r/wallstreetbets Sentiment Analysis

Shashank Nagaraja

IST 652 Final Project

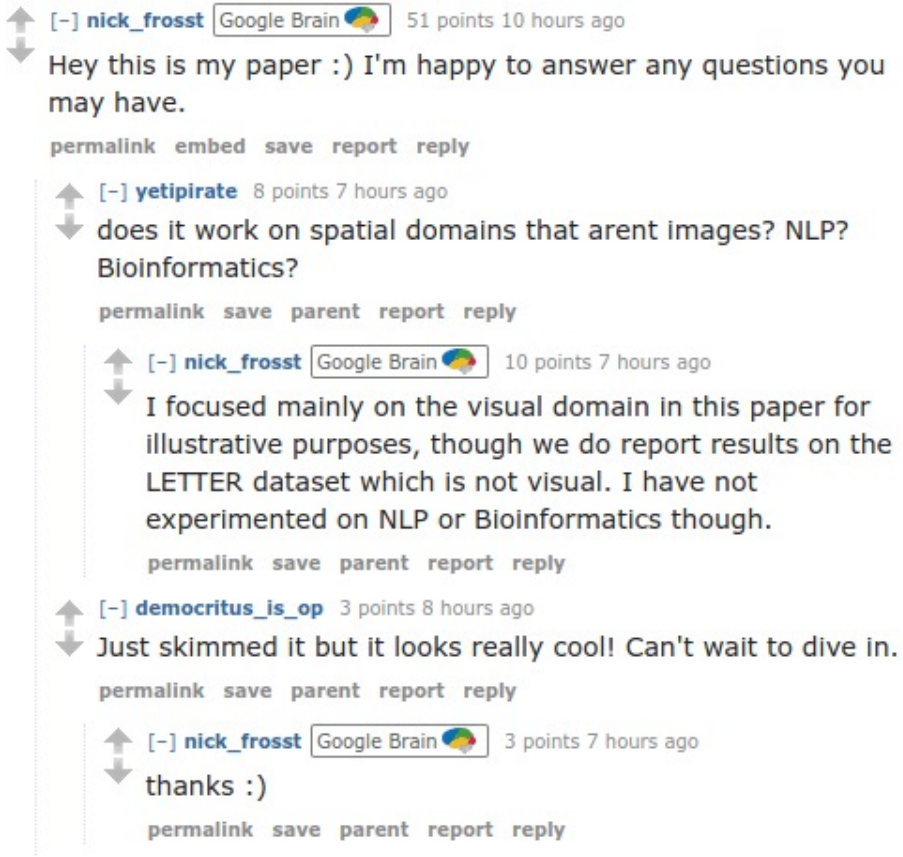
Reddit & WSB

Reddit:

- Social media platform
- Network of communities based on interest
- Largely text focused, ability to start "threads" with comments & upvotes

WSB:

- Group for people to discuss trades, gains, losses
- Disseminate information about potential portfolio positions
- Show off gains and losses
- Have fun



[–] nick_frosst Google Brain 51 points 10 hours ago

Hey this is my paper :) I'm happy to answer any questions you may have.

permalink embed save report reply

[–] yetipirate 8 points 7 hours ago

does it work on spatial domains that arent images? NLP? Bioinformatics?

permalink save parent report reply

[–] nick_frosst Google Brain 10 points 7 hours ago

I focused mainly on the visual domain in this paper for illustrative purposes, though we do report results on the LETTER dataset which is not visual. I have not experimented on NLP or Bioinformatics though.

permalink save parent report reply

[–] democritus_is_op 3 points 8 hours ago

Just skimmed it but it looks really cool! Can't wait to dive in.

permalink save parent report reply

[–] nick_frosst Google Brain 3 points 7 hours ago

thanks :)

permalink save parent report reply

Data Sources

- PRAW: API wrapper for Reddit’s API
  - Wraps around Reddit API
  - Easy to use interface to pull Reddit posts
  - Cannot query posts from a particular date
  - Need to re-acquire user/client keys every hour

