CIS 600: Principles of Social Media & Data Mining

SMDM PROJECT PRESENTATION

Reddit WallStreetBets Sentiment Analysis Data

SYRACUSE UNIVERSITY · FOUNDED 1870 ·
SUOS CULTORES SCIENTIA CORONAT

Team Members:

1. Sameer Balkawade	SUID: 200774310
2. Shivani Neharkar	SUID: 320953757
3. Ryan D’mello	SUID: 689654361
4. Mitil Jawale	SUID: 754778650
5. Atharvaa Rane	SUID: 275976322

Project Overview

This endeavor entails a thorough investigation into the Reddit community r/wallstreetbets, a notable online hub for stock trading discussions. The primary goals encompass:

Sentiment Analysis: Utilizing natural language processing techniques to evaluate sentiment within r/wallstreetbets posts and comments. This analysis aims to reveal the community's overall market sentiment, influencing stock prices and trading strategies.

Influential Stocks: Identifying stocks frequently discussed in the subreddit to pinpoint equities of interest. Tracking changes in mentions over time will uncover emerging trends and illustrate the impact of community sentiment on specific stocks.

Trading Behaviors: Quantifying and exploring the risk tolerance, investment horizons, and preferred strategies discussed by users on r/wallstreetbets. These insights hold value for both retail and institutional investors.

Market Impact: Evaluating how discussions and recommendations on r/wallstreetbets align with stock price movements, trading volumes, and options activity. This analysis aims to uncover the community's potential influence on broader financial markets.

Data Visualization: Employing visualization tools to generate interactive charts and graphs for effectively conveying project findings. The goal is to present intricate data in an accessible and engaging manner.

About the Data

This data is used to understand the trends initiated by Reddit educated crowd. WallStreetBets (r/wallstreetbets, also known as WSB), is a subreddit where participants discuss stock and option trading,The dataset contains posts from Reddit starting from 29th January 2021. Be aware that data don't start before GME stock hype. Moreover, this dataset is regularly updated and total number of rows increases.

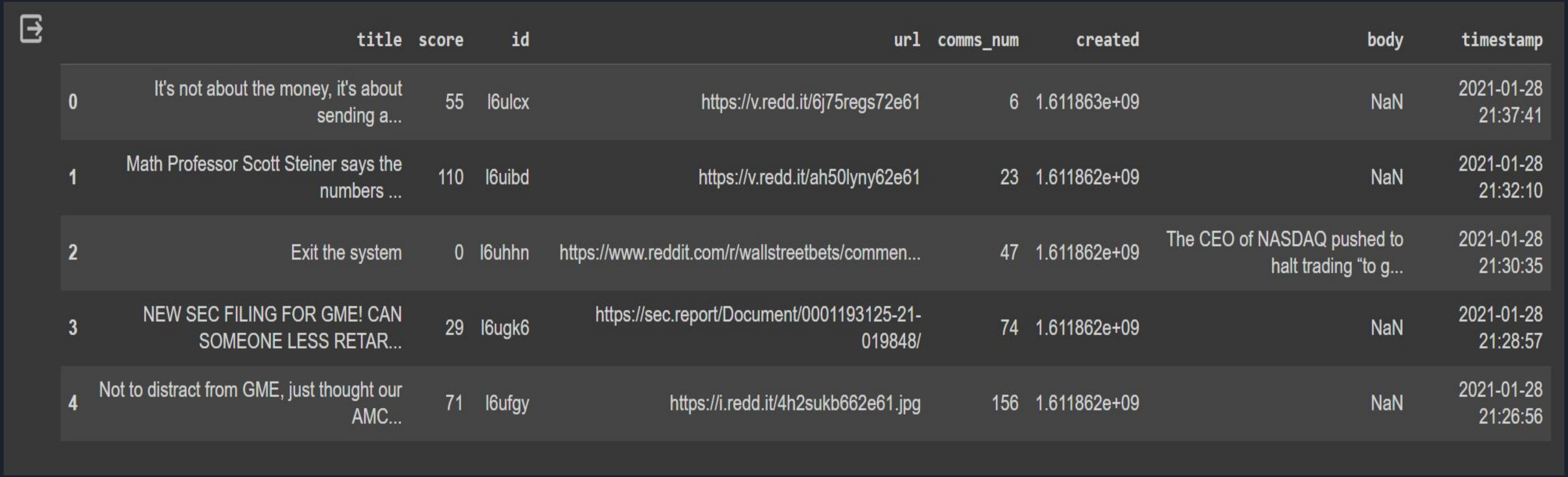
Data Sources

- Reddit Posts and Comments: posts and comments from WallStreetBets
- MetaData: Includes information like post titles, author, timestamp, upvote, downvotes and any attached links or images

