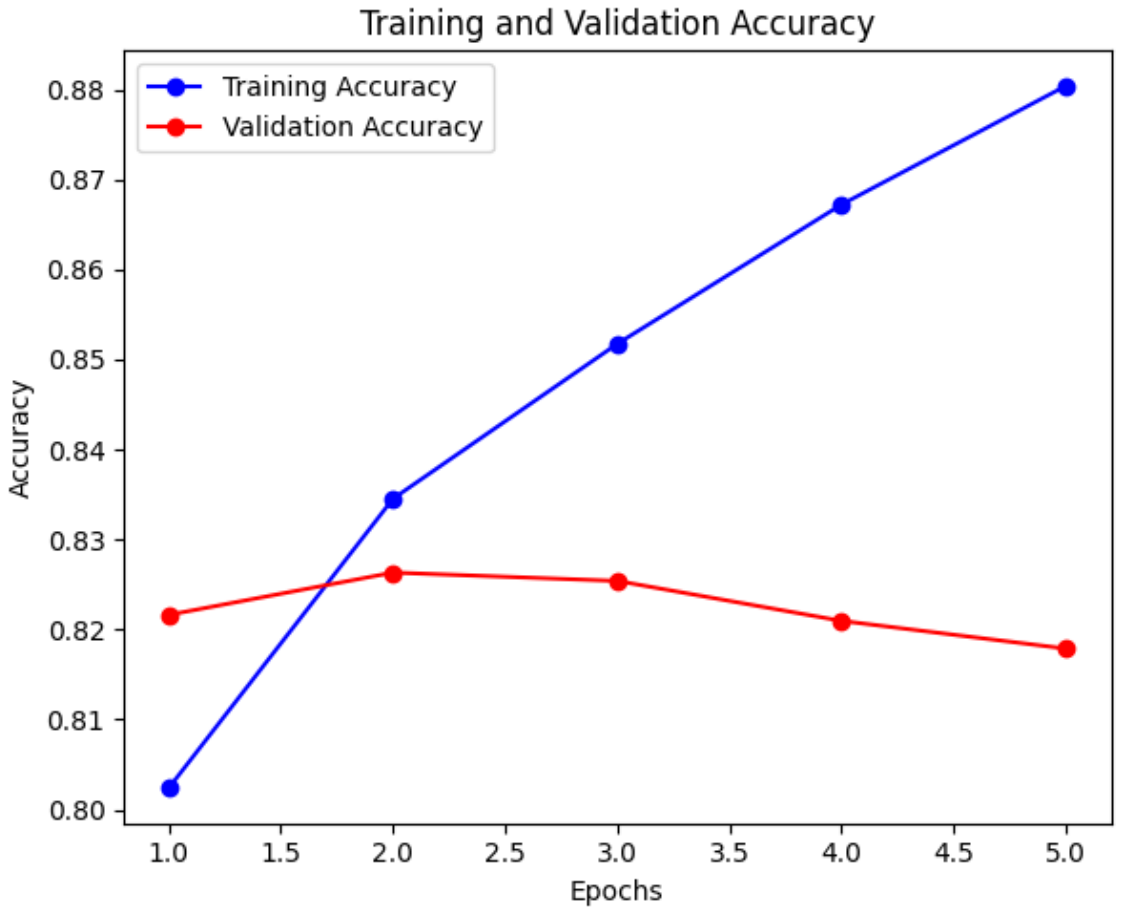


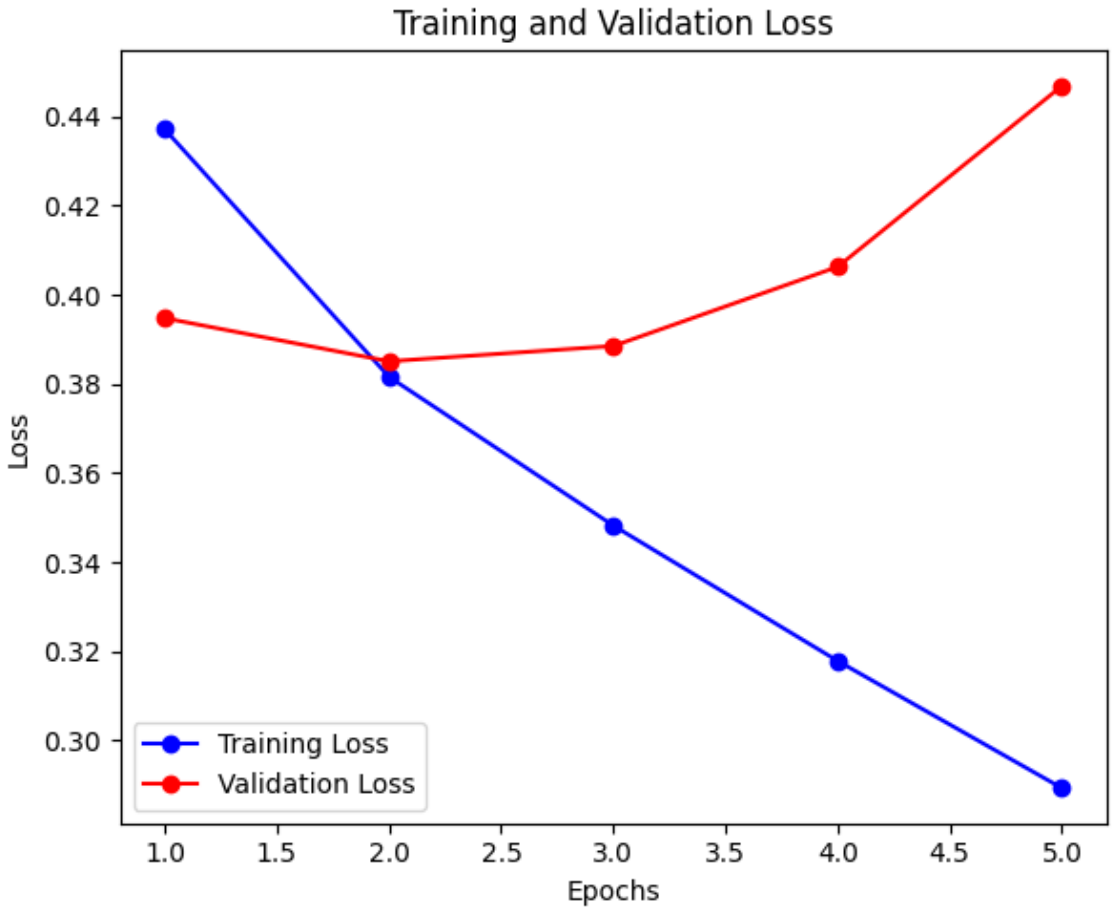
Model Performance

Model	F1 Score	Accuracy	True Pos	True Neg
Naïve Bayes	0.77	0.77	126,594	119,824
Logistic Regression	0.81	0.81	130,303	127,417
Naïve Bayes Random Search	**0.78**	0.77	126,583	119,853
Logistic Regression Random Search	0.81	0.81	130,303	127,417
RNN with LSTM	**0.82**	**0.82**	**132,752**	**130,583**

Model Performance - Cont.

Lowest Training & Validation Loss Achieved at Epoch 2





Sample Predictions

```
Tweet: I love the product, it's fantastic!
Predicted Sentiment: Positive

Tweet: Terrible experience, will not buy again.
Predicted Sentiment: Negative

Tweet: Service was average, not great.
Predicted Sentiment: Negative

Tweet: Worst purchase ever.
Predicted Sentiment: Negative

Tweet: Very satisfied with the quality.
Predicted Sentiment: Positive

Tweet: I just LOVE waiting on hold for 10 HOURS!
Predicted Sentiment: Positive
```

Challenges

Preprocessing Delays:

Took longer than expected

Sarcasm Handling

“Oh great, the Random Forest is STILL training”

Random Forest Training

Features not optimized

Long Runtime:

3 minutes per tree (100 trees, 5-fold CV)

Conclusion

Best Model: RNN

- F1 score of 0.81

Future Improvements:

- Sarcasm Detection
- Random Forest Failures
- Explore State-of-the-Art Models
  - e.g BERT
- Expand to other languages