- Kaggle
  - Pre-digested dataset
  - Readily Accessible

PRAW API Wrapper: https://github.com/praw-dev/praw

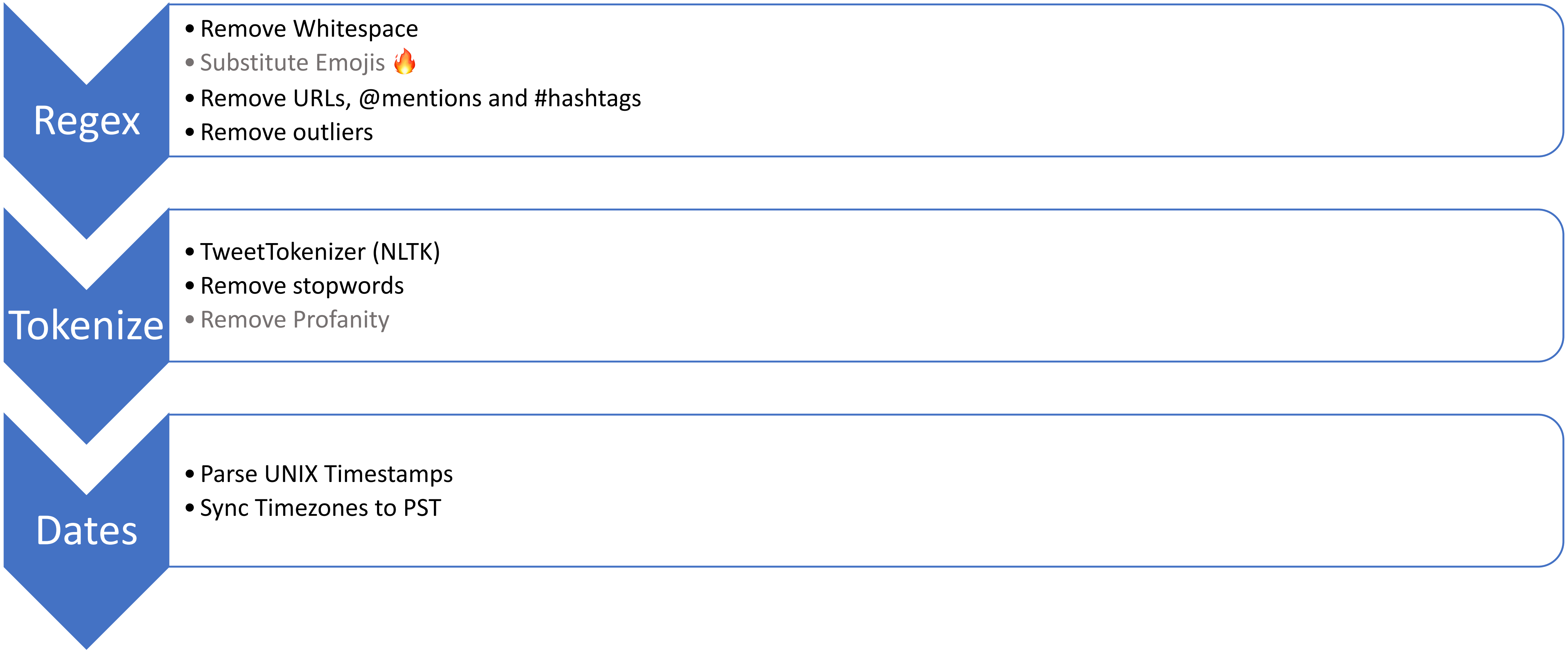
Kaggle Dataset: https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts

Data

~ 42k Posts in total

Field	Type	Information
title	Text	Post Title content
score	Float	Upvote Score (corrected for number of comments % impressions)
id	String	Unique identifier
url	URLs	URLs mentioned in post
comms_num	Integer	Number of direct comments
created	UNIX Timestamp	Post Time
body	Text	Post Body content

Preprocessing



Regex

- Remove Whitespace
- Substitute Emojis 🔥
- Remove URLs, @mentions and #hashtags
- Remove outliers

