TWITTER SENTIMENTAL ANALYSIS

BY SAYDAIN SHEIKH, TAHREEM ARIF &BRENDON WEISKOTT

AGENDA

PROBLEM STATEMENT

DATA DESCRIPTION

IMPLEMENTATION

ML MODELLING

CONCLUSION

PROBLEM STATEMENT

- The focus of our machine learning project is to develop a sentiment analysis model for tweets
- Sentiment analysis plays a crucial role in understanding public opinion and can be applied in various domains, including business, politics, and social media monitoring
- Our goal is to create a model that accurately classifies tweets into positive and negative sentiments

DATA DESCRIPTION

Context

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the Twitter API. The tweets have been annotated (0 = negative, 4 = positive) and can be used to detect sentiment.

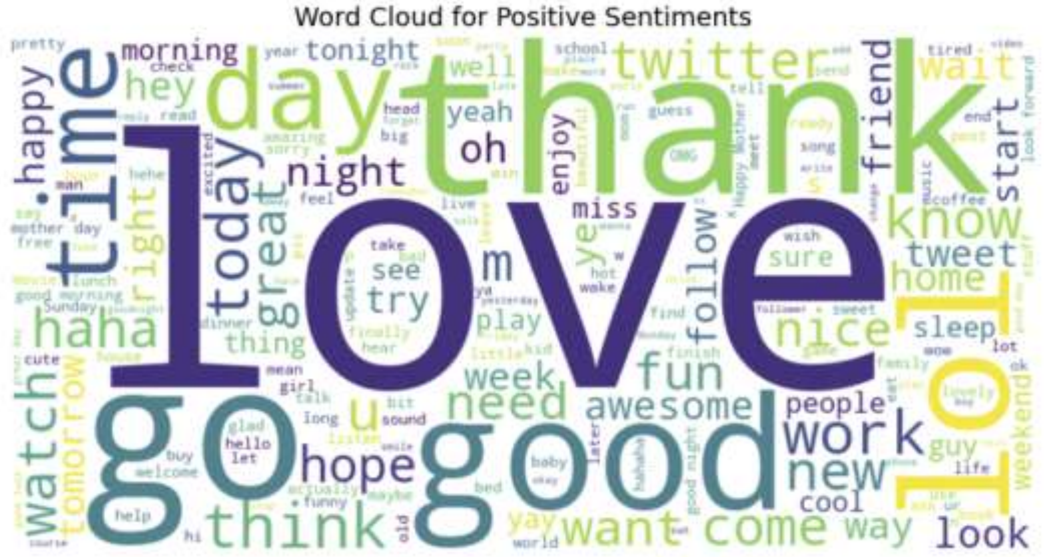
Content

It contains the following 6 fields:

- target: the polarity of the tweet (0 = negative and 4 = positive)
- ids: The id of the tweet (2087)
- date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
- flag: The query (lyx). If there is no query, then this value is NO_QUERY.
- user: the user that tweeted.
- text: the text of the tweet.

