Twitter Sentiment Analysis of Public Reaction to COVID-19 News

Prepared and Presented by Alexandra Yakovleva

Project Overview

Motivation:

By leveraging Natural Language Processing (NLP) and sentiment analysis models, we aim to gain insights into online conversations surrounding the pandemic.

Project Overview

Potential Stakeholders:

Public Health Officials, Media Outlets, Government Agencies, Researchers.

Business Problem and Objectives

***Problem Statement*:**

Potential Stakeholders need **better understanding** of how their announcements influence public sentiment on Social Media.

Business Problem and Objectives

Key Questions:

- What is the sentiment?
- Can we identify features related to specific sentiment?

Business Problem and Objectives

Project Objectives:

- Develop NLP pipeline for Social Media data analysis
- Deploy robust sentiment classification model
- Visualize and interpret results

Data Acquisition and Preparation

Covid-19 Twitter Dataset: A large collection of COVID-19-related tweets from Kaggle.

GloVe Embeddings: Pre-trained word embeddings from Stanford NLP Group.

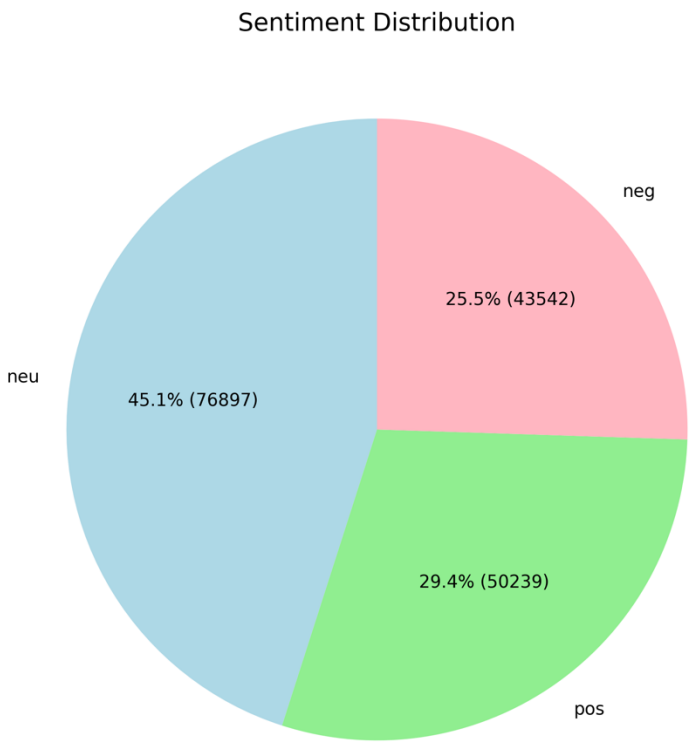
Data Limitations: Paywalled Access to Twitter Data

If you need higher levels of access, sign up for the Enterprise API today! Start here.

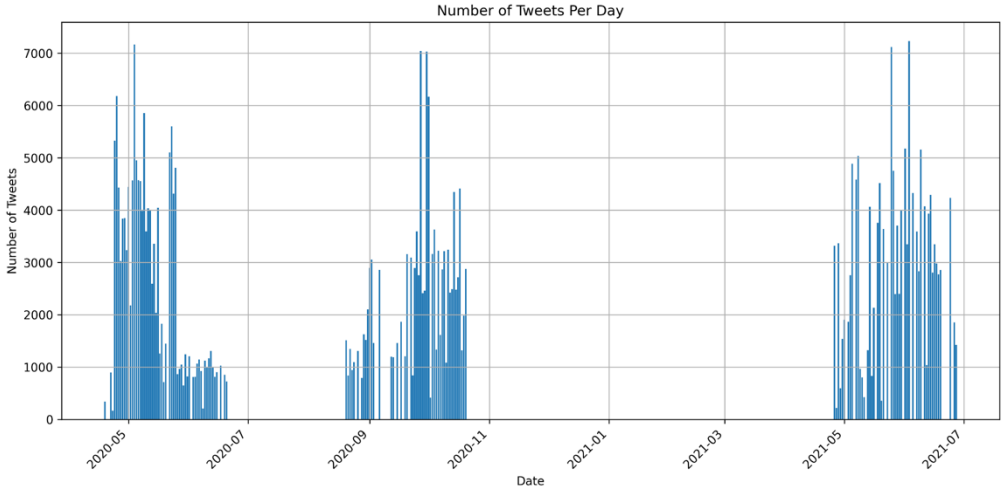
	Pro	Basic	Free
	Unleash the full potential of X's v2 API	Begin your journey with X's v2 API	Get limited access to X's v2 API
Apps	3 environments	2 environments	1 environment
DMs	Over 300K requests per month per user	75K requests per month per user	
Posts	Retrieve up to 1M Posts per month	Retrieve up to 15K Posts per month	Retrieve up to 100 Posts and 500 writes per month
Realtime Posts stream	Access to realtime posts stream	No access	
Users	Over 8M requests per month per user	50k requests per month per user	
	SAVE 10%	SAVE 12.5% - LIMITED TIME	
	$4500/month	$175/month	
	$54000 billed annually	$2100 billed annually	
	Upgrade now	Upgrade now	
	$5000/month	$200/month	
	$5000 billed monthly	$200 billed monthly	
	Upgrade in 1-Click	Upgrade in 1-Click	
			You have access to this plan
			Downgrade to free access?
			Downgrade

Data Limitations: Sentiment Labels

- Imbalanced Distribution.
- Unverified Labels.
- Unknown Labeling Algorithm.