Tokenize

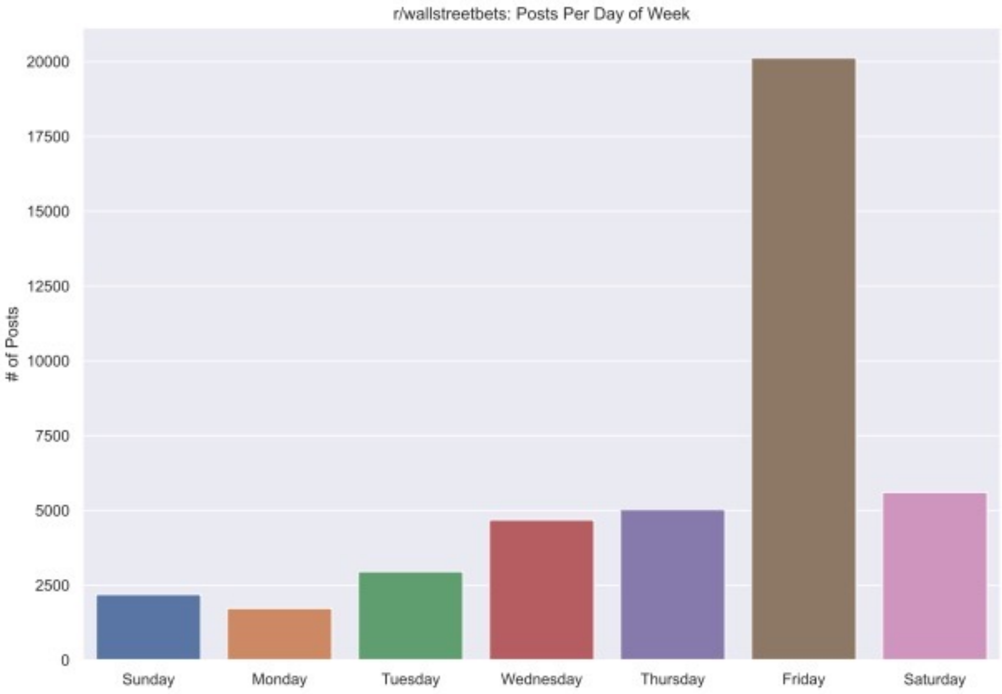
- TweetTokenizer (NLTK)
- Remove stopwords
- Remove Profanity

