Analyzing Market Trends through Sentiment Analysis

Keerthana, Parth, Prayut, Sirivanth & Shweta

THE UNIVERSITY OF CHICAGO

AGENDA

01

BUSINESS PROBLEM

- BACKGROUND
- BUSINESS PROBLEM

02

EXPLORATORY DATA ANALYSIS

- Data Definition
- Feature Engineering

03

METHODOLOGY

- Feature Engineering
- Modeling

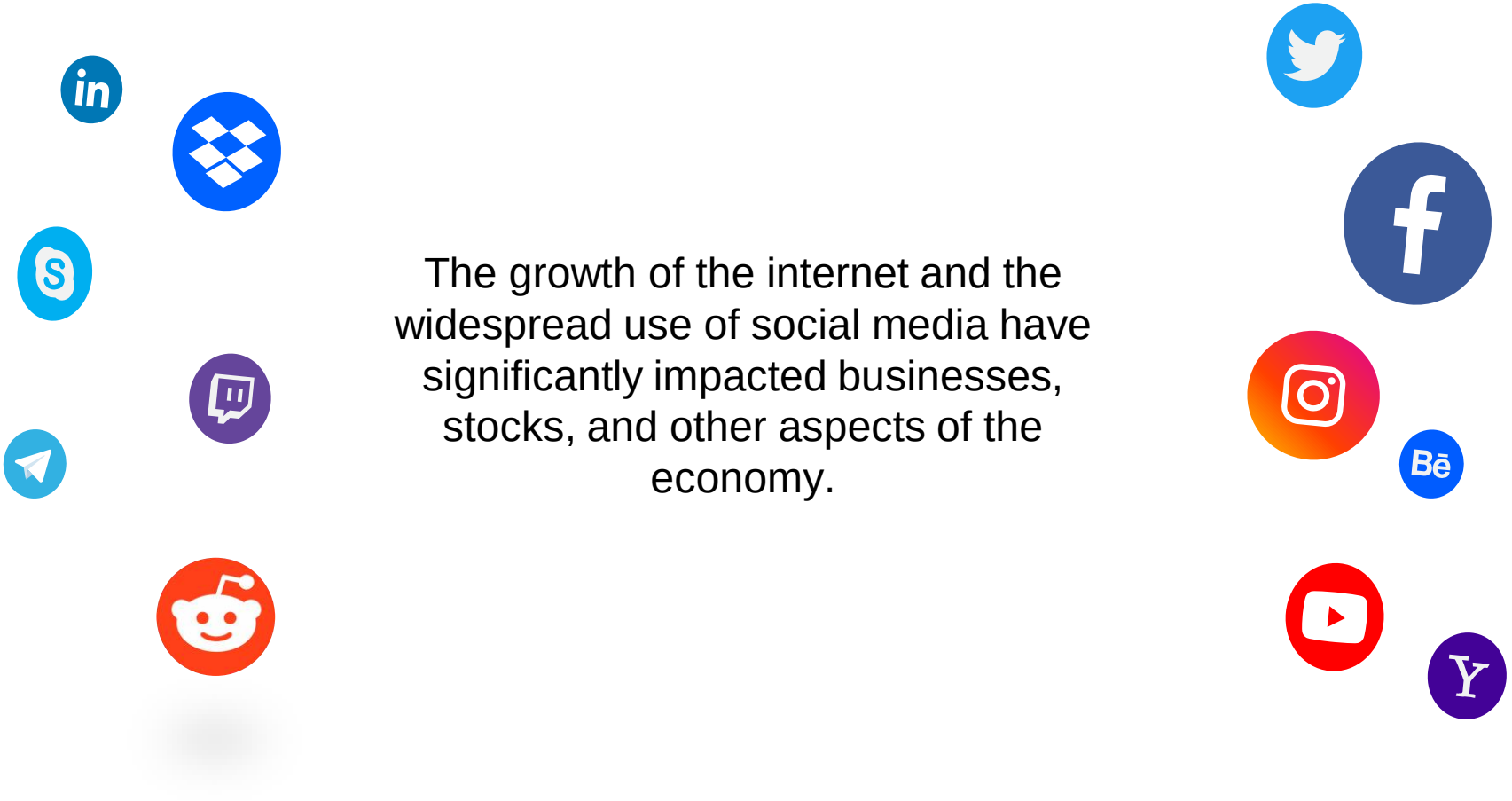
04

CONCLUSION

- Key Findings
- Future Implementations

THE UNIVERSITY OF CHICAGO

BACKGROUND



The growth of the internet and the widespread use of social media have significantly impacted businesses, stocks, and other aspects of the economy.

BACKGROUND

The growth of the internet and the widespread use of social media have significantly impacted businesses, stocks, and other aspects of the economy.

The sentiment on social media and online forums about the company often influences its stock price. For example, in 2018, Musk tweeted that he was considering taking Tesla private at $420 per share, which caused a surge in the stock price.

widespread use of social media have
significantly impacted businesses,

In 2017, a public relations crisis when a video showing a passenger being violently dragged off a flight went viral on social media. The negative sentiment on social media led to a sharp drop in United's stock price, and the company faced widespread criticism and boycotts.

