NLP + Binary Classification with Stock Market Tweets

Kris Rubiano

Data

This Kaggle dataset contains over 64,000 tweets for the top 25 most watched stock tickers on Yahoo Finance for the fiscal year of 2022.

Aim

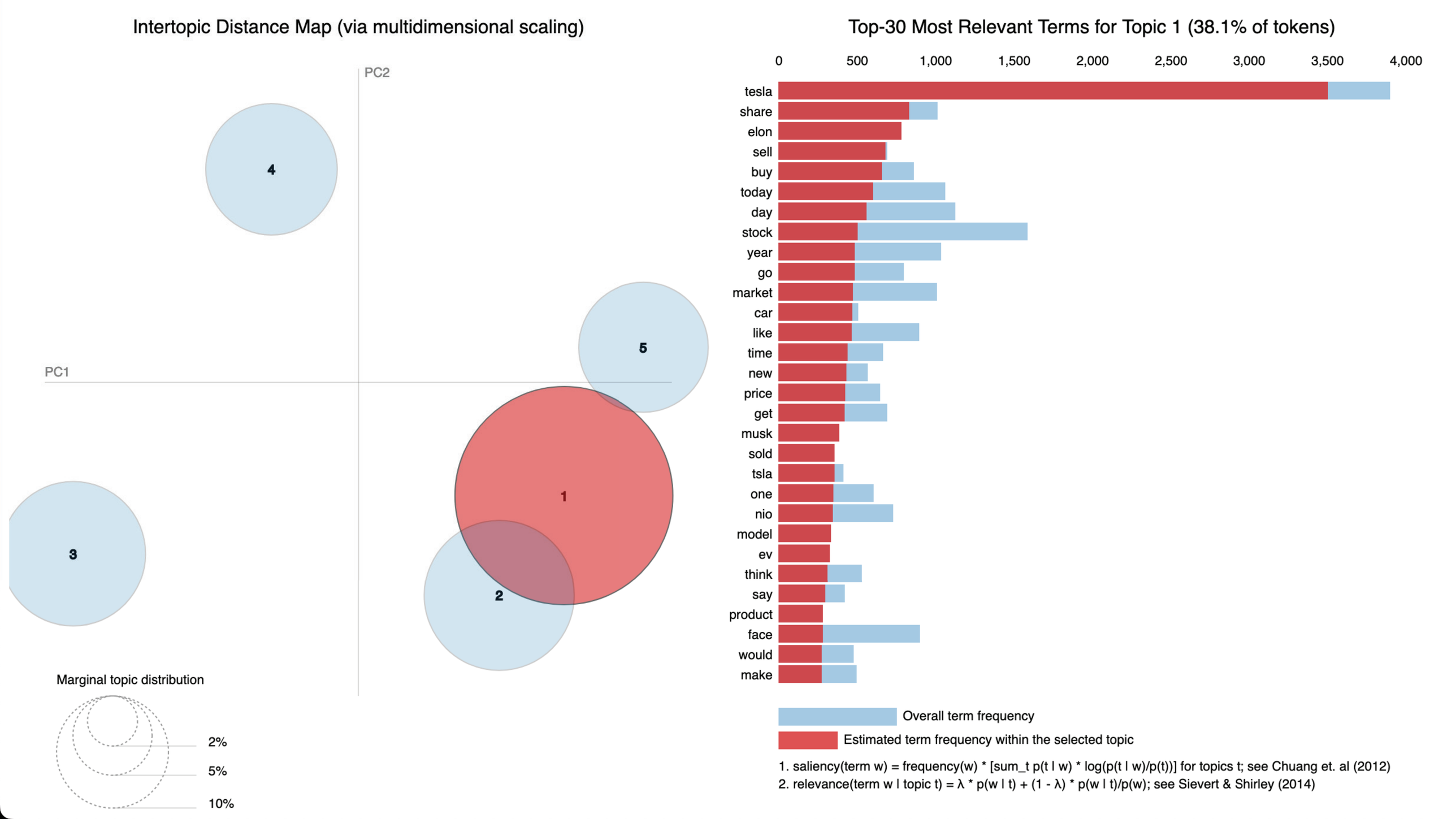
• Use NLP to classify and predict whether a tweet is regarding MAMAA (one of the top 5 tech stocks) or not, to measure social media engagement.

• Assess if Twitter is an effective platform for an a MAMAA company for audience interaction.

