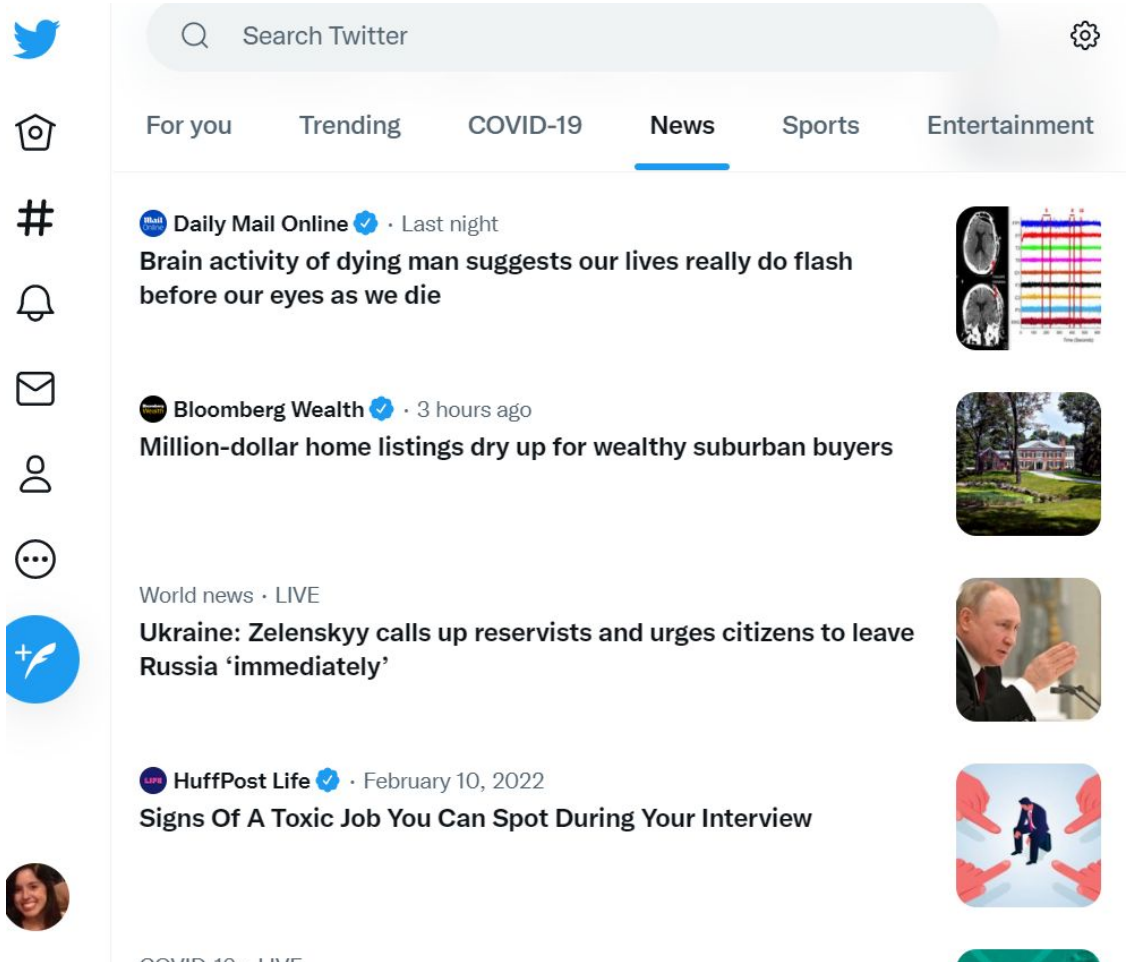
Case Study 1

By: Marcela Vasconcellos, Stokley Voltmer, Sirshendu Ganguly,



News on twitter:

- Fast-moving information
- Demand-oriented content
- Marketplace economy: high competition

↑Users

↑Subscribers and Advertisers

↑Profits

CNBC
481.2K Tweets

CNBC

Follow

CNBC
@CNBC

First in business worldwide.

Englewood Cliffs, NJ cnbc.com Joined February 2009

851 Following **4.5M** Followers

Not followed by anyone you're following

Search Twitter

You might like

Bloomberg
Bloomberg
@business

FT
Financial Times
@FT

REUTERS
Reuters Business
@ReutersBiz

Business question

- What types of news articles should we produce in order to engage users and increase profits?