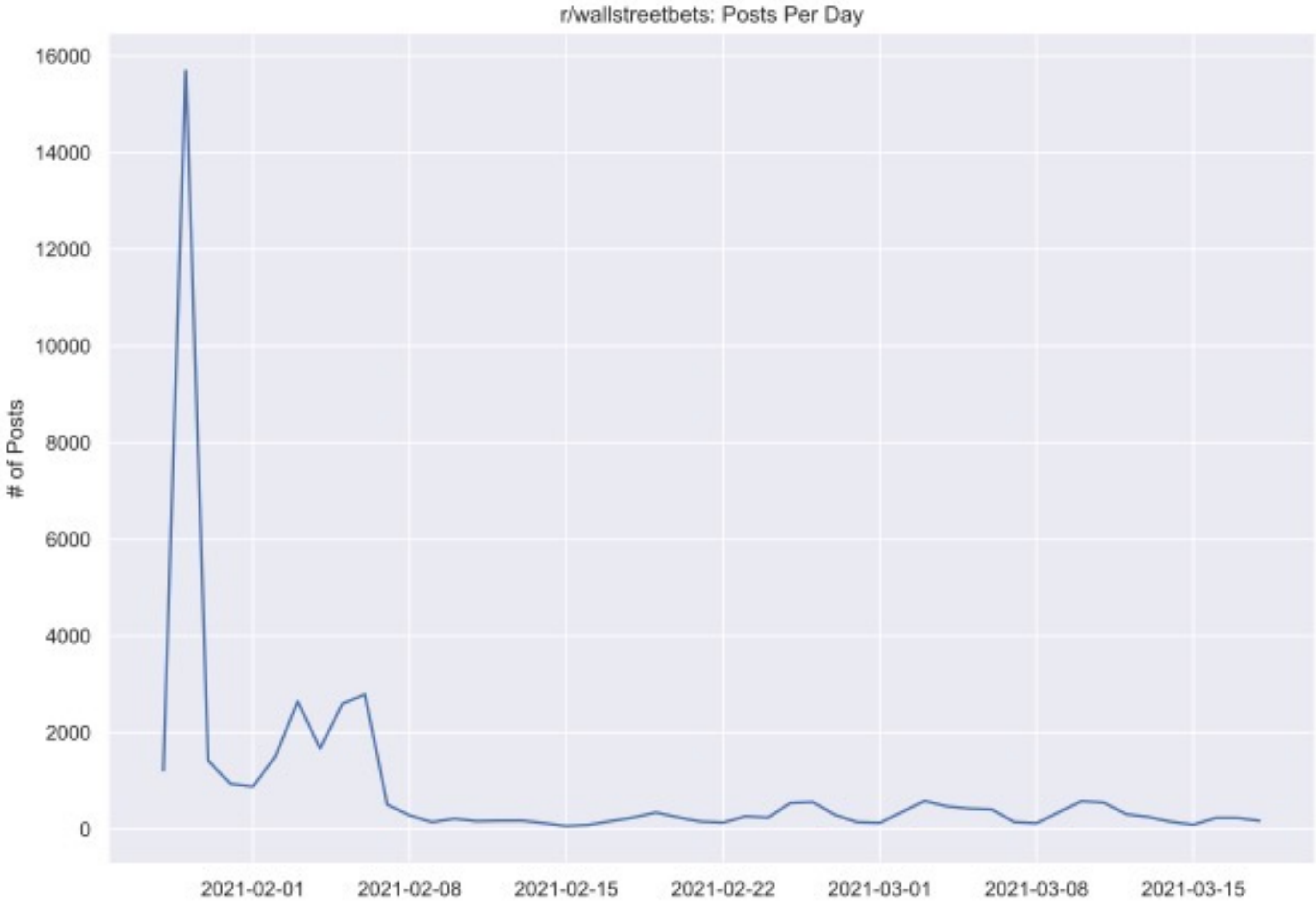
Dates

- Parse UNIX Timestamps
- Sync Timezones to PST

Data

- Initial frenzy followed by lull in posts
- Second wave of posts after initial bump
- Steady influx of posts after initial frenzy

