# DTSA 5510 Unsupervised Learning Final Project

# Health News Tweets Categorization

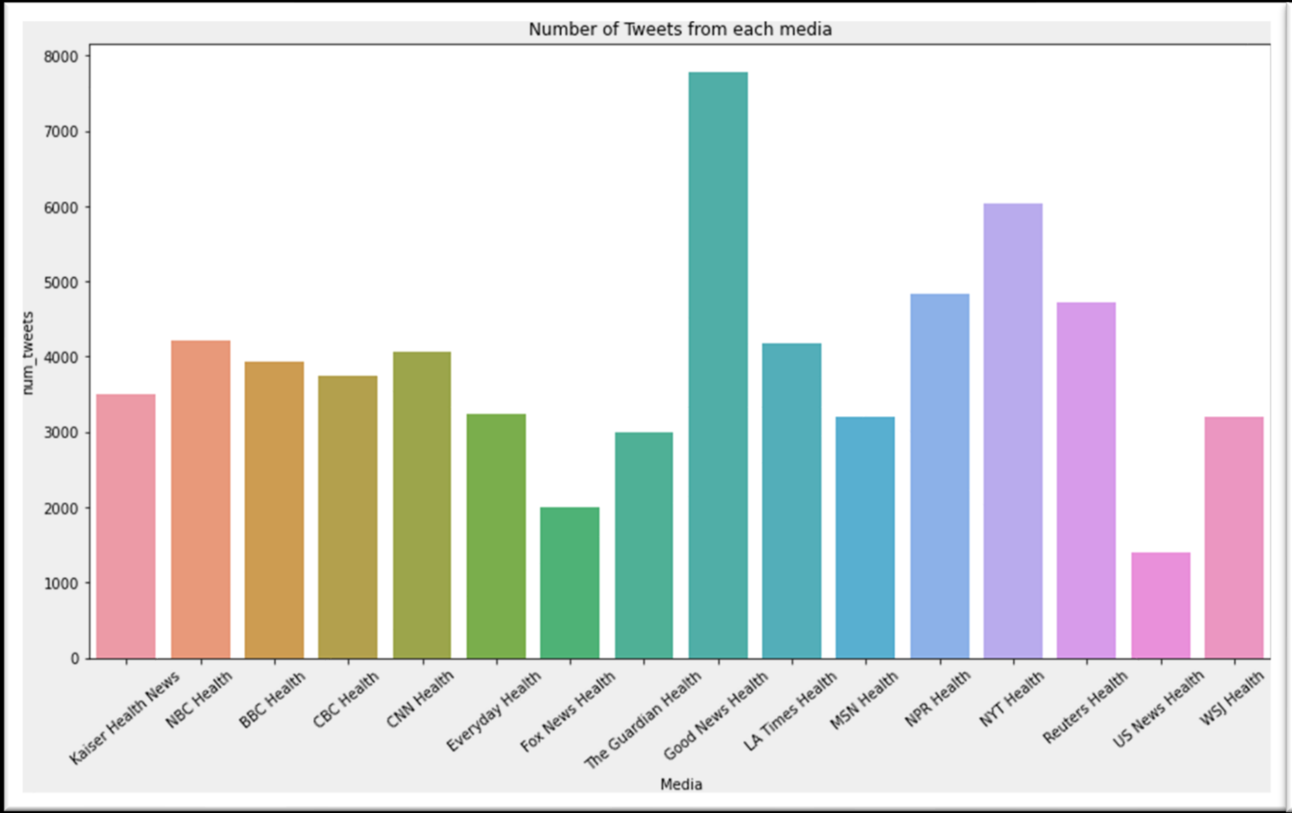
# Data and Problem

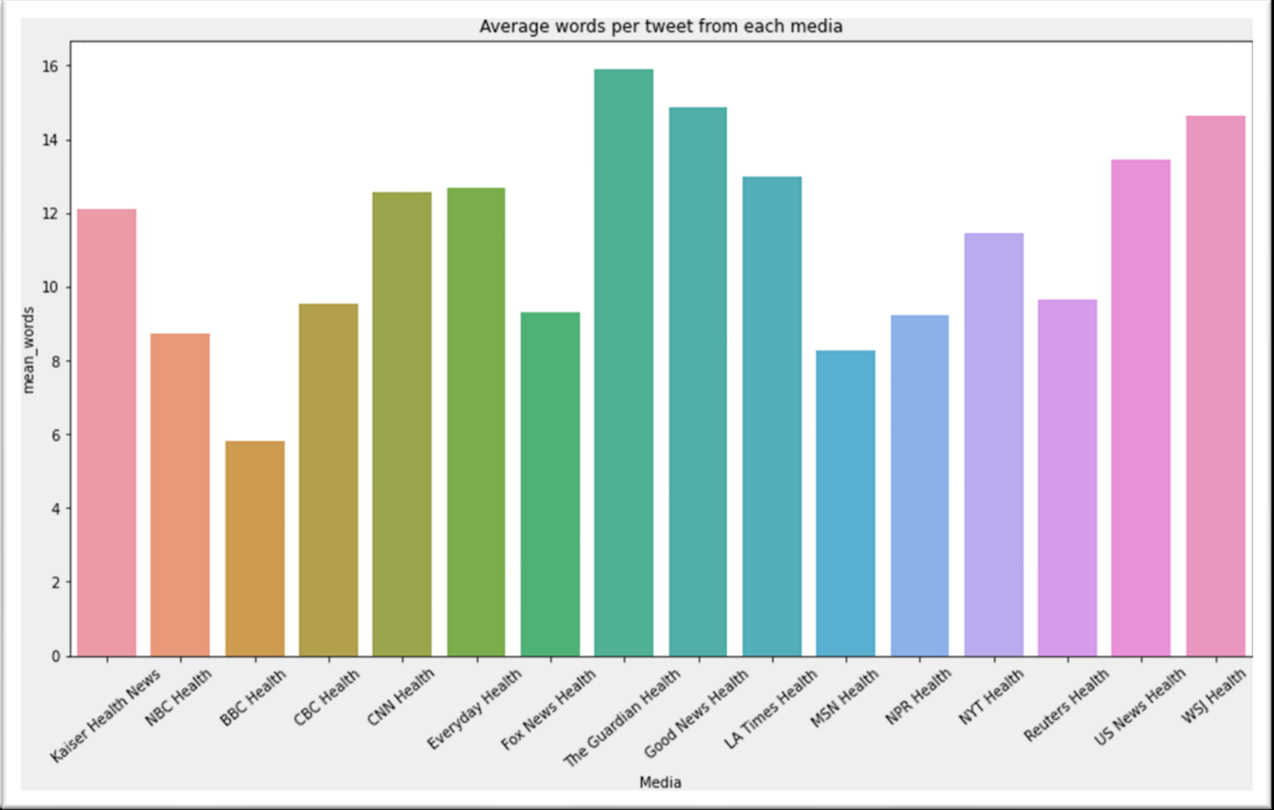
- We saw word vector based model worked well with news articles in module 4
- What if each document is much smaller, specifically Twitter text?
- Can we predict the attributions of tweets or the accounts from texts alone?
- Can we find meaningful latent factors from texts?

- The data was downloaded from UCI Machine Learning Repository
- The data contains 63K tweets from 16 health news accounts
- I will use 1,200 tweets from each media, except for the TFIDF analysis