Using Twitter

- Highly engaged customers: we know which articles were most popular
- Historic data: we can use Machine Learning to find patterns
- Objective: use unsupervised learning to group news posts by similarity and investigate which groups are most popular

Data Source

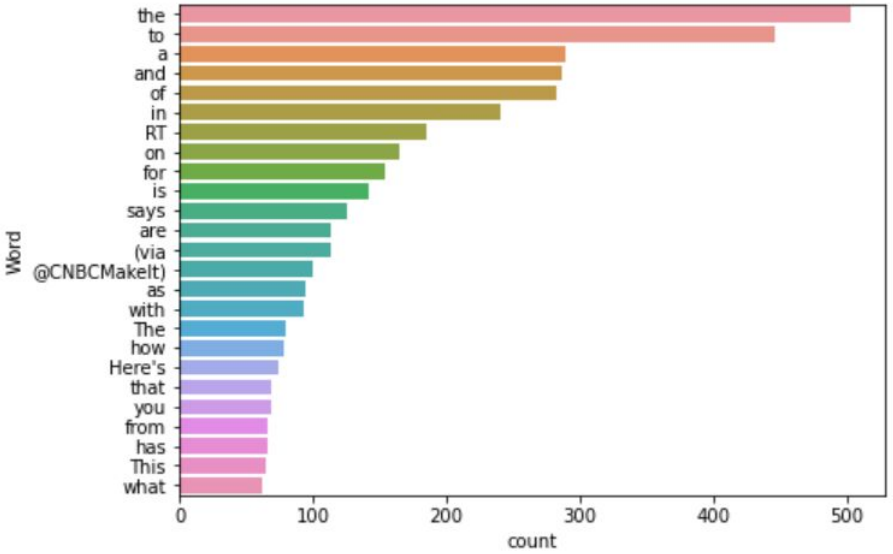
Attributes	Characteristics
text	The actual text from the tweets.
retweet_count	The number of times the tweets were retweeted

Methodology

- Exploratory Data Analysis
- Preprocessing
- TF-IDF indexing
- Clustering
- Analyzing Popularity

EDA

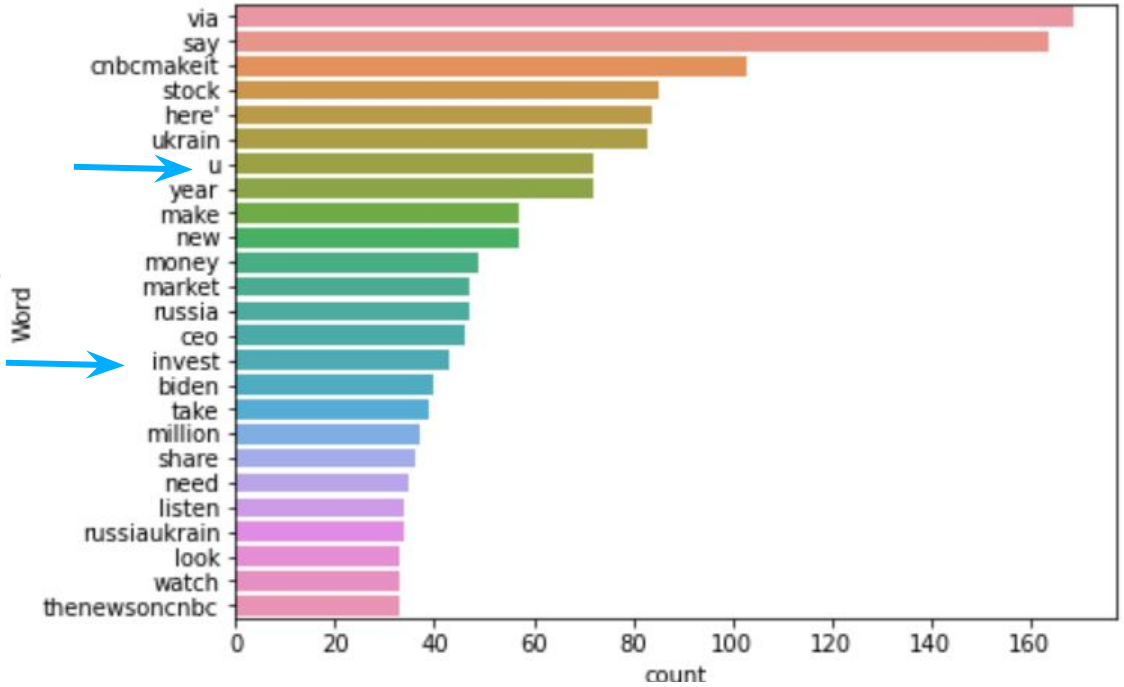
- Visualizing tweets
- 1000 tweets by @CNBC
- Word frequencies