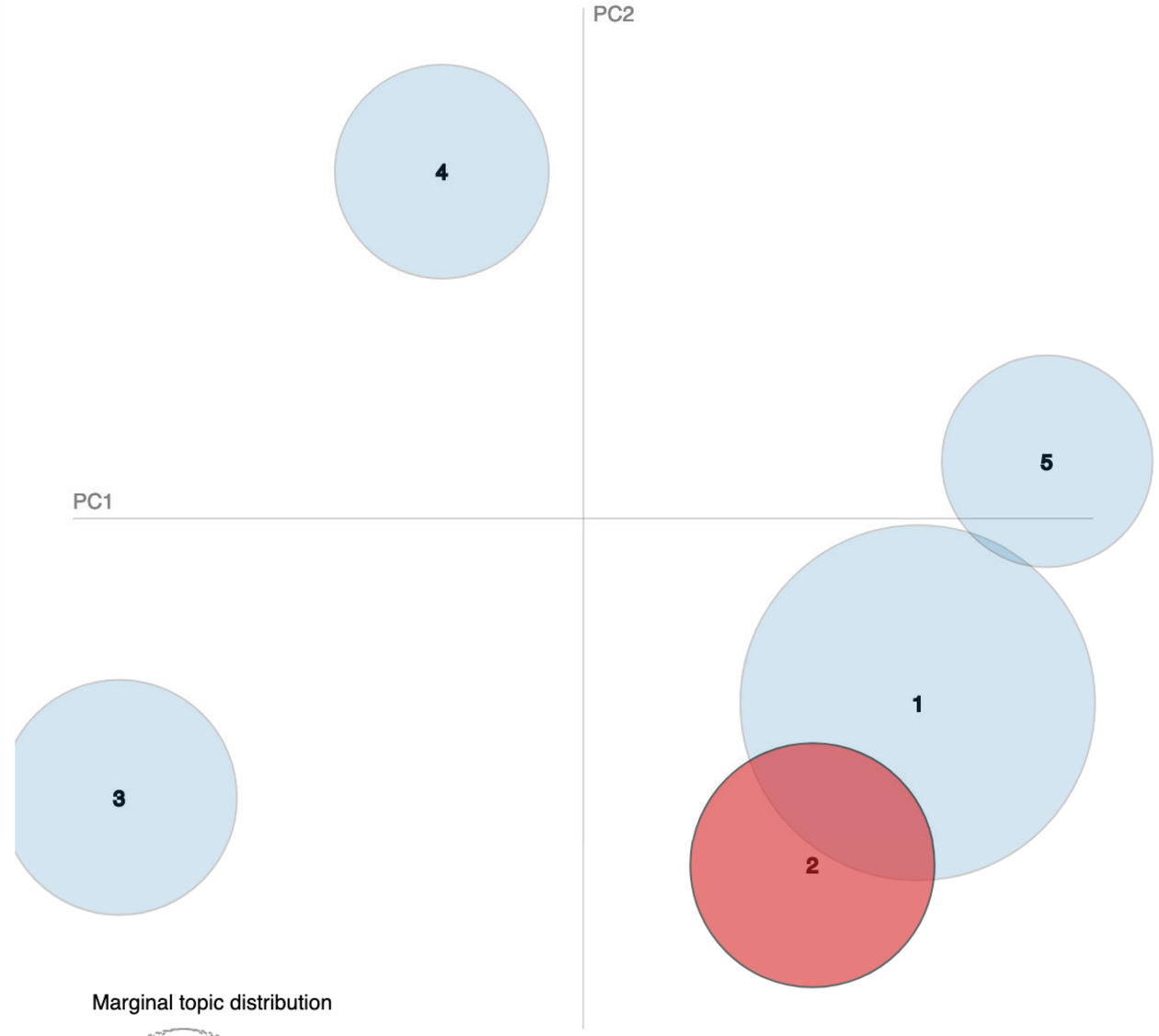
Preprocess

- TweetTokenizer
- stopwords
- SnowballStemmer
- Transformer for emoji

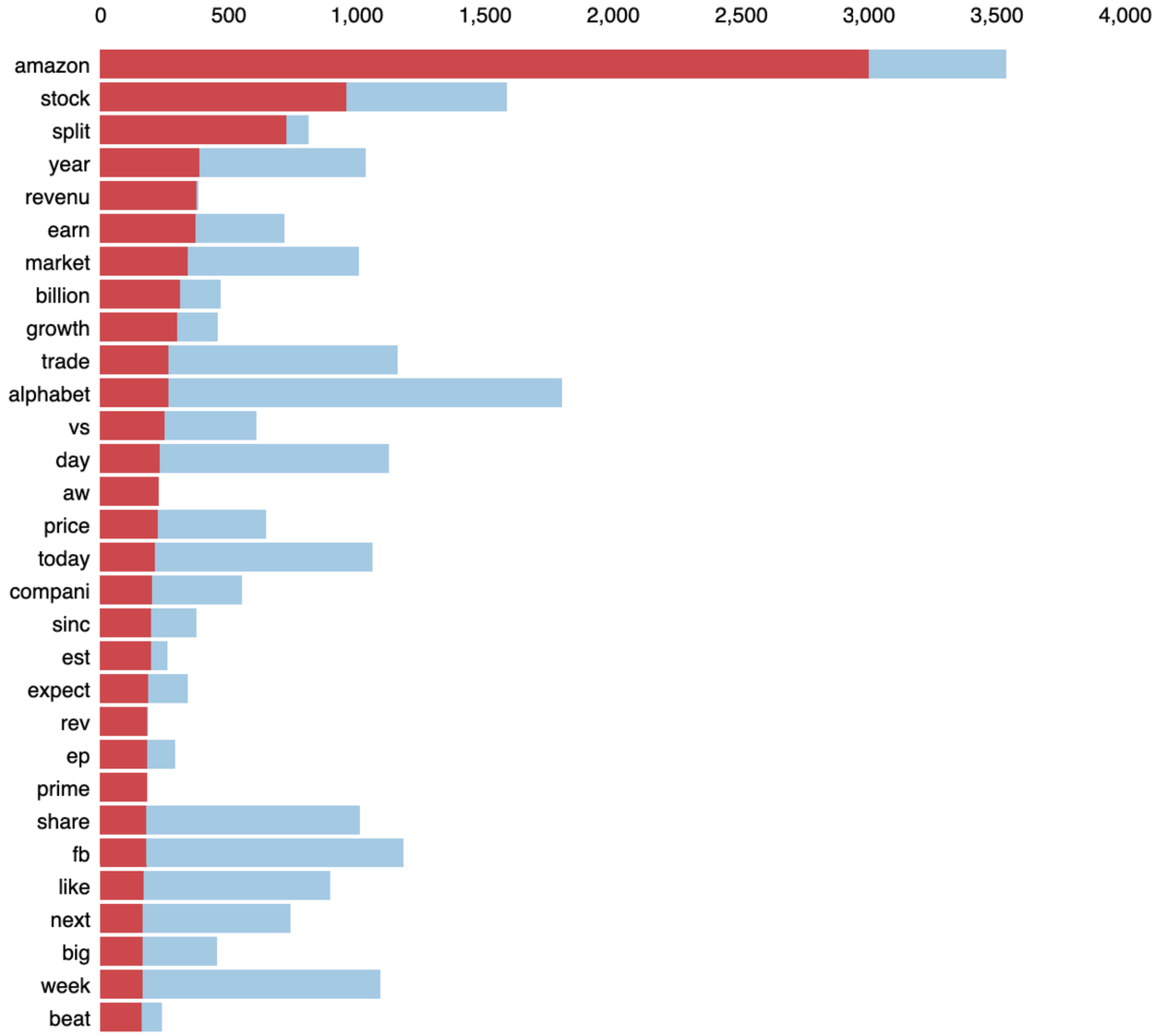
- TF-IDF Vectorizer
- Count Vectorizer


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
5
1
3
2
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (38.1% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
tesla
share
elon
sell
buy
today
day
stock
year
go
market
car
like
time
new
price
get
musk
sold
tsla
one
nio
model
ev
think
say
product
face
would
make
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