# Data Distribution

Number of tweets and average words per tweet



# TF/IDF 1-2 ngrams

10 words with the highest TFIDF scores from each news source

Many of them confirm our perceptions

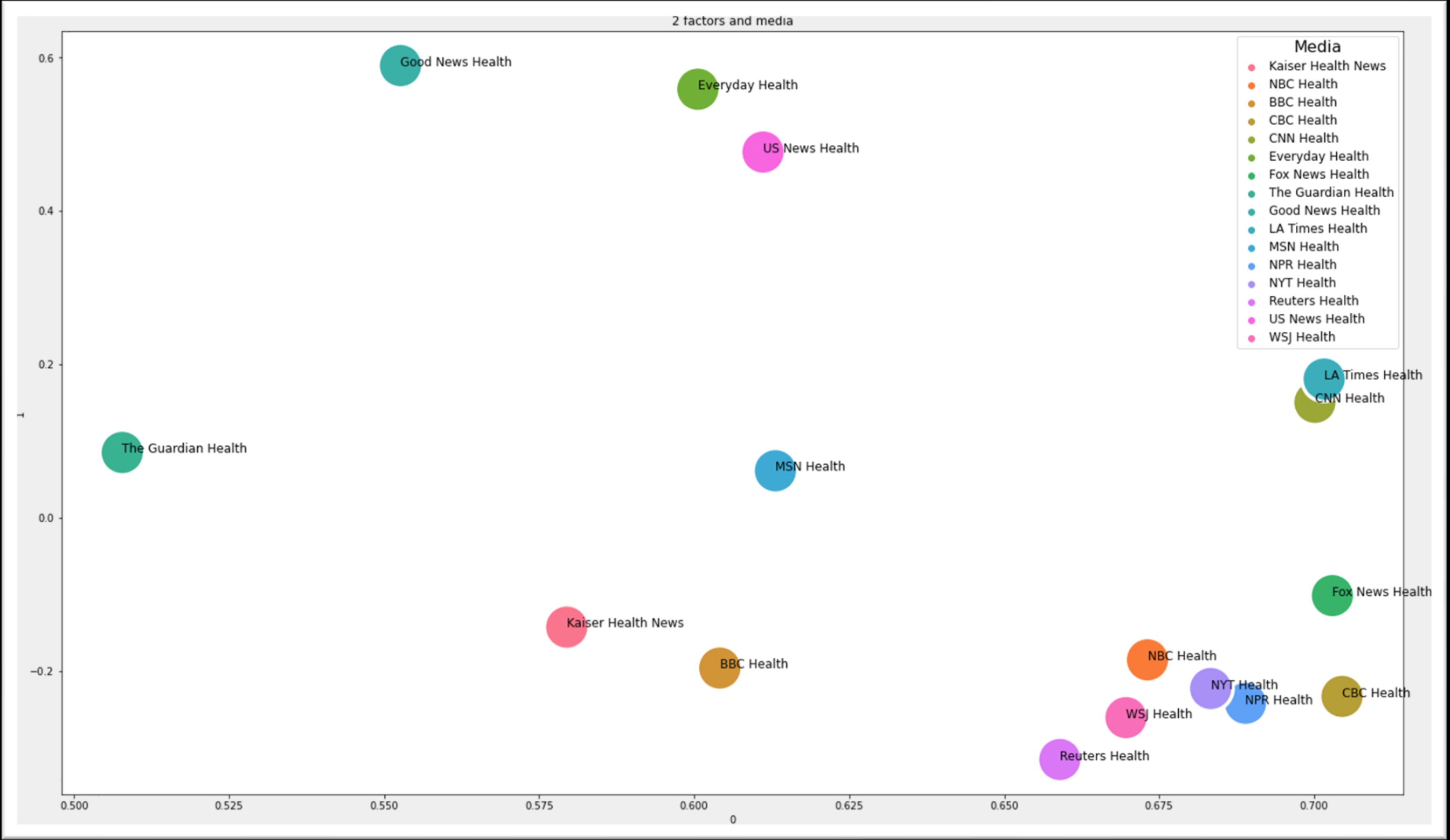
This could explain/predict news sources

Kaiser Health News:
['health law' 'headlines' 'cliff' 'law s' 'kaiserfamfound' 'marketplaces' 'reports today' 'cms' 'reports rt' 'fiscal cliff']
NBC Health:
['safety rules' 'genuine' 'nfl concussion' 'powerful painkiller' 'new safety' 'communion' 'fda proposes' 'bra' 'foster farms' 'november']
BBC Health:
['mps' 'gp' 'ebola video' 'video ebola' 'audio' 'uk ebola' 'care home' 'labour' 'ebola' 'centre']
CBC Health:
['canadian' 'canada s' 'b c' 'canadians' 'p e' 'quebec' 'ebola' 'ford' 'centre' 'facts ebola']
CNN Health:
['rt cnnhealth' 'cnn' 'lost pounds' 'triathlon' 'cnnhealth' 'weightloss' 'lost lbs' 'timehealthland' 'transformation' 'pls']
Everyday Health:
['everydayhealth' 'menshealth' 'jillianmichaels' 'digest' 'rt cspi' 'cspi' 'eatsmartbd' 'psoriasis' 'worldcancerday' 'rt everydayhealth']
Fox News Health:
['disneyland' 'weightloss' 'wakes' 'heart transplant' 'ebola' 'leone' 'sierra leone' 'georgia' 's' 'lodged']
The Guardian Health:
['midwife' 't miss' 'gp' 'day life' 'case missed' 'burnout' 'newsletter' 'andy' 'frontline' 'network s']
Good News Health:
['abs' 'rd' 'weightloss' 'recipes' 'healthyliving' 'fitfluential' 'vacuum' 'greatist' 'gorgeous' 'toned']
