
0.43%
-0.55%
24%
0.51%
0.1%
CREATED USING
POWTOON


Sentiment Analysis for Stock Price Prediction
Claire Sun
CREATED USING POWTOON

The main topics we are going to cover today

2020


U5ErkJggg==")base64,iVBORw0KG
0C1V9QD4NEOxs9xBQHQVCwSJF
hwVYBIdLn9vkLp79QcBCTDMiy3w2
QtfMBSO TRADING Uz4BDMle
MledhqOu/AzVSmzIUJKUz4BDMledI
/AzVSmzZ49CUjCC0yvim98iqtJT2L2
lQx8Q7hQYFek4AkixXFe1rsF84J/RT
CREATED USING
POWTOON

Everyone's a Day Trader Now -- WSJ

The #GameStop Saga Continues

-- Forb

The Stock Traders of Reddit and TikTok

-- The New York

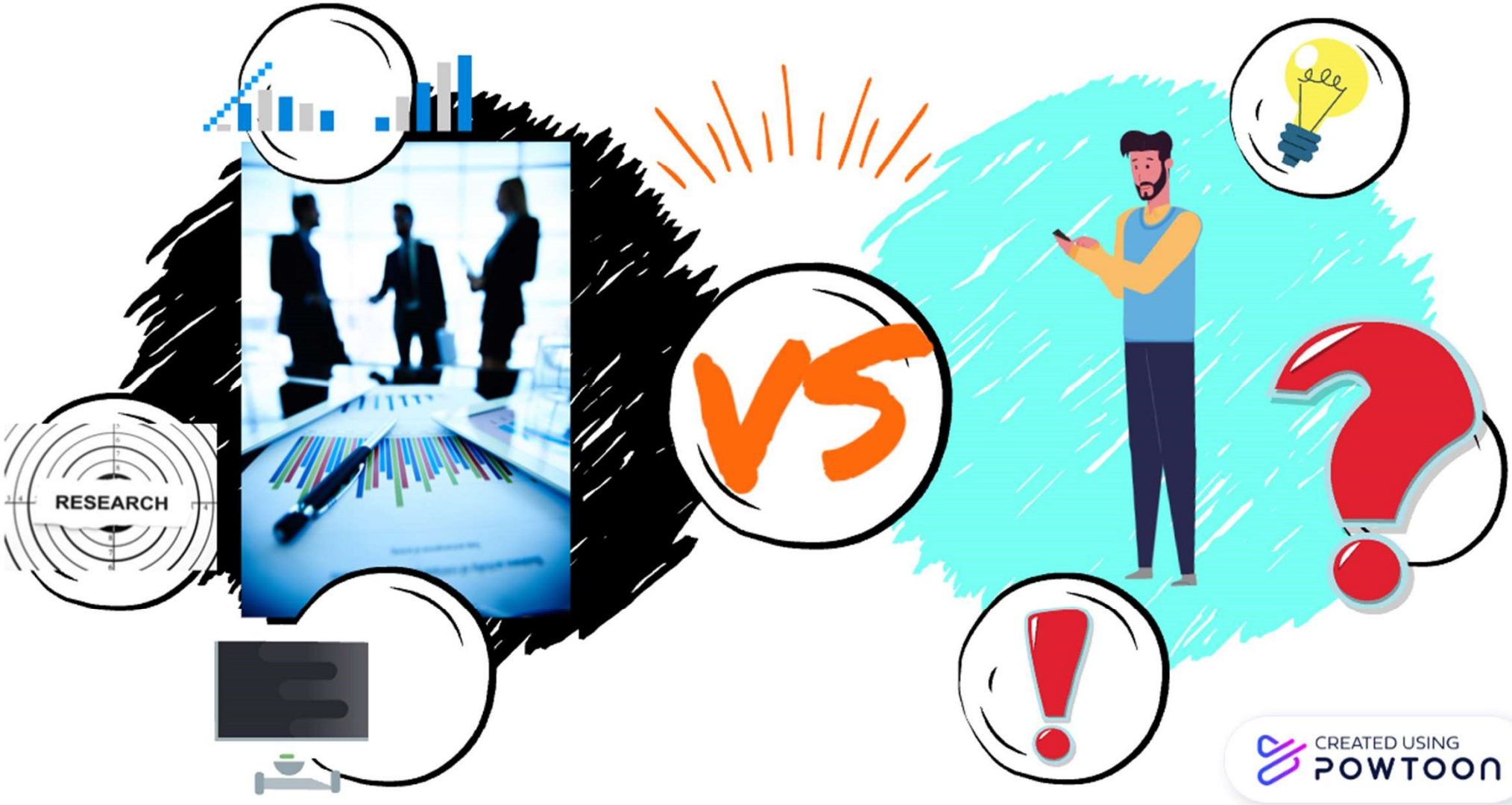
"Robinhood's Addictive App Made Trading a Pandemic Pastime"

-- Bloomberg

The

Wal

-- Fi


RESEARCH
VS
CREATED USING
POWTOON


Market Color
Investor Sentiment
Bullish
Panic
SELL
Strong Buy
Hold
CREATED USING
POWTOON

