Identifying precise forecasters

Liudmila Gavrilova

BrainStation Capstone Project – Sprint 1

Problem Statement

- Retail investment sentiment can be powerful
- Most likely wrong! Used as a contrarian indicator.

- But! Always look for exceptions from the rule!
- Whom to listen to, and whom to ignore

- Reading all the posts:
  - Extremely time-consuming
  - Subjective

- Solution: use ML and NLP techniques to look for "smart" social media accounts (the "Precise Forecasters").

Past Projects

Tried and tested:

- Identifying and measuring an aggregate investor sentiment
- Popular tickers trending on social networks
- Tweets by celebrities (Donald Trump, Elon Musk)

Novelty:

- Investment sentiment expressed by **specific identifiable social accounts**

Business Value

Value:

- Reduce noise to signal ratio. Curate content
- Augment decision-making process for various investors
- Allow machines to sift through information
- Build a real time trading strategy

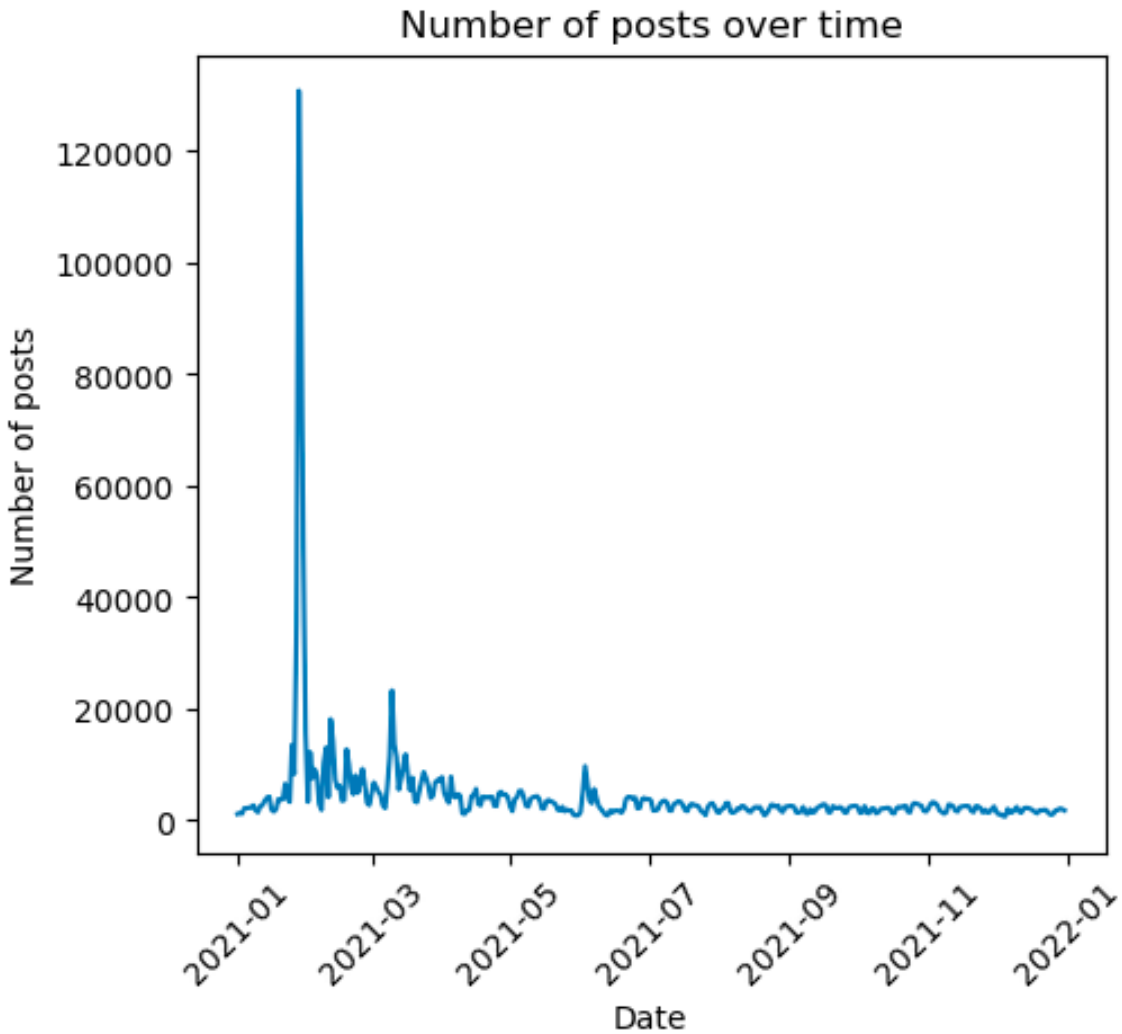
Result:

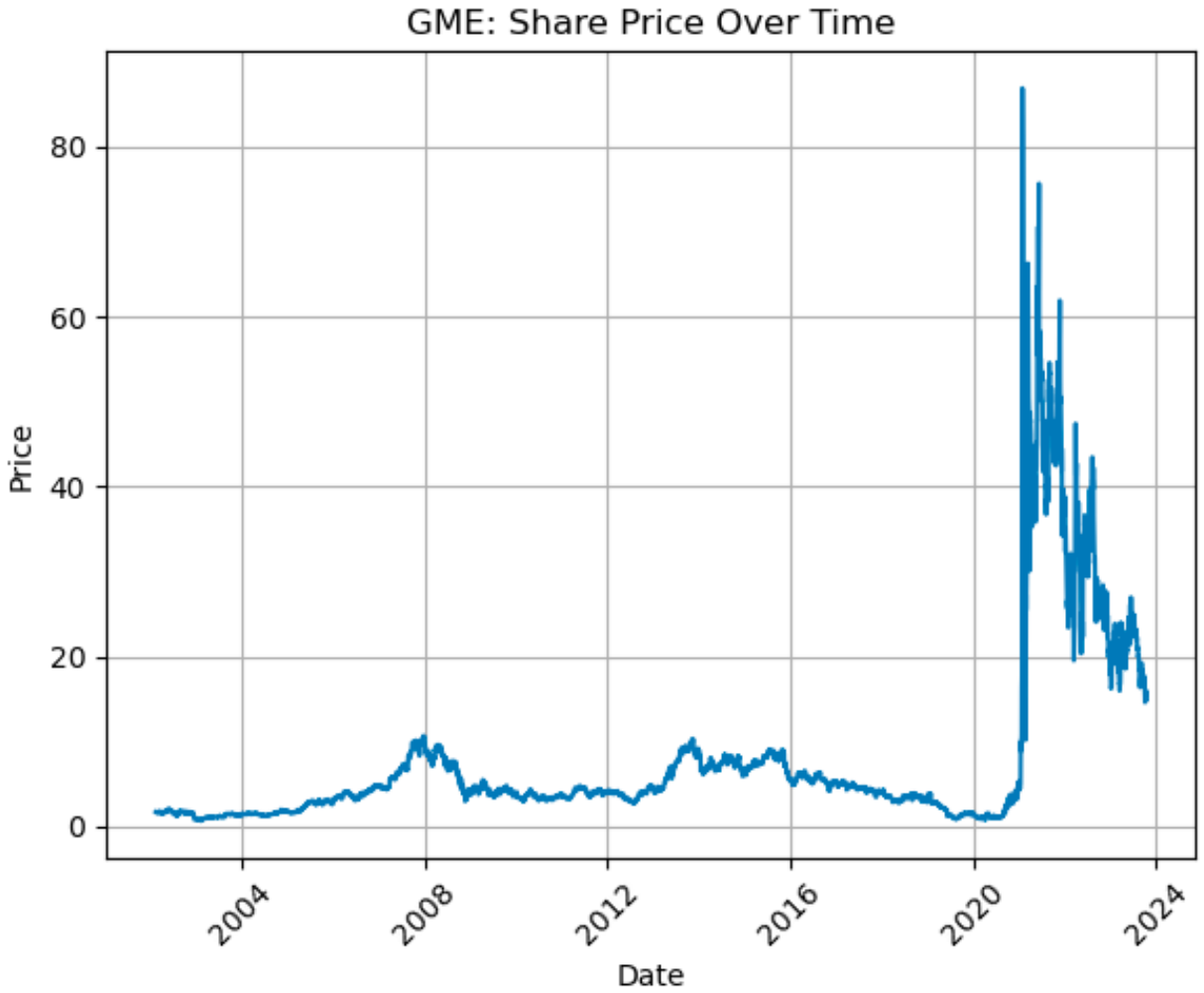
- Influence financial markets and make them more efficient in allocating capital

Beneficiaries:

- Investors, traders, financial analysts, all of us

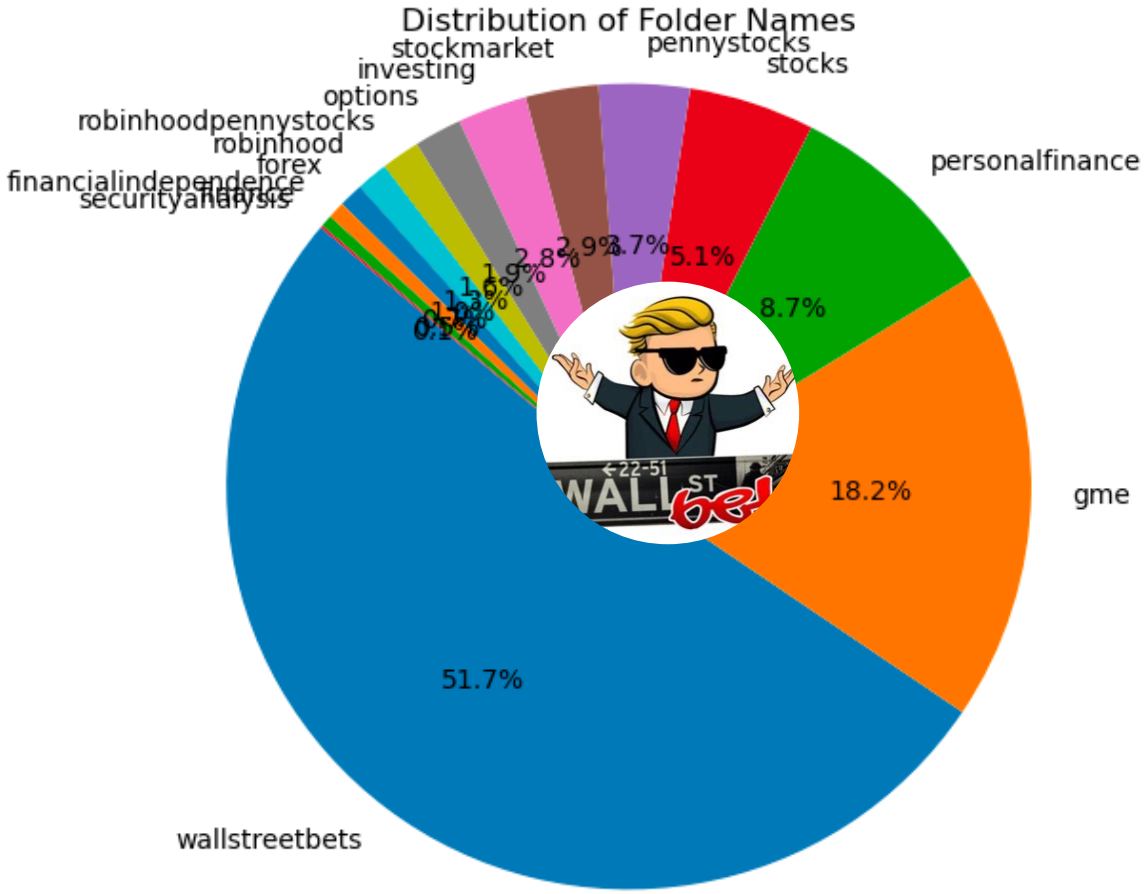
GameStop share price and Wallstreetbets subreddit





The Data Source

- Leukipp hosted on Kaggle
- 14 subreddits: WSB, GME, stocks, etc
- Overall: > 1.5m records
- WSB subreddit: 0.8m records, 24 columns
- 15% of authors names are deleted
- 50% of body text is missing
- Only 80% of the text is unique
- Unique authors: 600k
- Period: Jan-Dec 2021
- 303K posts with a proper text discussion, not just a picture or emoji posting.



EDA Preliminary Conclusions

- **No target variable. It needs to be extracted first.**
- 297k records available for long form textual analysis
- Top authors (0.01%) have **200-500 posts each** in 12m.
- Slang / own language (apes, diamond hands, paper hands, retards, HODL, degenerates).
- Emojis. 🚀, 🌙, 🦍, 🤞, 👐💎, 🚀🚀🚀🚀🚀🚀
- Ambiguity. Context.
- Topic extraction is an issue / SEC

Next Steps

Capstone Project - Wallstreetbets Oct02-Nov29

- Sprint 0 - Choose a topic Oct02-Oct06
- Find a dataset Oct09-Oct09
- Create GitHub, README, environment Oct16-Oct18
- Sprint 1 - EDA Oct18-Oct19
- Clean the text Oct20-Oct23
- Sentiment Extraction (VADER) Oct24-Oct26
- Vectorisation. Pre-processing Oct27-Oct30
- Logistic Regression Nov01-Nov03
- Portfolio of buys and sells Nov03-Nov07
- Sprint 2 - Baseline Modelling Nov07-Nov10
- Backtesting financial preformance Nov10-Nov13
- Refine Entity Resolution, Dependency, Part Of Speech etc. for each post Nov15-Nov20
- Instead of vectorise, try embeddings - word2vec Nov21-Nov23
- Sprint 3 - model optimization, evaluation and interpretation Nov24-Nov27
- Report, video and presentation Nov23-Nov29