0	phuck elonmusk this is a nice distraction for
1	rt marketrebels elon musk in email to cnbc bi
2	trump says he and capitol rioters wanted same
3	rt schwartzbcnbc exclusive elonmusk speaks to ...
4	rt danpriceseattle big mac prices are up in th...
5	rt jasuja how much money popular startups are ...
6	rt repmaloney i m requesting that the gsa cons...
7	rt fisa end ccp from cnbc international tv htt...
8	cnbc frankcnbc cnbc should start thinking abou...
9	aeharshada cnbc awaaz can nifty surpass
10	rt bruceduck jrubinblogger sanctions on russia...
11	did i miss something bond yieldspread cnbc htt...
12	rt sawyermerritt news musk says he would do th...
13	rt sawyermerritt news musk says he would do th...
14	rollingstone press traitortrump unforgivable a...

Preprocessing

- All lowercase
- Remove symbols and stopwords
- Lemmatizing
- Stemming
- New word count
- Example:




```
[['RT', '@CNBCTechCheck:', 'Despite', 'posting', 'Q4', 'results', 'that', 'mostly', 'beat', 'the', 'Street,', 'shares', 'of', 'cloud-based', 'platfor
m', '@mondaydotcom', 'still', 'plunged', 'to', 'ne…'],
```

↓

```
['cnbctechcheck despit post q4 result mostli beat street share cloudbas platform mondaydotcom still plung ne ',
```

TF-IDF

- Tweets x Vocabulary matrix
- Term Frequency
- Inverse Document Frequency

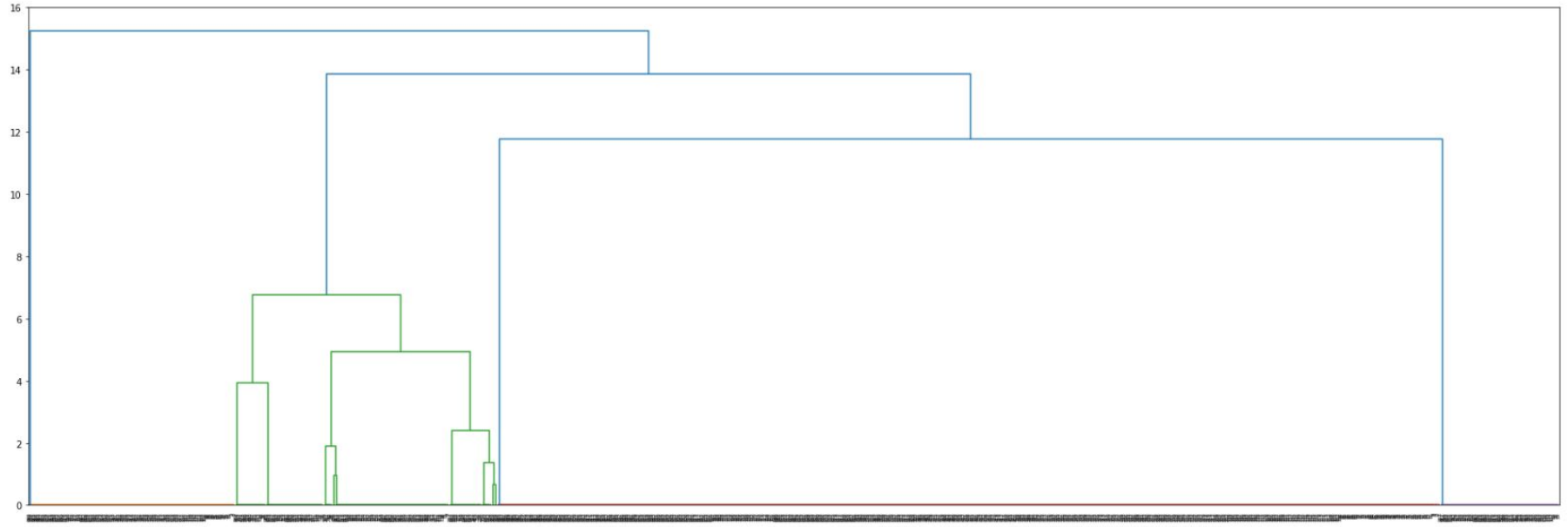
$$\mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