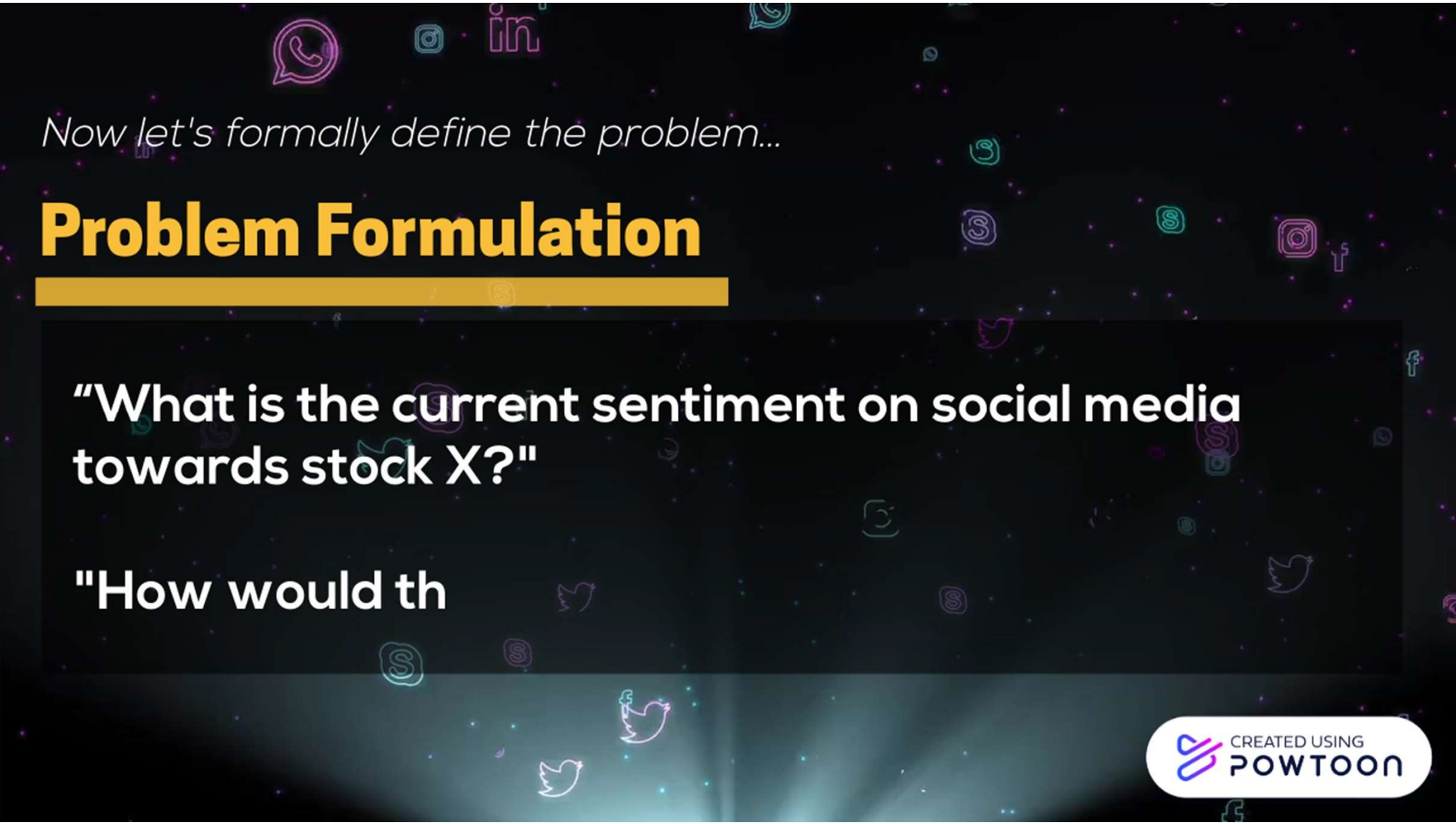
Two ~~Gaps~~

Opportunities

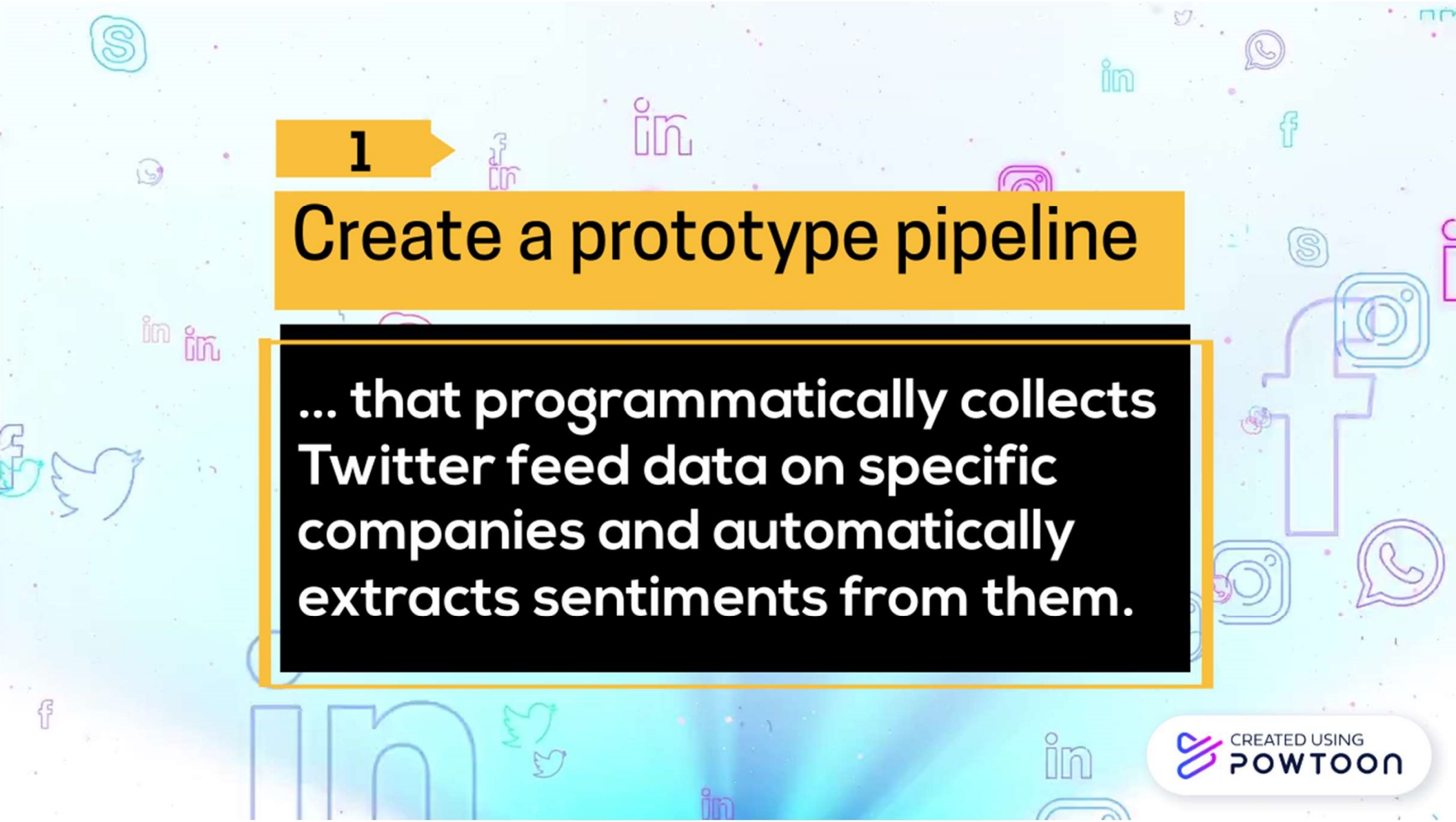
Identified...

