# STOCK PREDICTIONS WITH STREAMING DATA REPORT

With News, Twitter & PowerBI API

IE MBD Oct 2020
Group E
NLP

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

# Agenda

## This presentation shows the result of our SPRA Model

1. Why Streaming
2. Business Case
3. High level Architectures
4. Key Consideration
5. Demo
6. Concept
7. Growth Strategy
8. Conclusion & Next Steps


SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

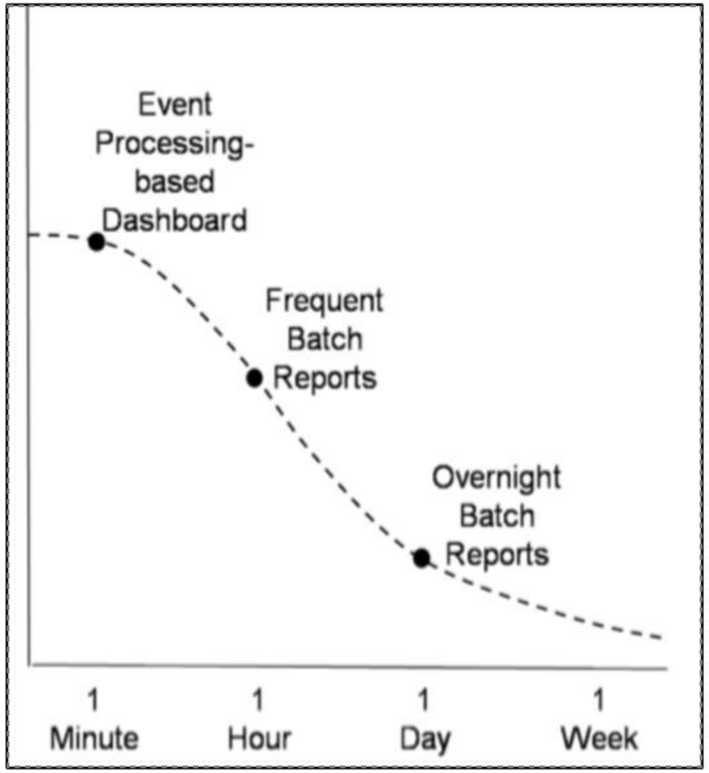
# 1. Why streaming

## Real time decision-making levels the value of information

### Information in real time

- Stock information is available at real time
- Fastest response time
- No need to store massive amounts of data & retain only useful parts

- Twitter, Newspaper, Reddit & much more info influences stock prices in real time
- 140 characters helps for quick analysis
- 130 million active users
- Personal data helps to segment




The value of information decreases over time, in this use case even **within 10 minutes**

### Information flood

**6.000**
tweets per seconds

**$8.2 Mn**
trading volume per day on Robinhood

**80% Daily Traders**
lose money on Etoro every day

# 2. Business case

## How will BELL monetize on its potential?

"In the stock market, information is money. Receiving the information first gives one a significant advantage over other traders. Thus, it makes sense that financial news have a great influence over the market."

### B2B

Integrations with banks or other relevant platforms depending on users + fixed cost

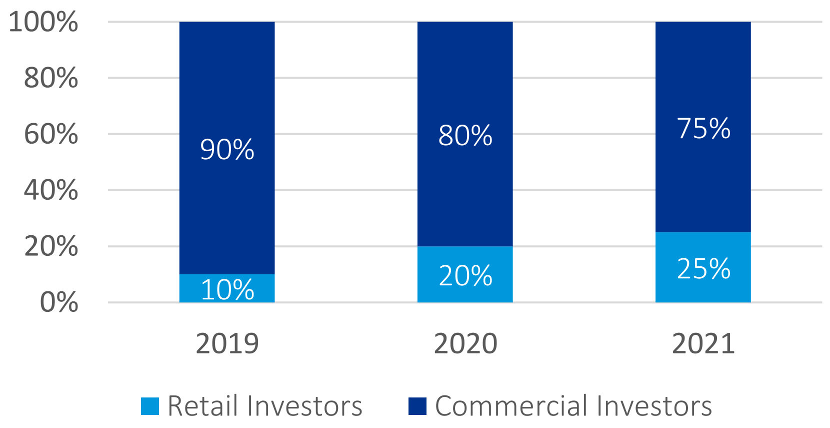
API Monetarization: depending on users in trading platform using it monthly subscription + the fixed cost