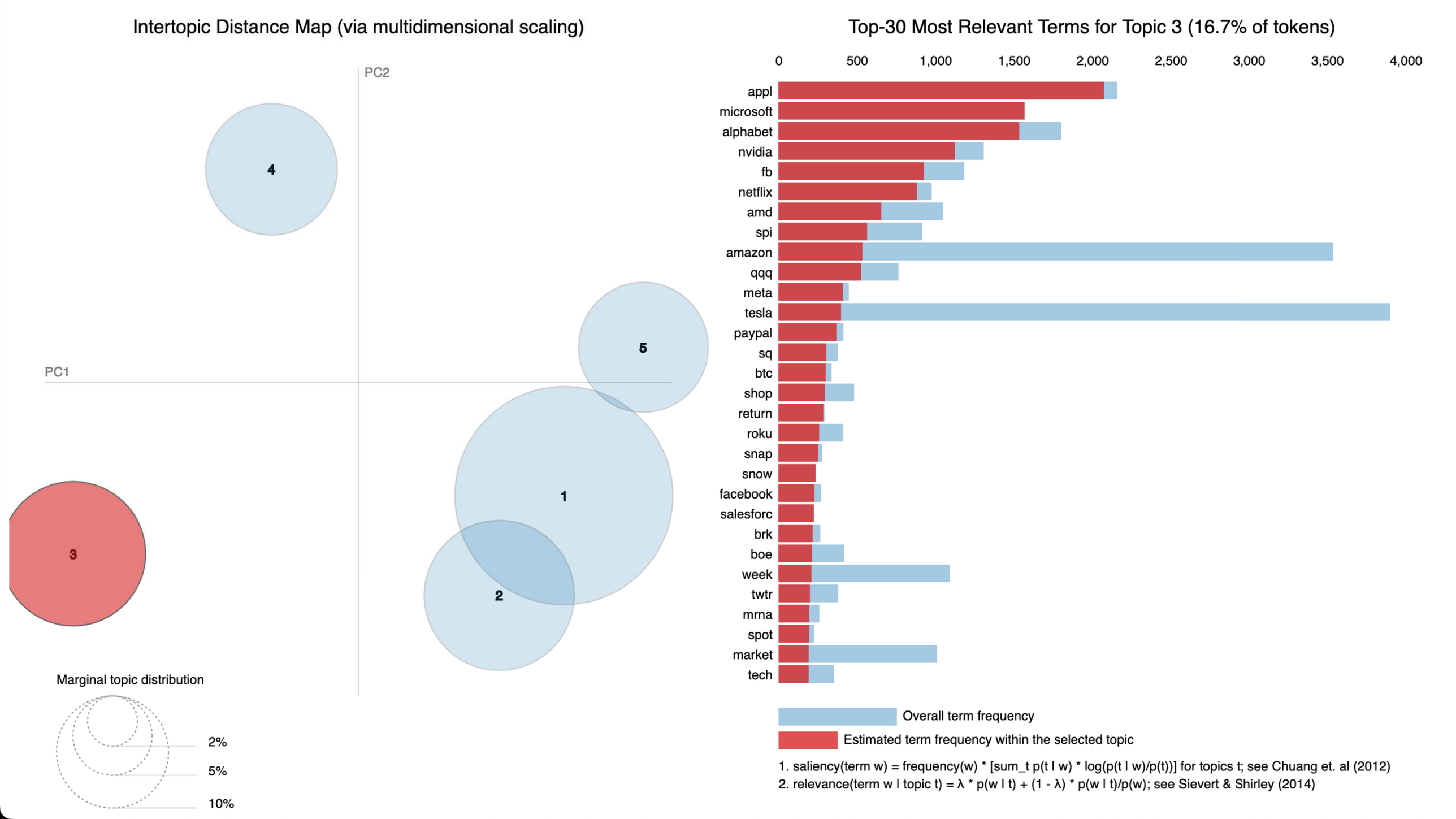
Intertopic Distance Map (via multidimensional scaling)