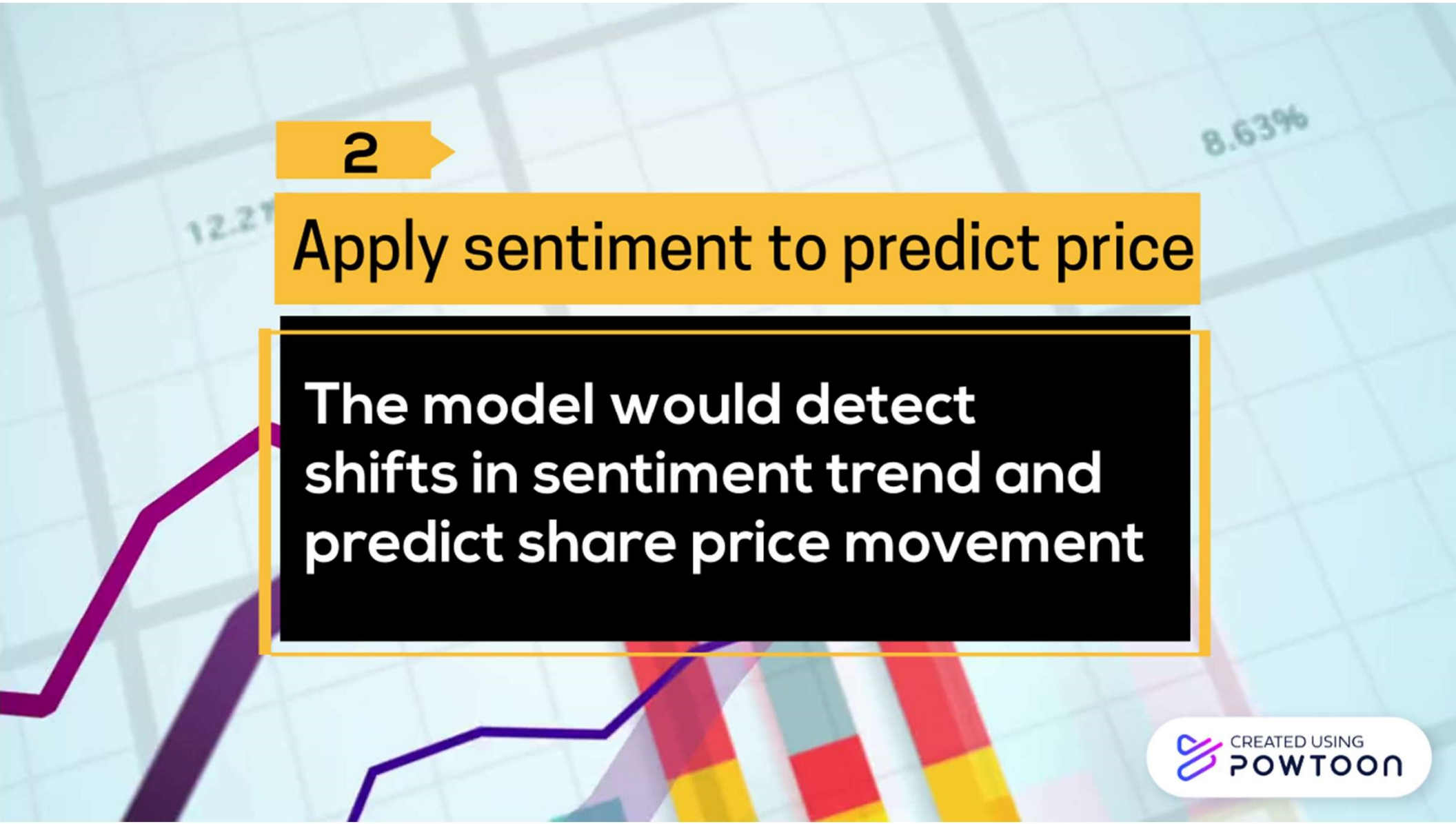

UNDERSERVED
RETAIL INVESTORS
CREATED USING
POWTOON

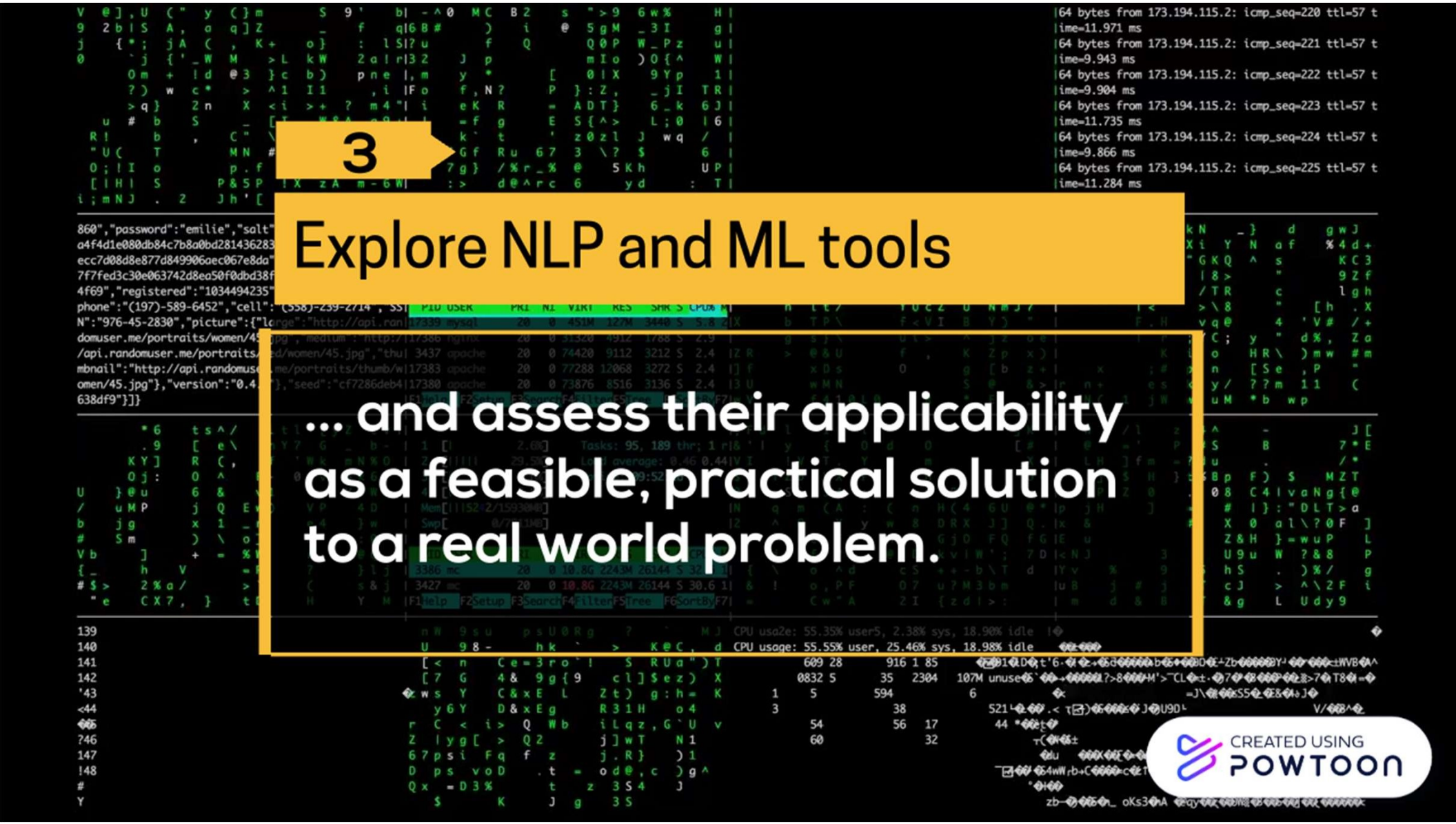

Institutional Investors
Social Media Sentiment
CREATED USING
POWTOON


Now let's formally define the problem...
Problem Formulation
“What is the current sentiment on social media towards stock X?"
"How would th
CREATED USING POWTOON

Project Objectives

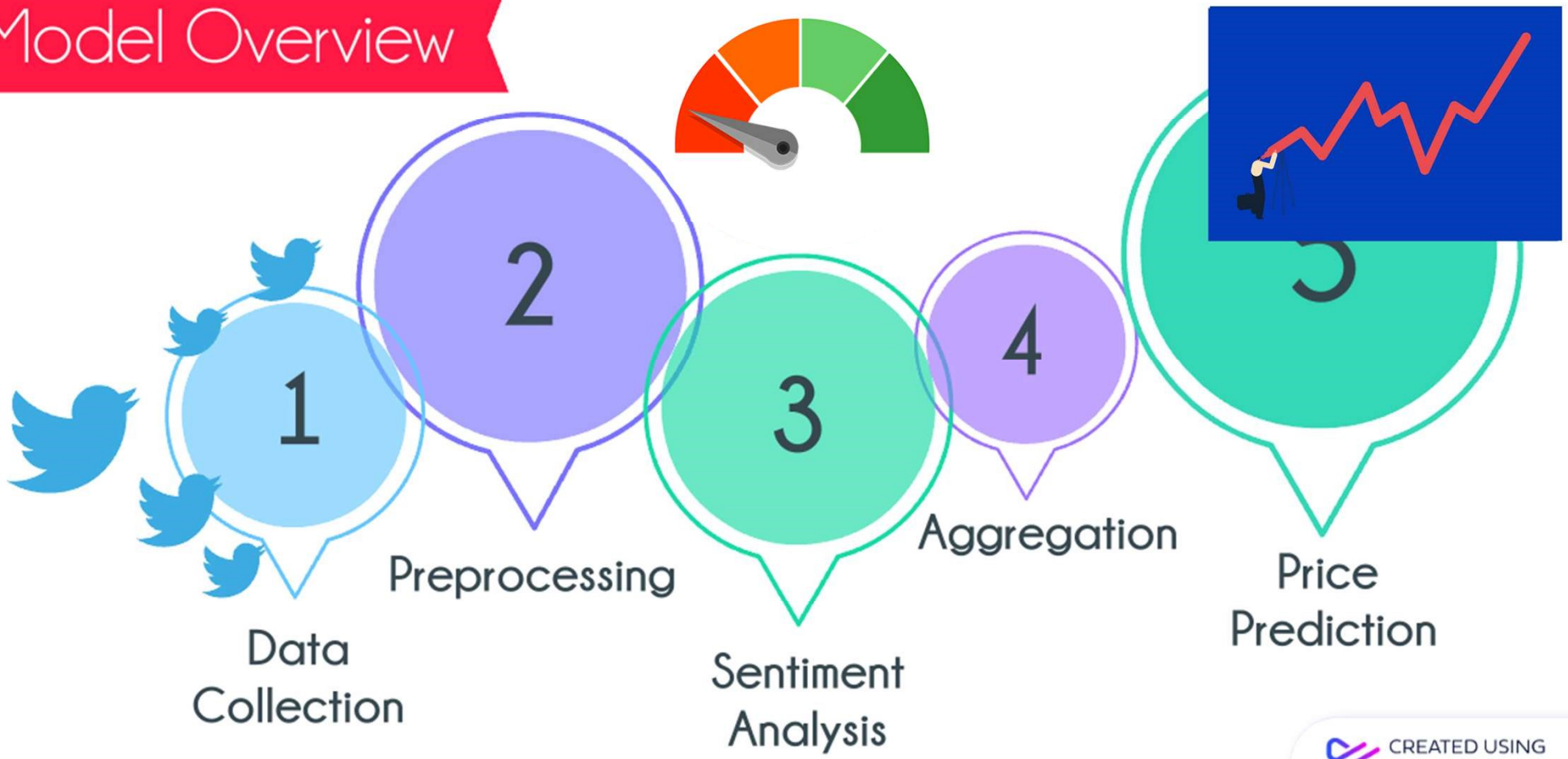

1
Create a prototype pipeline
... that programmatically collects Twitter feed data on specific companies and automatically extracts sentiments from them.
CREATED USING POWTOON


2
Apply sentiment to predict price
The model would detect shifts in sentiment trend and predict share price movement
8.63%
CREATED USING
POWTOON


3
Explore NLP and ML tools
... and assess their applicability as a feasible, practical solution to a real world problem.
CREATED USING POWTOON

Model Overview

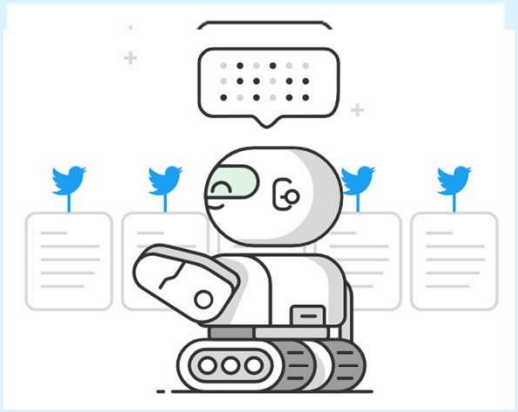
Model Overview



01

Data Collection

- Raw Tweets: Twitter API
- Annotated Dataset: SemEval 2016
- Share Price: Yahoo!Finance

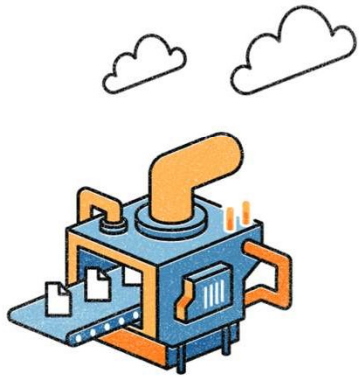


02

Preprocessing

- url, @username, #
- negation
- elongated words
- emojis and emoticons
- acronyms and slangs
- special characters (non-punctuation)

