Carnegie Mellon University

# Social Media's perception of Small Businesses

JAN 2022 - MAY 2022

Davidson Siga and Rajanikant Tenguria
MISM, CMU

Mentored by
Gabriela Gongora-Svartzman, Ph.D.
Assistant Teaching Professor, CMU

# Agenda

1. **Motivation**
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. Technical Challenges
5. Sentiment Analysis
6. Exploratory Data Analysis
7. Results
8. Conclusion
9. Next Steps

# Agenda

**Motivation**

1. Covid-19 disrupts the world
2. Covid-19 disrupts Small businesses
3. Question we’re trying to answer through Social Media

## Covid-19 disrupts the world

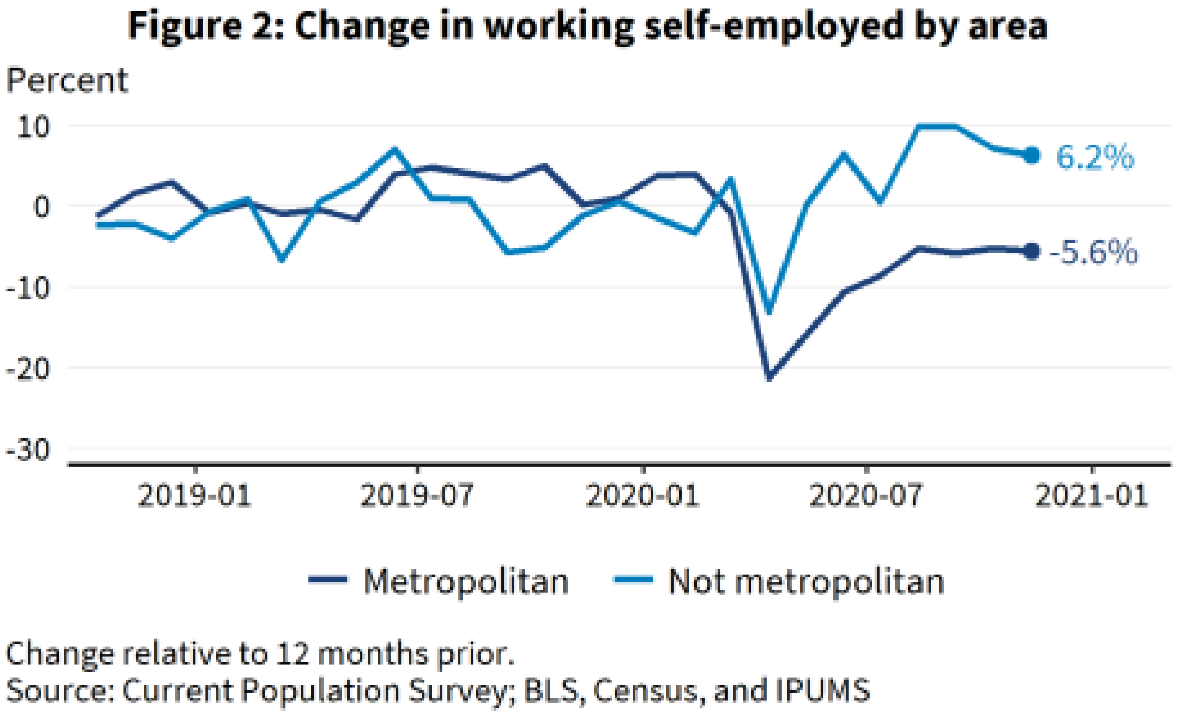
**An Additional 71 million** people are **pushed** into Extreme Poverty in 2020[1] due to COVID-19

*[1] : UN.ORG : https://www.un.org/sustainabledevelopment/wp-content/uploads/2019/07/E_Infographic_01.pdf*

# Covid-19 disrupts Small businesses

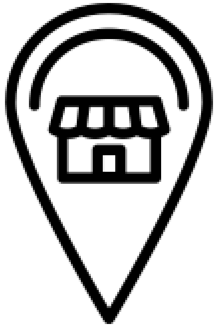
Self employment has steeply increased in covid times.

Location makes a difference in the decision making of SMB owners[2]





[2] : https://cdn.advocacy.sba.gov/wp-content/uploads/2021/03/02112318/COVID-19-Impact-On-Small-Business.pdf



# Question we're trying to answer through Social Media


- Have people started supporting small businesses differently due to covid?

- Does location play a role in this support?

# Agenda

1. Motivation
2. **Where do we get the Data?**
3. Data Cleaning and Transformation
4. Technical Challenges
5. Sentiment Analysis
6. Exploratory Data Analysis
7. Results
8. Conclusion
9. Next Steps

# Agenda

**Where do we get the Data?**

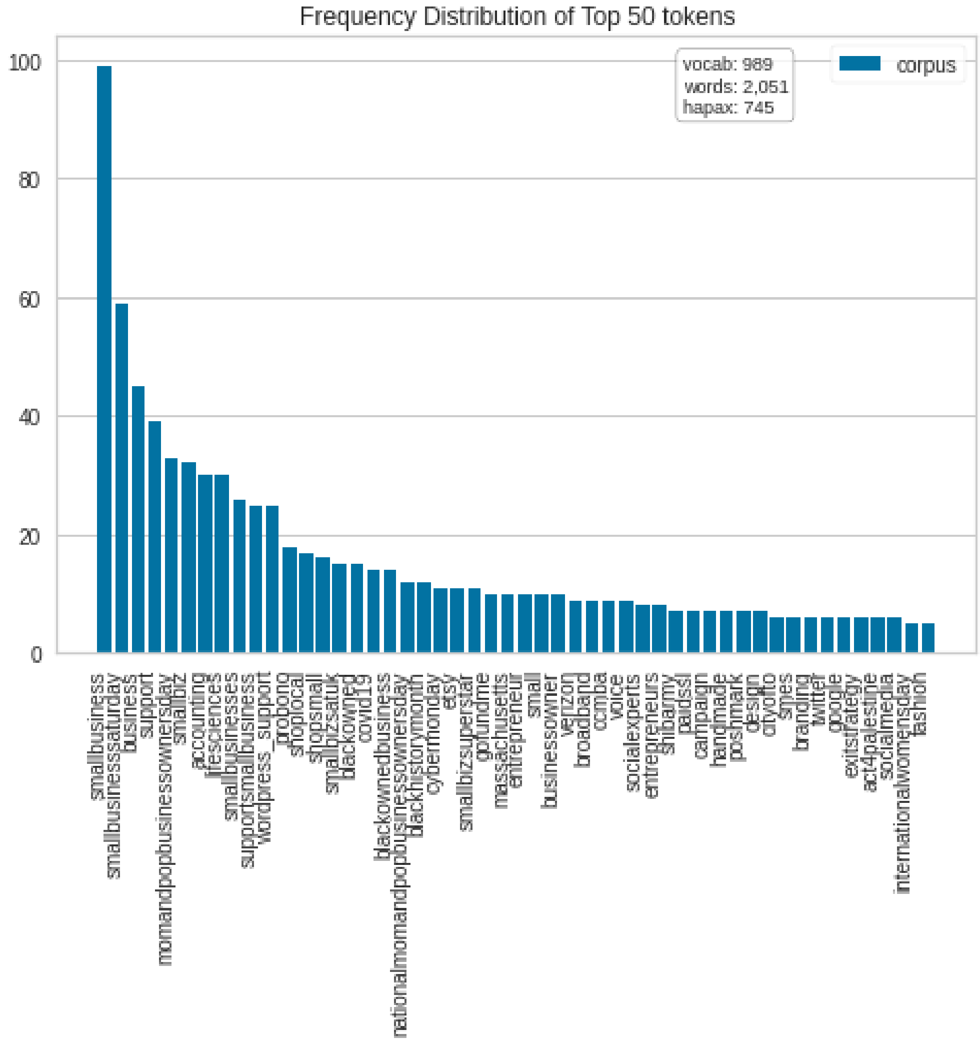
1. Twitter & API Pull
2. Keywords and Hashtag selections