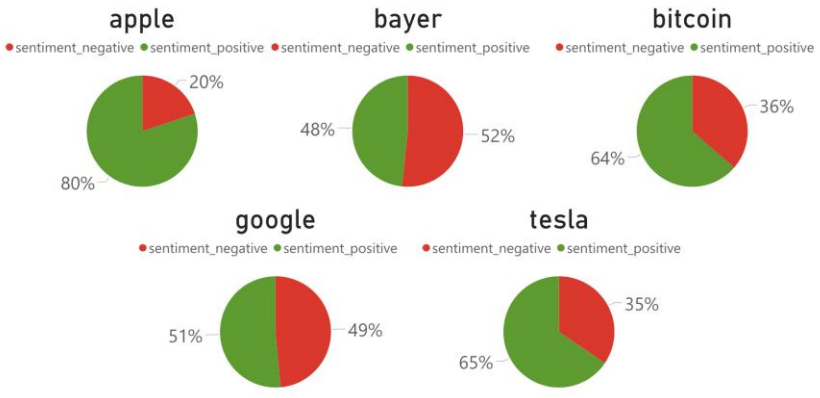
### B2C

3 stock information for free

Different bundles:

- Amateur (popular stocks: Nasdaq + Dax)
- Intermediate (amateur + 10 stocks of choice)
- Expert (all available stocks, unlimited stock price prediction access)