In early 2021, a group of retail investors on Reddit's WallStreetBets forum coordinated a buying spree of shares of GameStop, a struggling video game retailer. This caused a sharp rise in GameStop's stock price, triggering a short squeeze that resulted in huge losses for several hedge funds that had bet against the stock.

THE UNIVERSITY OF CHICAGO

BUSINESS PROBLEM & IMPACT

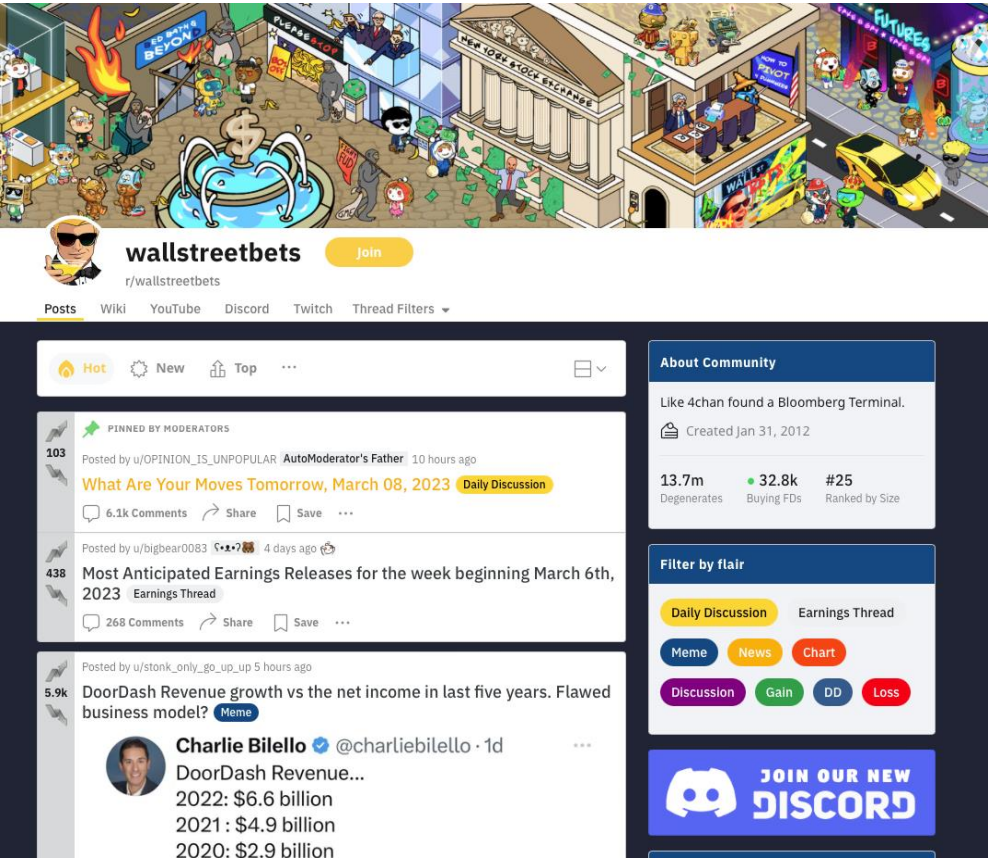
BUSINESS PROBLEM

Determine the sentiment of a community towards companies and their effect on the company in the stock market by analyzing their online presence on various social media platforms.

SOLUTION

Provide valuable insights into market sentiment and investor behavior by building Natural Language Processing models over **Reddit WallStreetBets**, which can be used to understand a companies' social media standing.

WHY REDDIT WALLSTREETBETS?



- The **GameStop Frenzy** incident highlighted the power of social media and its effect towards a particular stock or company.
- WallStreetBets, the **25th largest Reddit community** with 13.7 million followers, is a significant platform for finance and investing discussions.
- **Sentiment analysis** can help investors identify trends and fluctuations in real-time and understand the impact of real-world events on the stock market.

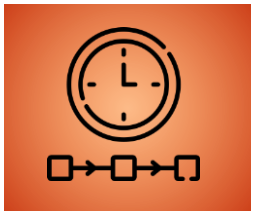
THE UNIVERSITY OF CHICAGO

DATA PROFILE

r/WallStreetBets & Stock Data from Yahoo Finance

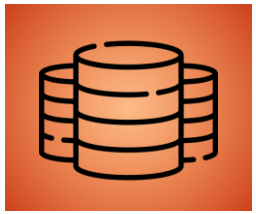
r/WallStreetBets Data Source

TIMELINE



2012-2022

SIZE



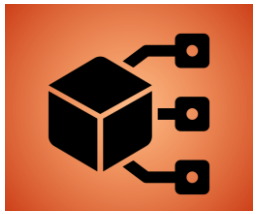
~800 MB

DATA POINTS

2,218,243

(Before Preprocessing)

FEATURES



14

+ 6 stock trends

(Before Preprocessing)

THE UNIVERSITY OF CHICAGO

DATA PROFILE

Snapshot of Data Set

id	title	body	upvote_ ratio	author	created_utc	score	num_comments	permalink
11fc79h	Easy 50k on ▽ NVAX goin Bankrupt		1.0	SkylerPhoenixx	1677696713	1	3	/r/wallstreetbets/comments/11fc79h/easy_50k_on...
11fbkiv	Branson collecting my WSB entry fee	My last post with this report seemed to do wel...	1.0	mytendies	1677695792	1	2	/r/wallstreetbets/comments/11fbhax/cheapest_pu.

