Distilling from Twitter: New Perspectives in Healthcare Organizations Using Association Rule Mining

Aditya Singhal

Group: 34

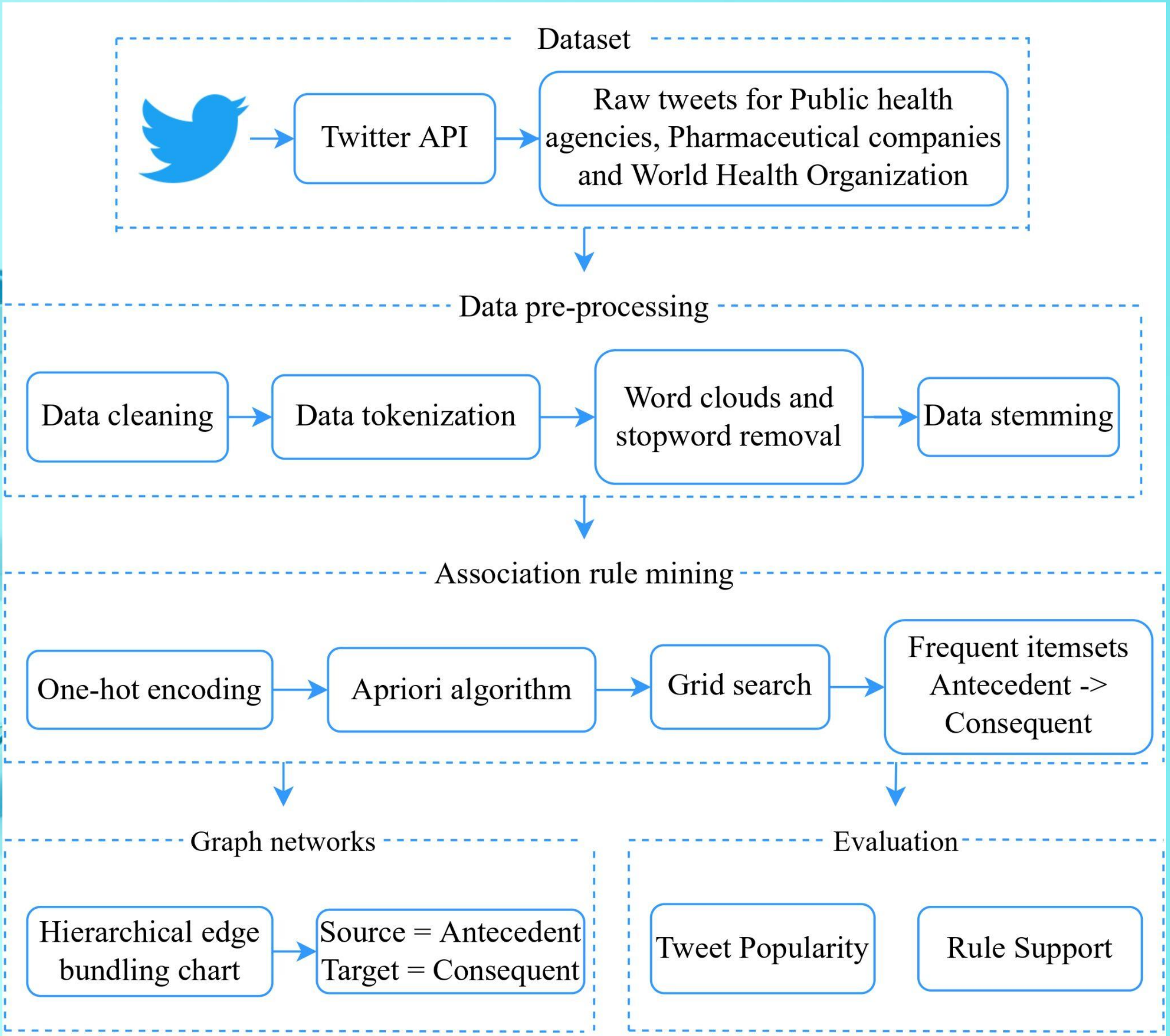
Course: (2023W) COMP-5112-WA - Research Methodology Comp Sci