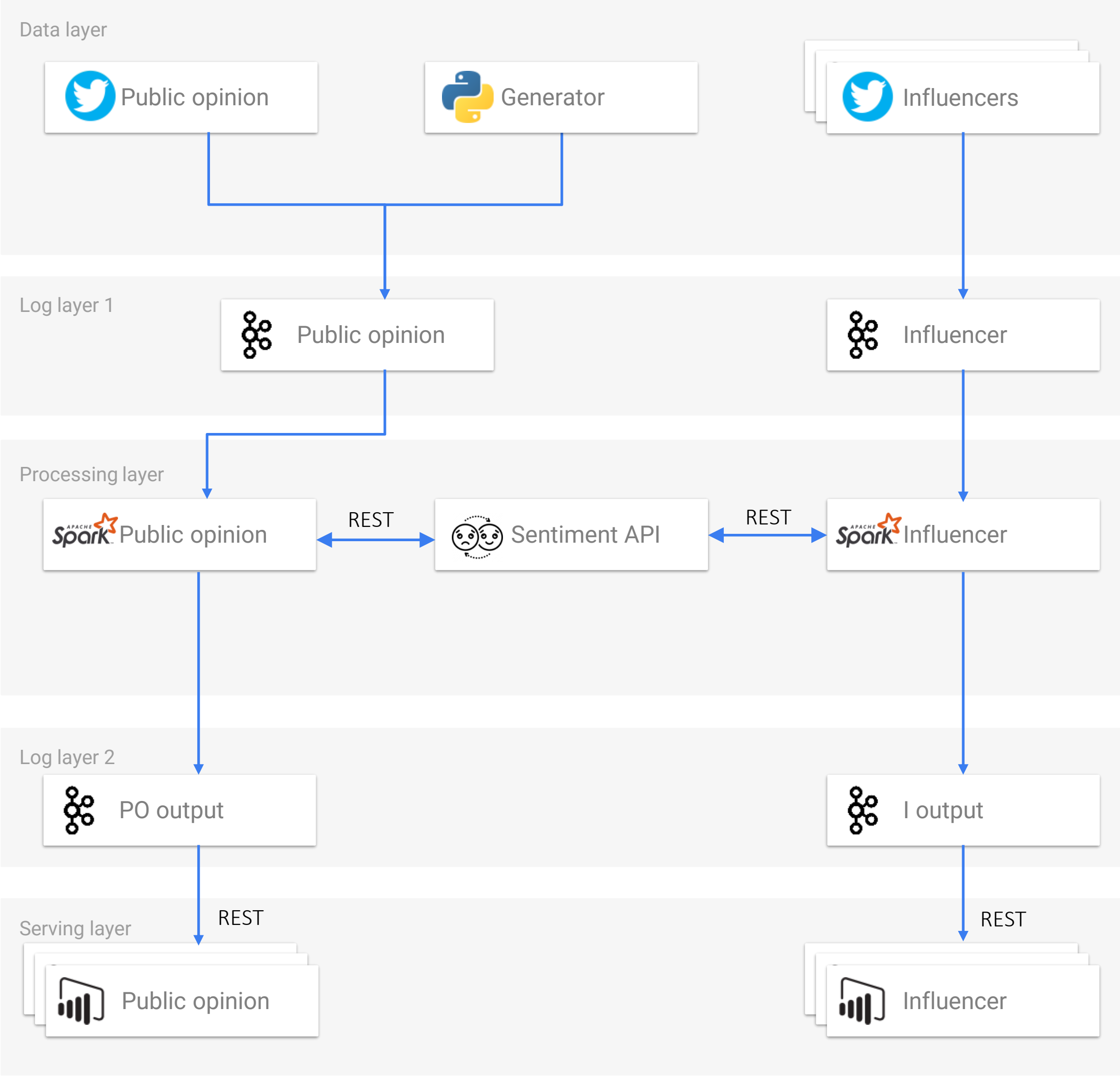




BELL's Dashboard

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

# 3. High level Architectures

## We created a Kappa architecture to benefit from less tools and no code duplication




- “All” twitter data is fetched to see what the Twitter community is talking about. Additionally, a generator is used to insert artificial tweets into the pipeline
- Tweets, responses and retweets of finance influencers (Elon Musk, Chamath) and a test user (Gianni) are fetched

- Tweets are filtered by equity (e.g. Tesla, Apple)
- Sentiment is extracted via Stanford sentiment API call
- Count, average sentiment are aggregated via a window functions

- Python consumer send data via REST request to Power BI

- Sentiment analysis is visualized
- Count and contribution is visualized
- Evolution of KPIs is visualized

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

# 4. Key implementation considerations & next steps

## We use many well-known streaming concepts. An additional implementation iteration could help us to refine parameters

### Data layer

- Public opinion
- Influencers
- Generator

**Technical aspects**

We receive 1% of tweets posted about Apple, Bitcoin, Bayer, Google, Tesla **via REST requests**

We created a tweet generator, so we don’t have to rely on the 1% tweets we get from Twitter

We created three producers for influencers to show case the **multi insert capabilities of a Kafka topic**

**Next steps**

- We want to insert real-time stock data to find correlations between these to streaming sources

### Processing layer

- Spark Public opinion
- Spark Influencer
- Sentiment API

**Technical aspects**

We use the **Spark Structured Streaming** API to benefit from an easy to write and read concepts

We use a **non-sliding 10 seconds window** to **aggregate** the number of tweets and the average sentiment which leads to **information dependency data.** Windows are based on **processing time**.

We use the **output mode “append”** to be able to work with aggregated windows

We send tweets to the Stanford **Sentiment API via REST request** to outsource the sentiment analysis

**Next steps**

- Spark streaming is only capable pf processing data in micro batches (<1 second). We should evaluate other tools like Storm or Flink (<100ms) to be sure that our latency is low enough for our use case
- We want to check if sliding windows and another window length will improve application
- We want to evaluate an in-house sentiment analysis to increase processing speed

# 5. Demo – public opinion

## We clean, transform, aggregate and visualize tweets to provide key social media insights for our customers

### Under the hood

**After Data layer**

We only keep the tweets itself…

```
tweet_str {"text": "@BitcoinMagazine The biggest advantage $BTC has
over $ETH is that you don't have to trust Ethereum people.
You don'\u2026 https://t.co/2L87z3GN5n"}
tweet_str {"text": "RT @icospeaks: Bitcoin PR Services and everything
related to crypto marketing.\nhttps://t.co/BqDmoWkSRk\n\n#cryptocurrency
#crypto #Cryptocurr\u2026"}
```

**After processing**

… We aggregate tweets …

```
+-------+------------------+------------------+-----------+--------------------+
|company| sentiment_positive|sentiment_negative|tweet_count|                time|
+-------+------------------+------------------+-----------+--------------------+
|  tesla|               0.4|               0.6|          5|2021-05-10 23:53:...|
| google|               0.0|               1.0|          9|2021-05-10 23:53:...|
|bitcoin|0.3571428571428571 5|0.6428571428571429|         14|2021-05-10 23:53:...|
|  apple|0.0833333333333333 3|0.9166666666666666|         12|2021-05-10 23:53:...|
|  bayer|               0.0|               1.0|          1|2021-05-10 23:53:...|
+-------+------------------+------------------+-----------+--------------------+
```