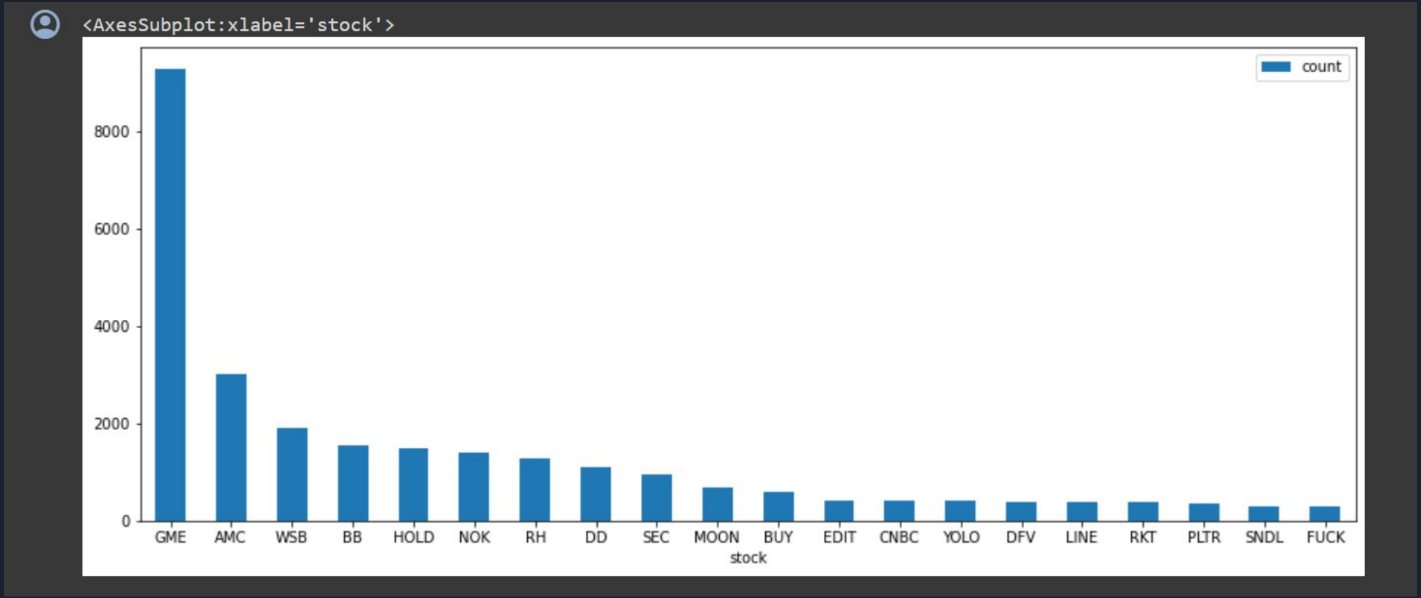
Data Collection

- API- Reddit’s API

Reddit WallstreetBets dataset



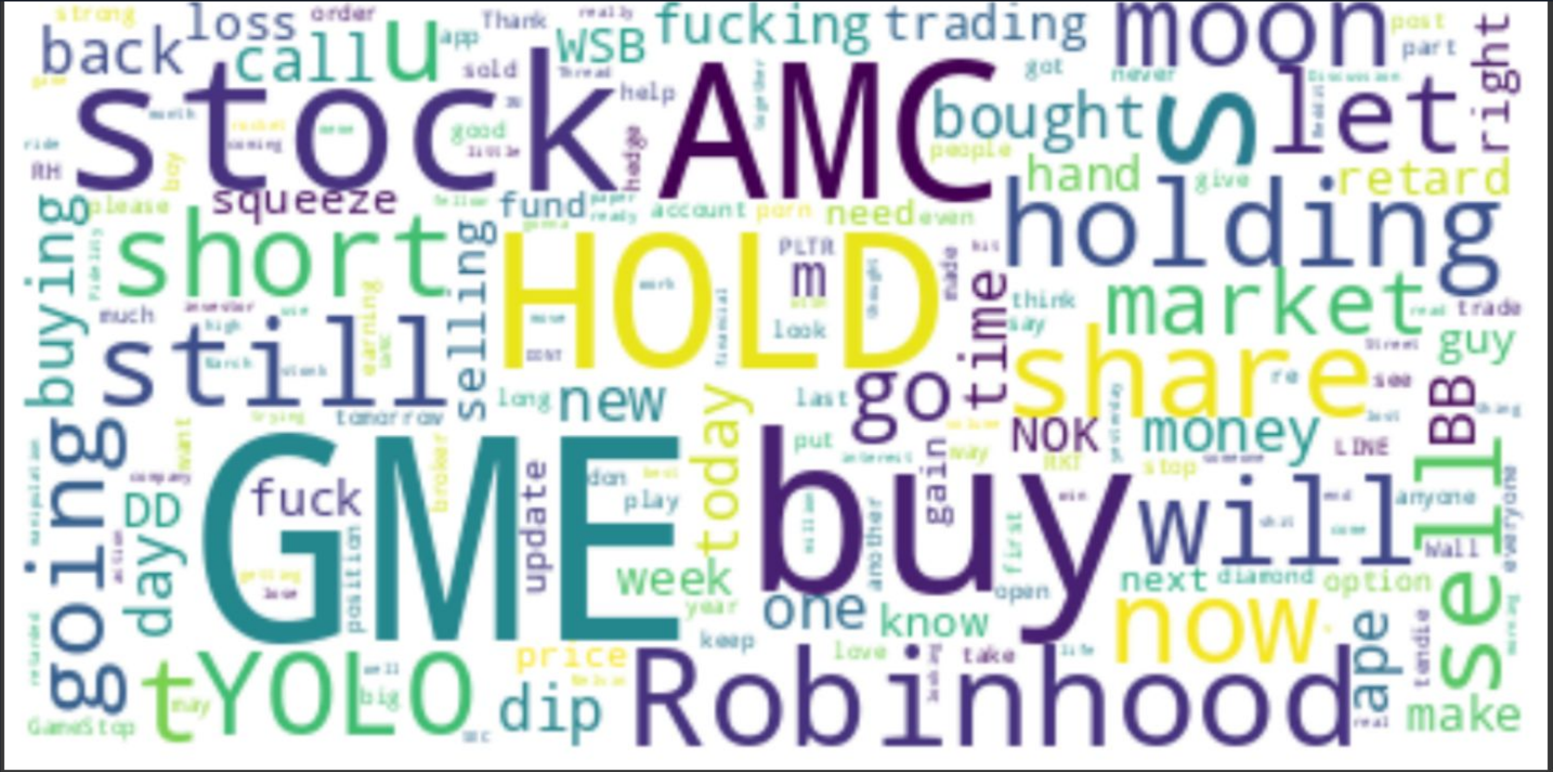
	title	score	id	url	comms_num	created	body	timestamp
0	It's not about the money, it's about sending a...	55	l6ulcx	https://v.redd.it/6j75regs72e61	6	1.611863e+09	NaN	2021-01-28 21:37:41
1	Math Professor Scott Steiner says the numbers ...	110	l6uibd	https://v.redd.it/ah50lyny62e61	23	1.611862e+09	NaN	2021-01-28 21:32:10
2	Exit the system	0	l6uhhn	https://www.reddit.com/r/wallstreetbets/commen...	47	1.611862e+09	The CEO of NASDAQ pushed to halt trading “to g...	2021-01-28 21:30:35
3	NEW SEC FILING FOR GME! CAN SOMEONE LESS RETAR...	29	l6ugk6	https://sec.report/Document/0001193125-21-019848/	74	1.611862e+09	NaN	2021-01-28 21:28:57
4	Not to distract from GME, just thought our AMC...	71	l6ufgy	https://i.redd.it/4h2sukb662e61.jpg	156	1.611862e+09	NaN	2021-01-28 21:26:56