Unique Post ID

Post Title

Post Body

 # Likes on the post

Reddit User ID

Time Posted

Post Score

 # Comments on the post

Permalink

THE UNIVERSITY OF CHICAGO

FEATURE ENGINEERNG

POST TITLE

- Removed Stopwords, punctuations and lemmatized
- Removed emojis/emoticons
- Dropped rows which were null (only 4 present)

POST BODY

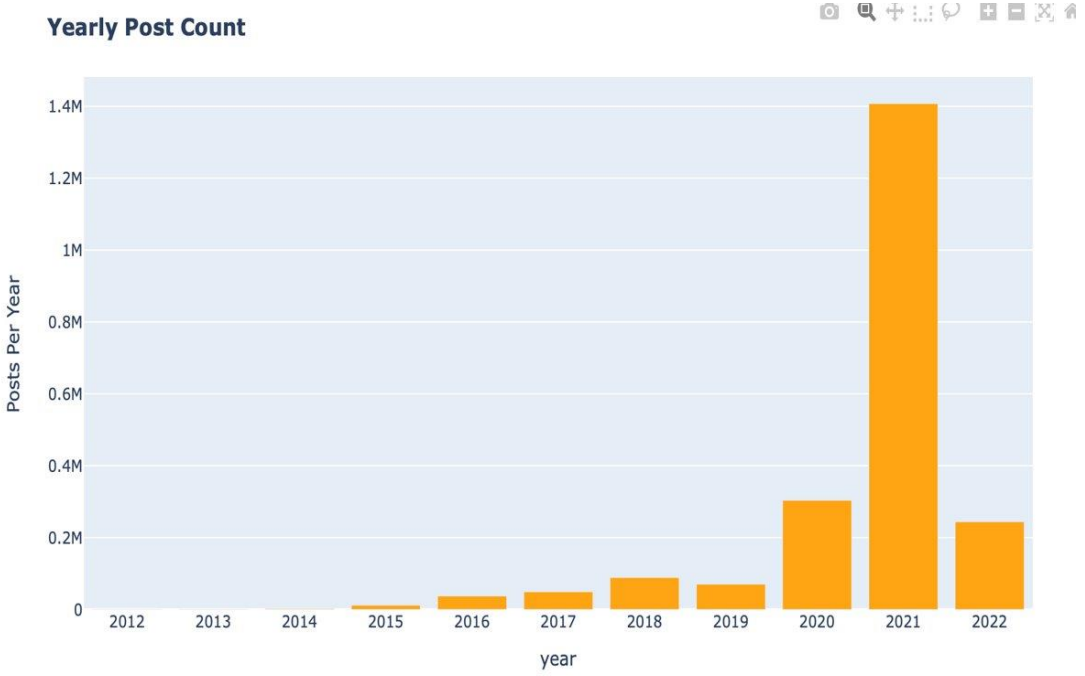
- Removed StopWords, punctuations, tabs, new lines and lemmatized
- Removed redundant post data
- Removed hyperlinks, tags and mentions
- Removed daily trading discussion threads and AutoModerator Warnings

YAHOO DATA

- Forward filled for missing values
- Calculated daily stock returns and polarity returns
- Standardization

DATA LIMITATIONS

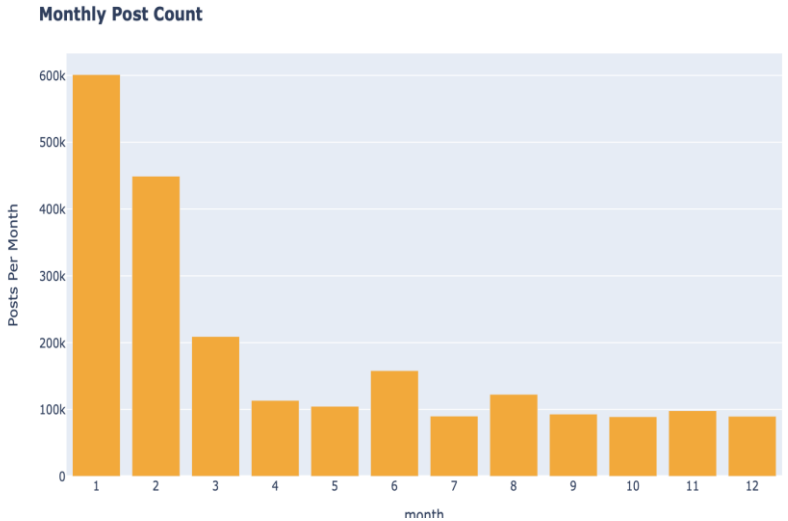
- Unnecessary/ Redundant columns
- Missing values
- Data is not balanced over the years (More posts after 2020)
- Slang and informal language
- Biased sample
- Contextual ambiguity due to limited context. Reddit posts lack in nuance in terms of context
- Misspellings and acronyms