**Before data gets to PowerBI**

… We send the results via REST request to Power BI

```
[{'company': 'bitcoin', 'sentiment_positive': 0.25, 'sentiment_negative': 0.75,
'tweet_count': 16, 'time': '2021-05-10T23:58:34.711+02:00'}]
<Response [200]>
[{'company': 'google', 'sentiment_positive': 0.0625, 'sentiment_negative': 0.9375,
'tweet_count': 16, 'time': '2021-05-10T23:58:34.711+02:00'}]
```

### Visualization

- **Table**
- Sentiment Overview
- SO_Detailed
- S_Time
- Tweet Distribution
- TC_Time
- TC_Time/Company

| company | sentiment_positive | sentiment_negative | tweet_count | %tweet_count |
|---|---|---|---|---|
| apple | 0.78 | 0.23 | 125.00 | 19.53% |
| tesla | 0.65 | 0.35 | 153.00 | 23.91% |
| bitcoin | 0.63 | 0.37 | 169.00 | 26.41% |
| google | 0.51 | 0.49 | 124.00 | 19.38% |
| bayer | 0.49 | 0.51 | 69.00 | 10.78% |

# 5. Live Demo – financial influencers

## Finance Influencers highly influence market dynamics

Visualization

**Influencer_Overview**

Influencer_Detailed

| | ELON MUSK | | CHAMATH | | GIANNI | |
|---|---|---|---|---|---|---|
| | Tweet Count | Sentiment | Tweet Count | Sentiment | Tweet Count | Sentiment |
| APPLE | 469 | 0.70 / 0.30 | 129 | 0.34 / 0.66 | 1 | 0.66 / 0.34 |
| BAYER | 591 | 0.70 / 0.30 | 121 | 0.51 / 0.49 | 1 | 0.14 / 0.86 |
| BITCOIN | 176 | 0.53 / 0.47 | 136 | 0.43 / 0.57 | 1 | 0.87 / 0.13 |
| GOOGLE | 468 | 0.70 / 0.30 | 128 | 0.49 / 0.51 | 1 | 0.79 / 0.21 |
| TESLA | 445 | 0.65 / 0.35 | 125 | 0.42 / 0.58 | 1 | 0.64 / 0.36 |

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

# 6. Growth Strategy

## Who will we target?

### Value Proposition

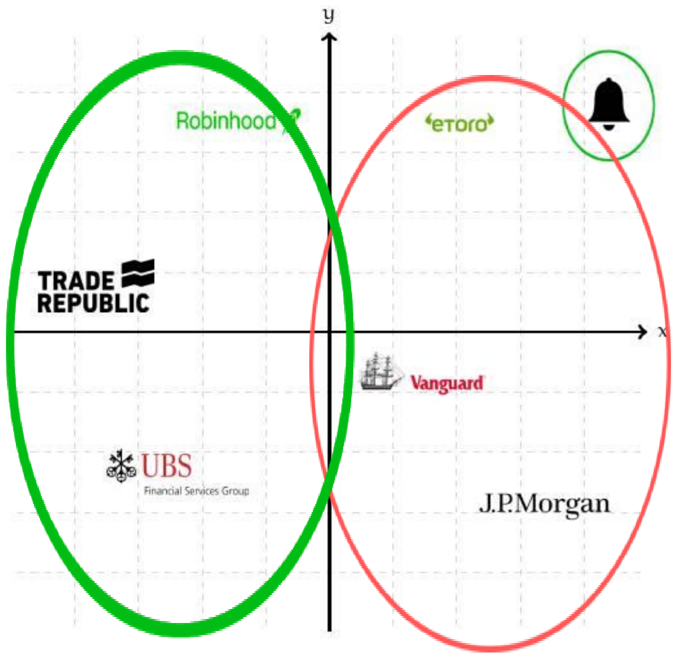
**Retail Investors**:

- Give more **live information** to the users to base their trades off
- Create a **benchmark score** which can help investors makes buy and sell decisions

**Institutional**:

- Create a **new category** of trading for large corporations
- Help them develop their presents in the **social media space**
- **Minimize losses** when large social media events take place