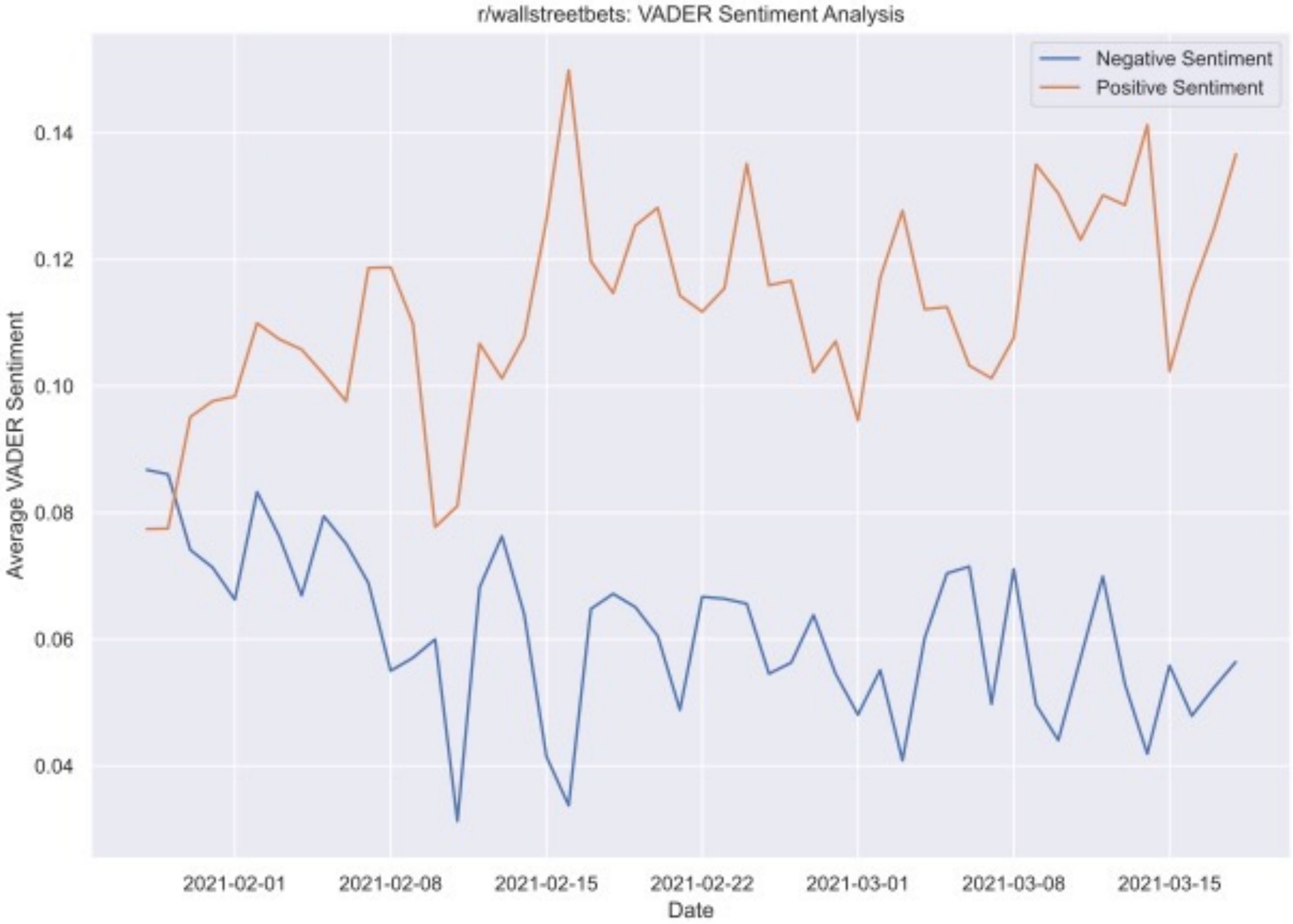

r/wallstreetbets: Posts Per Day of Week


r/wallstreetbets: Posts Per Day

VADER Sentiment Analysis

Valence Aware Dictionary and sEntiment Reasoner is a lexicon and rule-based sentiment analysis tool that is *specifically attuned to sentiments expressed in social media*.

Eg.: “Awesome” > “Good”, “Terrible” < “Meh”



Named Entity Recognition

Information extraction to locate and classify named “entities” form unstructured text.

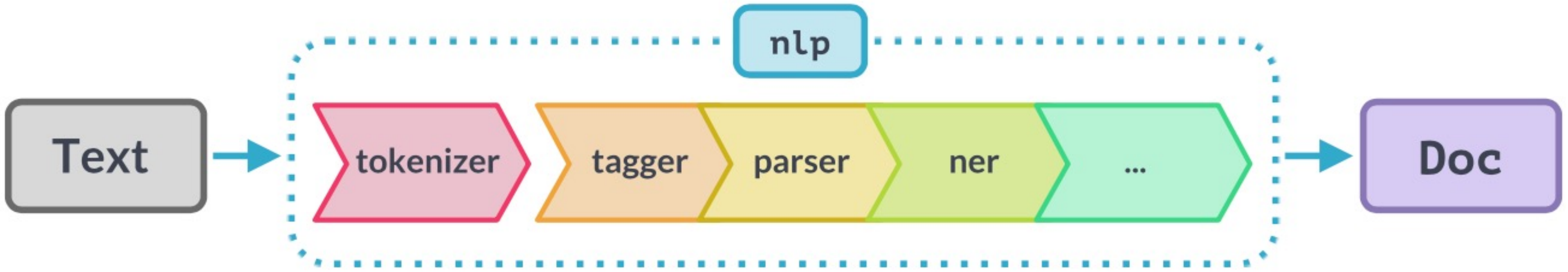
The economy has been showing signs of emerging from the pandemic crisis with renewed vigor, with spending picking up, manufacturing strengthening and employers adding workers. Hiring increased in February, with 379,000 added jobs — more than double January's total.

Credit card data from JPMorgan Chase showed that consumer spending jumped last week as the $1,400 checks that are going to most adults under President Joe Biden's $1.9 trillion emergency aid package began to be paid out. The Treasury says it has so far distributed 127 million payments worth $325 billion.

NER pipelines can recognize People, Organizations, Dates, Monetary Figures and more…