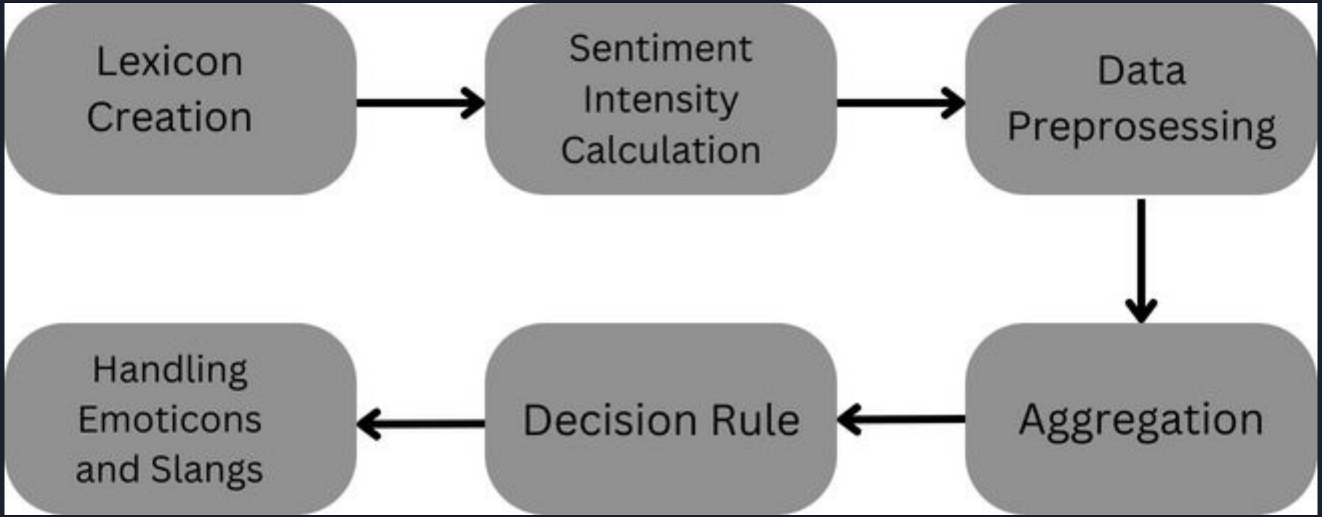
```
<AxesSubplot:xlabel='stock'>
```

Above graph lists the number of posts mentioning certain stocks.

Most Common Words Used In WSB Title

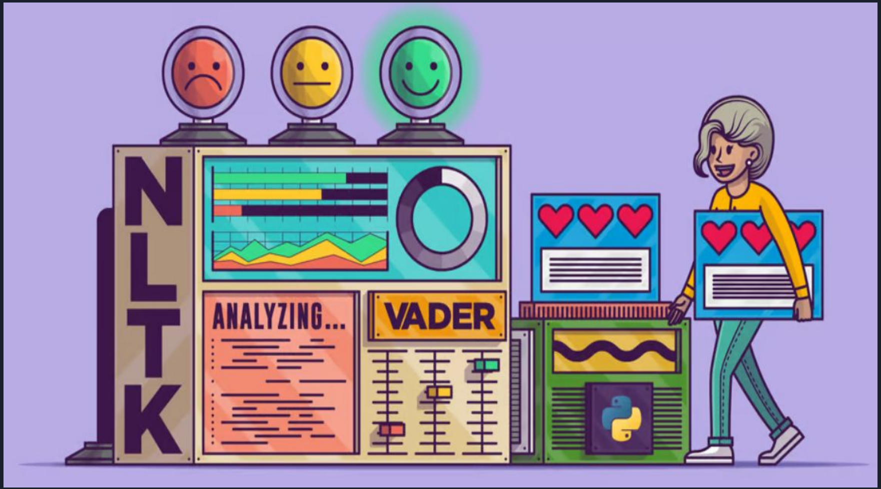


Methodology


Lexicon Creation
Sentiment Intensity Calculation
Data Preprosessing
Aggregation
Decision Rule
Handling Emoticons and Slangs

Training Model

NLTK's Sentiment Intensity Analyzer typically uses a pre-trained model.