# Twitter & API Pull

- We have pulled 10k original and unique tweets
  - Immediate future work : scale results to a million tweets
- We used the tweepy to pull tweets
- Use of Twitter Research API Access
- No retweets were pulled from Twitter so as to have unique tweets

# Keywords and Hashtag selections

- Initially we got some tweets and extracted hashtags from these tweets.
- These hashtags were later chosen based on the frequency and context count



# Agenda

1. Motivation
2. Where do we get the Data?
3. **Data Cleaning and Transformation**
4. Technical Challenges
5. Sentiment Analysis
6. Exploratory Data Analysis
7. Results
8. Conclusion
9. Next Steps

# Agenda

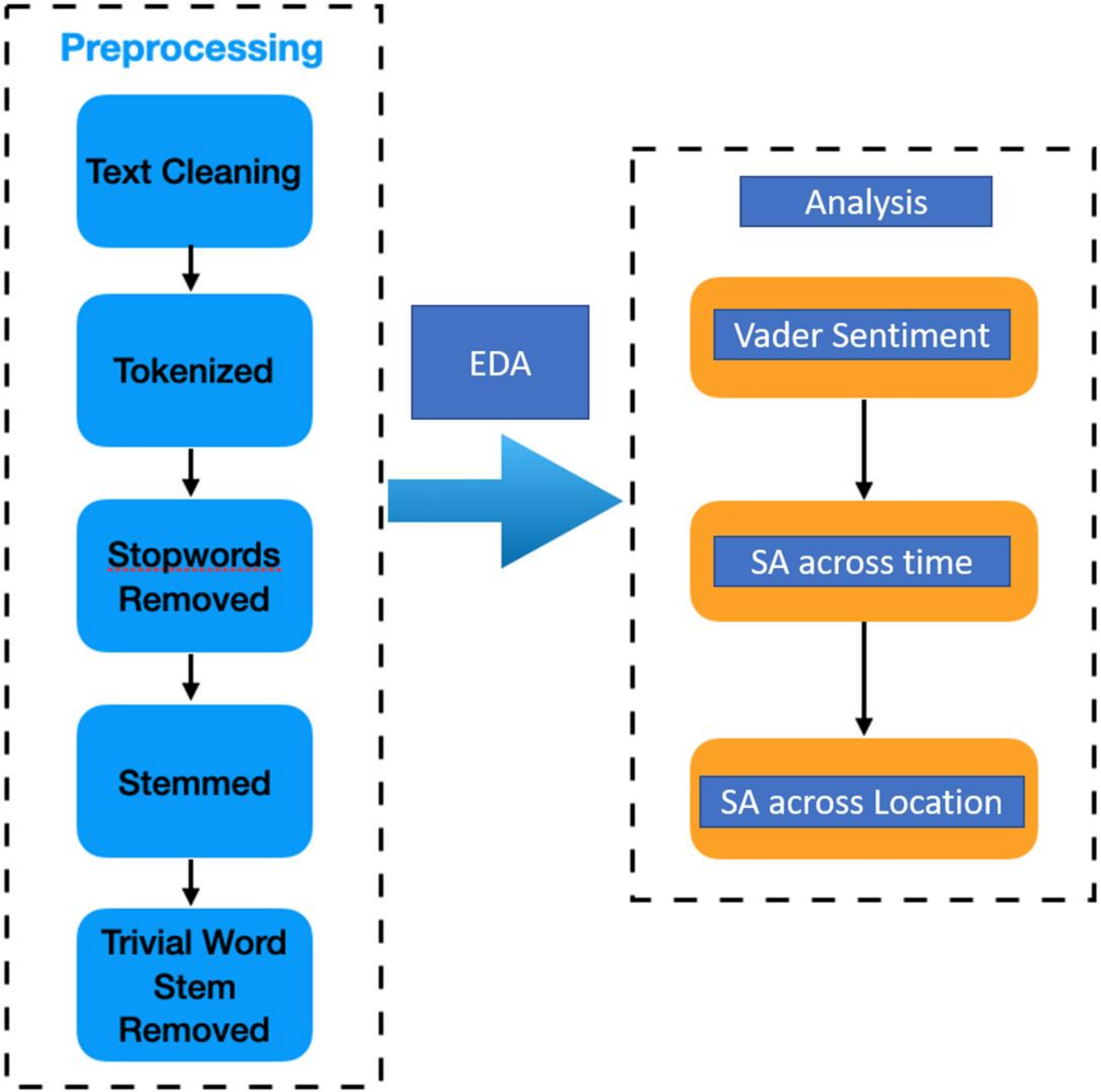
## Data Cleaning and Transformation

1. Data Cleaning Process Flow
2. Location Cleaning Approach
3. Final Dataset for Analysis

# Data Cleaning Process Flow

EDA: Exploratory Data Analysis

SA : Sentiment Analysis




[3] : https://towardsdatascience.com/how-to-build-a-real-time-twitter-analysis-using-big-data-tools-dd2240946e64

# Location Cleaning Approach

Use of GeoTag to clean the location tag.

Challenging task on big data, needs alternative approaches

```
It's coming from inside your tweet!
nan
Takoma Park, MD
Lagos, Nigeria
Davenport, Iowa
Berkeley, CA
Australia
Calvert County , Maryland
Stlmo.,Hinesville,Dallas...
California
Atlanta-ish
she/her || 17
Abuja only, for now
804
619✈228✈817✈301✈812✈540✈804
Hollywood, FL
under your bed
San Diego, CA
Spacecoast, FL
Cairo, EGYPT
卍
Washington, DC
```

Original Data

```
[None,
 Location(Nanno, Ville d'Anaunia, Comunità della Val di Non, Provinc
 Location(Takoma Park, Montgomery County, Maryland, United States, (
 Location(Lagos, Lagos Island, Lagos, 100242, Nigeria, (6.4550575, 3
 Location(Davenport, Scott County, Iowa, 52801, United States, (41.5
 Location(University of California, Berkeley, Milvia Street, North B
 Location(Australia, (-24.7761086, 134.755, 0.0)),
 Location(Calvert County, Maryland, United States, (38.5288529, -76.
 None,
 Location(California, United States, (36.7014631, -118.755997, 0.0))
 None,
 None,
 None,
 Location(804 봉, 산동면, 구례군, 전라남도, 57602, 대한민국, (35.33126
 None,
 Location(Hollywood, Broward County, Florida, United States, (26.011
 None,
 Location(San Diego, San Diego County, California, United States, (3
 None
```

GeoTag Outputs

# Final Dataset for Analysis

Information ranging from User's Aggregated information

to

Tweet specific information were a part of the pull

We're interested in processes_texts (an outcome of our Data Cleaning)

```
Index(['Unnamed: 0', 'created_at', 'id', 'id_str', 'text',
       'display_text_range', 'source', 'truncated', 'in_reply_to_status_id',
       'in_reply_to_status_id_str', 'in_reply_to_user_id',
       'in_reply_to_user_id_str', 'in_reply_to_screen_name', 'user', 'geo',
       'coordinates', 'place', 'contributors', 'is_quote_status',
       'quote_count', 'reply_count', 'retweet_count', 'favorite_count',
       'entities', 'favorited', 'retweeted', 'filter_level', 'lang',
       'matching_rules', 'verified', 'favourites_count', 'user-screen_name',
       'user-location', 'hashtag_list', 'extended_tweet',
       'extended_tweet-full_text', 'extended_entities', 'possibly_sensitive',
       'quoted_status_id', 'quoted_status_id_str', 'quoted_status',
       'quoted_status_permalink', 'quoted_status-user-screen_name',
       'quoted_status-text', 'quoted_status-extended_tweet-full_text',
       'place-country', 'place-country_code', 'location-coordinates', 'scopes',
       'processed_texts', 'POS'],
      dtype='object')
```