LA Times Health:
['cdcgov' 'healthful' 'glutenfree' 'planned parenthood' 'komen' 'rt' 'medi' 'ameracadpeds' 'nejm' 's']
MSN Health:
['clot risk' 'study u' 'gene mutations' 'study' 'heart trouble' 'patients study' 'study study' 's' 'women study' 'blood thinner']
NPR Health:
['npr' 'nprhealth' 'health law' 'ebola' 'health exchanges' 'exchanges' 's' 'insurance website' 'health exchange' 'health']
NYT Health:
['aug' 'briefing' 'op ed' 'nytimes' 'like doctor' 'contributor' 'global health' 'nyt' 'new health' 'letters']
Reuters Health:
['study u' 'mid stage' 'roche' 'eu' 'ebola' 'sanofi' 'abbvie' 'stage study' 'gilead' 'amgen']
US News Health:
['usnews' 'leonardkl' 'rankings' 'rt leonardkl' 'kerigans' 'd love''healthyliving' 'fitness tracker' 'goredforwomen' 'hearthealth']
WSJ Health:
['pharma' 'valeant' 'health law' 'h rt' 'wsj' 'allergan' 'abbvie' 'good morning' 'headlines' 'htt rt']

# TF/IDF Based Clustering And Categorization

- Predictive model using the TF/IDF scores was not successful with 30.18% accuracy score even with the training data
- New tweets can be categorized into “closest” source media
  - It is interesting that the model tells what each tweet “looks like”
  - But it does not accurately predict the tweets from the accounts that were in the training set
- TF/IDF vectors categorizes tweets and source accounts, but not suitable for building predictive models