### Competitor and customer landscape




X-axis: The amount of information provided to clients

Y-axis: Barrier to enter

### Potential customer

**Robinhood:** Provide our service to the customers

**UBS:** Additional source of information for analysts

**Trade Republic:** Make our API available to their customers on their application

### Potential competitors

**Etoro:** Retail investing applications which allows user to copy portfolio in real time

**Vanguard:** ETF's which make it easy for investors to invest their retirement plans

**J.P.Morgan:** Recent acquisition of Etrade a retail investing platform

# 7. Strengths & Weaknesses

## What are the implications?

A lot of research papers highlighting there can be **price indicators** derived from chosen sources

POC: would have given correct prediction **in the past**

Give retail investors **more information** than they usually consume

Literature illustrates that Twitter data does have a **statistically significant impact** on volume and fluctuations in price of a stock

Multiple use cases for further development

Opportunity cost: Indicator might arrive **too late** and price already fluctuated

Market volatility: dependency on market behaviour, which is **unpredictable**

Wrong predictions can always happen

Meme stocks can be **further pushed, retail investors** can be harmed by following the crowd

**Trust** of customers need to be gained

More time needed for training and testing

# 8. Conclusions & Next Steps

## What did we find and how can it be improved?

*"Making educated decision easy"*

### Conclusion

Proof of Concept achieved

Timing is important for Twitter and stock data

Potential Customers & Monetization

### Next Steps

Implement more streaming sources

- Reddit and/or Facebook API
- Real-time stock data

Tech: Evaluate and improve tool set

Vertical Integration

- Other Revenue Streams
- New Marketing Strategies derivation

Thanks!

IE MBD Oct 2020
Group E
NLP

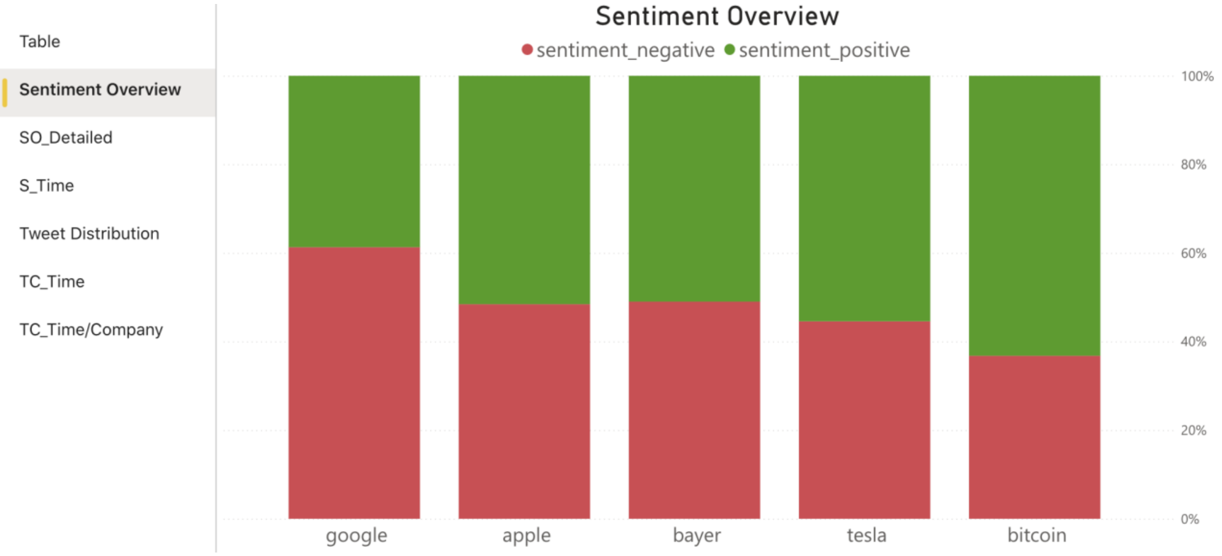
SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