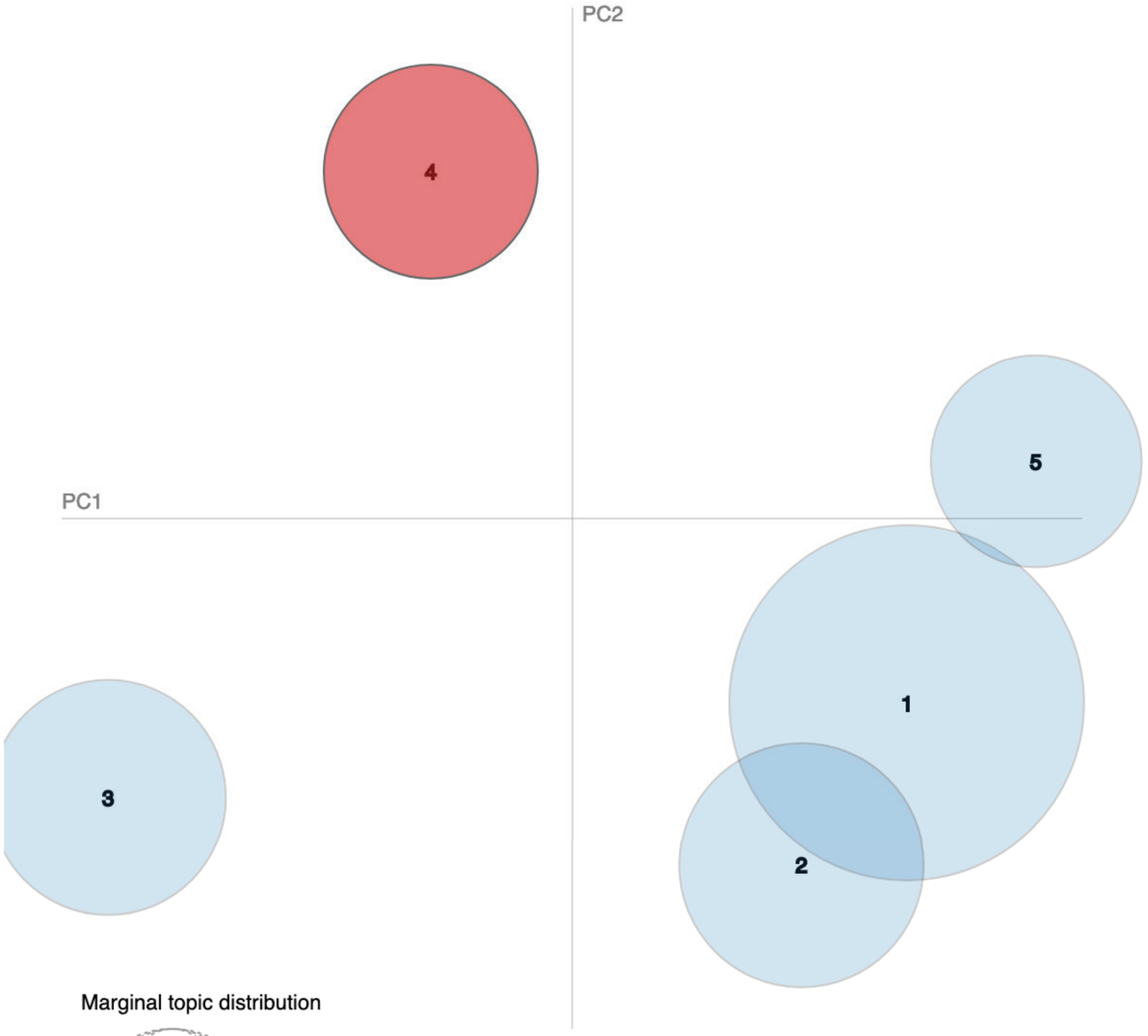
Top-30 Most Relevant Terms for Topic 2 (18% of tokens)



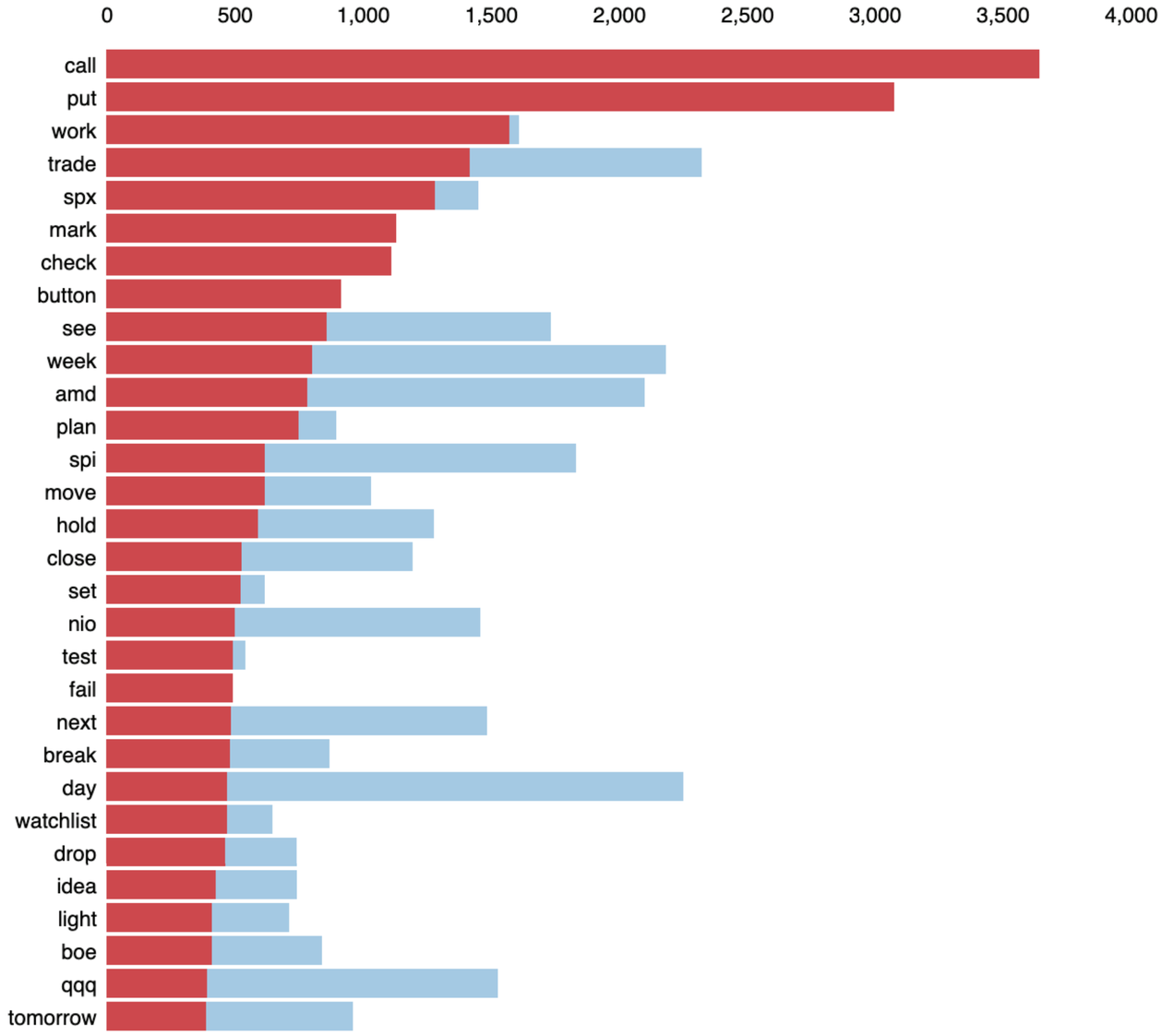
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
5
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (16.7% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
appl
microsoft
alphabet
nvidia
fb
netflix
amd
spi
amazon
qqq
meta
tesla
paypal
sq
btc
shop
return
roku
snap
snow
facebook
salesforc
brk
boe
week
twtr
mrna
spot
market
tech
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