# Agenda

1. Motivation
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. **Technical Challenges**
5. Sentiment Analysis
6. Exploratory Data Analysis
7. Results
8. Conclusion
9. Next Steps

# Agenda

## Technical Challenges

1. QA Log

# QA Document Gist

**Data Scarcity** - not many people talk about the topic

**Data pull** - continuous blocking by Twitter

**Location cleaning and grouping** for better and consolidated results is essential

**Detailed QA Log available here :**
https://docs.google.com/document/d/1fmC44xrybJrGUIL1wYVpO_tBWvqt3psnVFlLJ8Og7hY/edit?usp=sharing

# Agenda

1. Motivation
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. Technical Challenges
5. **Sentiment Analysis**
6. Exploratory Data Analysis
7. Results
8. Conclusion
9. Next Steps

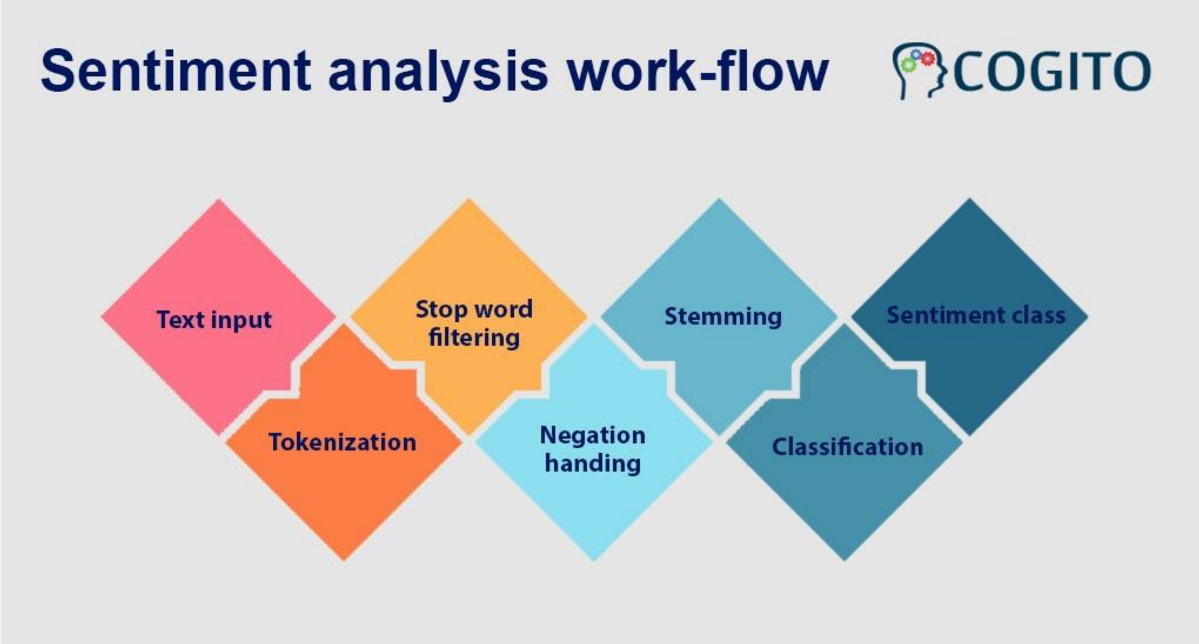
# Agenda

## Sentiment Analysis

1. What is Sentiment Analysis?
2. Vader Sentiment Analyzer

# What is Sentiment Analysis?

The background of what typically happens in a sentiment analyzer




 [3] : https://cogitotech.medium.com/sentiment-analysis-how-it-works-types-everything-you-need-to-know-822a1f2ddeaf

# Vader Sentiment Analyzer

We extracted tweet sentiments of people talking on twitter

This was performed using Vader.

| | label | review | scores | compound |
|---|---|---|---|---|
| **0** | pos | Stuning even for the non-gamer: This sound tra... | {'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co... | 0.9454 |
| **1** | pos | The best soundtrack ever to anything.: I'm rea... | {'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co... | 0.8957 |
| **2** | pos | Amazing!: This soundtrack is my favorite music... | {'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com... | 0.9858 |
| **3** | pos | Excellent Soundtrack: I truly like this soundt... | {'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com... | 0.9814 |
| **4** | pos | Remember, Pull Your Jaw Off The Floor After He... | {'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp... | 0.9781 |