NLTK's Sentiment Intensity Analyzer is a tool within the Natural Language Toolkit (NLTK) library, a popular Python library for natural language processing. The Sentiment Intensity Analyzer is designed to assess the sentiment expressed in a piece of text and assign a polarity score.

Tool that helps a computer figure out if a piece of text (like a paragraph or a sentence) sounds positive, negative, or somewhere in between. But instead of just saying "good" or "bad," it gives a score that shows how strong those feelings are.

It leverages a pre-trained model that takes into account the strength of words in conveying sentiment, providing a quantitative measure of the text's overall sentiment.

Filtering Criteria for VADER

To identify relevant content for sentiment analysis, we applied several filtering criteria to the dataset. These criteria included:

- Token case (lowercase or uppercase): To consider stock symbols or terms regardless of their case.
- Part-of-speech (POS) tagging: To focus on proper nouns (PROPN) commonly associated with stock symbols and company names.
- Token length: To filter out very short or very long words unlikely to be stock symbols.
- Emoji recognition: To exclude emojis, which are not stock symbols.
- Entity type recognition: To identify organizations (ORG) and persons (PERSON) associated with stock discussions.
- Stop word removal: To eliminate common words that do not carry significant sentiment.

Sentiment Analysis using VADER

The sentiment analysis was conducted using the following approach:

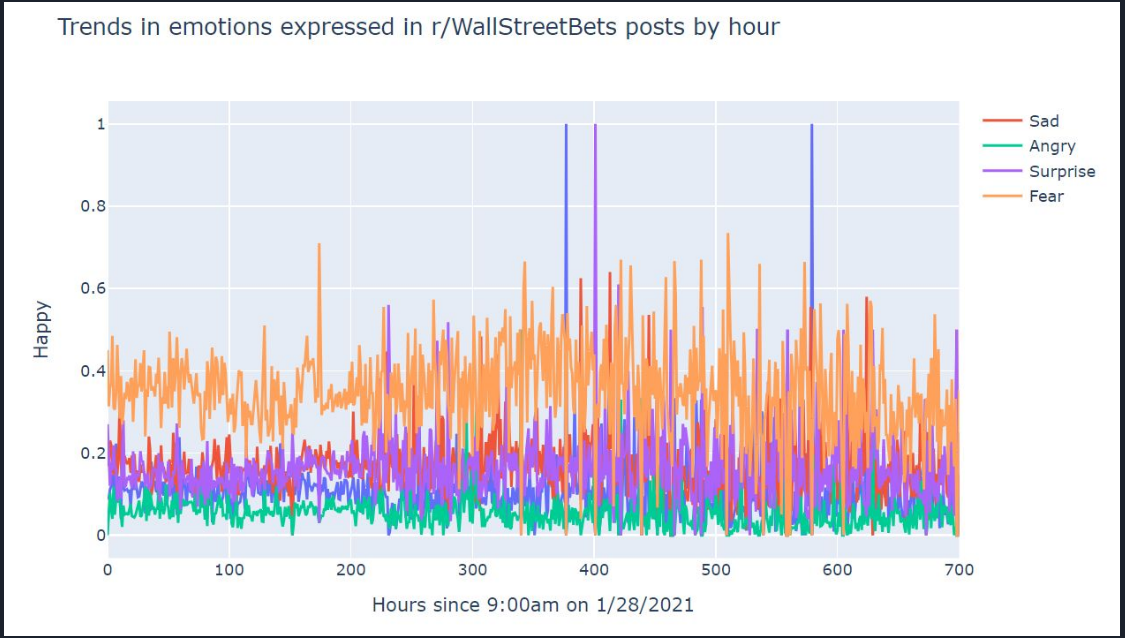
- Sentiment Scoring: Each tokenized word was assigned a sentiment score using NLTK's Sentiment Intensity Analyzer, which provides a compound score representing the overall sentiment (positive, negative, or neutral).

- Aggregating Sentiment Scores: For each post and comment, the individual word scores were aggregated to yield an overall sentiment score.

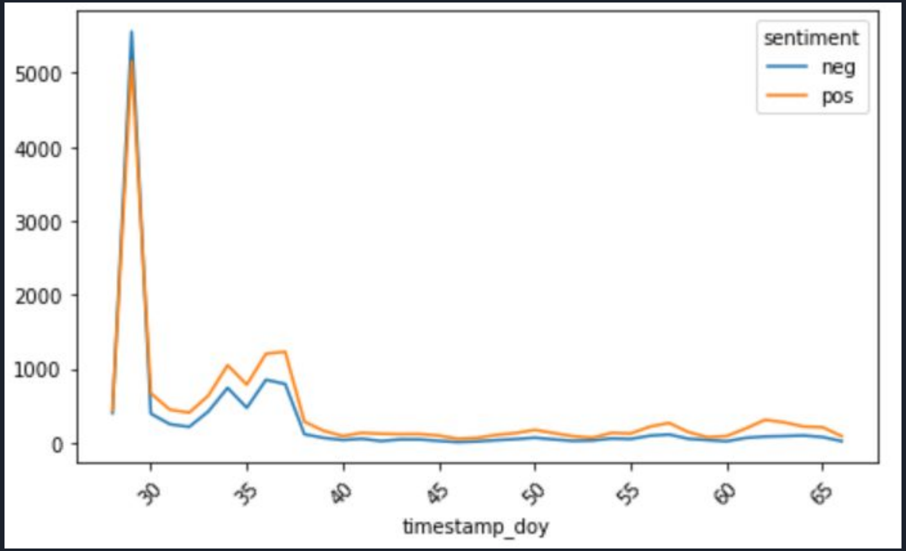
- Categorizing Sentiment: Based on the compound score, the texts were categorized into three sentiment categories: Positive, Negative, and Neutral. Thresholds for these categories were determined empirically.