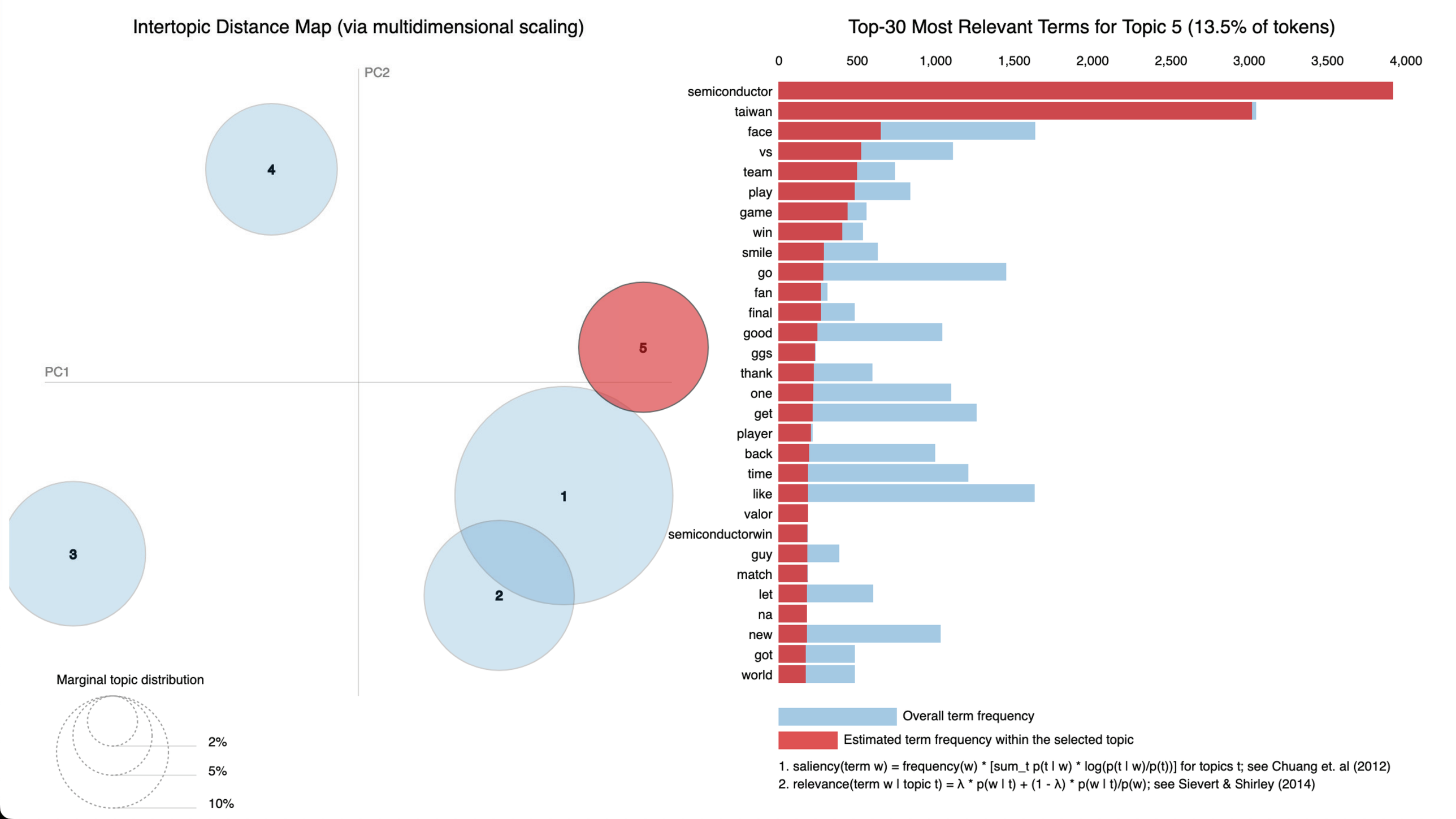
Intertopic Distance Map (via multidimensional scaling)