Before Preprocessing

After Preprocessing



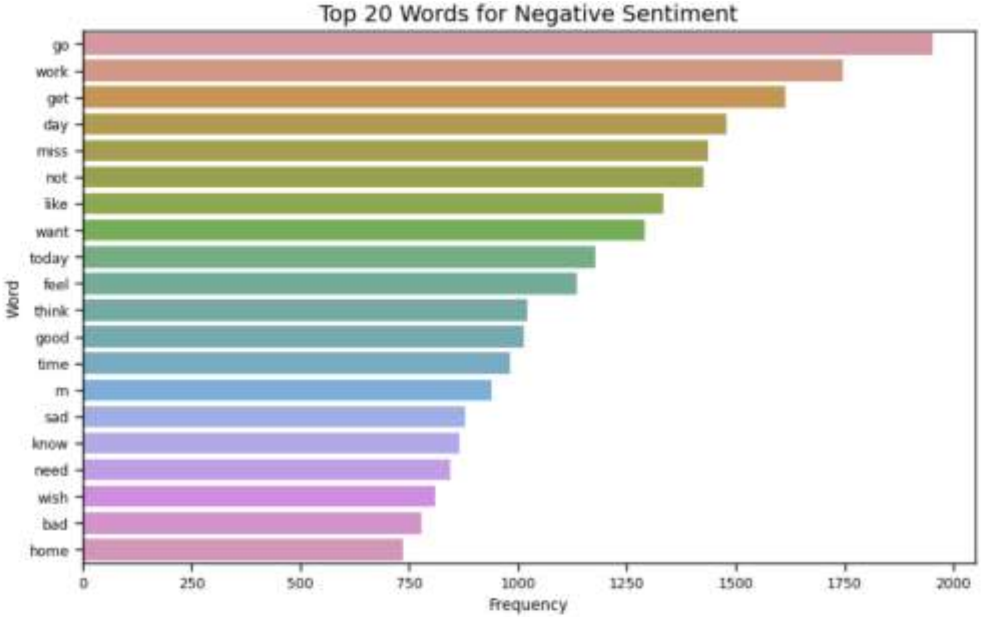
IMPLEMENTATION- DATA PREPROCESSING

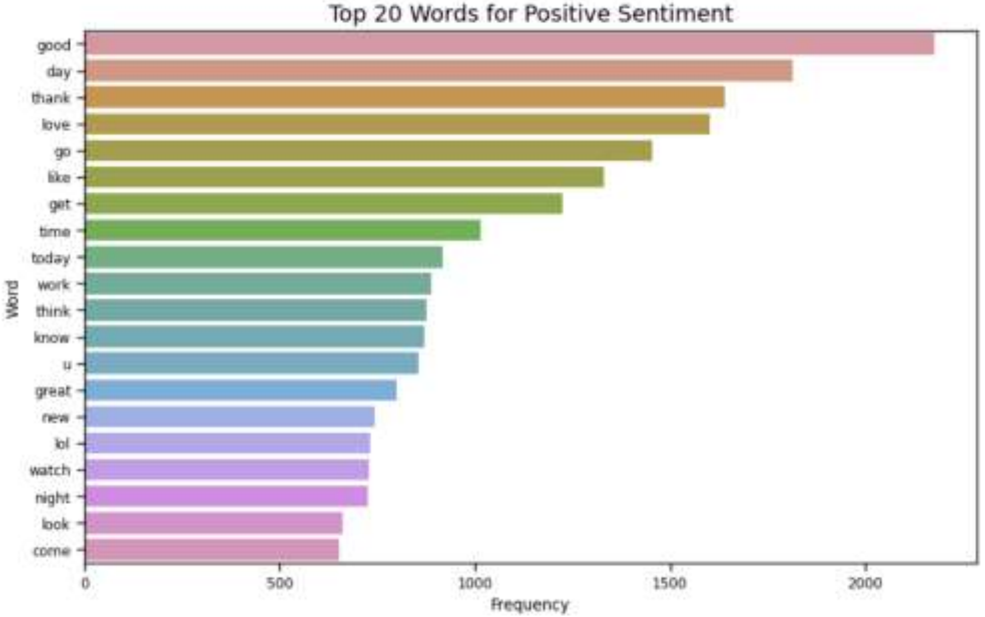
Used SpaCy library for preprocessing and perform these steps:

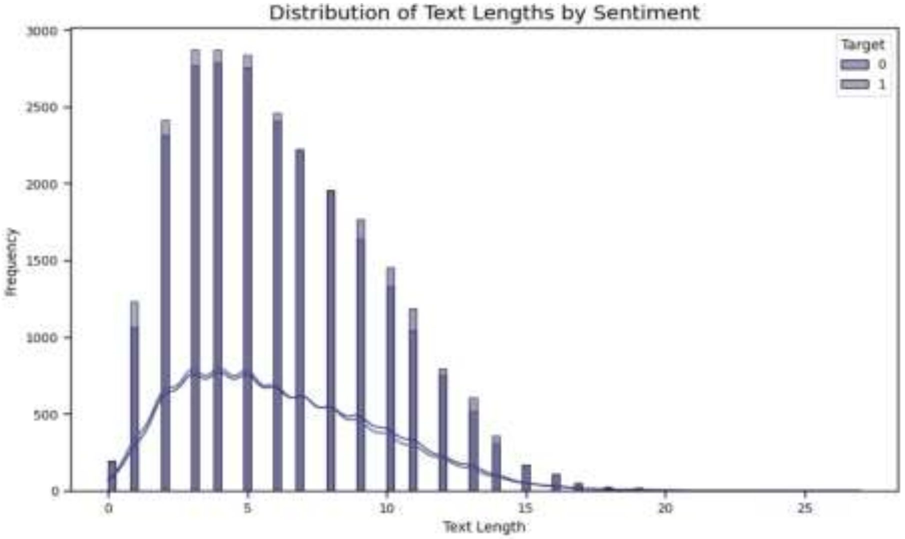
- Stop Words Removal
- URLs and @mentions removal
- Removal of HTML Character e.g. &quot

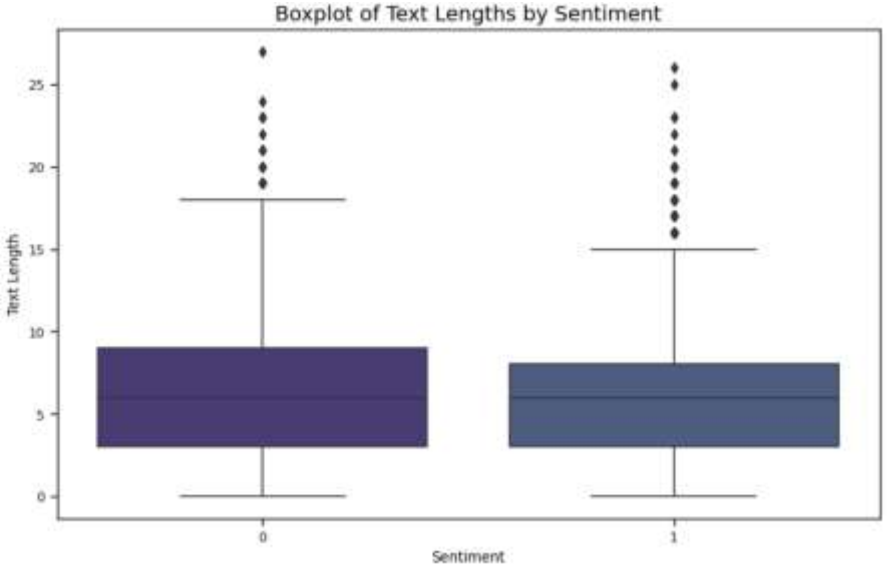
	Text	tokenized_text	Target
0	@Msfab1988 u so lucky @lamborghinibow answer ...	[u, lucky, answer, u]	0
1	@chrismusick didn't work still got em	[work, get, em]	0
2	Not doing to good....I hurt my knee last night...	[good, hurt, knee, night, dance]	0
3	@dulani247 Yep, I do.	[Yep]	1
4	@capemaybooks i see kitteh fwendz at the #bund...	[kitteh, fwendz, bunday, celeration]	1