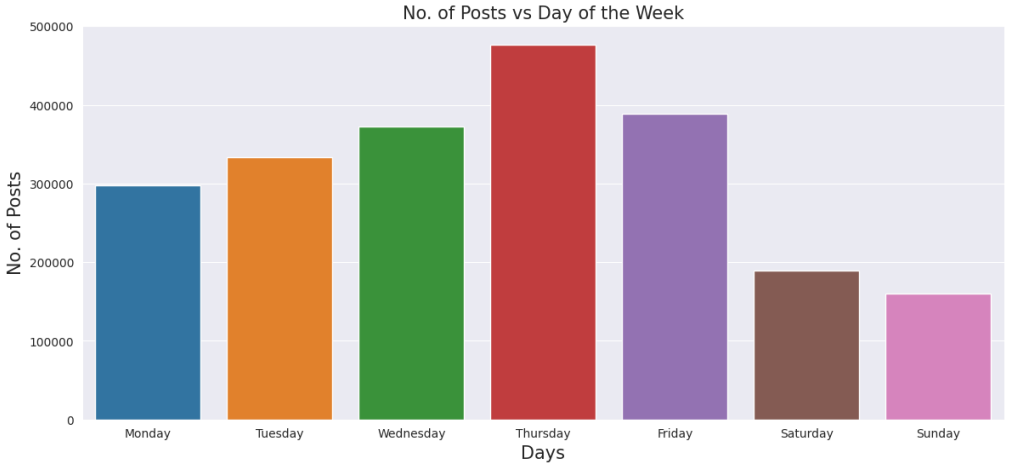


THE UNIVERSITY OF CHICAGO

EXPLORATORY DATA ANALYSIS

Post distribution in the Reddit WallStreetBets dataset

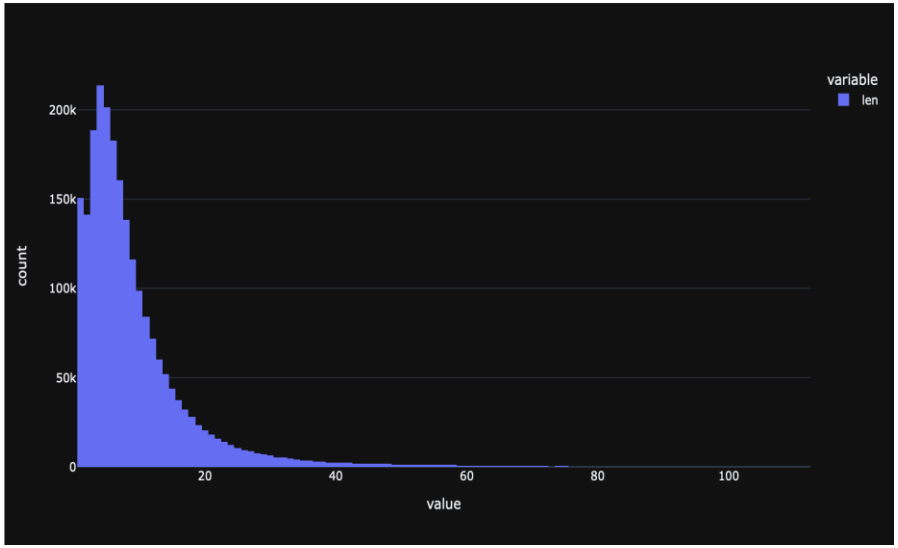




THE UNIVERSITY OF CHICAGO

EXPLORATORY DATA ANALYSIS

TOTAL AND AVERAGE WORD COUNT FOR TITLES



19,844,206

TOTAL NUMBER OF WORDS

8.95

AVERAGE NUMBER OF WORDS

MAX NUMBER OF WORDS IN A TITLE

112

EXPLORATORY DATA ANALYSIS

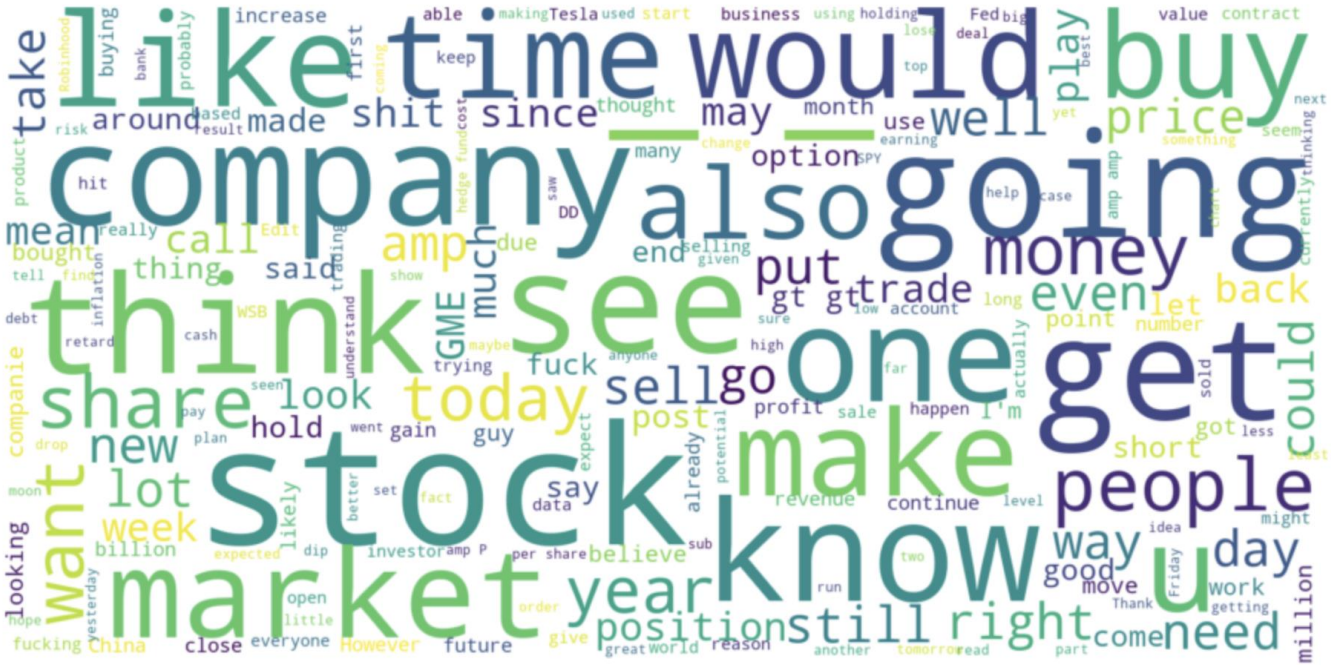
Word Cloud Based on Titles

Prevalent words in Titles

THE UNIVERSITY OF CHICAGO

EXPLORATORY DATA ANALYSIS