# TF/IDF and MF for Sources




The TFIDF vectors are decomposed with 2 components Matrix Factorization

See if the two factors are latent factors identifying the source medias

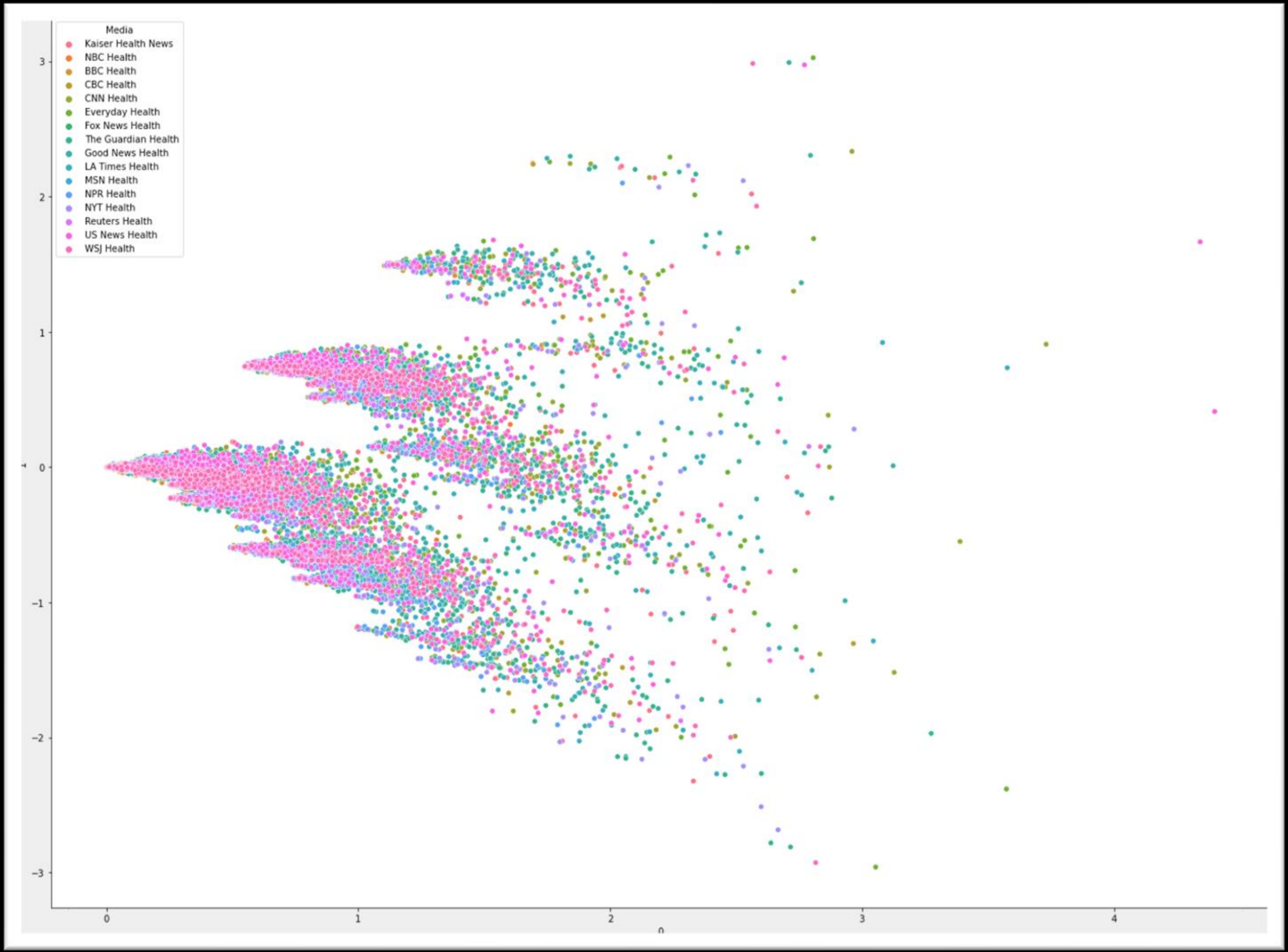
Looks to me that
- traditional vs unconventional horizontally (f-0)
- Sensational vs dried vertically (f-1)

# Word Count Vector and MF for Tweets

Created word count vectors and decomposed with 2 components Matrix Factorization on each tweet

Does not look like it can distinguish the source accounts, but it does cluster tweets