CONTENTS

- Main Objectives
- Research Overview
- What is Association Rule Mining?
- Data Set
- Pre-processing
- Itemsets and Association Rules
- Graphical Visualizations
- Quantitative Analysis
- Principal findings and Conclusions

Objectives:

- ***Examine Twitter usage by US and Canadian health agencies and pharmaceutical companies***
- ***Identify text patterns influencing the tweets' content.***


Dataset
Twitter API
Raw tweets for Public health agencies, Pharmaceutical companies and World Health Organization
Data pre-processing
Data cleaning
Data tokenization
Word clouds and stopword removal
Data stemming
Association rule mining
One-hot encoding
Apriori algorithm
Grid search
Frequent itemsets Antecedent -> Consequent
Graph networks
Hierarchical edge bundling chart
Source = Antecedent Target = Consequent
Evaluation
Tweet Popularity
Rule Support

Association rule mining is a rule-based machine learning method for discovering interesting if/then statements that help uncover relationships between variables in large databases.

Data Set

Name of organization (Twitter handle)	Total tweets, N
Public health agencies	
Centers for Disease Control and Prevention (CDCgov)	7,511
Indian Health Service (IHSgov)	1,632
Health Canada and PHAC (GovCanHealth)	52,695
Government of Canada for Indigenous (GCIndigenous)	3,725
Total	**65,563**
Pharmaceutical companies	
AstraZeneca (AstraZeneca)	1,284
Glaxo SmithKline (GSK)	2,359
Johnson & Johnson (JNKNews)	2,368
Novartis (Novartis)	715
Pfizer (pfizer)	2,474
Total	**9,200**
Non-governmental organization	
World Health Organization (WHO)	24,581

Table 1: Number of tweets for each organization.

https://creation.co/knowledge/top-50-pharma-tracker-hcps-share-mixed-opinions-on-covid-19-vaccines-and-treatment-data/

Pre-processing

```python
PORTER_STEMMER = PorterStemmer()
```

```python
def clean_tweets(x, STOPWORDS):
    # Lowercase
    sentence = x.lower()

    # Remove all non-alphabets (punctuation, numbers, new-line characters and extra-spaces)
    sentence = re.sub('http[s]?://\S+', '', sentence)
    sentence = re.sub(r'([^a-zA-Z ]+?)', '', sentence)
    #print(sentence)
    #sentence = sentence.replace('\n', '')
    # Remove URLs
    sentence = sentence.replace("world health organization", "who")
    #print(sentence)
    # Remove double spacing
    #sentence = re.sub('\s+', ' ', sentence)
    tokenized_tweet = [word for word in word_tokenize(sentence) if word not in STOPWORDS]
    tokenized_tweet = [PORTER_STEMMER.stem(word) for word in tokenized_tweet]
    return tokenized_tweet
```

```python
# Plotting the wordcloud
# you can specify fonts, stopwords, background color and other options
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS

# Creating the custom stopwords
customStopwords=list(stopwords_df)

wordcloudimage = WordCloud(
                          max_words=100,
                          max_font_size=500,
                          font_step=2,
                          stopwords=customStopwords,
                          background_color='white',
                          width=1000,
                          height=720
                          ).generate(Tweet_Texts_Cleaned)

plt.figure(figsize=(15,7))
plt.axis("off")
plt.imshow(wordcloudimage)
wordcloudimage
plt.show()
```