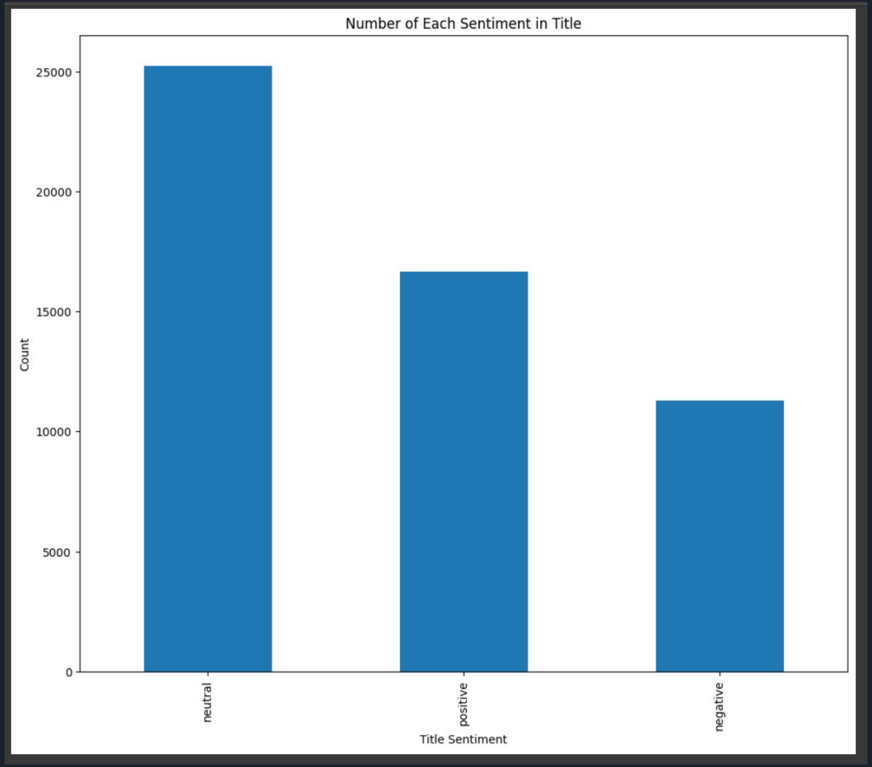
Emotion Analysis

- We used the NRC Emotion Lexicon to perform emotion analysis on our dataset.
- We created a dictionary to map words to their emotion
- Then we analyzed the dataset to match the emotions associated with the words and assigned scores to identify the emotion.