- spelling, stopwords, stemming



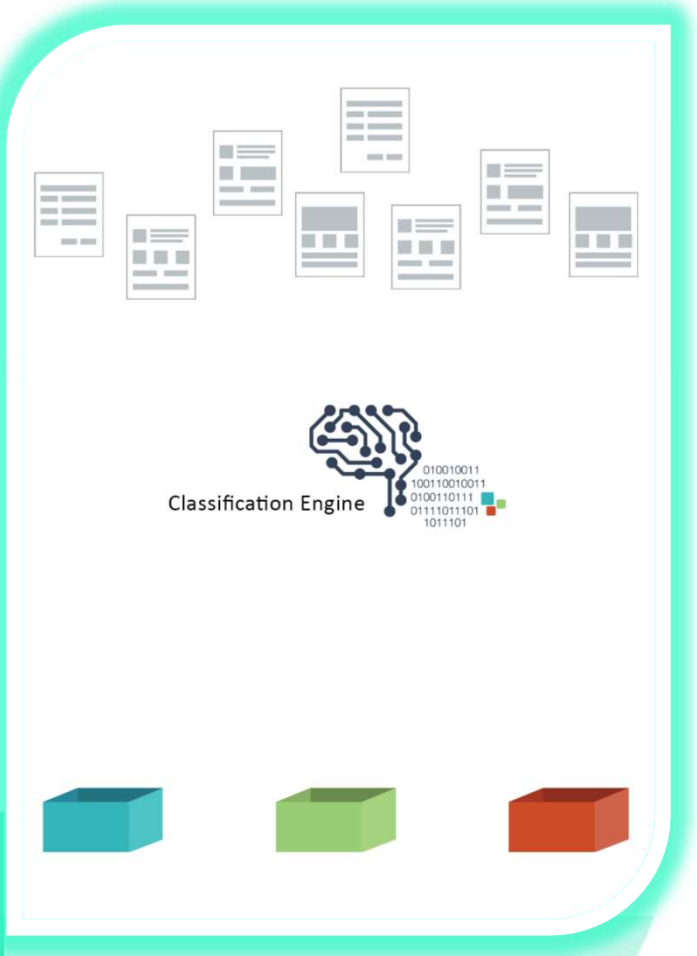
03

Sentiment Analysis

- Lexicon / Rule-based

- Classical ML-based:
  * Naive Bayes
  * Logistic Regression
  * Support Vector Machine

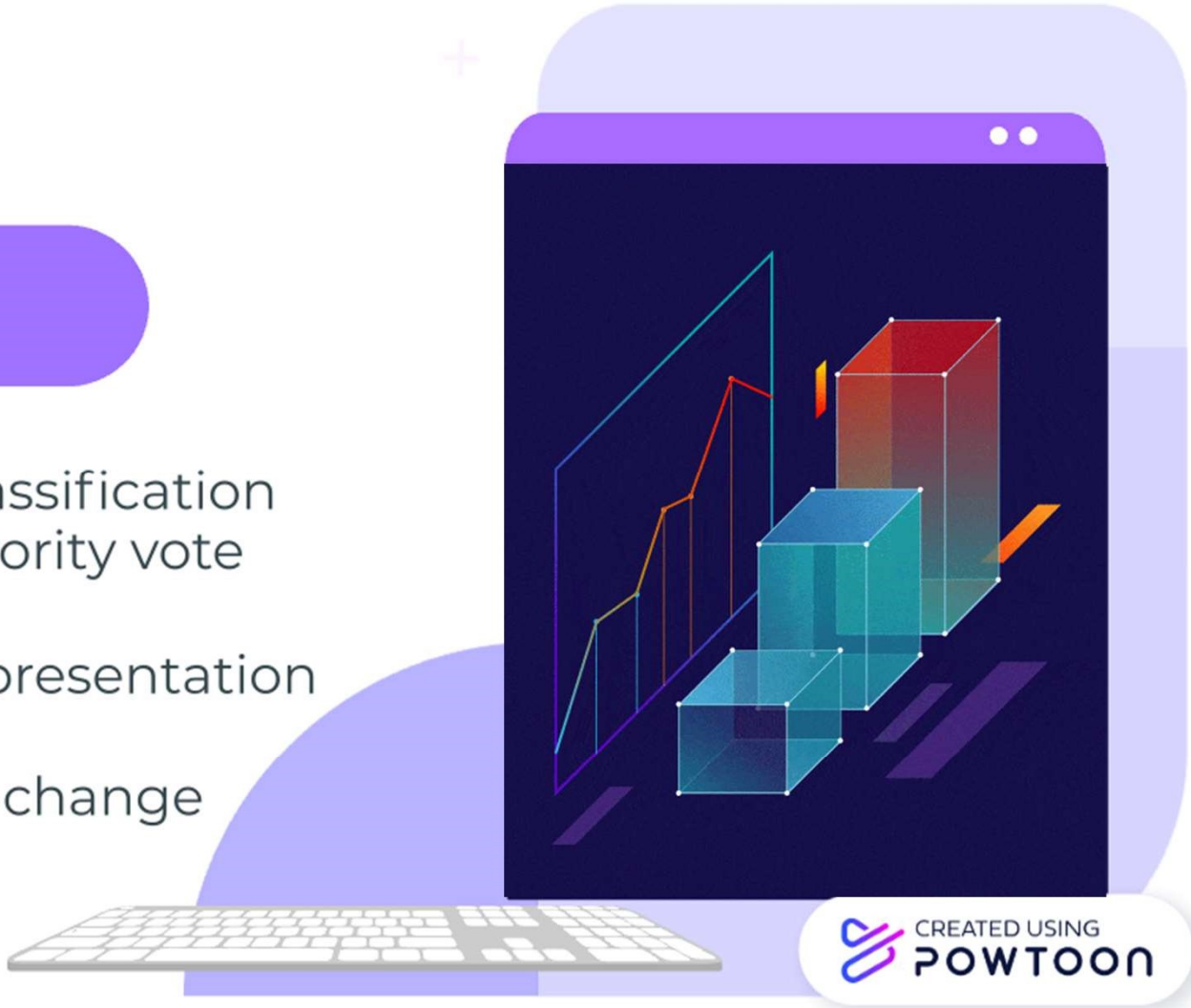
- BERT with fine-tuning



04

Aggregation

- Select best performing classification models; ensemble by majority vote
- Sentiment vector, shift representation
- Relative percentage price change