 [4] : https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

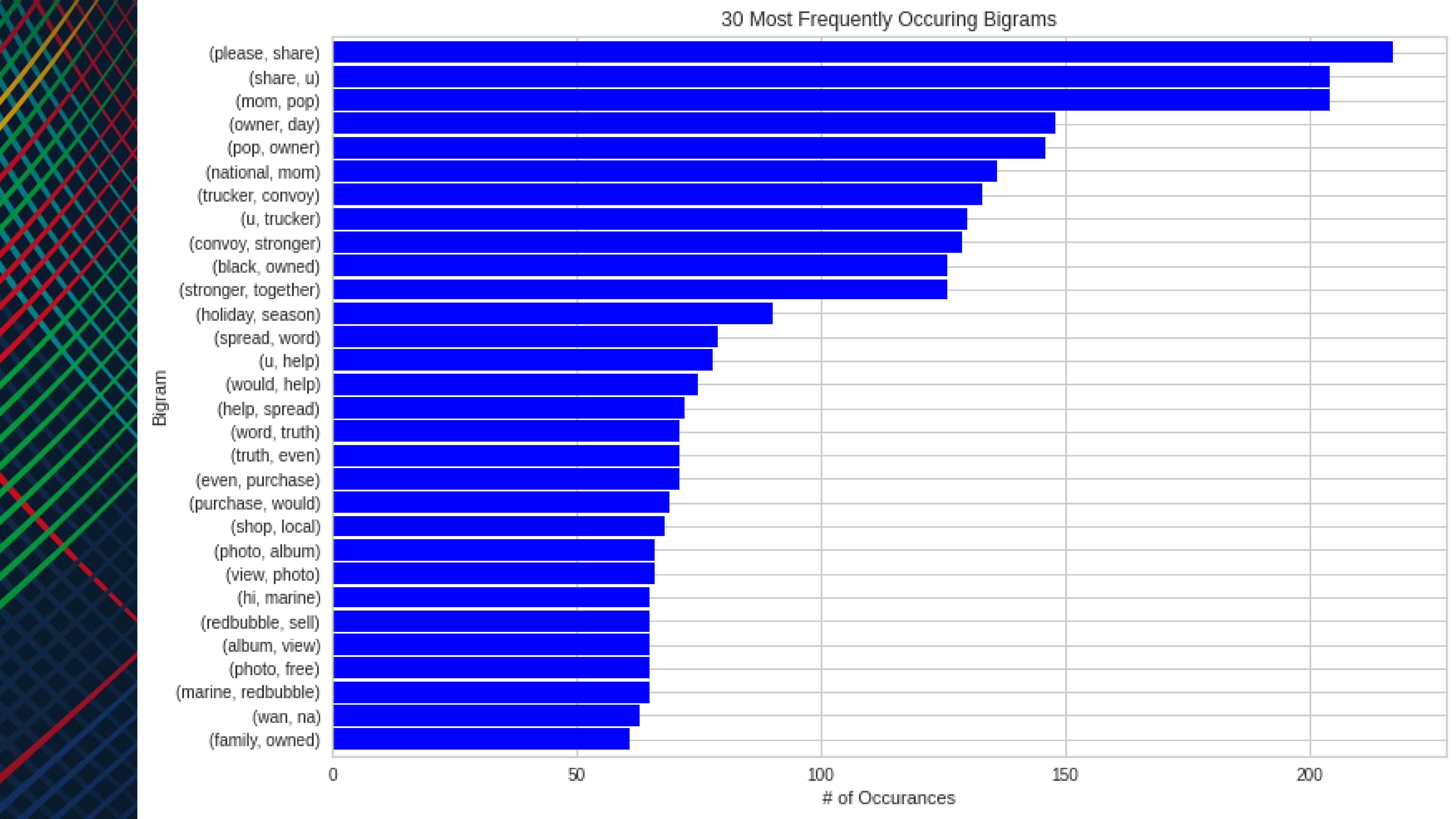
# Agenda

1. Motivation
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. Technical Challenges
5. Sentiment Analysis
6. **Exploratory Data Analysis**
7. Results
8. Conclusion
9. Next Steps

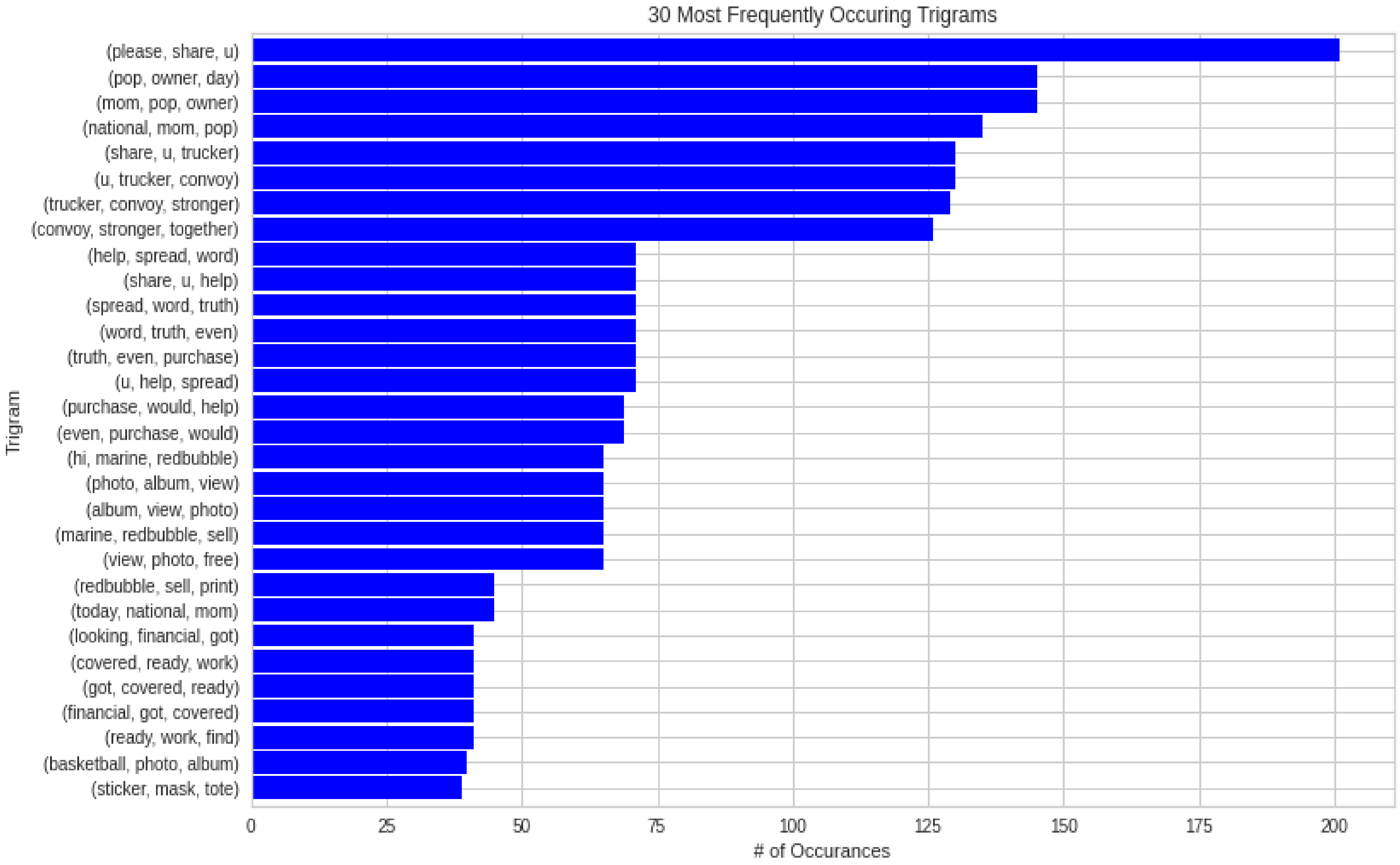
# Agenda

**Exploratory Data Analysis**

1. N-Gram Analysis
2. LDA



30 Most Frequently Occuring Bigrams
(please, share)
(share, u)
(mom, pop)
(owner, day)
(pop, owner)
(national, mom)
(trucker, convoy)
(u, trucker)
(convoy, stronger)
(black, owned)
(stronger, together)
(holiday, season)
(spread, word)
(u, help)
(would, help)
(help, spread)
(word, truth)
(truth, even)
(even, purchase)
(purchase, would)
(shop, local)
(photo, album)
(view, photo)
(hi, marine)
(redbubble, sell)
(album, view)
(photo, free)
(marine, redbubble)
(wan, na)
(family, owned)
Bigram
0
50
100
150
200
# of Occurances