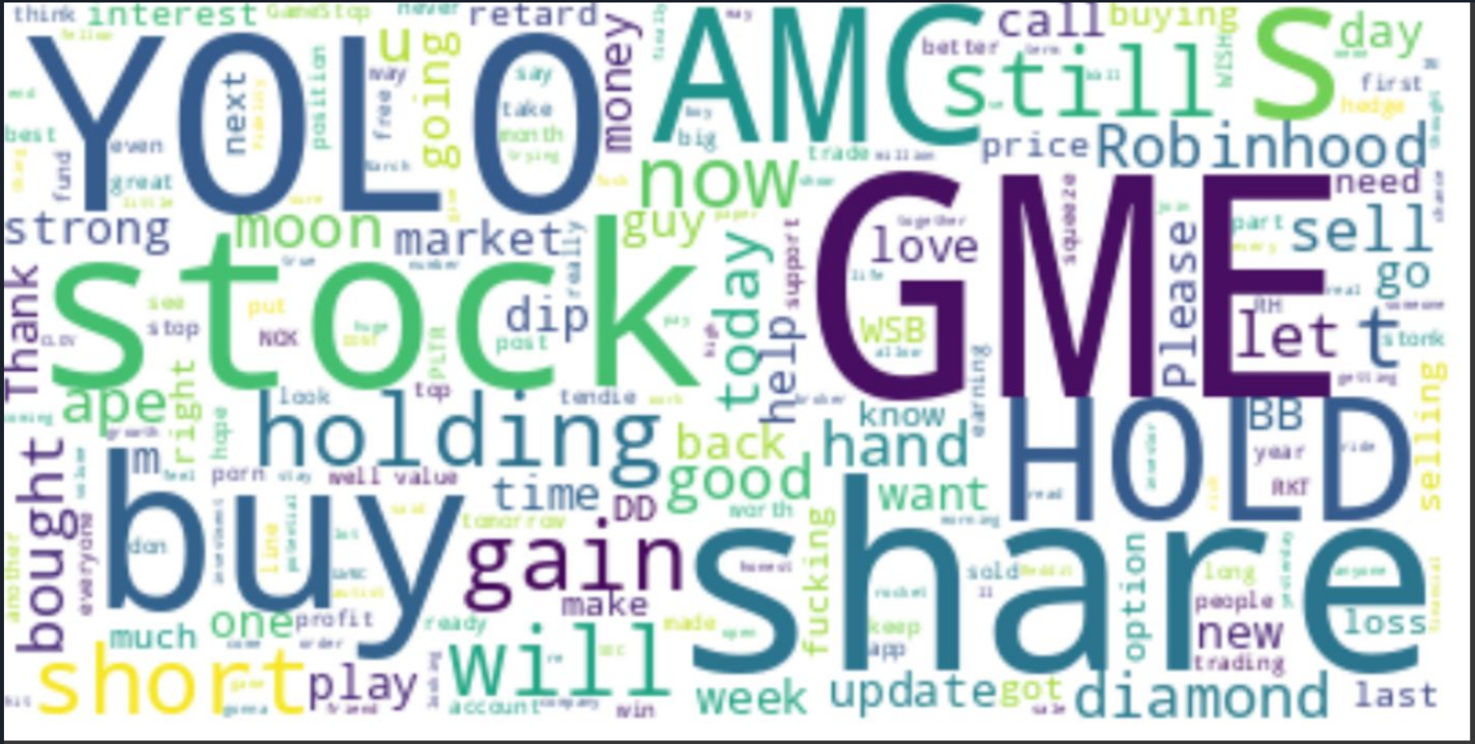


Observed Trends





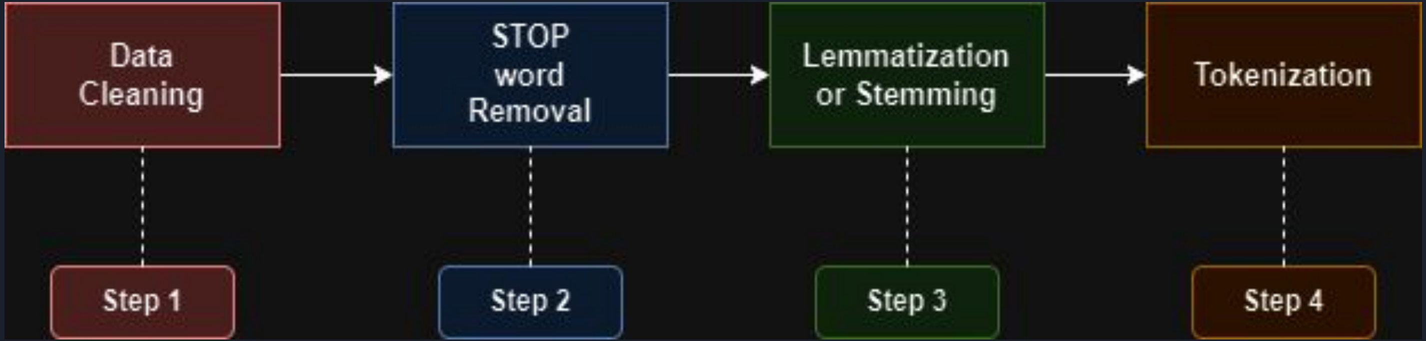
Most Common Words Used In Positive WSB Title



Approach 2: GloVe + LSTM

- We were not satisfied with Vader so we decided to build our own LSTM model to analyze the sentiment.
- One hurdle was that there is no labelled dataset.
- So what we’ve done is, we’ve trained our model on twitter data to learn to identify sentiment.
- We’ll then be using this model to analyze the sentiment of the WSB subreddit.

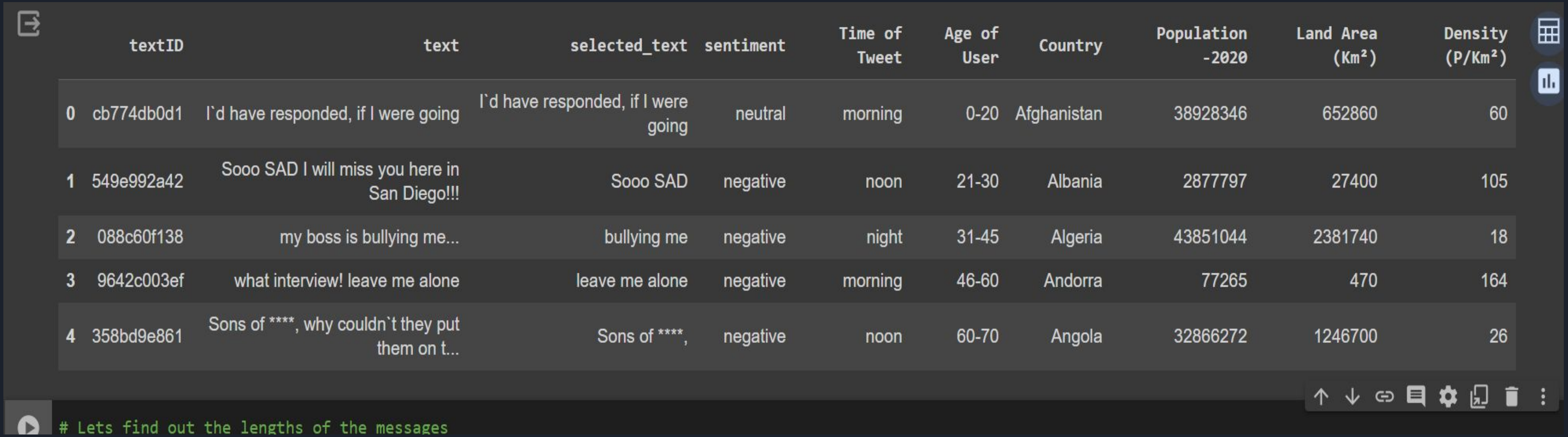
Data Processing for LSTM



Text to Vector Conversion -

- Count Vectorization
- TF-IDF vectorization
- N-gram modeling
- GloVe embeddings