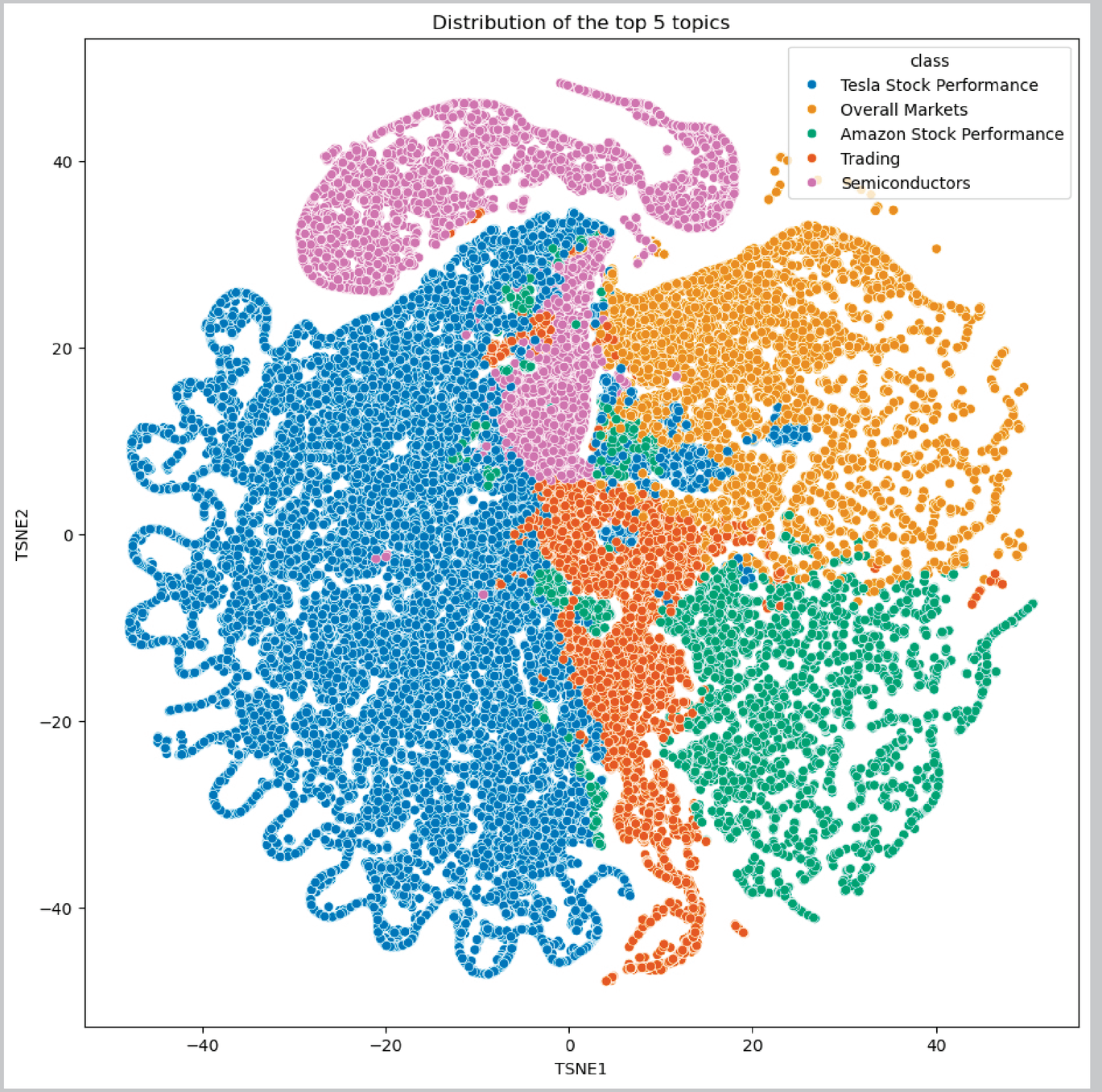


Top-30 Most Relevant Terms for Topic 4 (13.8% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

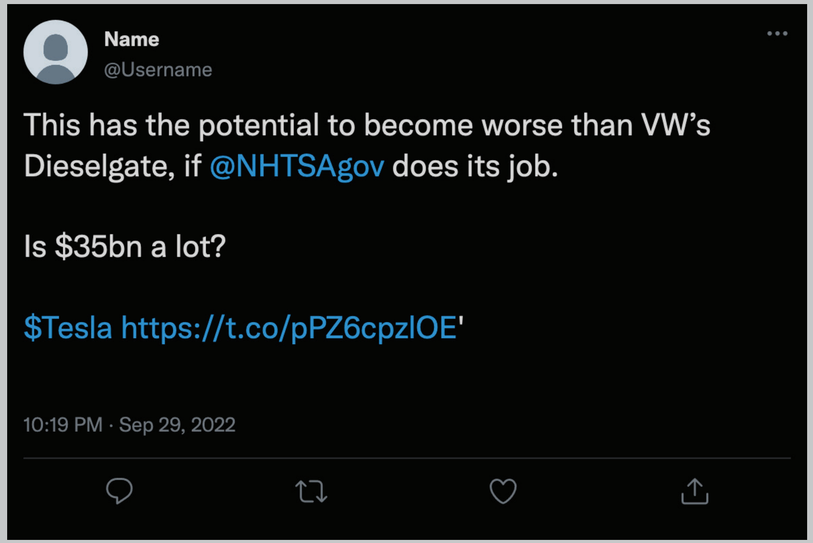

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
5
1
3
2
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 5 (13.5% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
semiconductor
taiwan
face
vs
team
play
game
win
smile
go
fan
final
good
ggs
thank
one
get
player
back
time
like
valor
semiconductorwin
guy
match
let
na
new
got
world
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Distribution of the top 5 topics
class
Tesla Stock Performance
Overall Markets
Amazon Stock Performance
Trading
Semiconductors
TSNE2
TSNE1
40
20
0
−20
−40
−40
−20
0
20
40