Word Cloud Based on Body



Prevalent words in Posts

THE UNIVERSITY OF CHICAGO

MODELING

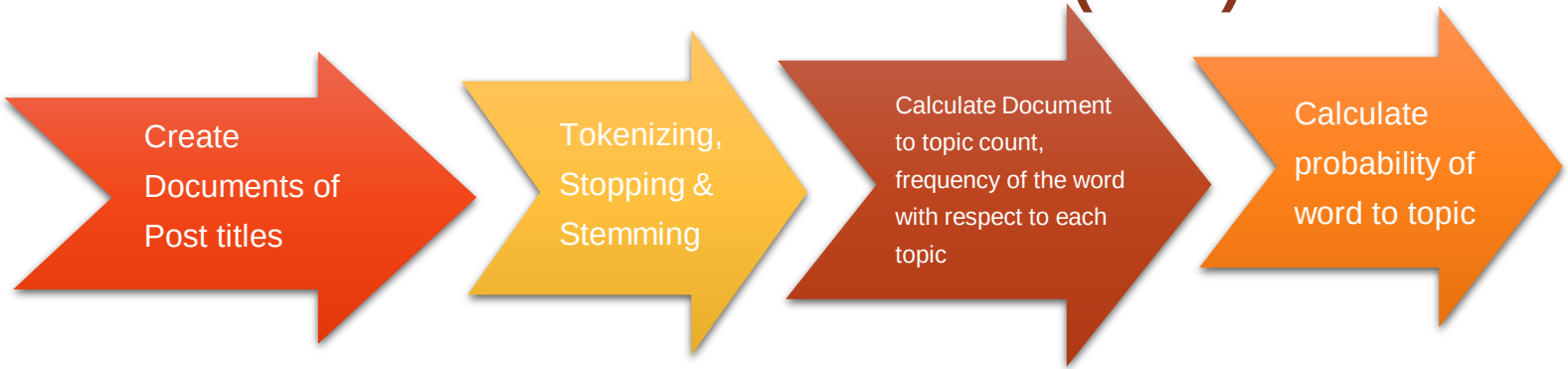
Models Applied

Latent Dirichlet Allocation	LDA	Topic Modelling
spaCy NER	Named Entity Recognition	Organization labelling
Valence Aware Dictionary for Sentiment Reasoning	VADER	Sentiment Compounding
TextBlob Pattern Analyzer		Polarity and Subjectivity
Aspect Based Sentiment Analysis	ABSA	Aspect based Sentiment
Naive Bayes		Probabilistic algorithm

THE UNIVERSITY OF CHICAGO

MODELING

Latent Dirichlet Allocation (LDA)