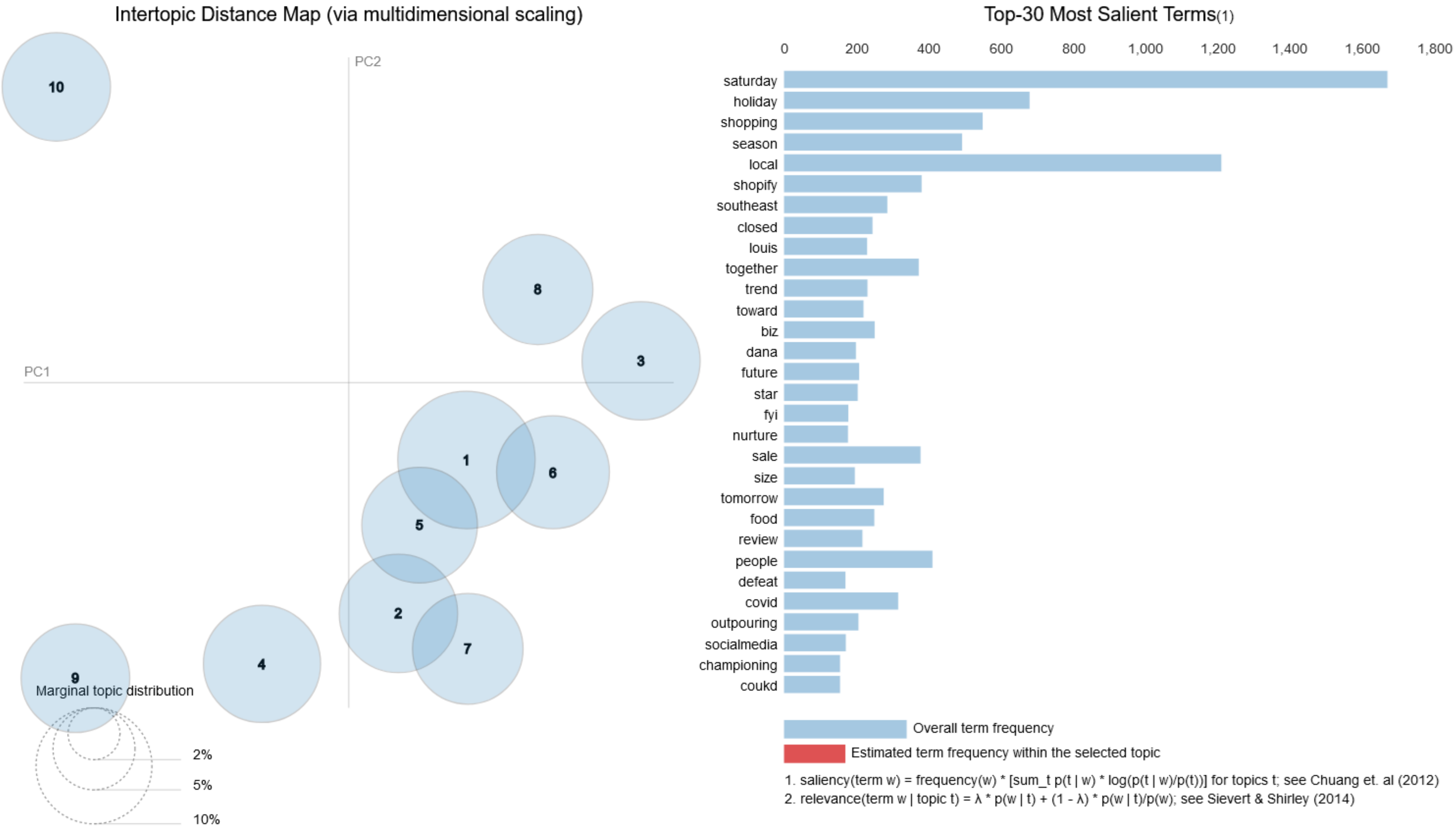


30 Most Frequently Occuring Trigrams
Trigram
(please, share, u)
(pop, owner, day)
(mom, pop, owner)
(national, mom, pop)
(share, u, trucker)
(u, trucker, convoy)
(trucker, convoy, stronger)
(convoy, stronger, together)
(help, spread, word)
(share, u, help)
(spread, word, truth)
(word, truth, even)
(truth, even, purchase)
(u, help, spread)
(purchase, would, help)
(even, purchase, would)
(hi, marine, redbubble)
(photo, album, view)
(album, view, photo)
(marine, redbubble, sell)
(view, photo, free)
(redbubble, sell, print)
(today, national, mom)
(looking, financial, got)
(covered, ready, work)
(got, covered, ready)
(financial, got, covered)
(ready, work, find)
(basketball, photo, album)
(sticker, mask, tote)
0
25
50
75
100
125
150
175
200
# of Occurances

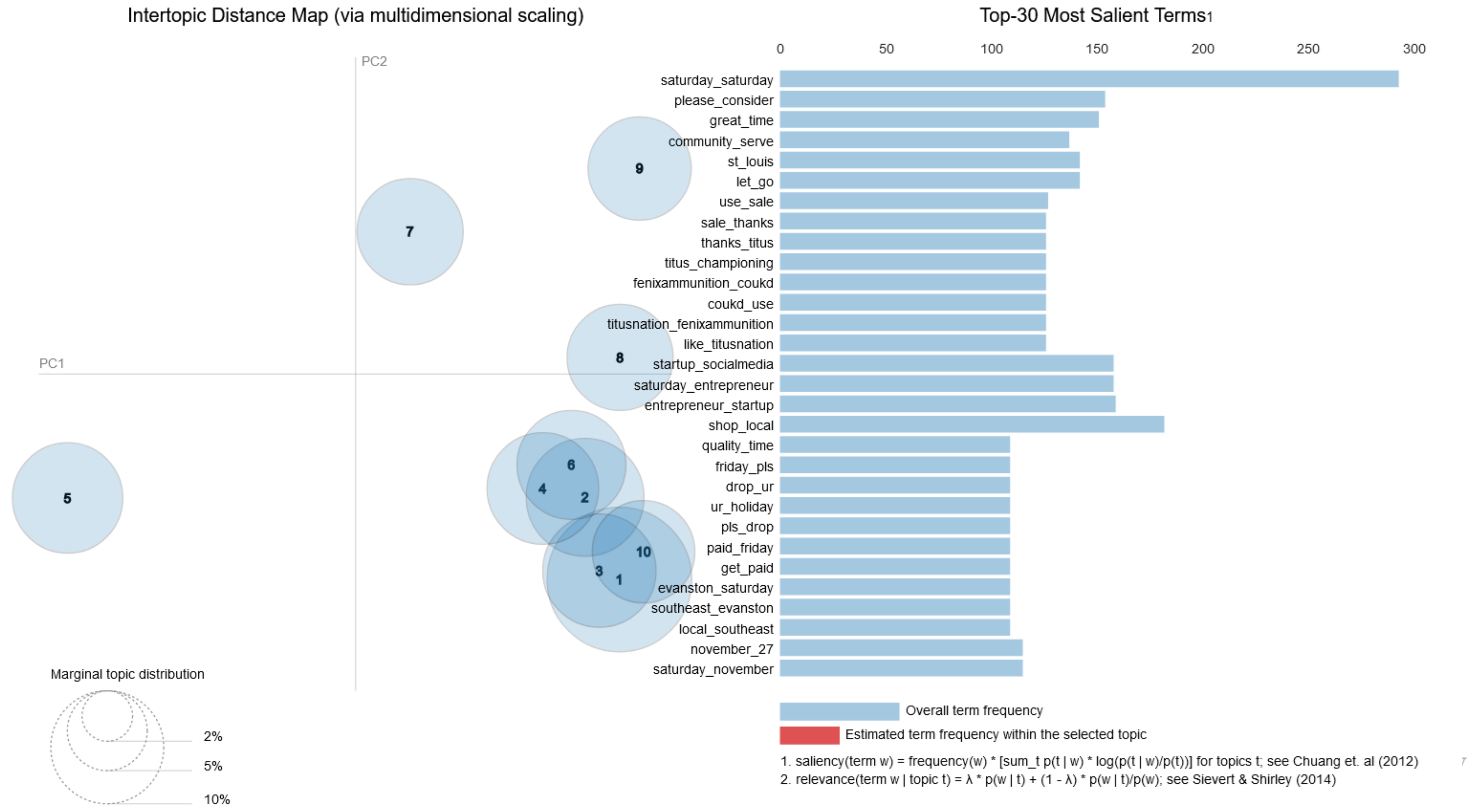
# LDA (Latent Dirichlet Allocation)

- LDA is a topic modeling technique which we used to extract most talked about topics.
- Here we used n-grams that were used to get relevant results.
- We observed trigrams and bigrams better results as higher n values give more context
- However, four-grams did not perform increase relevance and was time-intensive