IMPLEMENTATION-EDA

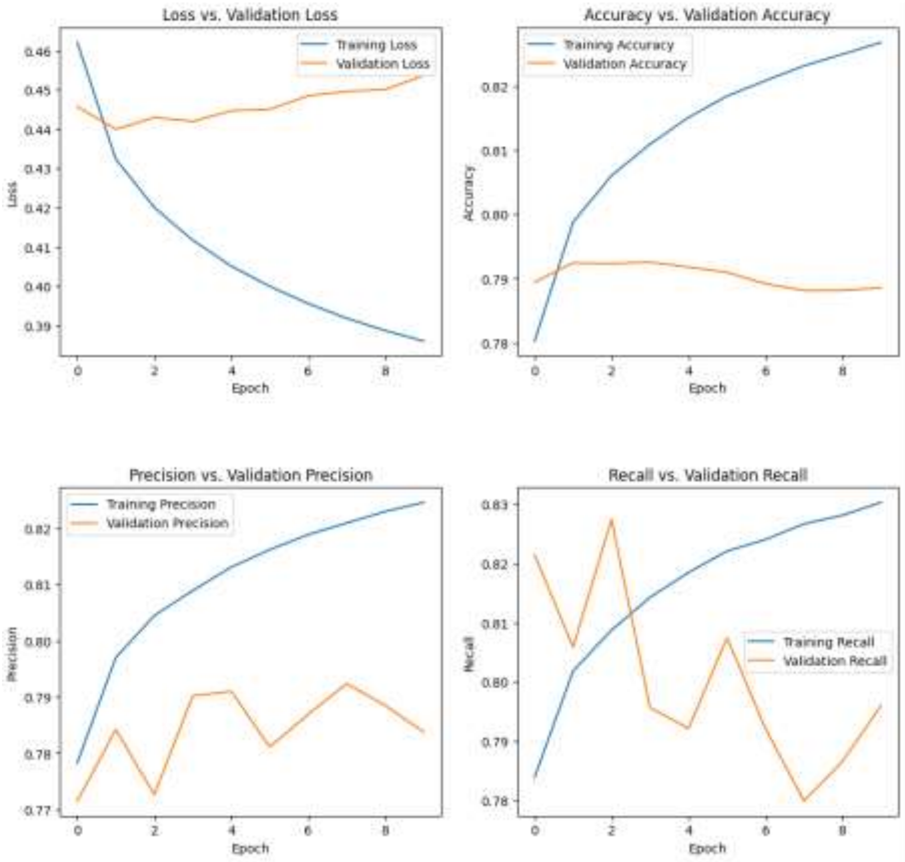






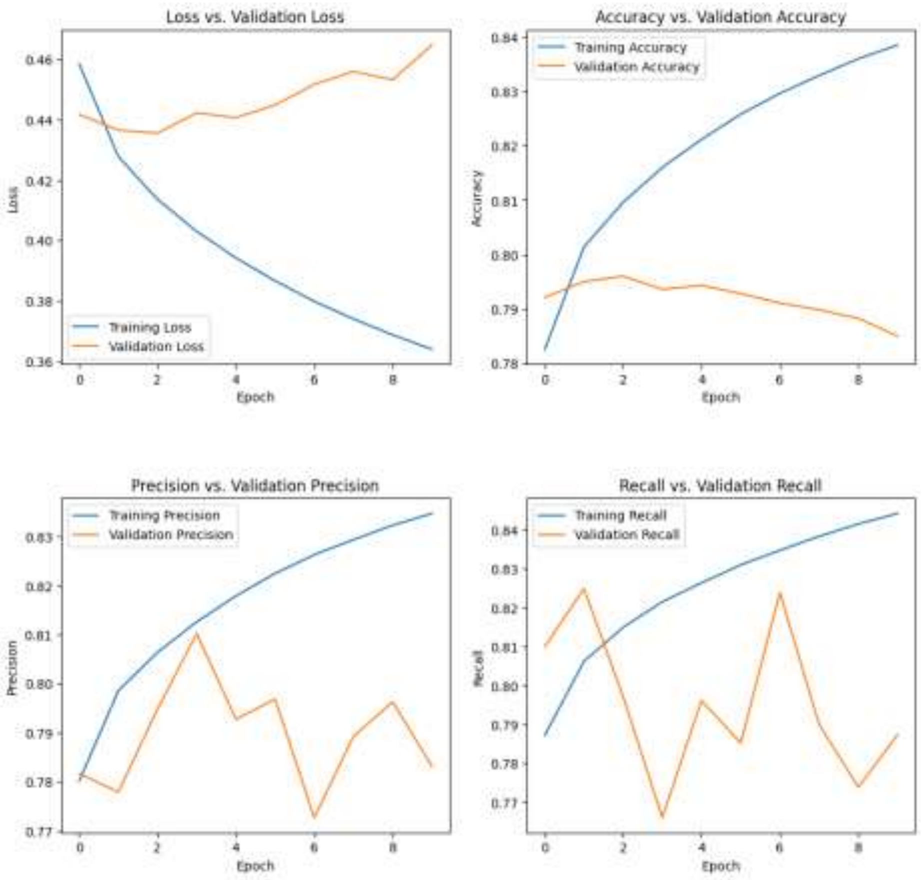


RNN



- Embedding size of 20
- RNN with 15 units
- Global Max Pooling
- Dense Layer with 32 units and RELU
- Dense Layer with 1 unit with Sigmoid for Binary Classification
- Adam Optimizer
- Binary Cross entropy Loss

LSTM



- Embedding size of 20
- LSTM with 15 units
- Global Max Pooling
- Dense Layer with 32 units and RELU
- Dense Layer with 1 unit with Sigmoid for Binary Classification
- Adam Optimizer
- Binary Cross entropy Loss

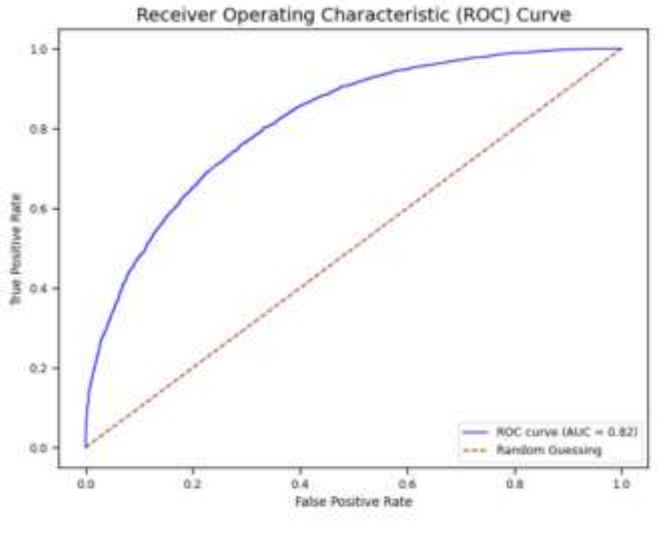


Fig. 5. ROC Curve

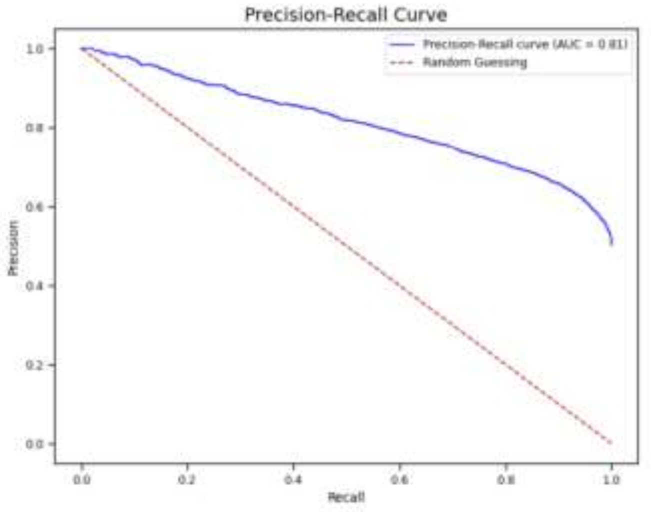


Fig. 6. Precision-Recall Curve

NAÏVE BAYES CLASSIFIER

- After Hyperparameter tuning, got alpha =10 as the best parameter for naïve bayes classifier

COMPARISON OF 3 MODELS

TABLE II
EVALUATION METRICS

Model	Accuracy	Precision	recall	F1-score
LSTM	0.79	0.78	0.80	0.79
RNN	0.79	0.78	0.79	0.79
Naive Bayes	0.74	0.77	0.69	0.73

CONCLUSION

- The Neural Network models outperformed the Naïve Bayes Model
- Given the difficulty of sentiment analysis we were able to achieve only such good results
- The large vocab and infrequency of many words made this task even harder
- Given that the results are reasonably strong

THANK YOU