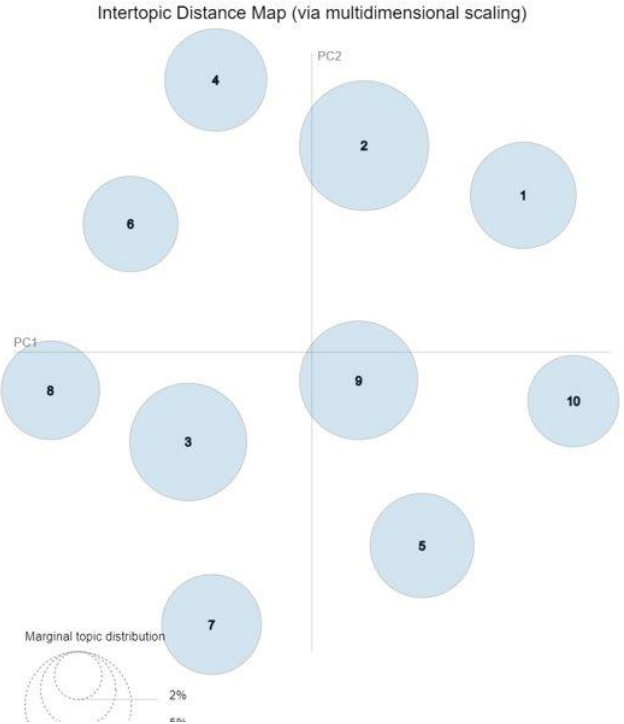


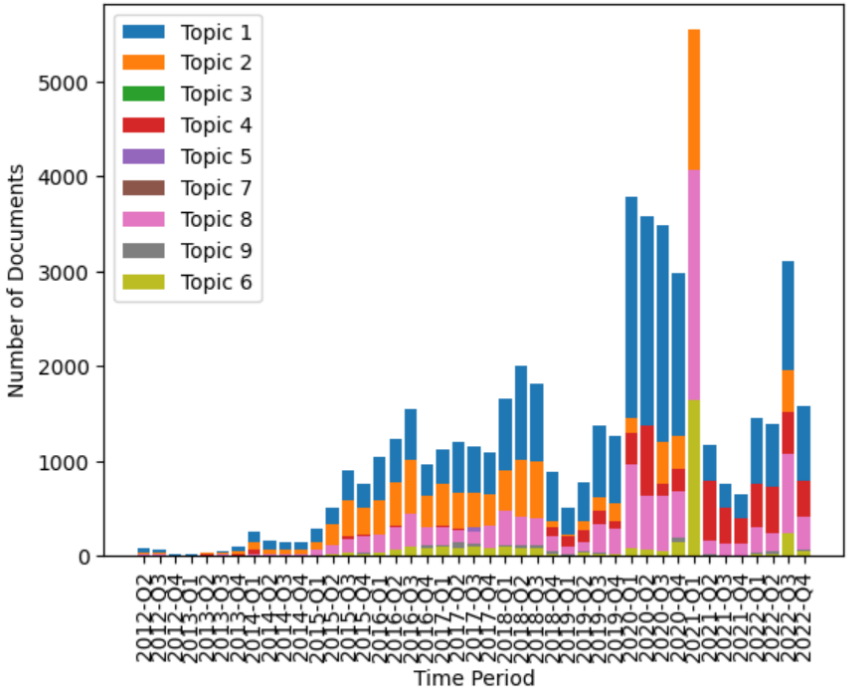
- **Robinhood** app, stock trading, and related terminology like "dd" (due diligence), "rh" (Robinhood), and "stonks".
- Short selling, stock prices, and related terms like "**squeeze**" (short squeeze), "**tendies**" (profits)
- Hedge funds, news, and various other topics like the SEC (Securities and Exchange Commission), Donald Trump, and **cryptocurrencies**.
- Buying, holding, and selling stocks, as well as various other related terms like "dip" (a temporary drop in stock prices) and "earnings".
- Specific stocks that have gained a lot of attention on WSB, including **AMC**, **Blackberry**, **Nokia**, and **Dogecoin**, as well as related terms like "moon" and "rocket".
- Discussions related to investing, including the WSB community itself, as well as related terms like "retard" (used humorously) and **"gain"**.
- **GameStop** (GME), a stock that gained a lot of attention on WSB in early 2021, as well as related terms like "next" and "buy".