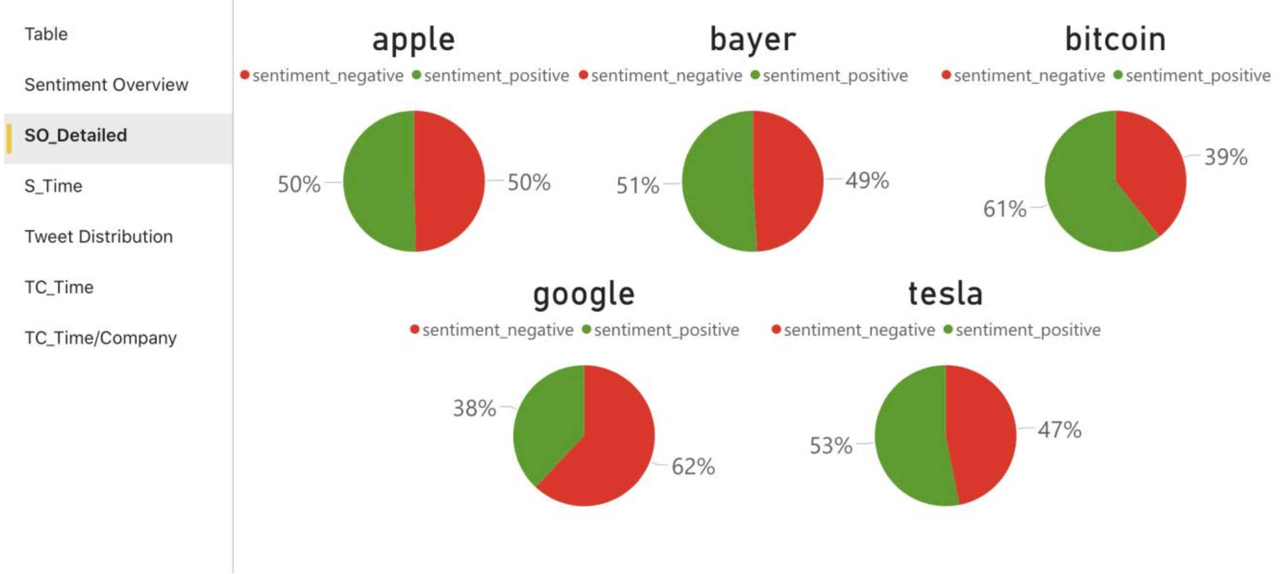
# Appendix A

## Public Opinion

Table
Sentiment Overview
SO_Detailed
S_Time
Tweet Distribution
TC_Time
TC_Time/Company

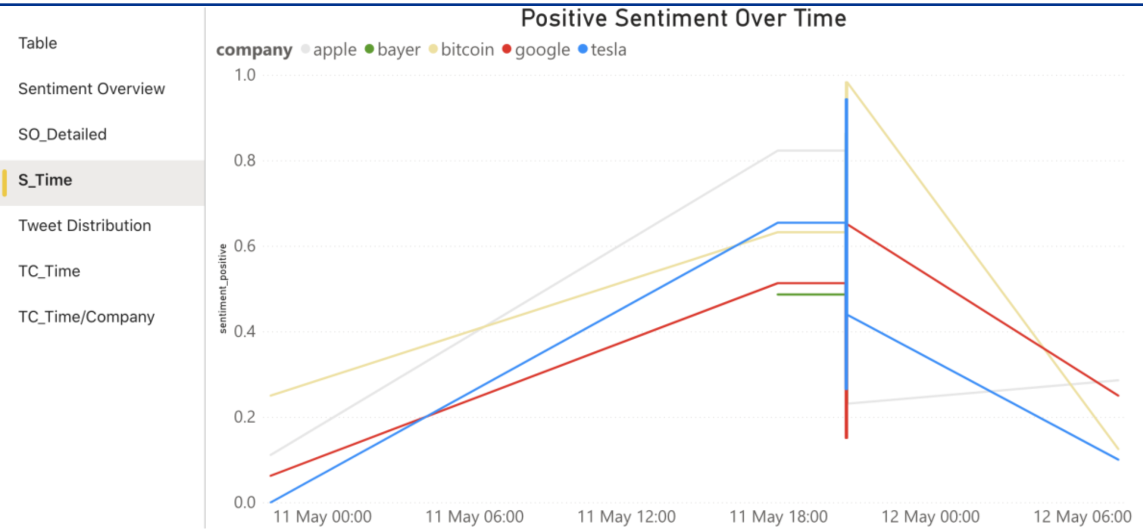
| company | sentiment_positive | sentiment_negative | tweet_count | %tweet_count |
|---|---|---|---|---|
| bitcoin | 0.63 | 0.37 | 295.00 | 36.20% |
| tesla | 0.55 | 0.45 | 153.00 | 18.77% |
| bayer | 0.49 | 0.51 | 96.00 | 11.78% |
| apple | 0.40 | 0.60 | 105.00 | 12.88% |
| google | 0.35 | 0.65 | 166.00 | 20.37% |