Decided to build a model to cluster tweets based on word count vector and MF

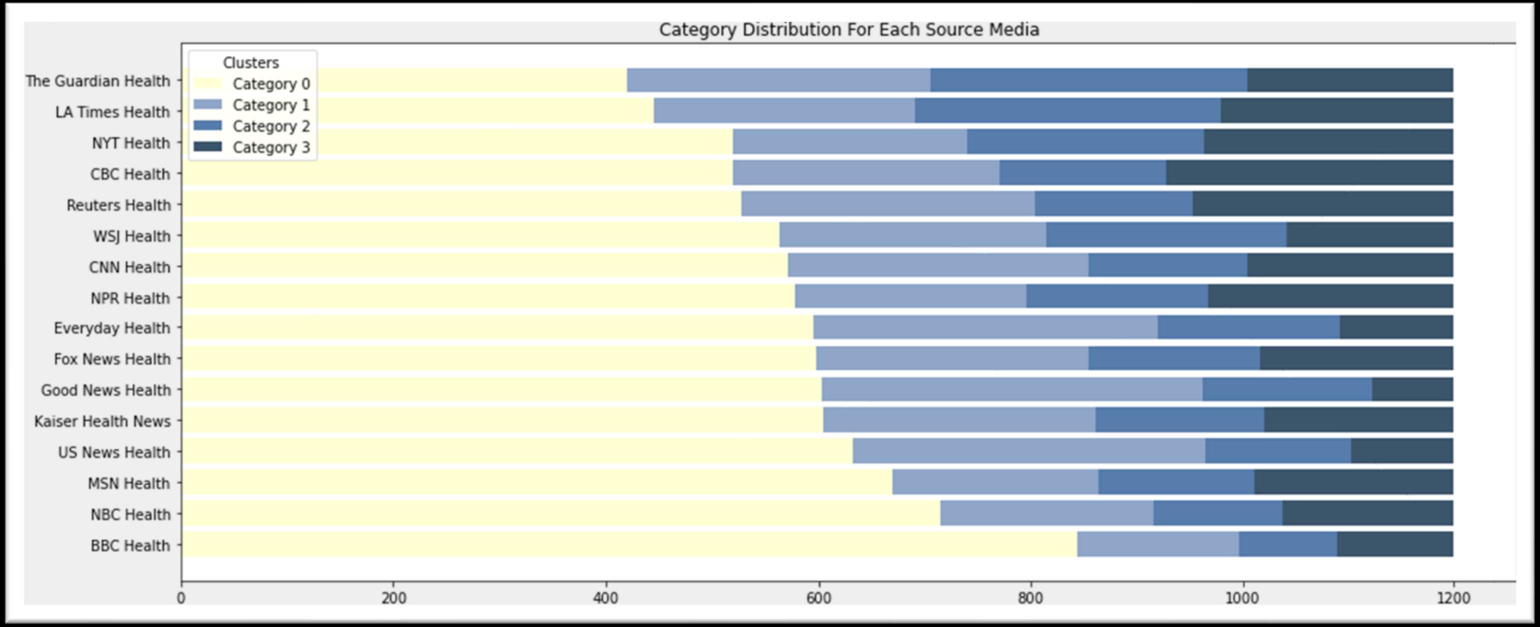


# Tweet Categorization

- Built a model to categorize tweets into 4 categories
- Used Word Count Vectorization, 5 components Matrix Factorization and Kmeans (k=4)
- Tweets from each source account are categorized as shown in the bar chart

- It gives certain evaluations to the source accounts, and categories to the tweets

- Tweets categorization intuitively make sense



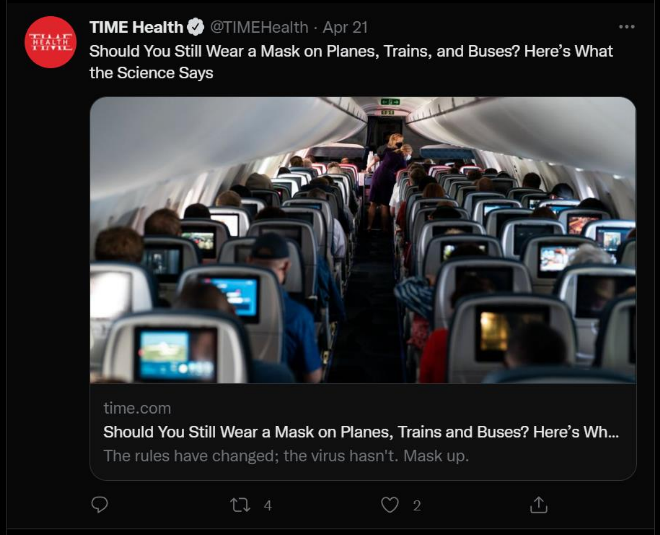
# Evaluation and Conclusion

- Tweet texts are too short to make predictive models
- Tweet analysis should include linked pages, images, and hashtags
- Vectorization/Factorization clustering still yields interesting insights
  - It could give certain labels to twitter accounts
  - It could categorize tweets – could make a basis for recommendation or curation engines
- For validation of unsupervised model, manual review by multiple individuals is mandatory