THE UNIVERSITY OF CHICAGO

MODELING

LDA Distribution and Intertopic distance

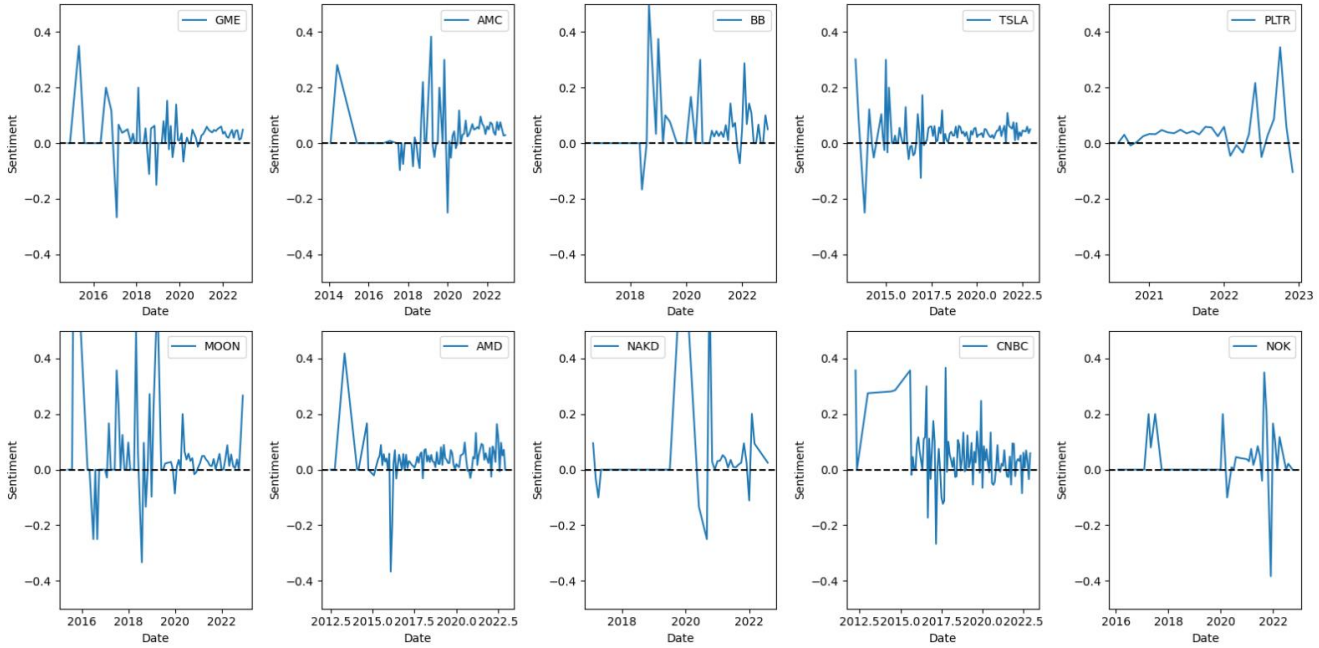




THE UNIVERSITY OF CHICAGO

MODELING

Sentiment analysis - Top 10 stocks



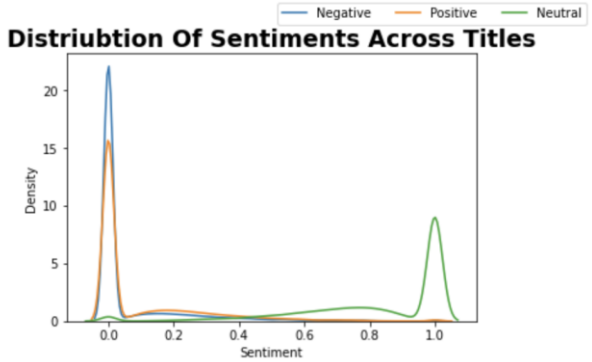
- Game Stop
- AMC Theatres
- BlackBerry
- Tesla
- Palantir Tech.
- Direxion Moonshot Innovators ETF
- AMD
- NAKD Brand
- CNBC
- Nokia

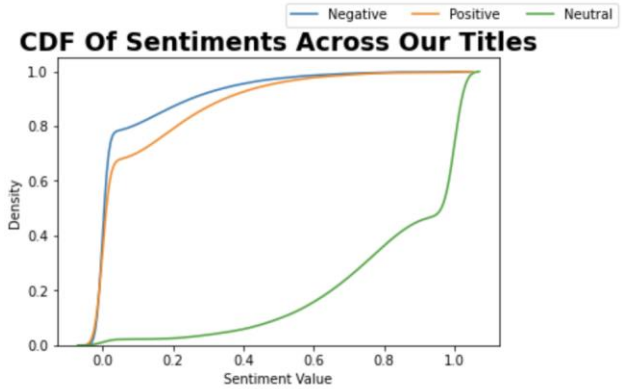
*retrieved using spaCy NER

THE UNIVERSITY OF CHICAGO

MODELING

VADER & Pattern Analyzer





Rule-based sentiment analysis tool specifically designed for sentiment analysis of social media texts.

Tokenizing & cleaning the data

Pre-built lexicon used to calculate sentiment

Aggregates polarity, intensity, and context of the words to get overall sentiment score between -1 to 1

Score is post-processed to produce a binary classification of the sentiment as either positive, negative, or neutral

THE UNIVERSITY OF CHICAGO

MODELING

What were the positive topics?

- "**GameStop stock hype**": **gme**, moon, help, bb, friend, need, wsb, **amc**, cnbc
- "Short squeeze and interest": **gme**, short, **amc**, squeeze, interest, love, ape, please, someone
- "Excitement and party atmosphere": **gme**, like, **amc**, look, 💎, stock, late, hold, moon
- "Rocket emojis and money-making": 🚀, **gme**, make, **amc**, money, like, stock, cnbc,
- "Future speculation and hype": **gme**, **pltr**, moon, **amc**, bb, go, ready, wish, next, 🚀 🚀 🚀
- "**AMC** stock discussion": amc, **gme**, good, ape, stock, bb, buy, hand, diamond
- "Investment options": yolo, tesla, **gme**, call, **tsla**, amd, **pltr**, stock, update, put

THE UNIVERSITY OF CHICAGO

MODELING

What were the negative topics?