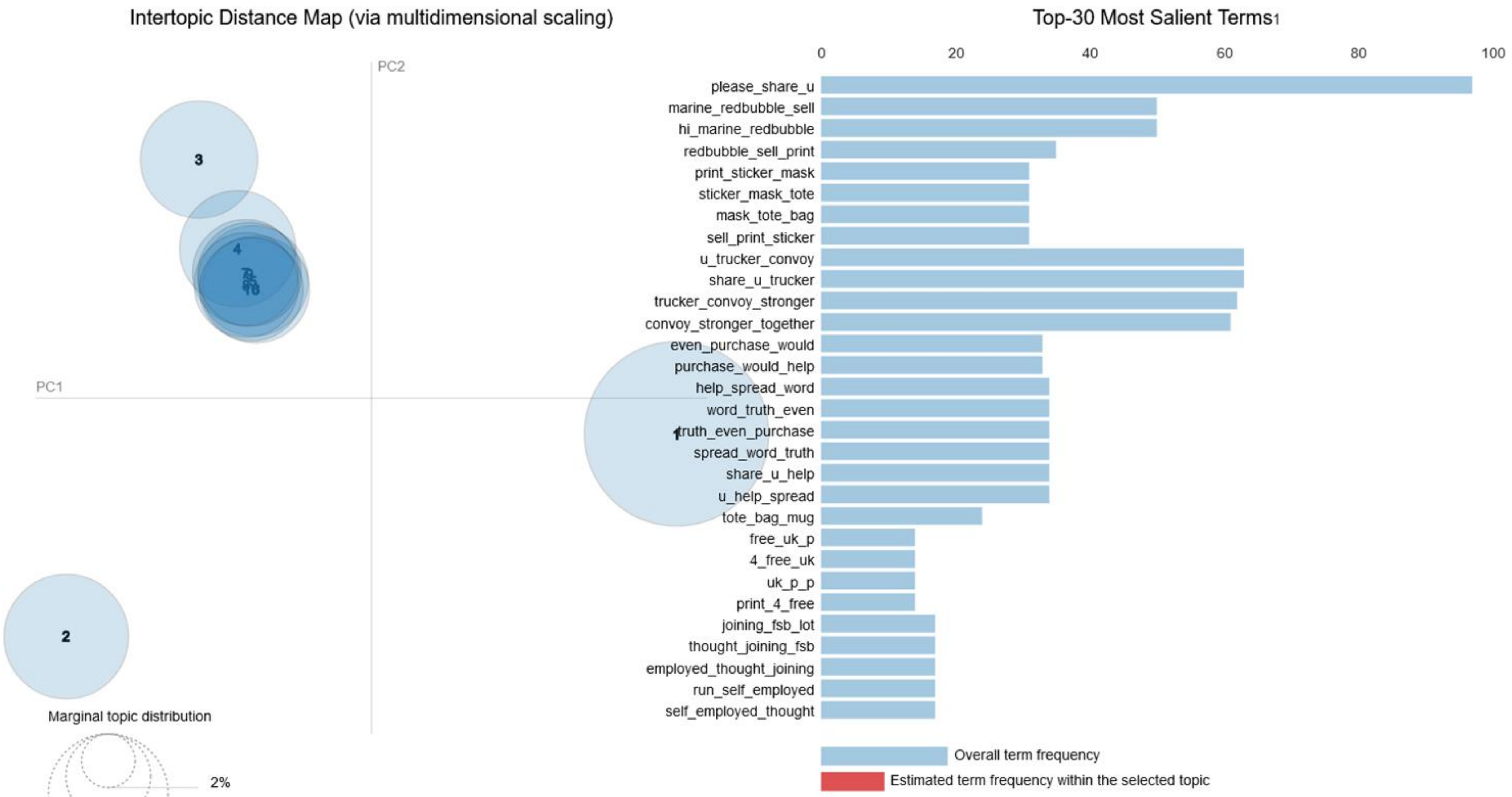
# LDA (Latent Dirichlet Allocation) - unigrams



# LDA (Latent Dirichlet Allocation) - bigrams



# LDA (Latent Dirichlet Allocation) - trigrams



# Agenda

1. Motivation
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. Technical Challenges
5. Sentiment Analysis
6. Exploratory Data Analysis
7. **Results**
8. Conclusion
9. Next Steps