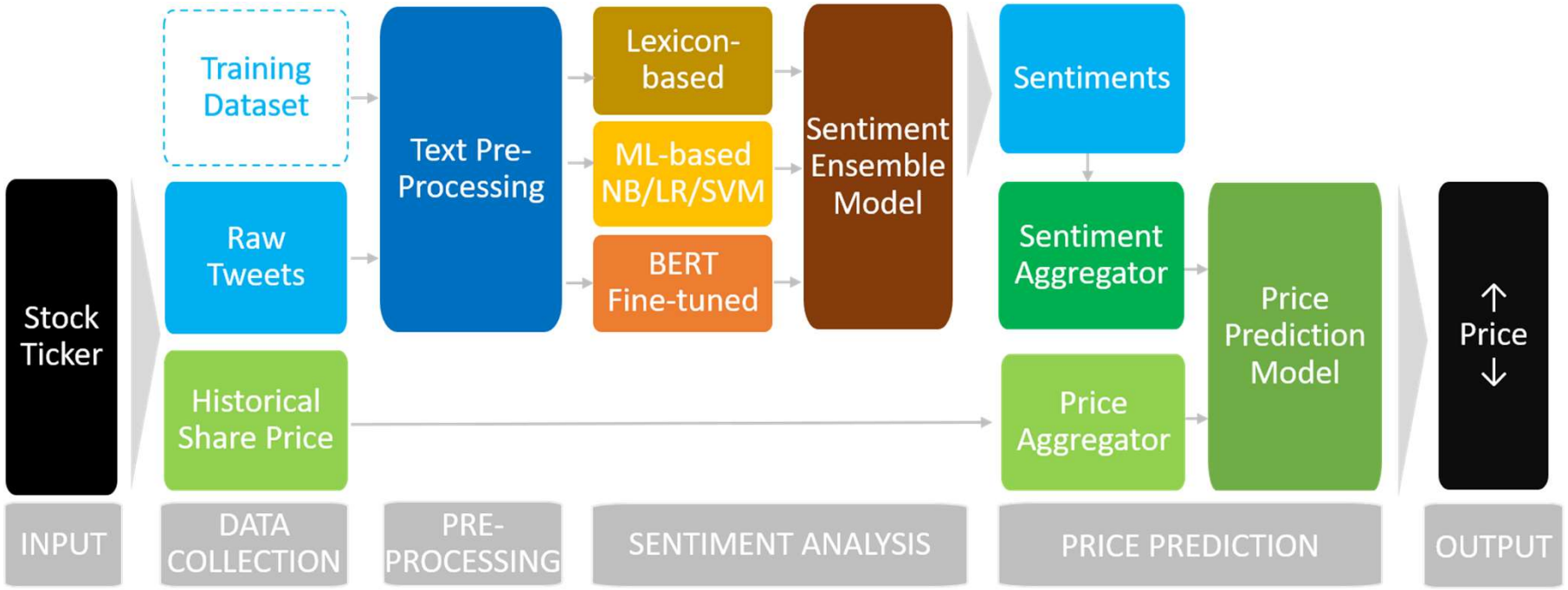
05

Price Prediction

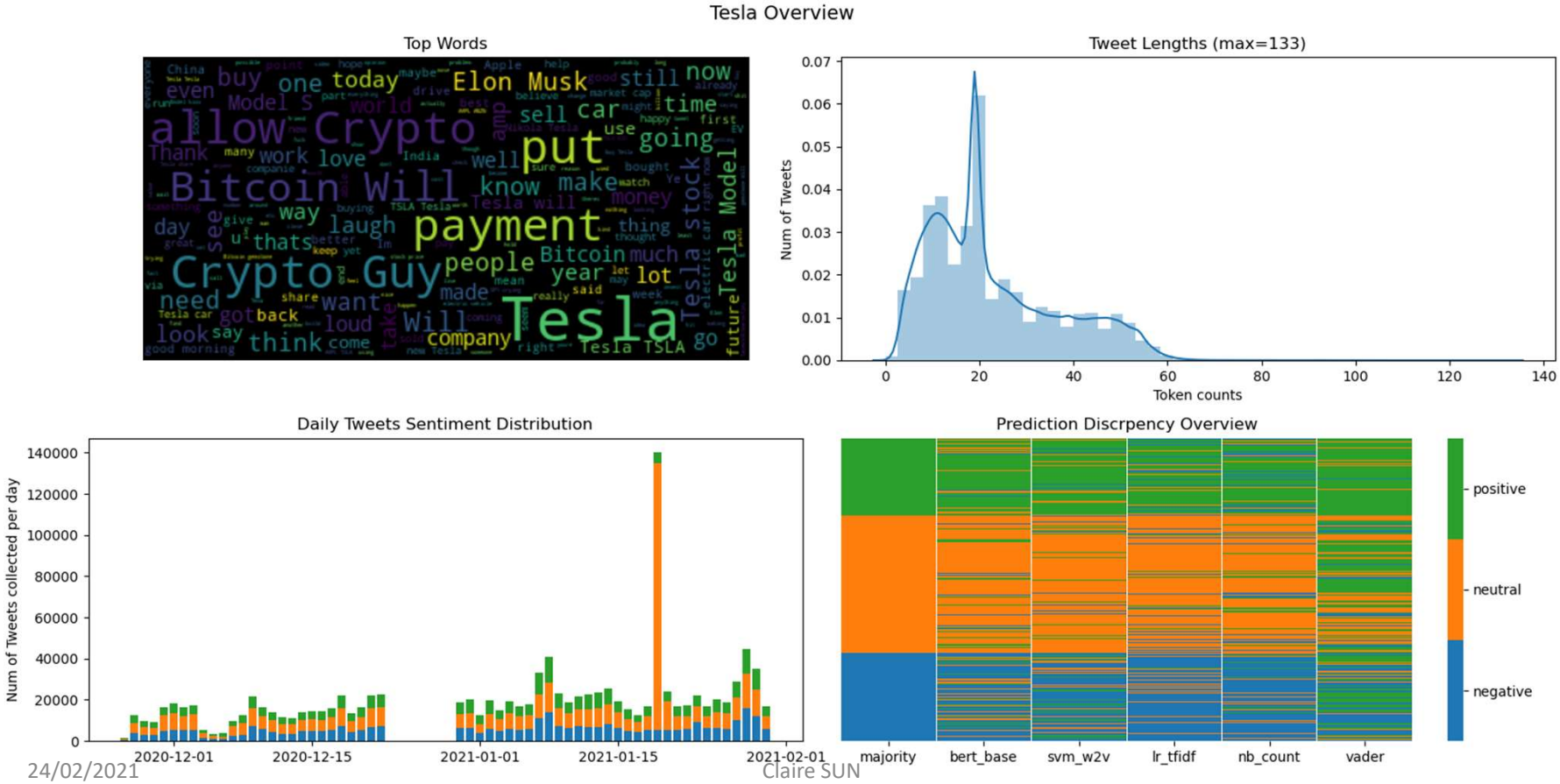
- RNN / LSTM / CNN
- Work-in-Progress
- Only 39 trading days

Experiment Setup

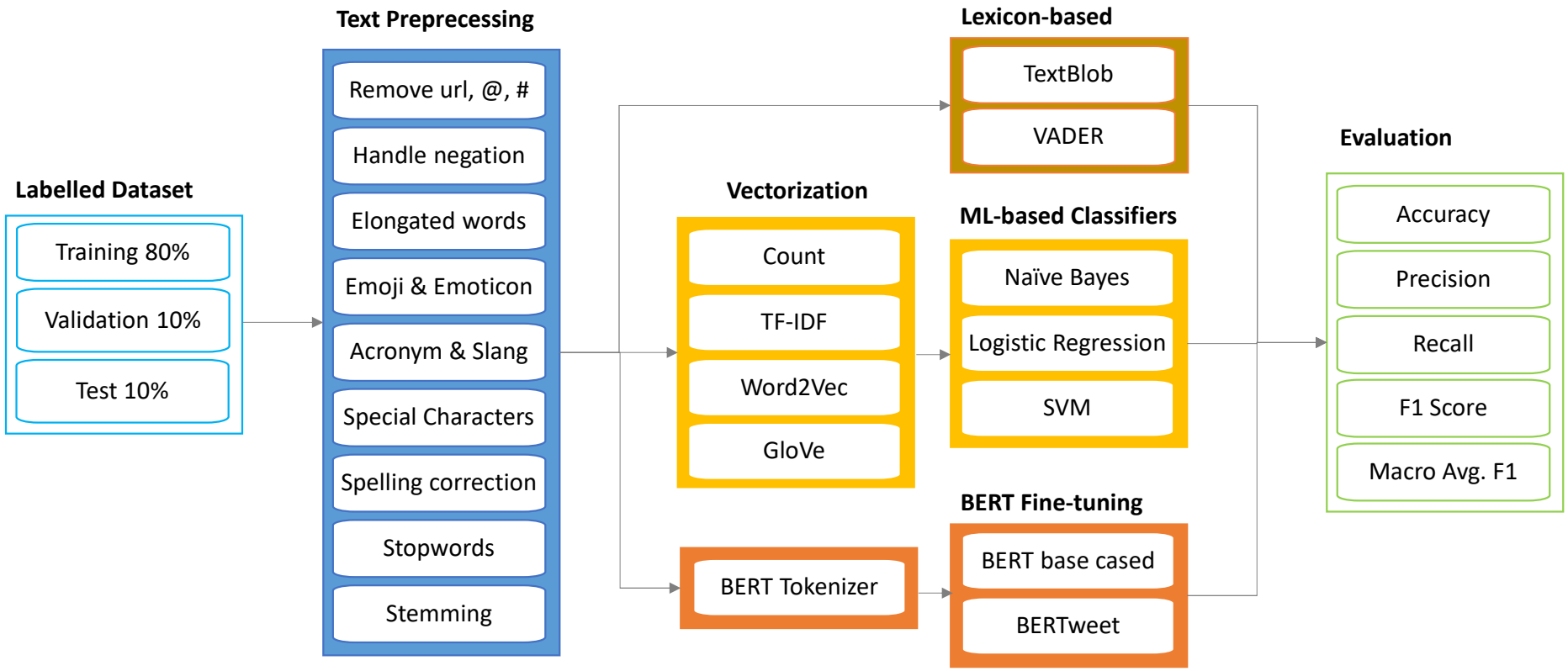
Project Planning



Twitter Data Visualization



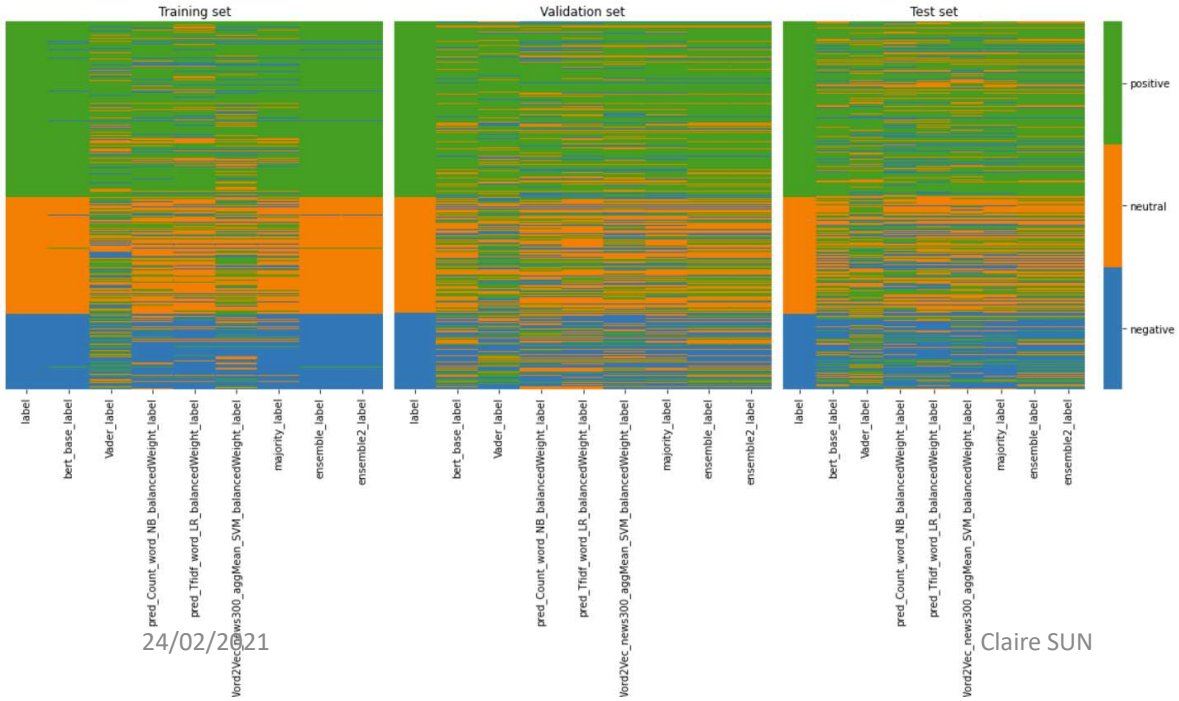
Experiment Setup



Model Training & Selection

Sentiment	Count	%Total
POSITIVE	9,388	47.8%
NEUTRAL	6,232	31.7%
NEGATIVE	4,033	20.5%
TOTAL	19,653	100.0%