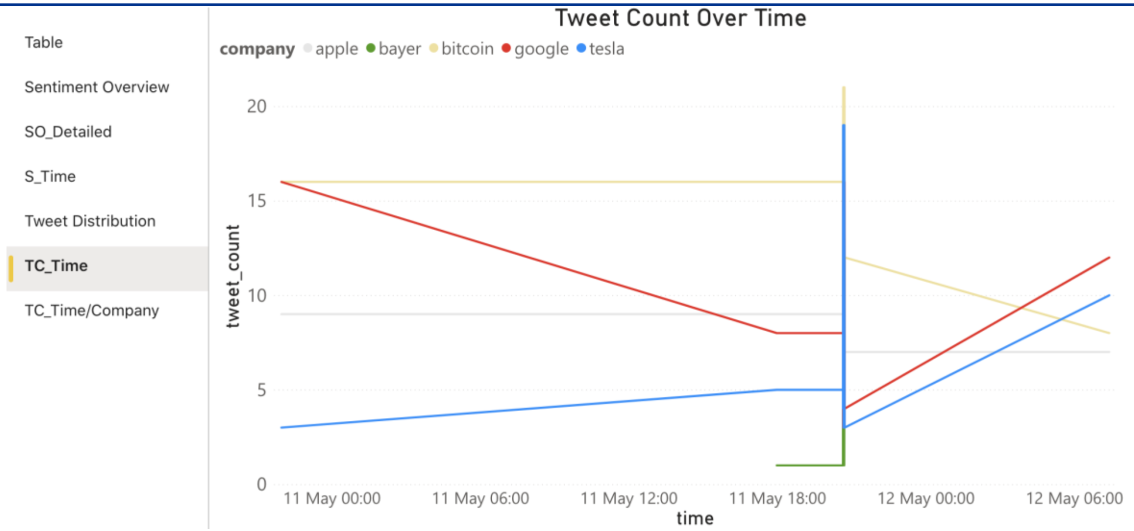


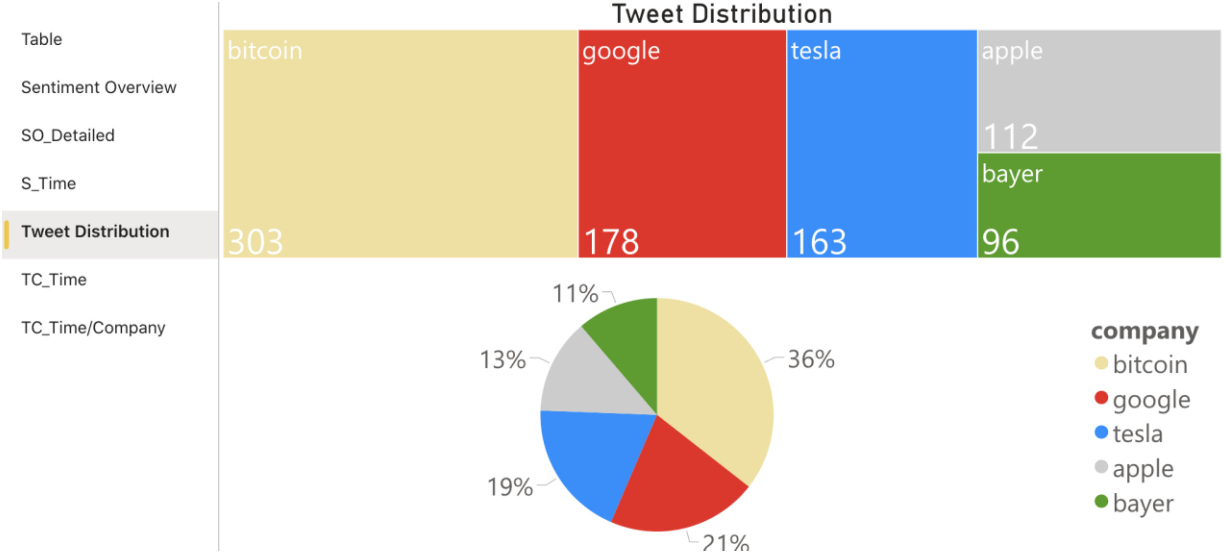



# Appendix A

## Public Opinion




# Appendix B

## Financial Influencer