	Word1	Word2	…
Tweet 1		(Tfidf value)	
Tweet 2			

```
Number of features (words): 212
Sample indexed tweet:
  (0, 168)      0.6355892105048007
  (0, 162)      0.5111822990026225
  (0, 31)       0.57854905814138
Features in sample tweet: [array(['share', 'street', 'cnbctechcheck'], dtype='<U14')]
```

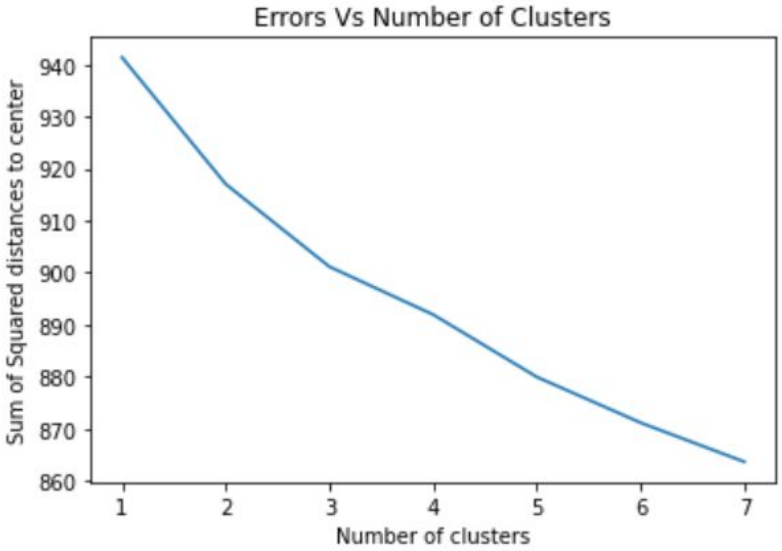
Clustering

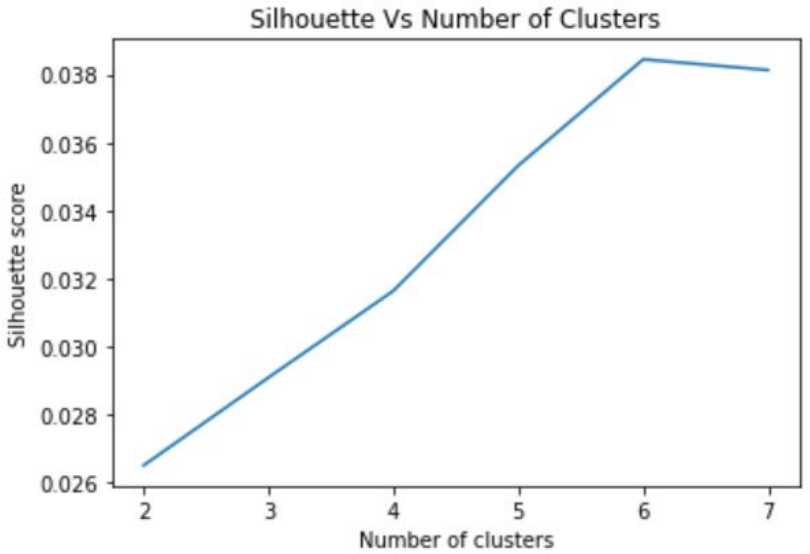
- Dendrogram



Clustering

- K-means
  - Tested 1 to 7 clusters





Top Words Cluster 0	Count
via	119
cnbcmakeit	103
here	22
year	21
money	19
make	13
home	12
top	12
21	9
biggest	9

Top Words Cluster 1	Count
stock	82
market	26
say	23
russiaukrain	20
tension	19
week	9
fed	9
listen	8
ukrain	8
biggest	8

Top Words Cluster 2	Count
here	58
invest	36
million	28
year	25
much	21
money	19
pay	17
spend	17
save	16
need	16