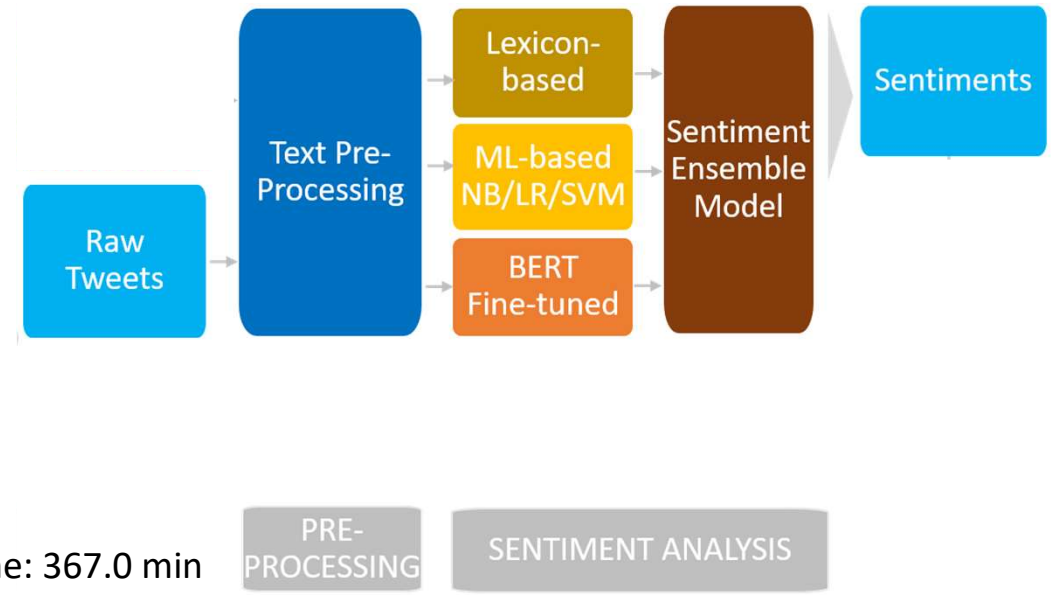
id	text	sentiment
634633252020555781	"b""HERE today, gone tomorrow- but still here! A short note on Nokia's patent deals with @Microsoft and @Alcatel_Lucent http://t.co/y5wFgUygFD"""	neutral
637308387353432064	b'@joebelfiore @GabeAul @Microsoft @Lumia @satyanadella plz add L1520 in the 1st wave of windows10 phones release.plz dont hurt ur diehardfans'	neutral



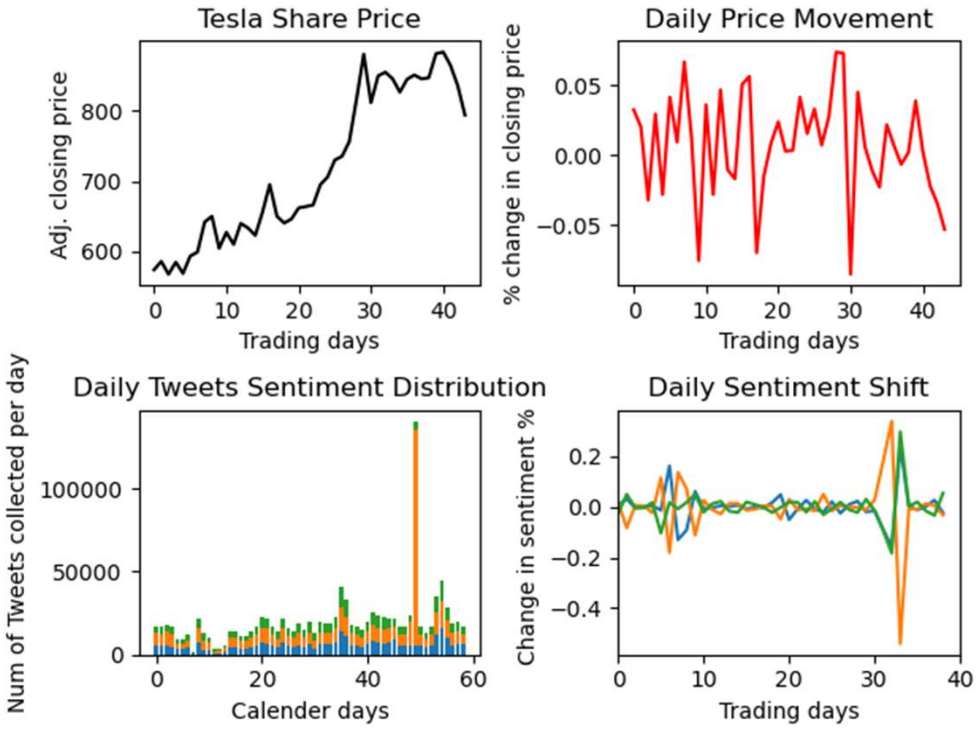
Classification Method	Macro Avgerage Precision	Recall	F1-score
Vader	**54%**	**54%**	**53%**
TextBlob	*48%*	*48%*	*48%*
Count_NB (SMOTE)	*58%*	*60%*	*59%*
Count_LR (SMOTE)	*55%*	*56%*	*55%*
Count_SVM (SMOTE)	*53%*	*53%*	*53%*
Count_NB (balanced_weight)	**59%**	**60%**	**59%**
Count_LR (balanced_weight)	*57%*	*58%*	*57%*
Count_SVM (balanced_weight)	*52%*	*53%*	*53%*
Count_NB_unigram,bigram,trigram	*57%*	*58%*	*57%*
Count_LR_unigram,bigram,trigram	*57%*	*57%*	*57%*
Count_SVM_unigram,bigram,trigram	*53%*	*52%*	*52%*
TFIDF_NB (balanced_weight)	*59%*	*58%*	*58%*
TFIDF_LR (balanced_weight)	**60%**	**61%**	**60%**
TFIDF_SVM (balanced_weight)	*57%*	*57%*	*57%*
TFIDF_NB_bigram	*52%*	*53%*	*53%*
TFIDF_LR_bigram	*52%*	*53%*	*52%*
TFIDF_SVM_bigram	*50%*	*50%*	*50%*
GloVe-Twitter-25_NB	*48%*	*49%*	*44%*
GloVe-Twitter-25_LR	*51%*	*53%*	*51%*
GloVe-Twitter-25_SVM	*51%*	*52%*	*51%*
GloVe-Twitter-200_NB	*51%*	*51%*	*48%*
GloVe-Twitter-200_LR	*55%*	*57%*	*55%*
GloVe-Twitter-200_SVM	*55%*	*56%*	*56%*
Word2Vec-News-300_NB	*51%*	*52%*	*48%*
Word2Vec-News-300_LR	*60%*	*62%*	*60%*
Word2Vec-News-300_SVM	**61%**	**62%**	**61%**
BERTweet	*29%*	*30%*	*20%*
BERT Base Cased	**64%**	**64%**	**64%**
Ensemble_FC_Softmax	*64%*	*64%*	*64%*
Ensemble_MajorityVote	**64%**	**66%**	**64%**

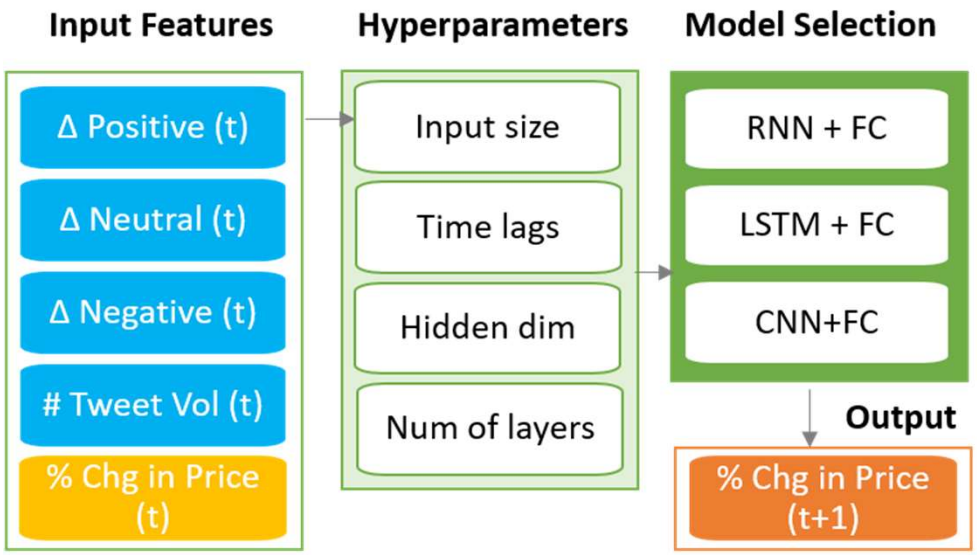
Processing Time

- ### Tesla ###
- Loading raw tweets from 9 files:
  >> loading: 4.896 s
- Start processing 1181205 raw tweets:
  >> text preprocessing: 1414.389 s
- Start sentiment analysis on 1181205 tweets:
  >> vader: 125.264 s
  >> nb_count: 18.192 s
  >> lr_tfidf: 18.299 s
  >> svm_w2v: 114.864 s
  >> bert: 19473.192 s (cuda)
  >> majority: 346.132 s
- Sentiment analysis completed! Total processing time: 367.0 min