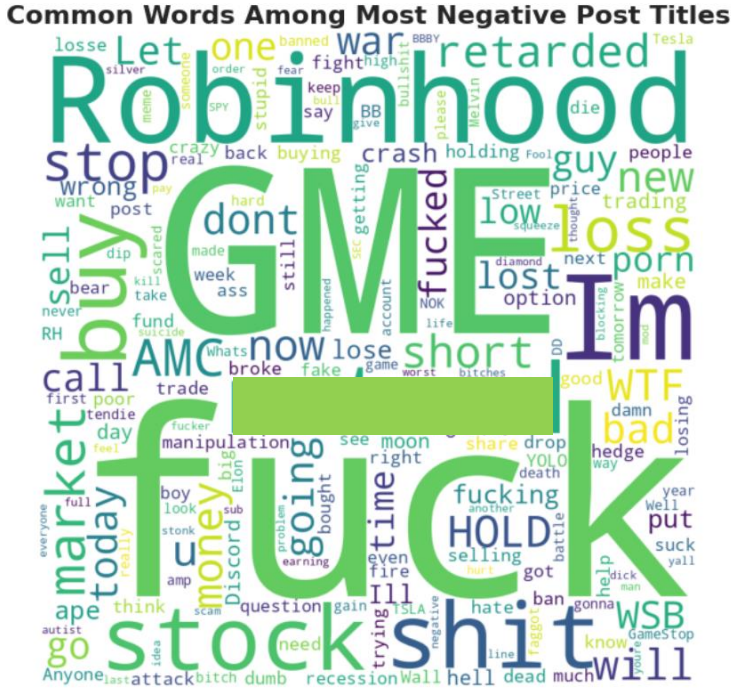
GAMESTOP

That's all people talked about and not everyone was happy...

THE UNIVERSITY OF CHICAGO

MODELING

Word Cloud Based on Negative & Positive Sentiment



THE UNIVERSITY OF CHICAGO

MODELING

Aspect Based Sentiment Analysis (ABSA)

ABSA is a natural language processing technique that focuses on identifying the aspects or attributes of a product, service or entity that are being discussed in a text, and analyzing the sentiment associated with each aspect.

- **Aspect extraction:** This is done using techniques such as rule-based parsing, named entity recognition, or machine learning models.
- **Sentiment analysis:** Using a variety of sentiment analysis techniques, such as lexicon-based methods, machine learning models or hybrid approaches.
- **Aspect-level sentiment aggregation:** Once the sentiment score for each aspect has been calculated, ABSA aggregates the scores to produce an overall sentiment score for the product, service or entity at the aspect level.
- **Document-level sentiment analysis:** Combines the aspect-level sentiment scores to produce an overall sentiment score for the entire document.

"Amazon (AMZN) and Microsoft (MSFT). Amazon has been dominating the e-commerce space for years! But unfortunately, Microsoft has been lagging"

Symbol	pos_sentiment	neu_sentiment	neg_sentiment
AMZN	1.87197	0.11727	1.01076
MSFT	0.08516	0.09648	2.81836

{'label': 'neutral', 'score': 0.8062297701835632}

THE UNIVERSITY OF CHICAGO

MODELING

Is there a correlation between the sentiment expressed on WallStreetBets and the actual stock price movements?

There was a very weak positive correlation between bullish sentiment expressed on the forum and stock price movements, but no significant correlation between bearish sentiment and stock price movements.

date	polarity_title	TSLA_return	diff
2013-05-13	0.466667	0.143834	NaN
2013-05-15	0.136364	0.019283	-0.330303
2013-07-16	0.272222	-0.143093	0.135859
2013-07-31	-0.093750	0.019242	-0.365972
2013-11-08	-0.500000	-0.012986	-0.406250
...	...	...	...
2022-12-27	0.118065	-0.114089	0.129134
2022-12-28	0.032838	0.033089	-0.085227
2022-12-29	0.099619	0.080827	0.066780
2022-12-30	-0.048359	0.011164	-0.147977
2022-12-31	0.000000	NaN	0.048359

THE UNIVERSITY OF CHICAGO

KEY FINDINGS

- The dominant **sentiment** of WallStreetBets posts is **neutral**.
- The sentiments of both bodies and titles is not stationary over our timeline and it seems there is an **incline in the positivity** of the posts.
- Looking at the top 10 discussed organizations we see that the **dominant sentiment is positive**, there is also an example for a very discussed topic with dominant negative sentiment.
- Based on additonal external new sources, WallStreetBets has demonstrated the ability to influence stock prices through coordinated buying and selling activity.
- Only 450 words are needed to account for 79% of the variability in the text of WallStreet related posts titles.
- WallStreetBets posts have no significant correlation between its sentiment and stock price movements.

FUTURE SCOPE

- **Expand the analysis to other subreddits or social media platforms** - By analyzing data from other subreddits or social media platforms, investors could gain a more complete understanding of trends and sentiment.
- **Incorporate additional data sources** - In addition to analyzing posts and comments, investors could also incorporate data from news articles, financial reports, and other sources to get a more holistic view of the market.
- **Develop predictive models** - By analyzing historical data and market trends, investors could potentially develop predictive models that could help identify emerging trends or anticipate changes in the market.
- **Explore the impact of sentiment on specific stocks or sectors** - By focusing on specific stocks or sectors, investors could explore the impact of sentiment on price movements and potentially identify opportunities for profit.

THE UNIVERSITY OF CHICAGO

THANK YOU!