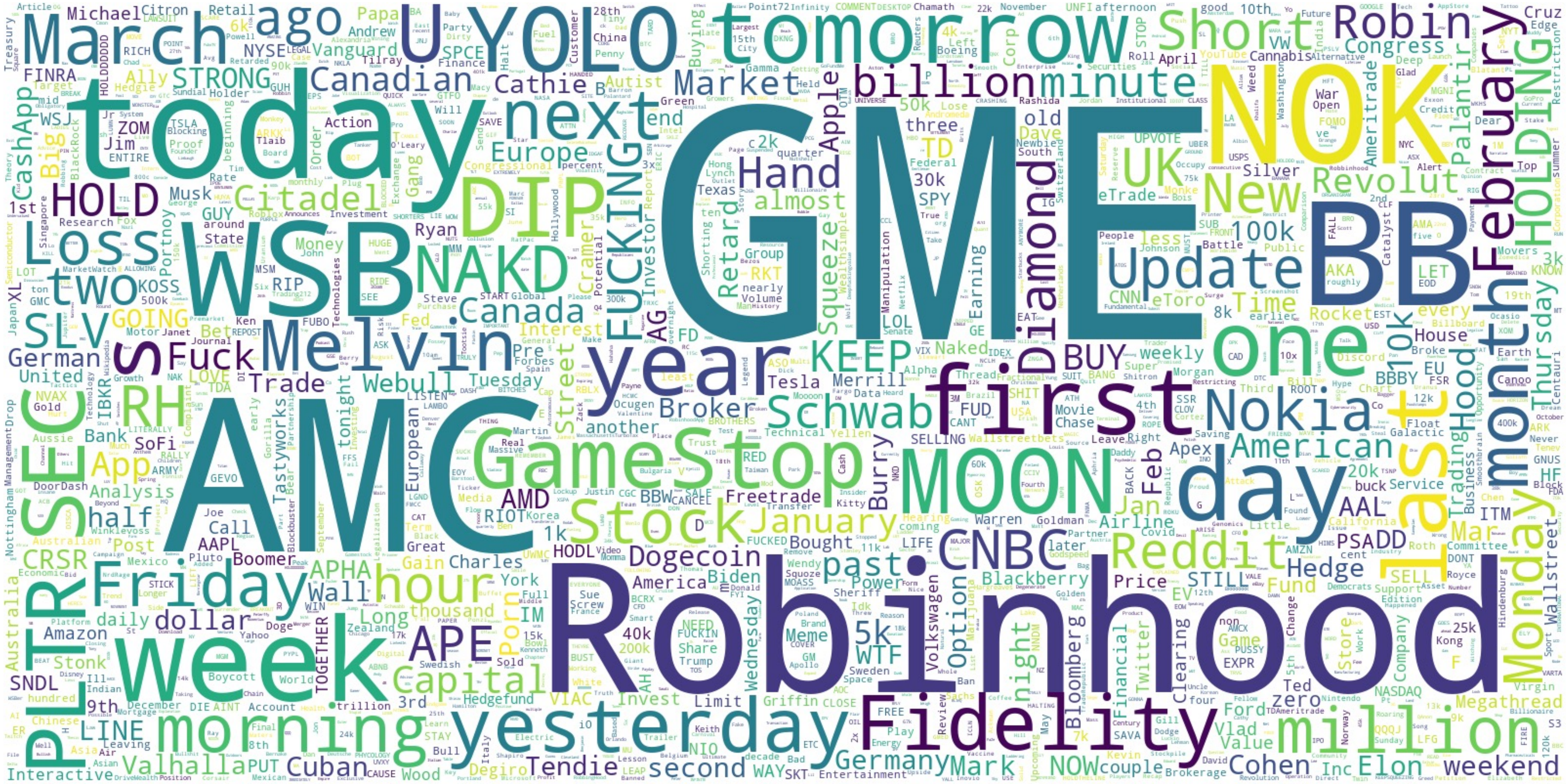
spaCy

- Open source NLP algorthms for Python
- Automated pipelines for multiple languages



- Provides pre-trained models to use on unstructured text

r/wallstreetbets: Named Entities



Results and Takeaways

Gamestop ($GME) was not the only security involved in the trading frenzy

Intense hype followed by lull in posts

Increasingly positive sentiment as “short squeeze” progressed

Cloud computing can make lengthy, impossible tasks into fast, expensive tasks

Challenges Faced

- So many posts, so little compute power
  - Pushing 42,000 through an advanced NLP pipeline costs time, or money
  - Tradeoff influences choice of hardware
  - 2014 Macbook Pro: 23 Days (Free)
  - Google Compute: 6 mins (200$)
- Reddit API
  - Forced to resubmit queries due to Reddit throttling requests
  - Inability to access historical posts, had to build script to run daily
- Specific trading lingo and sarcasm in posts

Questions?

Thank You!