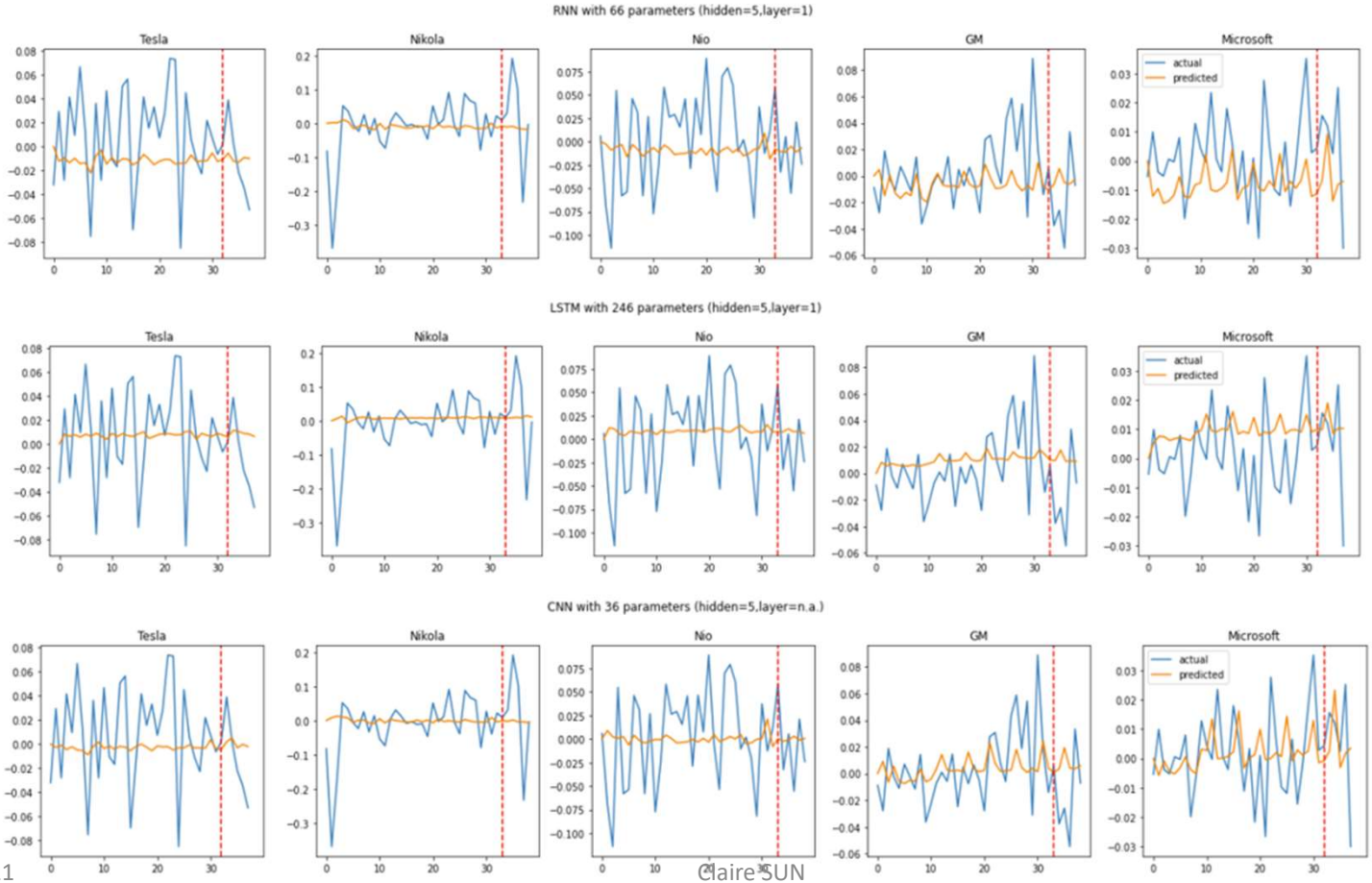


Price Prediction





Price Prediction



Demo

Input

Stock ticker: **TSLA** **Prediction date:** **2021-02-19**

Data Collection

- ✓ *Share price downloading completed in **1.8s***
- ✓ *Tweets downloading completed in **26.8 min***

Preprocessing

- ✓ *Start processing **50,159** raw tweets:*
- ✓ *Text preprocessing completed in **63.2s***

Sentiment Analysis

- ✓ *VADER Analyzer: **5.5s***
- ✓ *Naïve Bayes_CountVectorizer: **0.8s***
- ✓ *Logistic Regression_TFIDF: **0.8s***
- ✓ *SVM_Word2Vect: **17.4s***
- ✓ *BERT base cased: **915.7s***
- ✓ *Majority Vote: **17.1s***

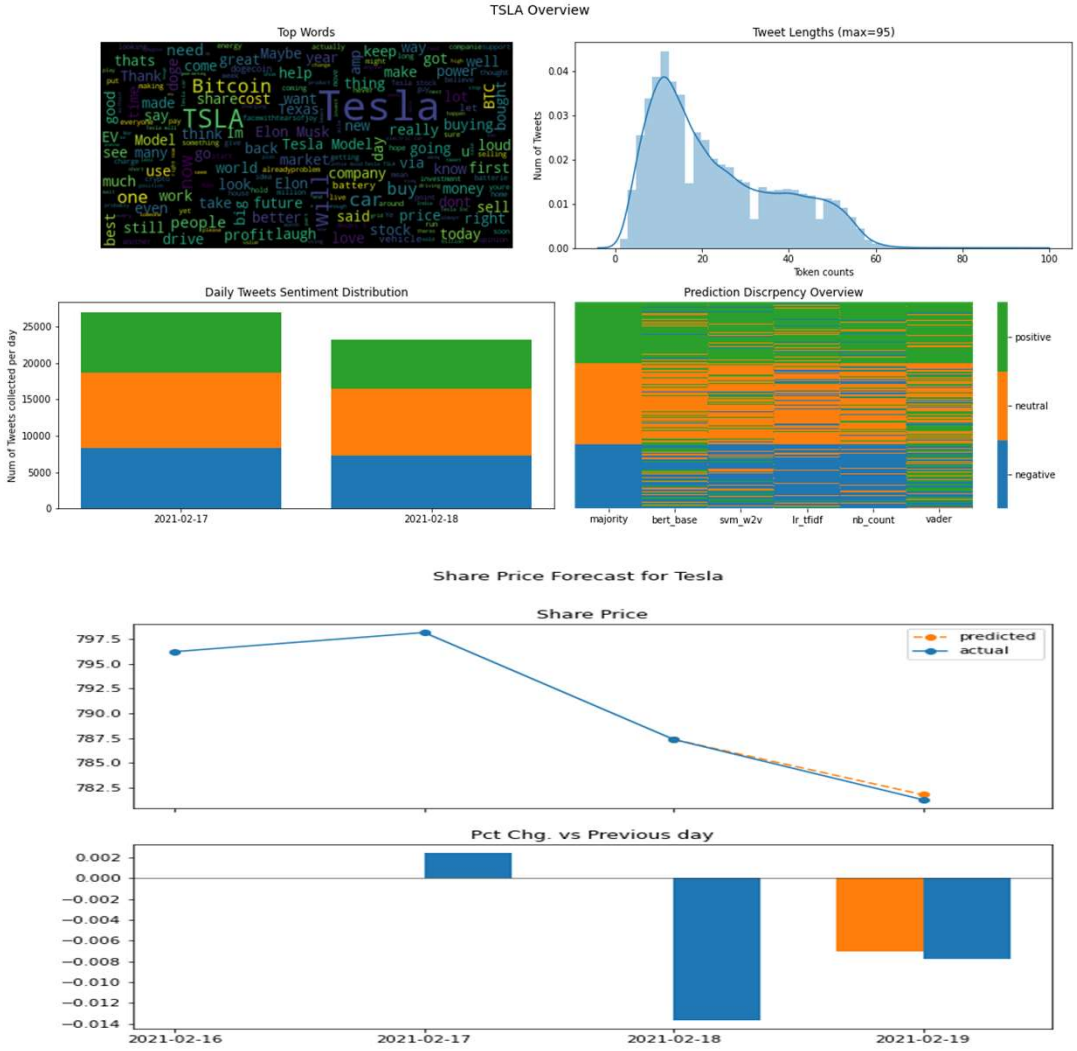
- **Sentiment analysis completed! Total processing time: 17.0 min**

