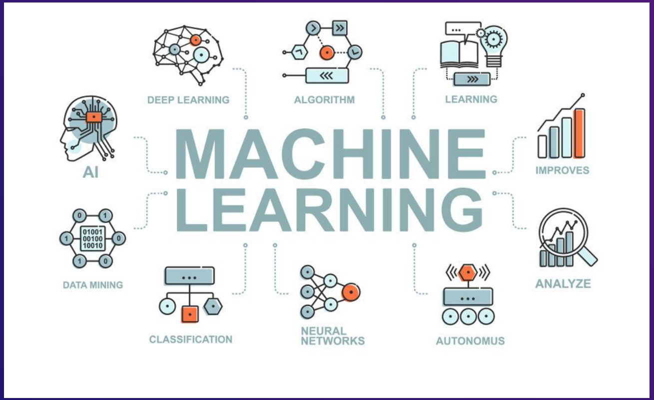
TWITTER SENTIMENT ANALYSIS PROJECT

Owen Andreasen & Justin Park

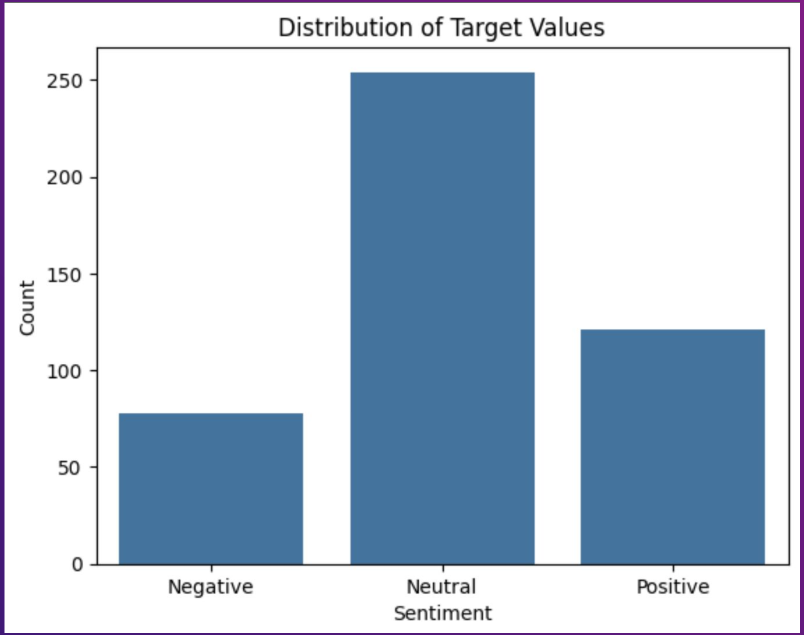
PROJECT DESCRIPTION



In this project, we aim to conduct a sentiment analysis on Twitter data. Our primary objective is to compare traditional statistical methods against modern AI/ML techniques to evaluate tweets. We plan to analyze the sentiment of tweets stored in the PostgreSQL Twitter database using different approaches and tools.

DATA SOURCE


```sql
Select distinct(status_id), text
from twitter.statuses s
where s.lang = 'en' and status_id not in (
    select status_id from twitter.statuses s2
    where text like '%https://%')
limit 1000
```



EVALUATION METRICS

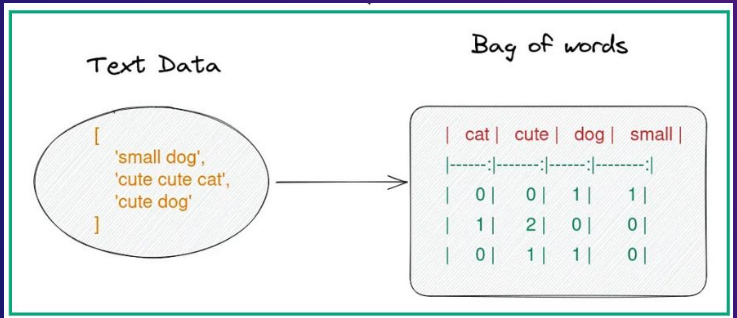
- Accuracy: The proportion of correctly classified tweets

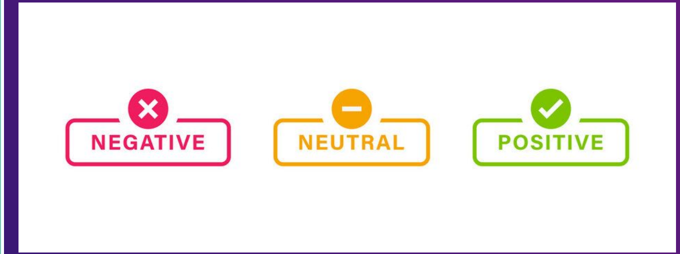
- Recall: The proportion of true positive tweets identified by the model