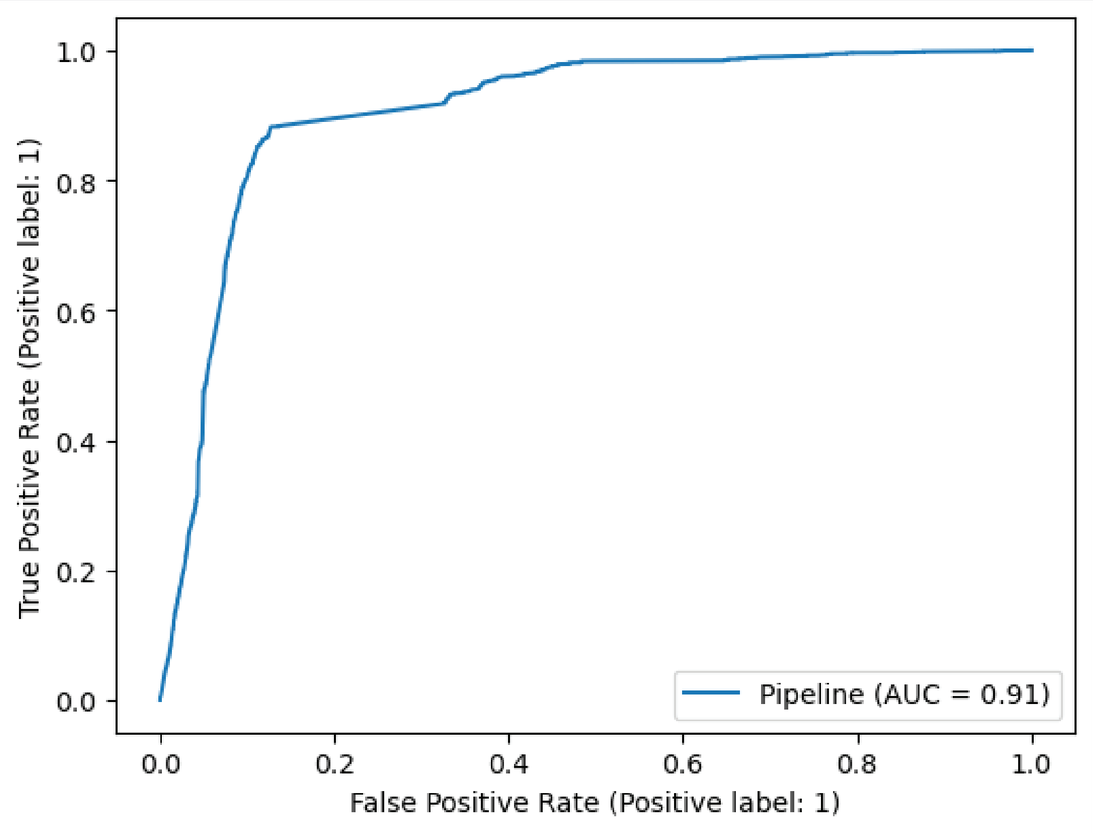
Sample Tweets from NMF

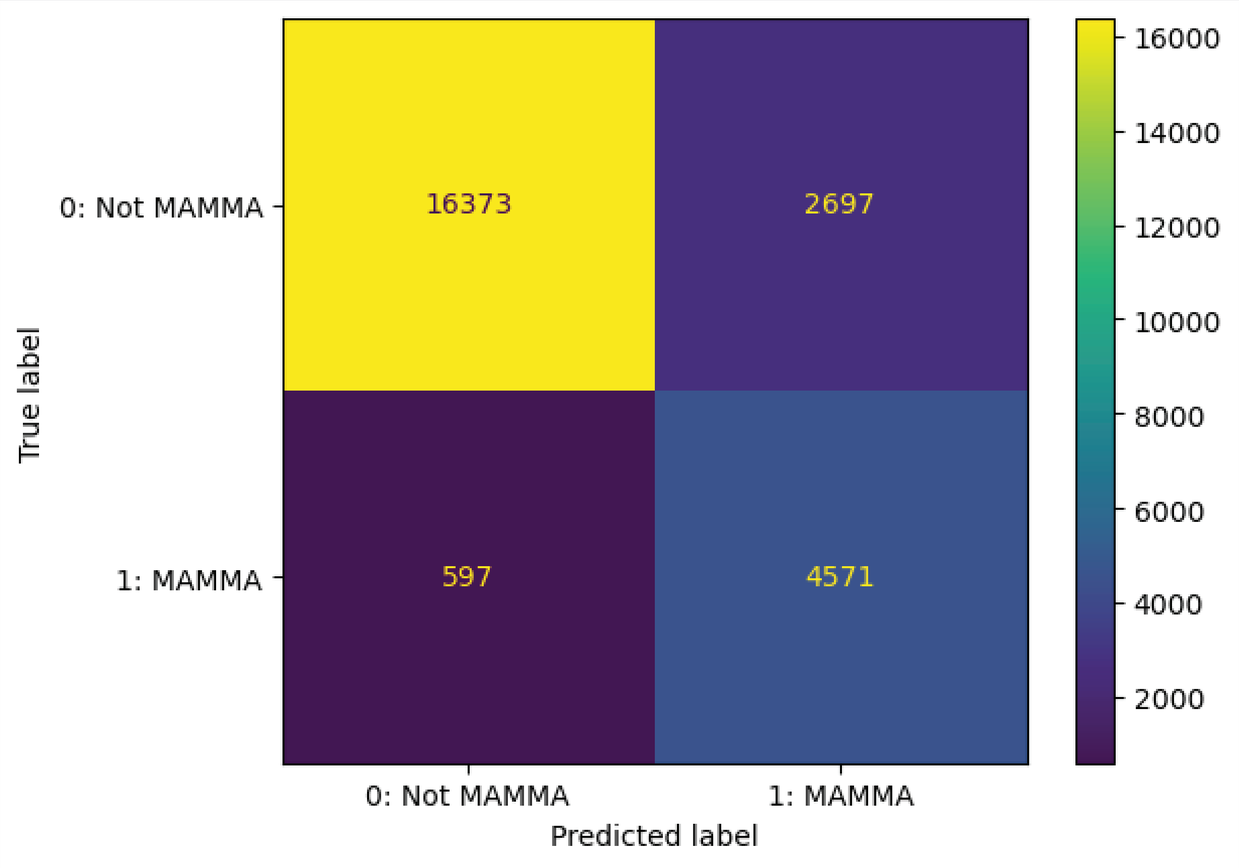
Amazon (MAMAA)

Tesla (not MAMAA)