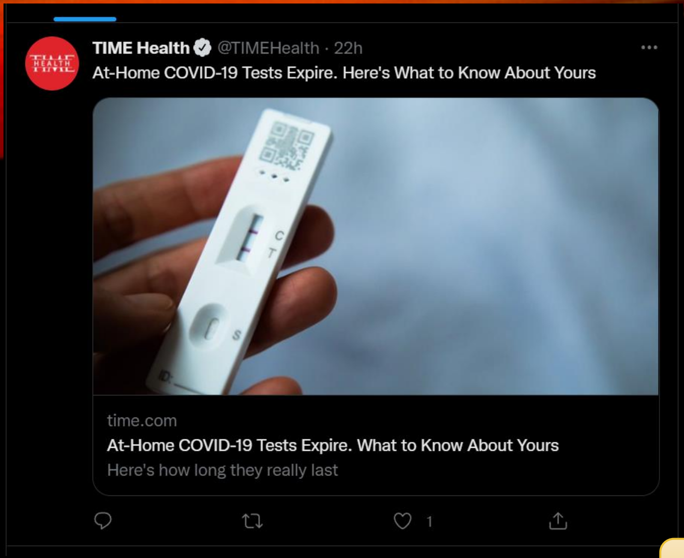
- I gave my personal interpretations on the 4 categories
  - Category 0: social matters and policies, Category 1: diseases and illness
  - Category 2: personal health such as diet, Category 3: scientific and medical
  - This was not too bad – see what the model said on some tweets today in next 4 slides

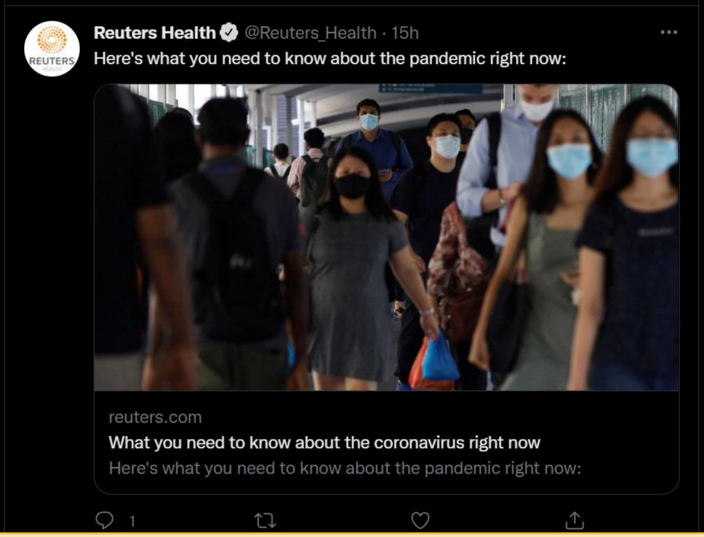
# Category 0: Social and Policies





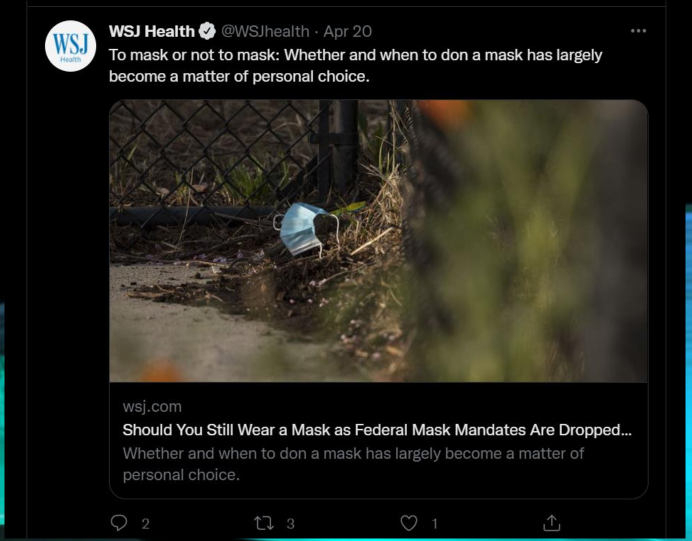

Category 1: Diseases and Illness









# Category 2: Personal Health

Reuters Health @Reuters_Health · 10h

Pfizer said on Friday it was voluntarily recalling five batches of its Accupril blood pressure tablets after finding elevated levels of a potential cancer-causing agent in the medicine.