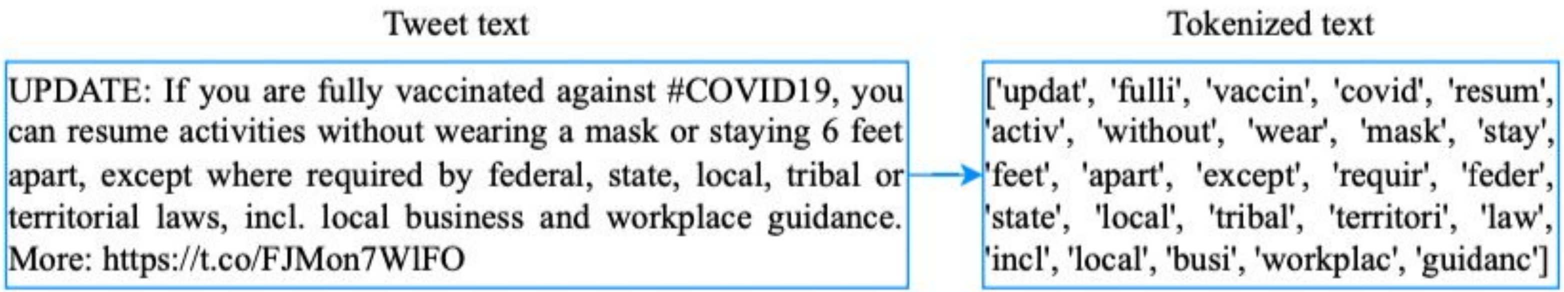


Figure 2: Sample tweet from the Centers for Disease Control and Prevention (CDC).

(a) Public health agencies (b) Pharmaceutical companies (c) World Health Organization

Figure 3: Word clouds for different Twitter groups. Public health agencies and WHO shared content focused on COVID-19.

Language use and Style using Itemsets and Association Rules

- *Mlxtend* python library for ARM
- Data set is encoded in the form of *Numpy* arrays using *TransactionEncoder*() API.
- Using *fit* and *transform* methods, the input data is transformed into a one-hot encoded *Numpy* boolean array

- Generate rules of the form X→Y by performing a grid search, where X is the antecedent and Y refers to the consequent. To find an interesting rule, we calculate the confidence metric

$$confidence(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)} \quad ; range : [0, 1] \tag{1}$$

where

$$support(A \rightarrow C) = support(A \cup C) \quad ; range : [0, 1] \tag{2}$$

- ***Lift*** **metric to filter rules having statistically independent antecedents and consequents, i.e., *lift ≥ 1*.**

$$lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)} \quad range : [0, \infty) \tag{3}$$

Twitter group	Support value			Confidence value		
	Start	End	Step size	Start	End	Step size
Public health agencies	0.1	0.5	0.0625	0.5	1.0	0.1
Pharmaceutical companies	0.01	0.1	0.00625	0.5	1.0	0.1
World health organization	0.01	0.1	0.00625	0.5	1.0	0.1

Table 2: Grid search parameters used for obtaining the number of relevant association rules for each Twitter group.

ARM (Rules)

```
matrix_df = pd.DataFrame(columns=['Threshold Support', 'Threshold Confidence', 'Count of rules'])
for min_support_initialize in np.arange(0.1, 0.5, 0.0625): #0.125, 0.5, 0.0625
  for min_threshold_initialize in np.arange(0.5, 1, 0.1):
    frequent_itemsets_temp = apriori(df, min_support=min_support_initialize, use_colnames=True)
    if(frequent_itemsets_temp.empty):
      continue
    rules = association_rules(frequent_itemsets_temp, metric="confidence", min_threshold=min_threshold_initialize)
  # rules = rules.sort_values(by='confidence', ascending =False)
  # print(rules)
    matrix_df.loc[len(matrix_df.index)] = [min_support_initialize, min_threshold_initialize, len(rules.index)]

print(matrix_df)
```

	Threshold Support	Threshold Confidence	Count of rules
0	0.1000	0.5	1705.0
1	0.1000	0.6	1590.0
2	0.1000	0.7	1316.0
3	0.1000	0.8	1057.0
4	0.1000	0.9	986.0
5	0.1625	0.5	89.0
6	0.1625	0.6	68.0
7	0.1625	0.7	57.0
8	0.1625	0.8	47.0
9	0.1625	0.9	41.0
10	0.2250	0.5	6.0

Twitter group	Support threshold	Confidence threshold	Number of rules
Public health agencies	0.1	0.8	1057
Pharmaceutical companies	0.01625	0.5	274
World health organization	0.01	0.8	1022

Table 3: The parameters used for obtaining association rules for each Twitter group.