LSTM Dataset



	textID	text	selected_text	sentiment	Time of Tweet	Age of User	Country	Population -2020	Land Area (Km²)	Density (P/Km²)
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral	morning	0-20	Afghanistan	38928346	652860	60
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	noon	21-30	Albania	2877797	27400	105
2	088c60f138	my boss is bullying me...	bullying me	negative	night	31-45	Algeria	43851044	2381740	18
3	9642c003ef	what interview! leave me alone	leave me alone	negative	morning	46-60	Andorra	77265	470	164
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative	noon	60-70	Angola	32866272	1246700	26

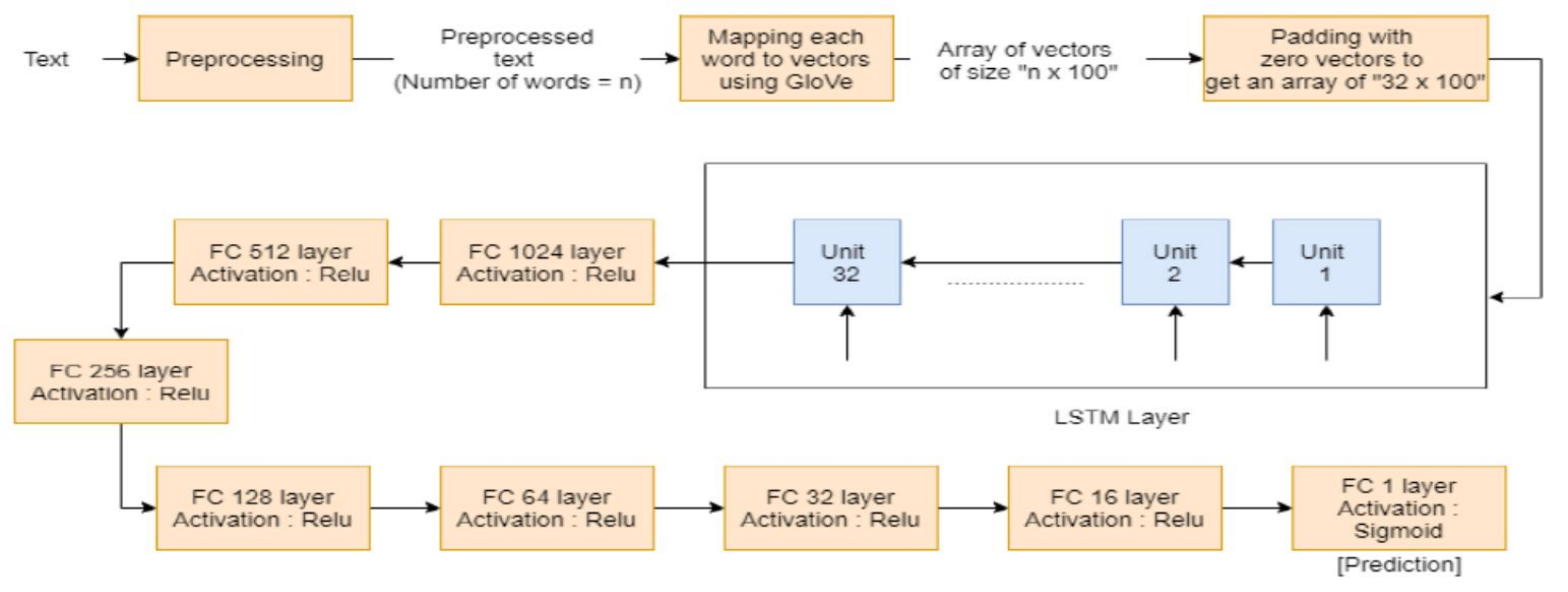
```
# Lets find out the lengths of the messages
```

GloVe Embedding + LSTM

- We used a pretrained GloVe embedding layer from Kaggle to represent each phrase as a 100-dimensional vector

- We kept the maximum length of a post to 32 characters. Posts were truncated or padded depending on the length.

- The output was routed through dense layers with the last layer producing the sentiment of the post as being positive, negative or neutral.



Text

Preprocessing

Preprocessed text (Number of words = n)

Mapping each word to vectors using GloVe

Array of vectors of size "n x 100"

Padding with zero vectors to get an array of "32 x 100"