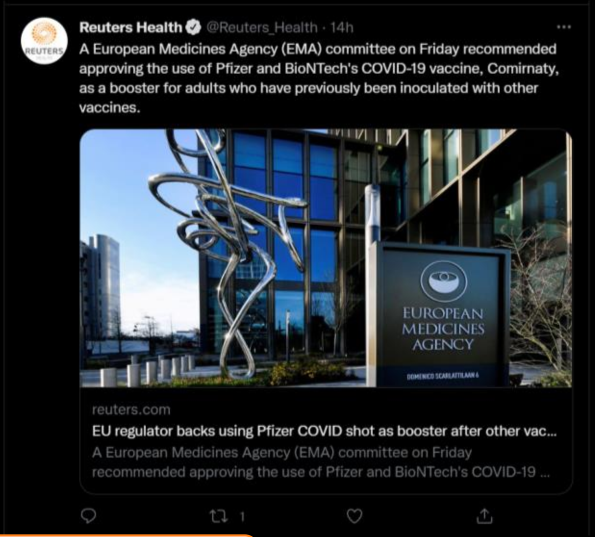


reuters.com

Pfizer recalls some batches of blood pressure drug over carcinogen pr...

Pfizer said on Friday it was voluntarily recalling five batches of its Accupril blood pressure tablets after finding elevated levels of a ...

1 4 7



Reuters Health @Reuters_Health · 14h

A European Medicines Agency (EMA) committee on Friday recommended approving the use of Pfizer and BioNTech's COVID-19 vaccine, Comirnaty, as a booster for adults who have previously been inoculated with other vaccines.



reuters.com

EU regulator backs using Pfizer COVID shot as booster after other vac...

A European Medicines Agency (EMA) committee on Friday recommended approving the use of Pfizer and BioNTech's COVID-19 ...

1

Everyday Health @EverydayHealth · 3h

A recent survey reported that during the COVID-19 pandemic, #anxiety was the No. 1 reason for using CBD; more than one-third of CBD users with anxiety surveyed said they used it five or more times a week to manage their symptoms. But does CBD really help?

everydayhealth.com

Anxiety, CBD, and the Endocannabinoid System

When anxiety gets the heart and mind racing, the ECS puts on the brakes. Could CBD have the sam...

1 2 2

TIME Health @TIMEHealth · Apr 21

Why the Rise of Rapid Tests Makes COVID-19 Case Counts Hard to Trust



time.com

Why the Rise of Rapid Tests Makes COVID-19 Case Counts Hard to Tr...

Countless cases are being missed

# Category 3: Medical Science

**U.S. News Health** @USNewsHealth · 1h
These are the Best Children's Hospitals in the nation for neurology and neurosurgery. #BestHospitals


health.usnews.com
These Are the Top Children's Hospitals in the National for Neurology a...
For treating everything from epilepsy to stroke in kids, these medical centers are the best of the best.

1 2

**WSJ Health** @WSJhealth · Apr 21
Amid rising marijuana use in the U.S., researchers are exploring risks to bystanders, children—and pets

wsj.com
Rising Marijuana Use Presents Risks to Pets, Bystanders
A variety of recent studies have examined the incidental effects of pot use, including one tying its legalization to increased cannabis ...

2 5

**BBC Health News** @bbchealth · 20h
Covid cases in Scotland fall by 30,000 in a week


bbc.com
Covid cases in Scotland fall by 30,000 in a week
It is the fourth week in a row the number of positives tests has dropped, according to official figures.

1 1 6

**BBC Health News** @bbchealth · 19h
Covid-19: Three more deaths and 390 in hospital


bbc.com
Covid-19: Three more deaths and 390 in hospital
The total number of deaths linked to Covid-19 in NI since the start of the pandemic is 3,405.

1 1

**U.S. News Health** @USNewsHealth · 2h
.@TexasChildrens is the top-ranked hospital in treating high-risk heart defects and helping children with severe congenital heart disease.

health.usnews.com
Best Hospitals for Children With Severe Congenital Heart Disease
These 26 hospitals are top-ranked in pediatric heart care and experienced in treating high-risk defects.

1