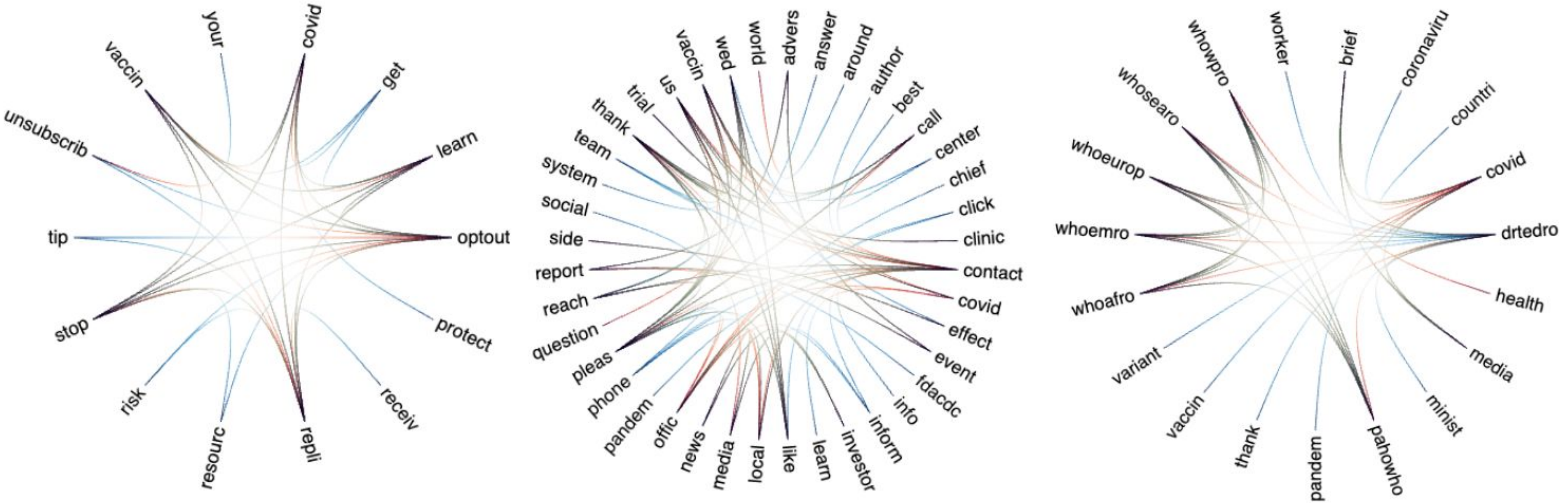
```
# 3            0.1000                    0.8          1057.0
frequent_itemsets_temp = apriori(df, min_support=0.1, use_colnames=True)

rules = association_rules(frequent_itemsets_temp, metric="confidence", min_threshold=0.8)
rules[rules['lift']>=1]
print(rules)
```

Twitter group	Antecedents	Consequents	Antecedent support	Consequent support	Overall support	Confidence	Lift	Leverage	Conviction	Rule Support
Public health agencies	(repli, optout, stop, covid)	(learn, vaccin)	0.122	0.179	0.122	1.0	5.580	0.100	∞	∞
Pharmaceutical companies	(click, offic)	(contact)	0.016	0.107	0.016	1.0	9.330	0.014	∞	∞
World Health Organization	(whoafro, whowpro, pahowho)	(whosearo)	0.019	0.023	0.019	1.0	42.454	0.019	∞	∞

Table 4: Top association rules and performance metrics obtained.

Graphical Visualizations

We create a hierarchical edge bundling chart (*d3js* library) where
Source = antecedents and Target = consequents



(a) Public health agencies (b) Pharmaceutical companies (c) World Health Organization

Figure 5: Graph networks showing Antecedent - Consequent pairs. Public health agencies and WHO generate sparse graphs focused on COVID-19, while pharmaceutical companies generate a denser graph with words from different topics.