Unit 1

Unit 2

Unit 32

LSTM Layer

FC 1024 layer Activation : Relu

FC 512 layer Activation : Relu

FC 256 layer Activation : Relu

FC 128 layer Activation : Relu

FC 64 layer Activation : Relu

FC 32 layer Activation : Relu

FC 16 layer Activation : Relu

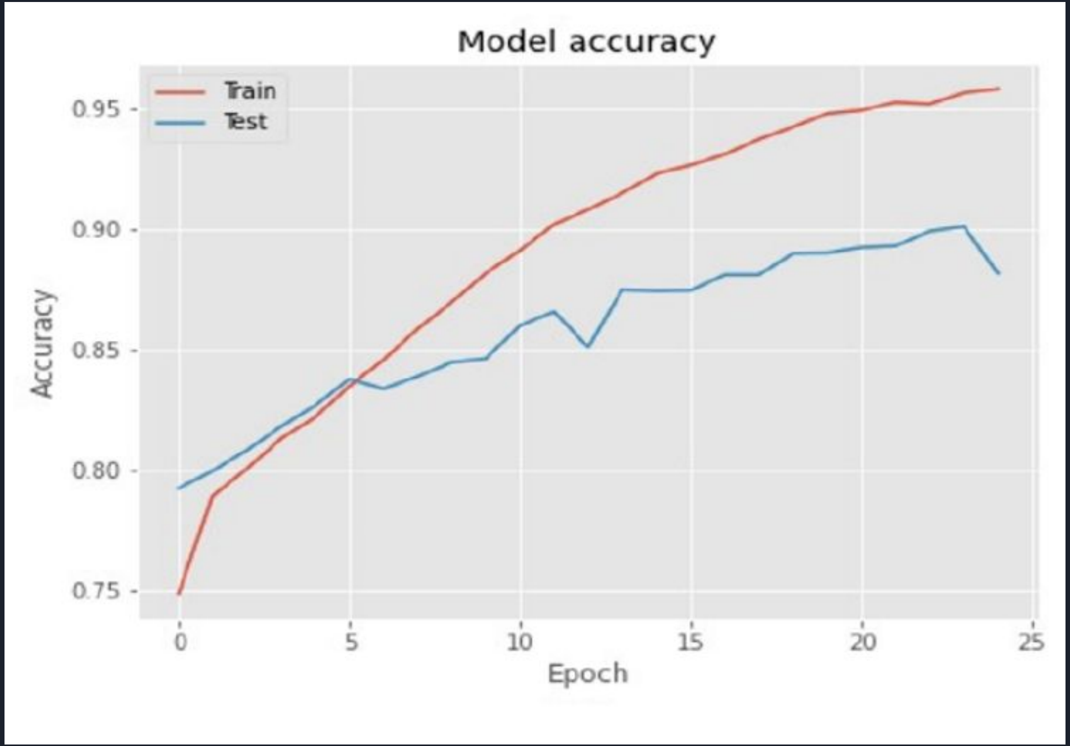
FC 1 layer Activation : Sigmoid

[Prediction]

Model Parameters

- Batch size = 64
- Epochs = 25
- Dropout rate = 0.2
- Learning rate = 0.01
- Optimizer = Adam
- Loss Function = Binary Cross Entropy

Model Accuracy



F1 score on Test data -

0.81

Comparison of both approaches

Architecture	# Positive posts	# Negative Posts	# Negative Posts
Vader	14012	7052	32123
Glove + LSTM	14894	6789	31504

Comparison of both approaches

Post Title	Vader	GloVe + LSTM
Feeling like exercising, dunno may delete this later	Positive	Negative
begging is cringe	Neutral	Negative
The Wallstreetbets War Museum	Negative	Neutral
No fear! This is going to be the new winner! $XSPA	Negative	Positive

Future Scope

Machine Learning Predictive Models: Develop and refine machine learning models to enhance predictive capabilities based on sentiment analysis. This would enable more accurate forecasting of market movements for traders and investors.

Real-Time Monitoring System: Implement a real-time monitoring system for frequently mentioned stocks and sentiment shifts within r/wallstreetbets. Users could receive timely alerts, facilitating swift responses to emerging trends and potential market disruptions.

Integration with Trading Platforms: Explore integration possibilities with popular trading platforms to provide direct access or alerts within trading interfaces. This would empower users to act on community sentiment in real-time as they execute trades.

Dynamic Risk Management Tools: Evolve the project to offer dynamic risk management tools that adapt to changing community sentiments. This could assist users in adjusting their risk strategies in response to evolving market conditions.

Global Community Impact Analysis: Expand the analysis beyond r/wallstreetbets to encompass global online financial communities. Understanding cross-cultural variations in market influences would provide a more comprehensive and nuanced perspective on the impact of online communities on financial markets.

Challenges

Data Quality and Integrity: The Kaggle dataset may exhibit variations in data quality and completeness. Inconsistencies or gaps in the data could impact the precision of sentiment analysis and trend identification, necessitating a thorough evaluation and resolution of any dataset-related issues.

Applicability to Different Platforms: Extending findings to other social media platforms presents a challenge as the Kaggle dataset is specific to Reddit's r/wallstreetbets. Each platform has distinct community dynamics, language nuances, and user behaviors that must be considered when generalizing conclusions.

Temporal Relevance Challenges: The Kaggle dataset may be constrained to a specific timeframe, while online community dynamics evolve over time. Ensuring alignment between the dataset's temporal scope and the research objectives is critical to obtaining insights that remain current and pertinent.

Incomplete Contextual Information: Kaggle datasets might lack certain contextual details, such as information about discussions, user backgrounds, or external events. Acknowledging and addressing these limitations in available context is vital for the accurate interpretation and analysis of sentiments and trends.

Conclusion

- Our sentiment analysis of the WallStreetBets subreddit provides valuable insights into the emotions and sentiment expressed within this influential online community.
- We found that WallStreetBets exhibits a dynamic emotional tone, closely tied to real-world financial events and market performance.
- Emotions within the subreddit range from exuberance to frustration, reflecting the diverse sentiments of retail investors.
- The data collected from WallStreetBets may be subject to biases, as it represents only a subset of user-generated content.
- Some members might use sarcasm or humor in their posts, which can be hard for sentiment analysis tools to understand correctly.