- Precision: The proportion of correctly classified positive tweets out of all predicted positive tweets
- F1-score: Combination of Recall and Precision

SAS

Data Import and Cleaning → Data Exploration → Text Tokenization and Vocabulary Creation → Feature Selection → Modeling → Testing





SAS RESULTS

Sentiment Metrics

Obs	sentiment	precision	recall	f_measure
1	Positive	0.56667	0.18889	0.28333
2	Negative	0.66667	0.13333	0.22222
3	Neutral	0.71948	0.95848	0.82196

Obs	_FREQ_	accuracy	total_count
1	424	70.75%	300

SAS RESULT WITH STOP WORDS

	sentiment	precision	recall	f_measure
1	Positive	1	0.4777777778	0.6466165414
2	Negative	0.7954545455	0.7777777778	0.7865168539
3	Neutral	0.8427299703	0.9826989619	0.9073482428

Obs	_FREQ_	accuracy	total_count
1	424	85.38%	362

MACHINE LEARNING MODELS

Read in Data

Preprocessing

- Emoji conversion
- Cleaning & Filtering
- Oversampling

Naive Bayes

- Bayes' Theorem
- Assumes feature independence
- Handles large data well
- Good starting point for baseline

78% Accuracy

BERT Transformer

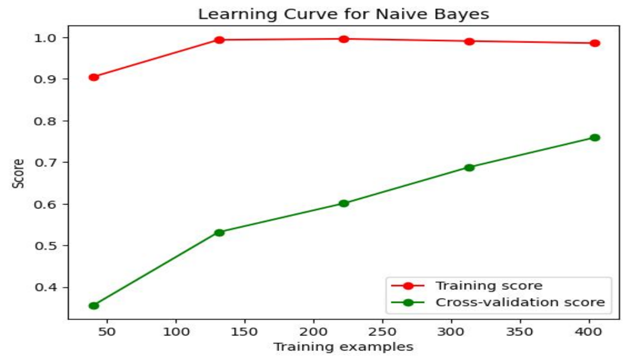
- Bidirectional Encoder Representations from Transformers
- Developed by Google
- Pre-trained on large amounts of text
- Reads text in both directions

XGBoost

- Extreme Gradient Boosting
- Designed for speed and performance
- Builds multiple models sequentially
- New models correct errors of old models

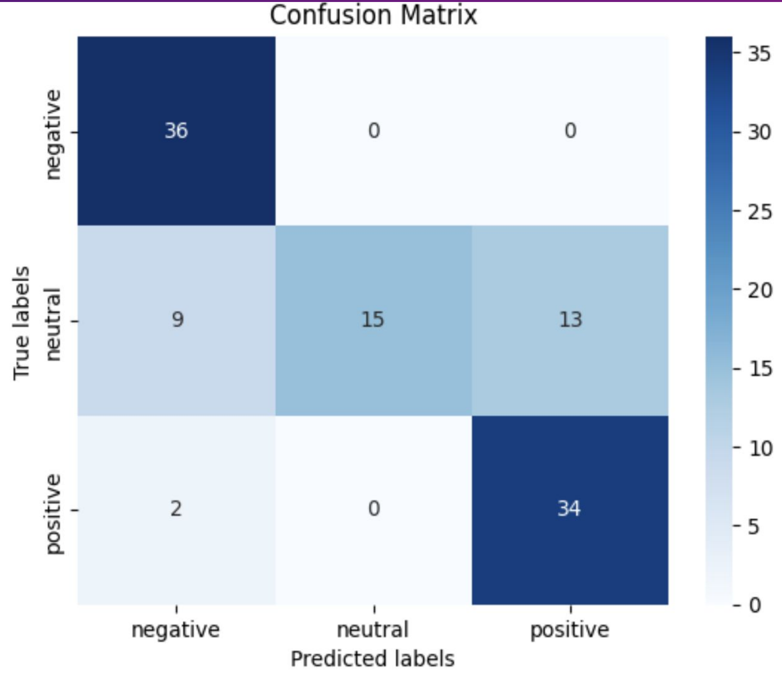
92% Accuracy

NAIVE BAYES RESULTS

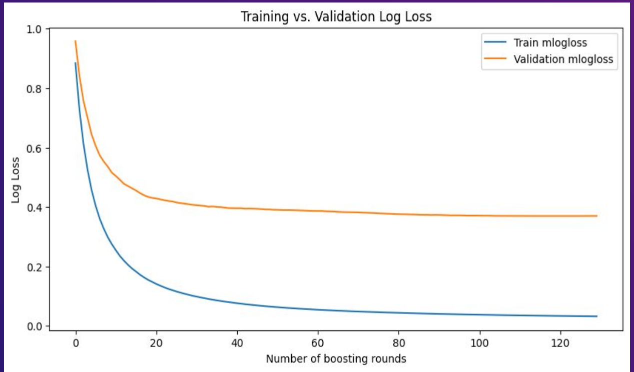


Classification Report for Naive Bayes:

	precision	recall	f1-score	support
negative	0.77	1.00	0.87	36
neutral	1.00	0.41	0.58	37
positive	0.72	0.94	0.82	36
accuracy			0.78	109
macro avg	0.83	0.78	0.75	109
weighted avg	0.83	0.78	0.75	109

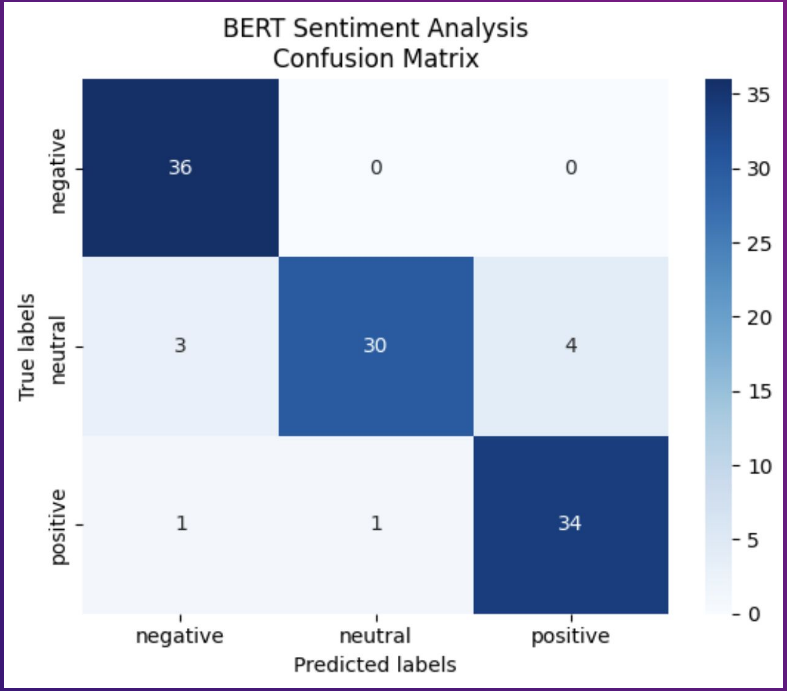


BERT+XGBOOST RESULTS



Classification Report for BERT:

	precision	recall	f1-score	support
Negative	0.90	1.00	0.95	36
Neutral	0.97	0.81	0.88	37
Positive	0.89	0.94	0.92	36
accuracy			0.92	109
macro avg	0.92	0.92	0.92	109
weighted avg	0.92	0.92	0.92	109



COMPARISON & DISCUSSION

"Muslim woman in Hijab makes valid points and audience members tell her to calm down Funny that Muslim woman who doesnt…"

NEGATIVE

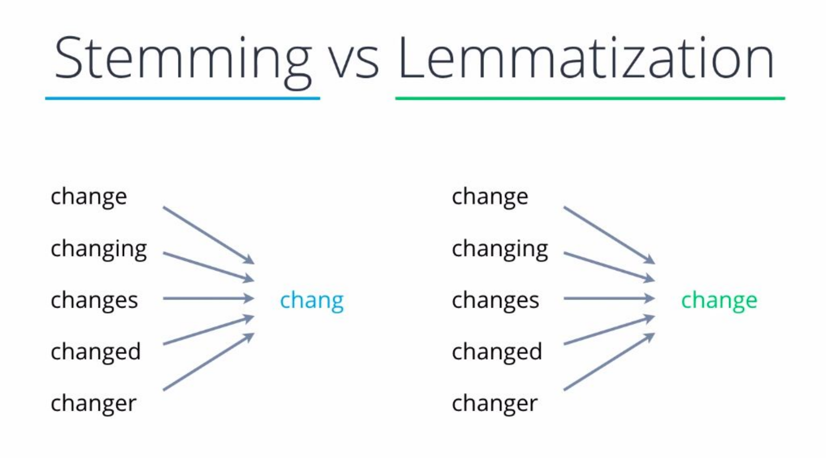
"Please trash me on Wikipedia, I'm begging you"

NEGATIVE

"Just spilled my coffee all over my white shirt. Great start to the day! Can't wait to see what other surprises are in store for me. "

NEGATIVE

IMPROVEMENTS



THANKS!

QUESTIONS?