Quantitative Analysis using Evaluation Metrics

$$Tweet\_popularity = likes + quotes + retweets + replies \tag{4}$$

$$\begin{aligned} Rule\_support = {} & antecedent\_support \\ & + consequent\_support \\ & + overall\_support + confidence \\ & + lift + leverage + conviction \end{aligned}$$

$$leverage(A \rightarrow C) = S(A \rightarrow C) - S(A) \times S(C)$$

where range: [-1,1], and $S = support$

$$conviction(A \rightarrow C) = \frac{1 - support(C)}{1 - confidence(A \rightarrow C)}$$

where range: [-1,∞)

Frequency of occurrence of association rules in the top 10% of tweets is calculated, and the results obtained are plotted:

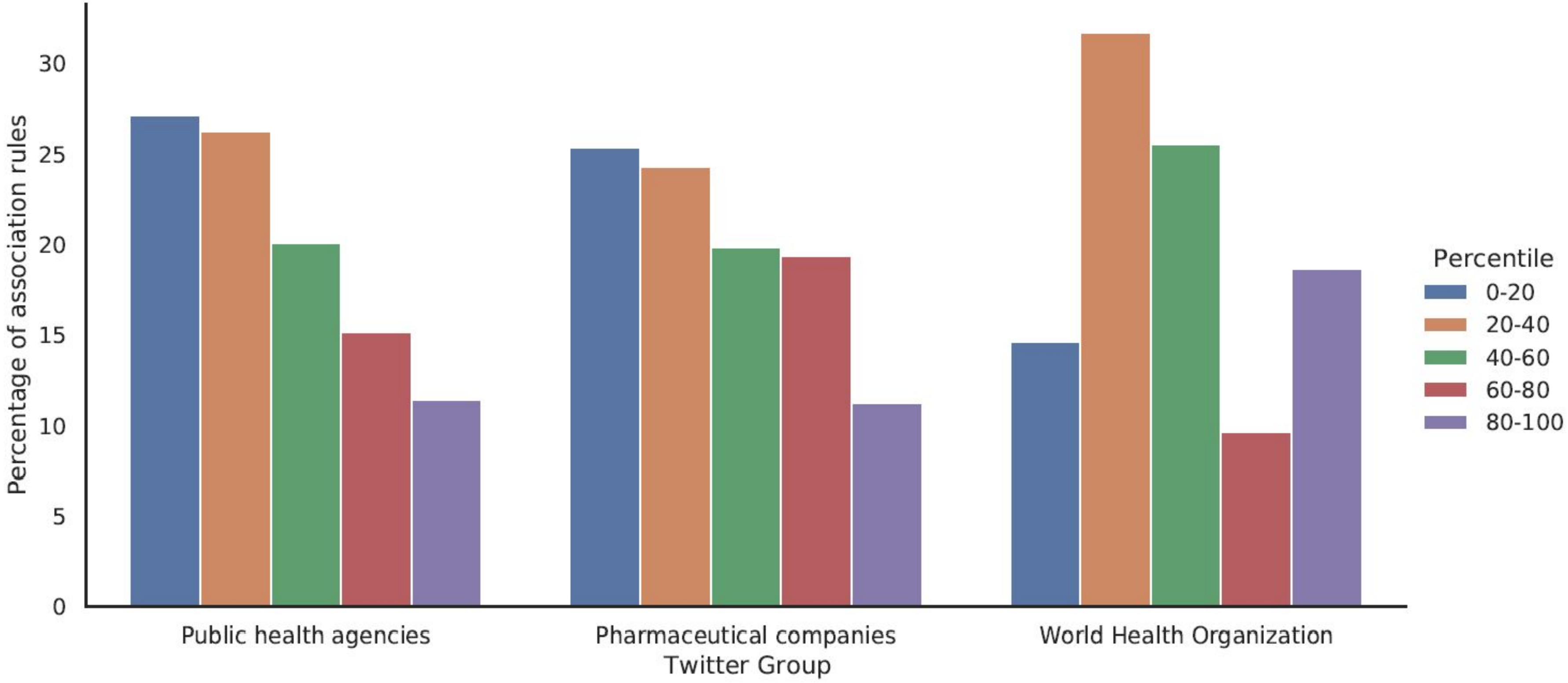


Figure 6: Occurrence of association rules in data set. The top ranked tweets from public health agencies and pharmaceutical companies contain a higher percentage of relevant association rules.

Principal Findings & Conclusion

- **Building a reputation goes beyond just evaluating a tweet’s popularity in the online sphere.**
- **Language use and style across the Twitter groups impacts public engagement.**
- **The association rules, which are mined from existing content, can be utilized to structure future tweets’ content to ensure maximum public engagement.**

Thank you!

Any questions?