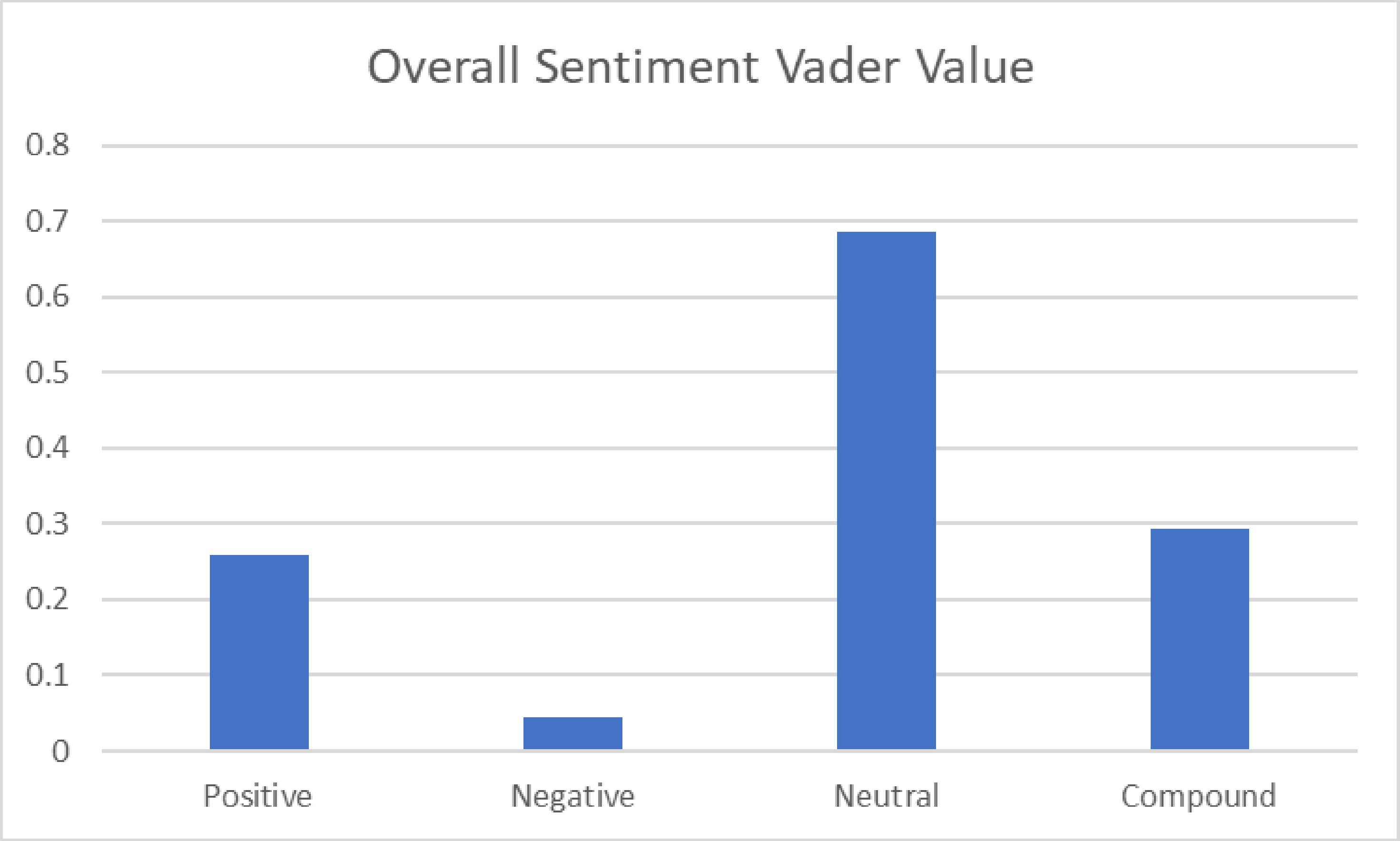
# Agenda

## Results

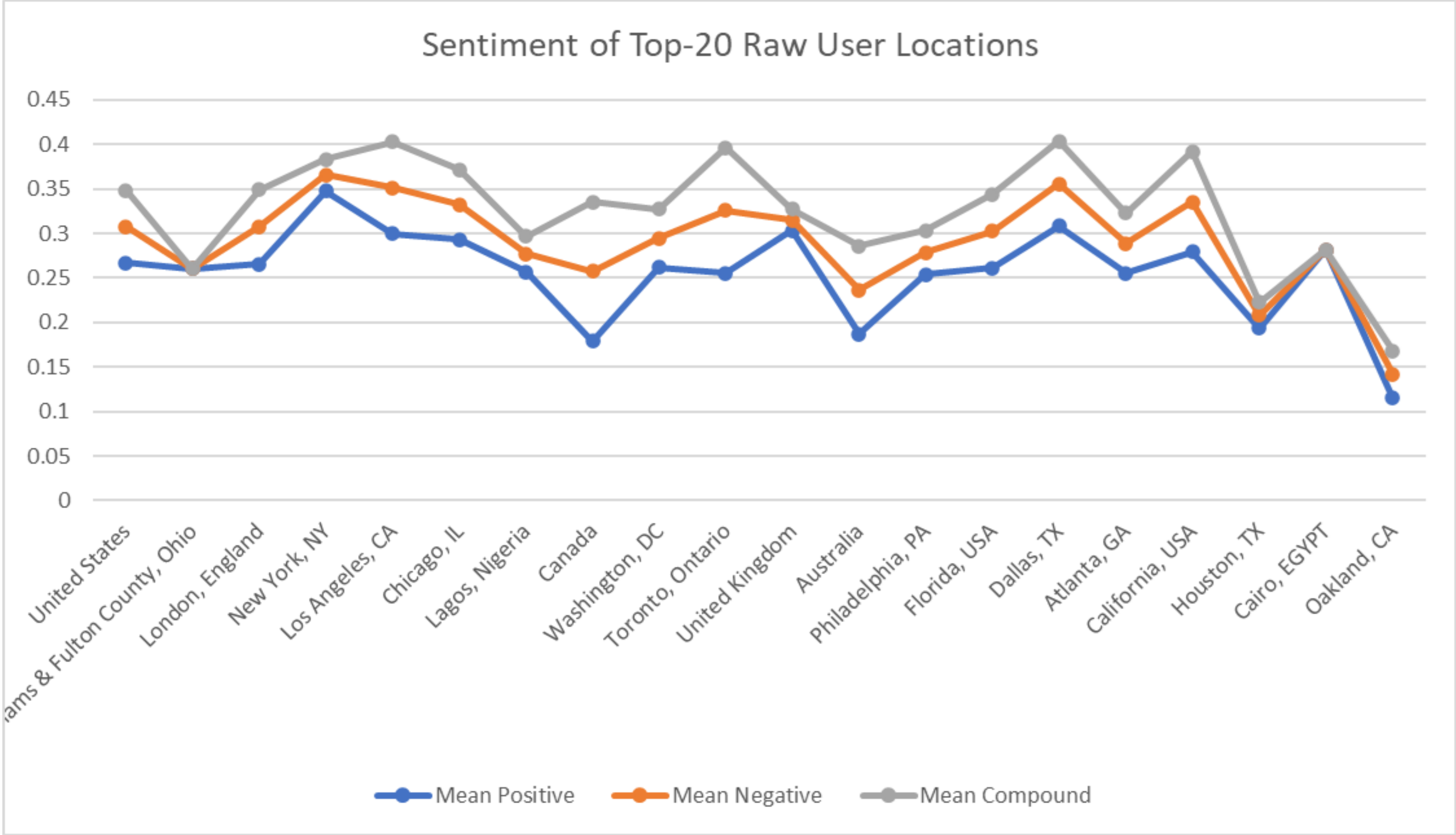
1. Overall Sentiment
2. Sentiment across Locations
3. Sentiment across Time
4. Sentiment across Locations and Time

# Overall Sentiment

| Overall Sentiment | |
|---|---|
| **Sentiment Type** | **Vader Value** |
| Positive | 0.257962179 |
| Negative | 0.043617503 |
| Neutral | 0.686645744 |
| Compound | 0.292655823 |