Count Vectorizer with Complement NB



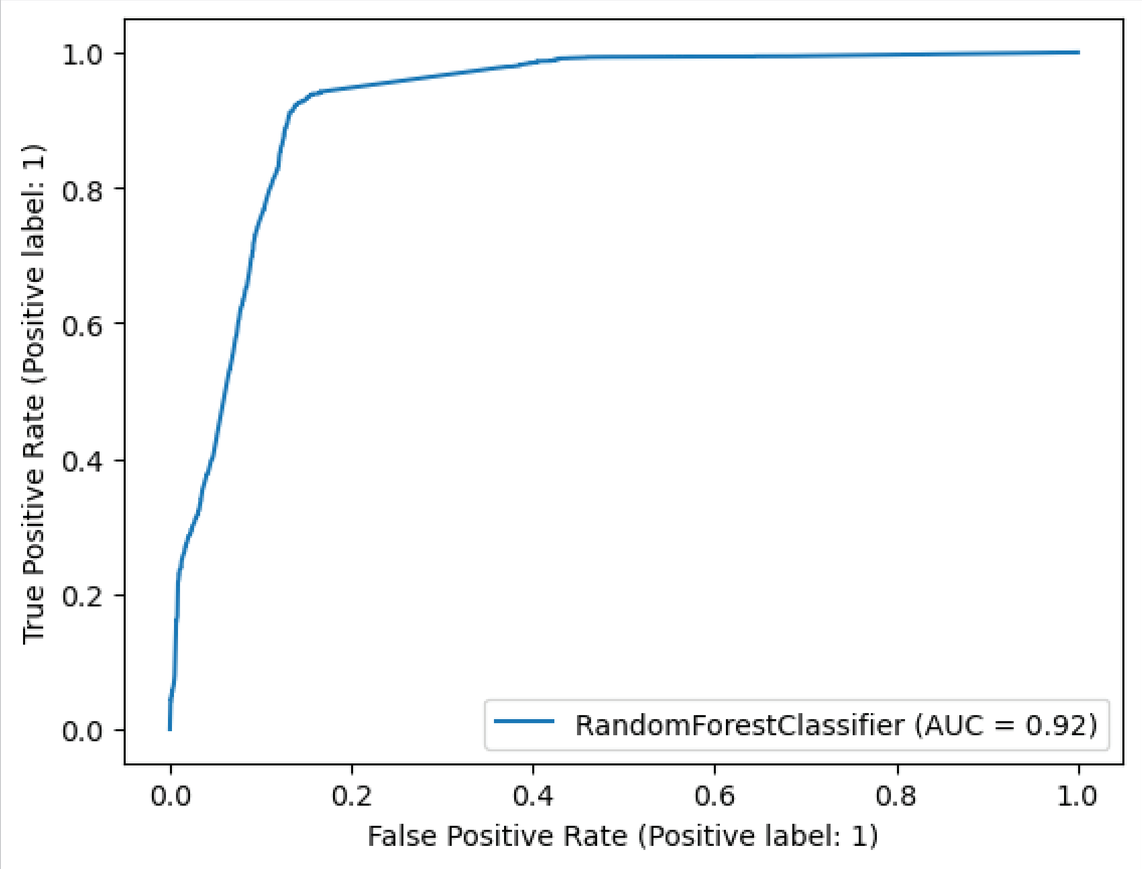


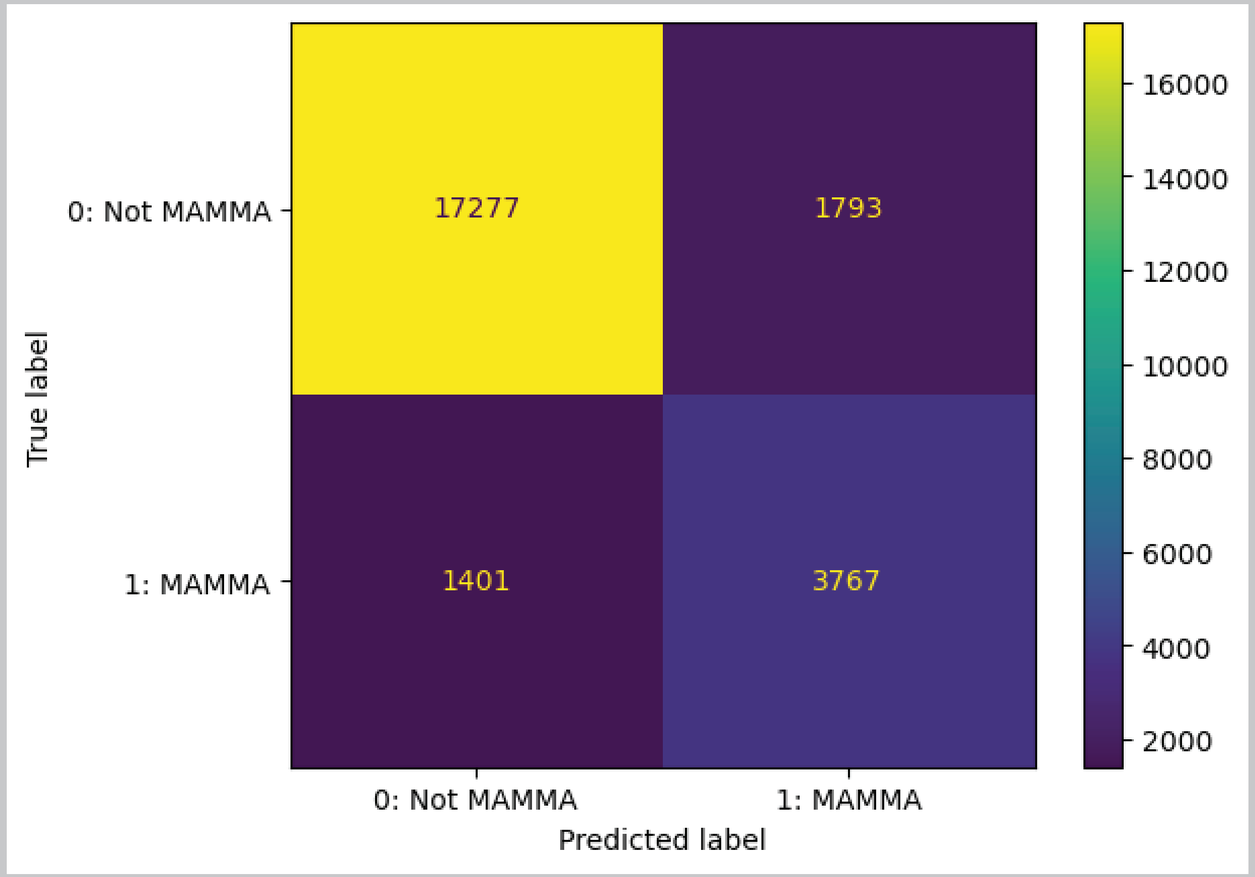
Predicts with 86% accuracy and a false positive rate of ~11%.

Count Vectorizer with Complement NB

	precision	recall	f1-score	support
0	0.86	0.96	0.91	16970
1	0.88	0.63	0.74	7268
accuracy			0.86	24238
macro avg	0.87	0.80	0.82	24238
weighted avg	0.87	0.86	0.86	24238

Random Forest





Predicts with 87% accuracy and a false positive rate of ~7%.

Random Forest

	precision	recall	f1-score	support
0	0.91	0.92	0.92	18678
1	0.73	0.68	0.70	5560
accuracy			0.87	24238
macro avg	0.82	0.80	0.81	24238
weighted avg	0.87	0.87	0.87	24238

Summary

• Random Forest was the best model as it had an increase in 1% accuracy compared to the best NB model.

• Dataset may need the current year tweets for better insight. Twitter appears to not be the best social media platform for MAMAA stock engagement.