Price Prediction

- ✓ *Share price prediction completed in **0.3s***

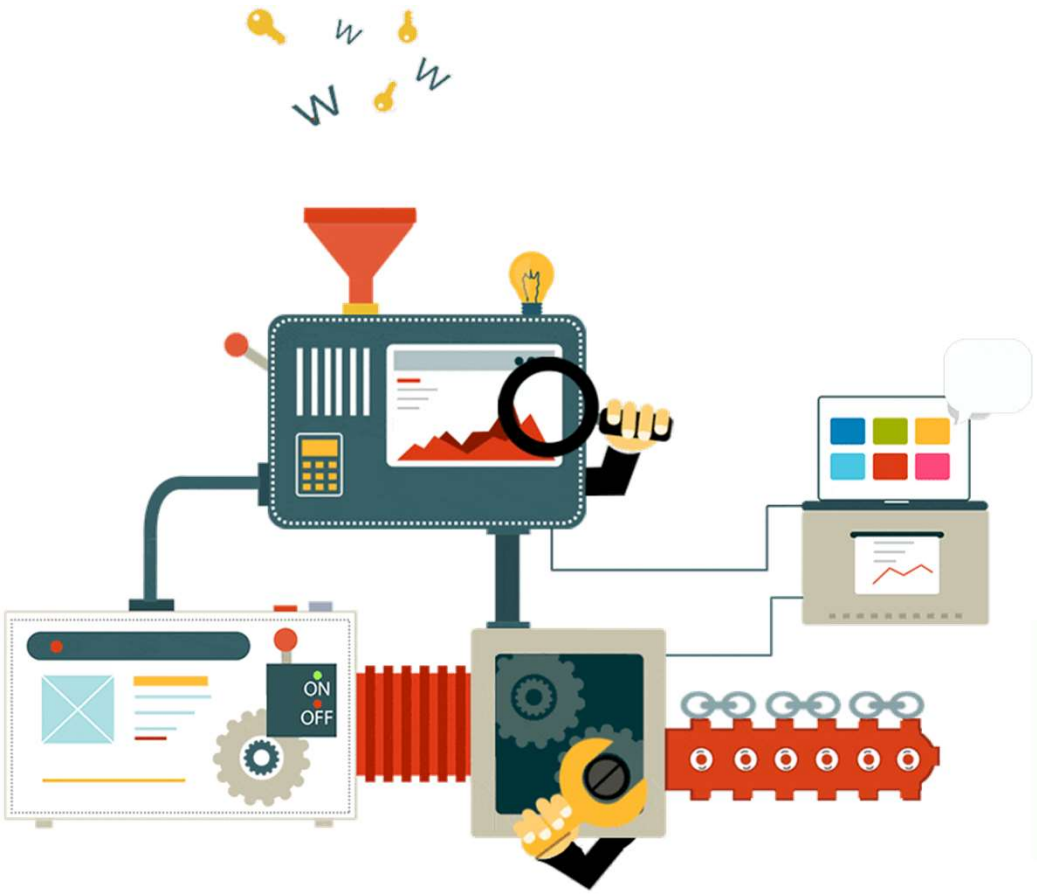
Output

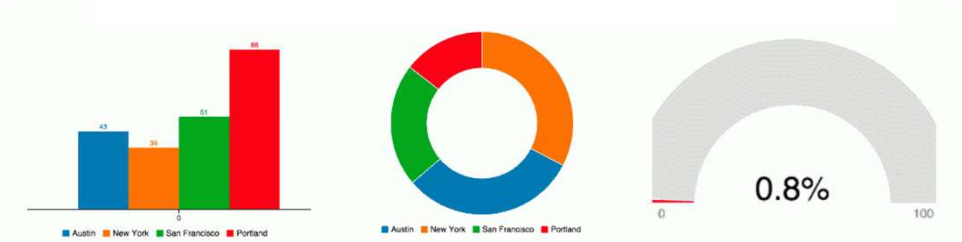
- **Predicted next day price movement for Tesla: *-0.71%***
- *Prediction: **$781.80**/share vs actual closing **$781.30**/share*



Challenges & Key Learnings

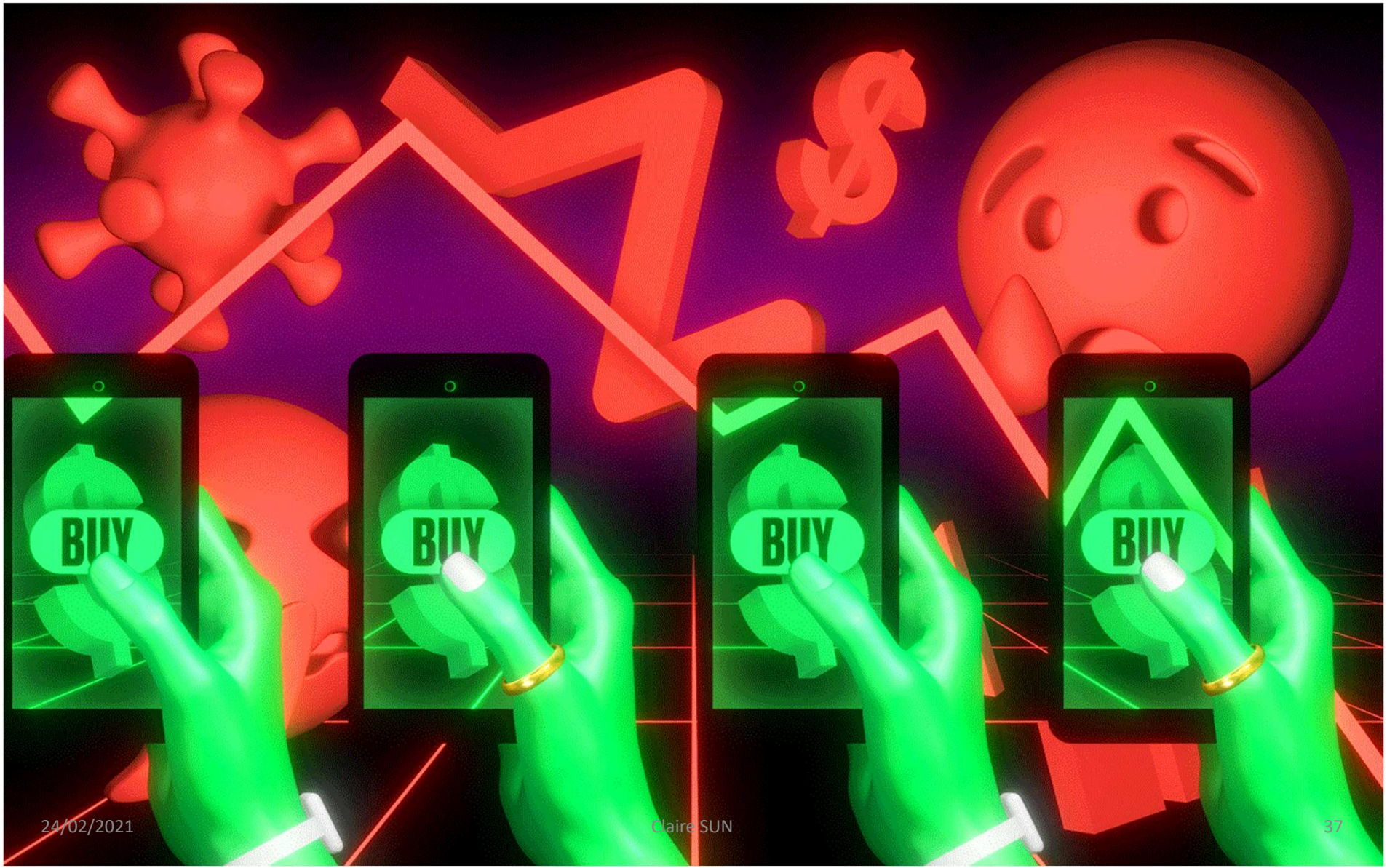
- **Pilot project to evaluate feasibility**
  - Demo is a toy example; price prediction model is still WIP
- **Domain-specific Training Dataset:**
  - Potential benefit from developing proprietary, annotated Twitter dataset vs. time/costs
- **Data collection over a longer period of time:**
  - Or subscription to Twitter's Enterprise API for some proper back-testing
- **Model speed vs accuracy:**
  - BERT model is c. 4-5% more accurate at sentiment prediction during training than NB/LR but c. 1000x slower at inferencing
  - Could potentially try other lighter-version of BERT such as DistilBERT
- **Potential improvement areas in the next iteration**:
  - Incorporating extra features: Not all collected data is consumed by the current model. For example, we have trading volume data as well as Twitter metadata that we could apply to potentially enhance the model architecture
  - Iteratively refining and optimizing the pipeline in order to accelerate processing speed and increase data-handling capacity
  - Improve design for user-interface


ON
OFF


43
34
51
86
0
Austin
New York
San Francisco
Portland
Austin
New York
San Francisco
Portland
0.8%
0
100

Conclusion

- From research findings into a real world problem solution
- End-to-end sentiment analysis and price prediction pipeline
- Key concepts and methodologies in NLP and ML
- Challenges of applying research techniques to a practical problem
- Improvement roadmap for the next iteration


BUY
BUY
BUY
BUY