Data Limitations: Date Ranges



Data Cleaning and EDA

Date: Temporal trends of tweet data are extracted.

Language: Relevant language is subset to increase relevant tweet ratio.

Location: Location data is standardized and processed for geocoding.

Source: Platforms (e.g., Twitter for Android) are identified.

Sentiment: Sentiment labels are explored and distributions analyzed.

Social Connections: Mentions and retweets are analyzed for network insights.

Text Preprocessing and Feature Engineering

Cleaning: *another day in paradise grinning face with big eyes*

Preprocessing: *another day paradise grinning face big eye*

Feature Extraction: *'another day', 'another day paradise', 'grinning face'*

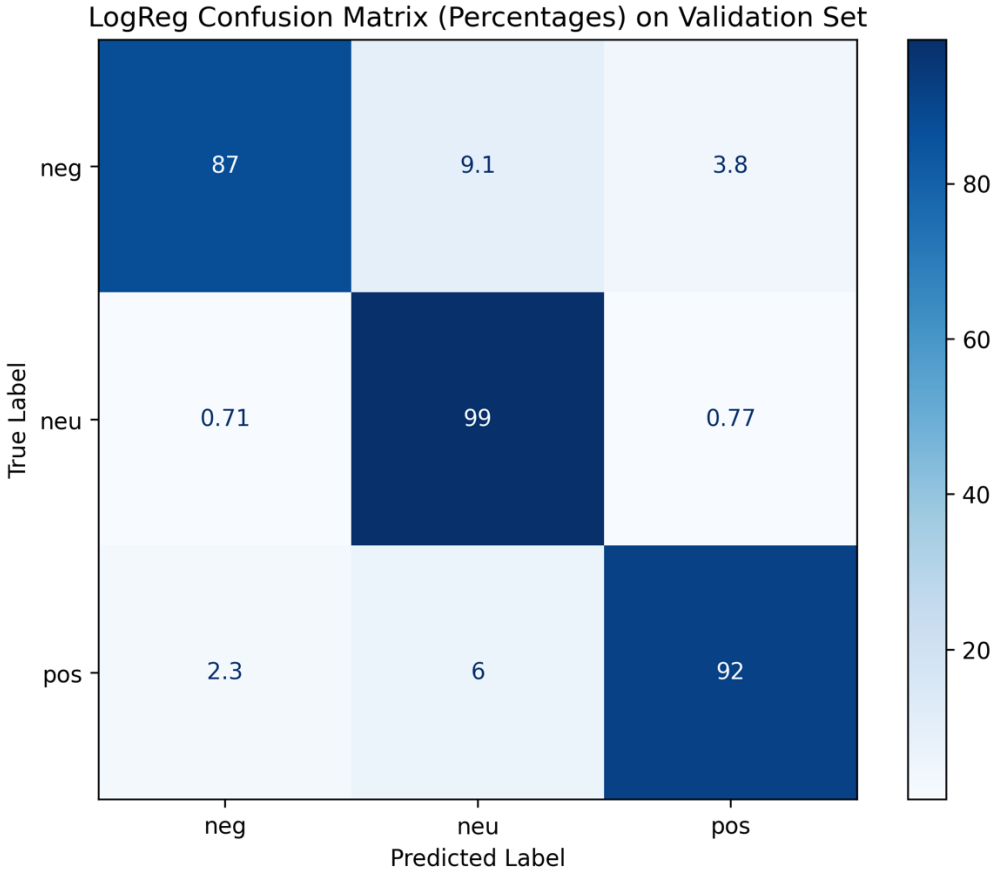
Sentiment Analysis with Supervised Machine Learning

Methods:

- **Features:** X (170679 x 10200)
- **Target:** y (sentiment categories)
- **Data Split**: Train (70%) – Test (15%) – Validate (15%)
- **Sentiment Classification:** Logistic Regression
- **Performance Metric:** balance of precision and recall.

Results and Visualization: Performance

Weighted Accuracy 94%



Results and Visualization: Predictive Features

Positive	*Great, Best, Free*
Neutral	*Number, Support, Death*
Negative	*Death, Ban, Dangerous*

Conclusion

- **Summary:** *Developed a system to classify sentiment and extract informative features.*
- **Strengths:** *Model performs well on new data and can provide actionable insights for the stakeholders.*
- **Limitations:** *Model is limited by a specific sentiment classification. Model covers extensive time-period and is not specific to a particular stakeholder.*
- **Future Work:** *Implement advanced sentiment classification and feature engineering. Re-train model on more specific subset of data.*