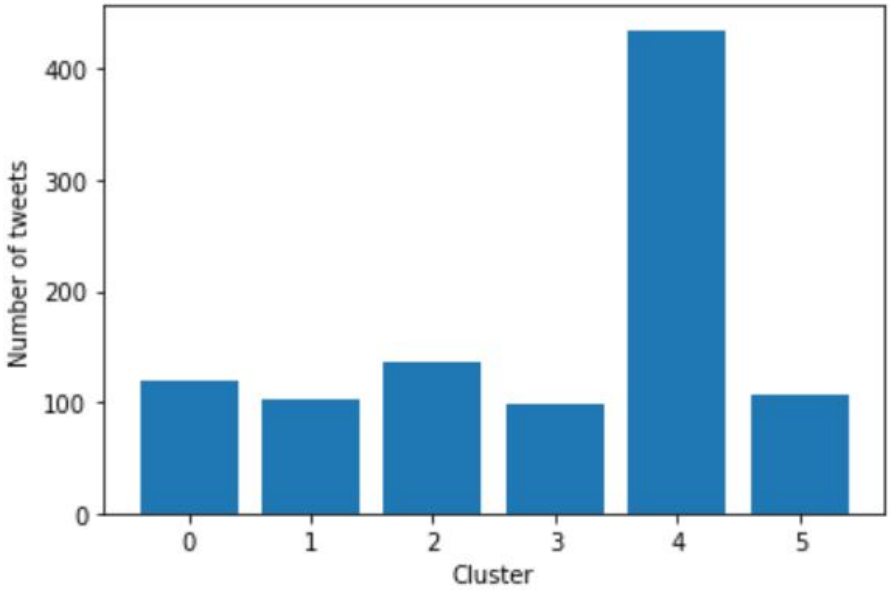
Top Words Cluster 3	Count
ukrain	73
russia	51
putin	34
sanction	26
biden	26
say	23
russian	20
presid	20
troop	15
invas	14

Top Words Cluster 4	Count
new	33
ceo	26
tesla	26
watch	24
busi	21
via	21
here	20
report	20
take	18
year	18

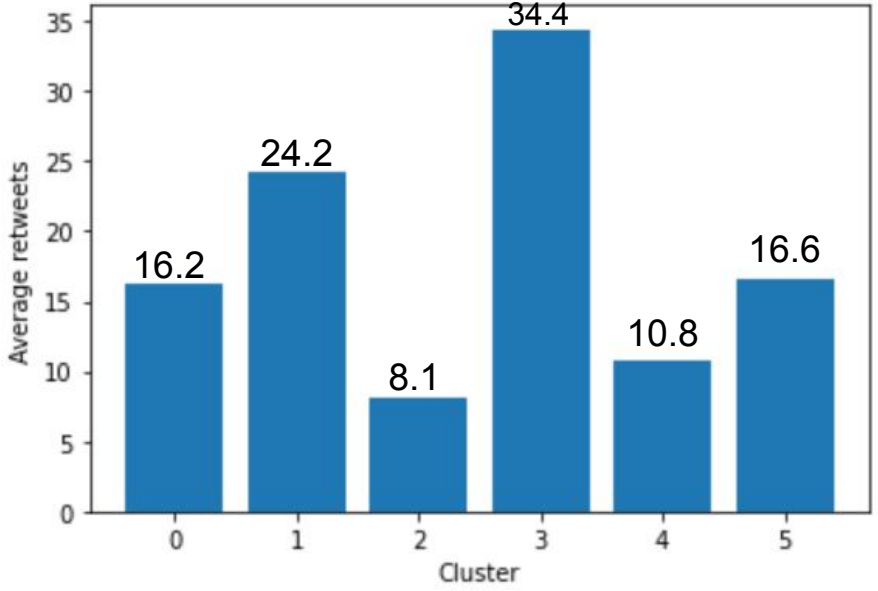
Top Words Cluster 5	Count
say	103
via	22
expert	15
harvard	13
make	13
avoid	12
ceo	8
covid	7
food	7
worst	7

Clustering

How many tweets are in each of the clusters



How many average retweets per cluster



Conclusion

- Our analysis will be helpful to the organization in categorizing the tweets before uploading so that they can tag their tweets with suitable keywords i.e. politics, stock market, etc.
- Thus allowing the readers of the tweet to easily scroll through the topics of their interest.
- This in return will prevent CNBC from losing their Twitter followers.