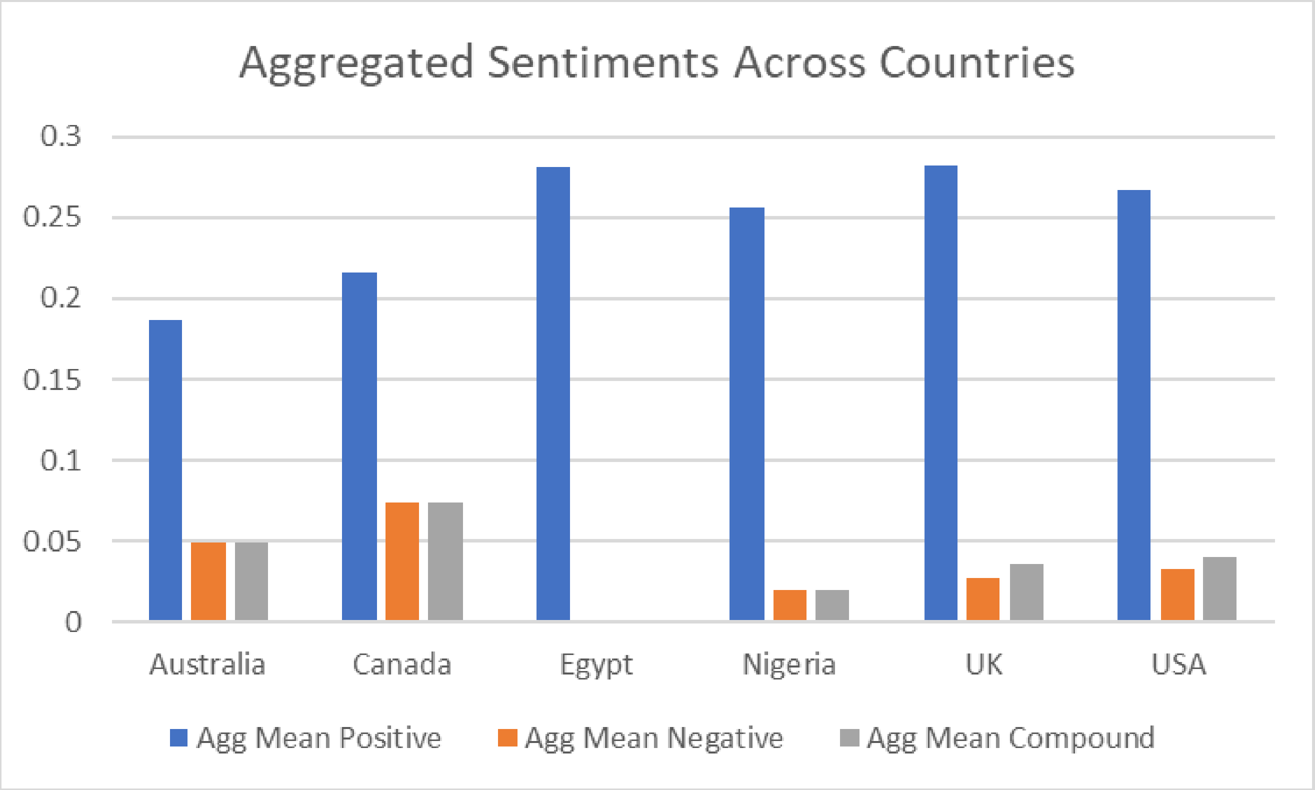
# Sentiment across Locations



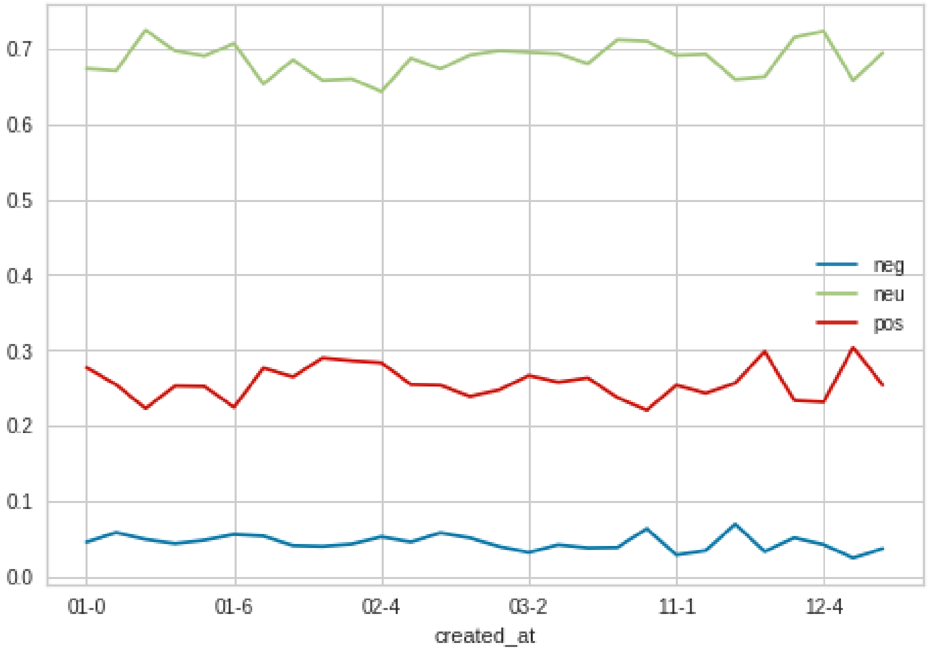
# Sentiment across Locations

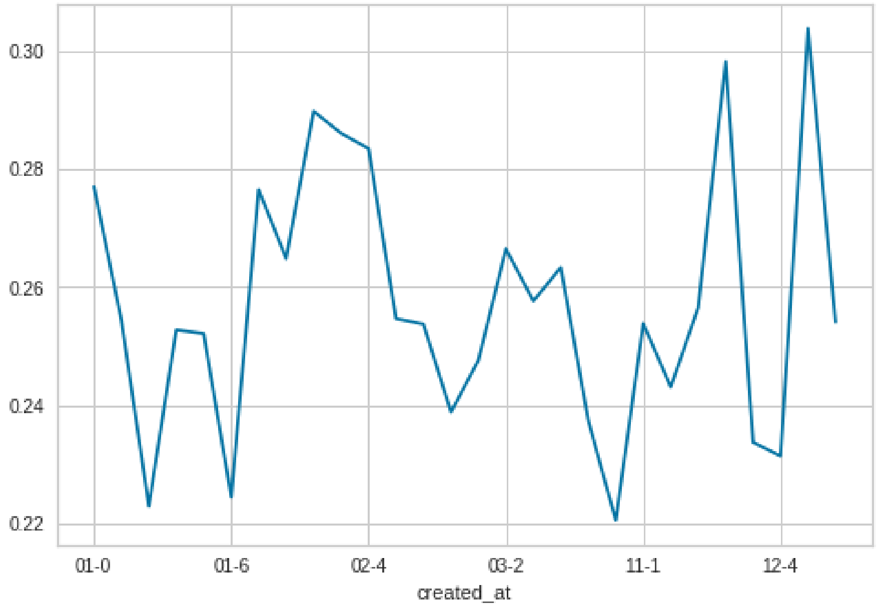
**Some Speculative Comments :**

- Maybe Egypt and Nigeria are amazing places to be an entrepreneur
- UK>USA>Australia ~ Canada
- USA has a better overall sentiment than Canada towards SMB



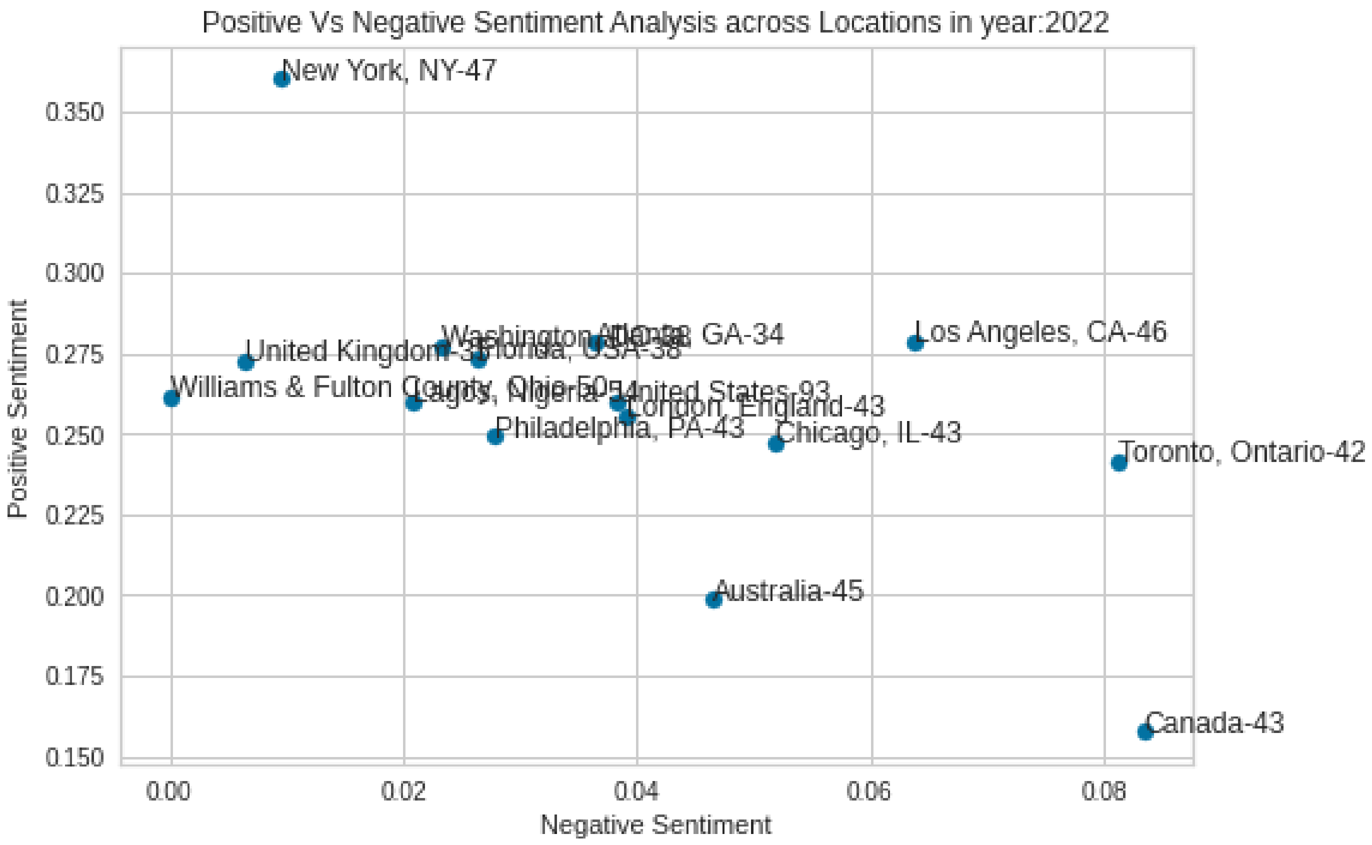
# Sentiment across Time



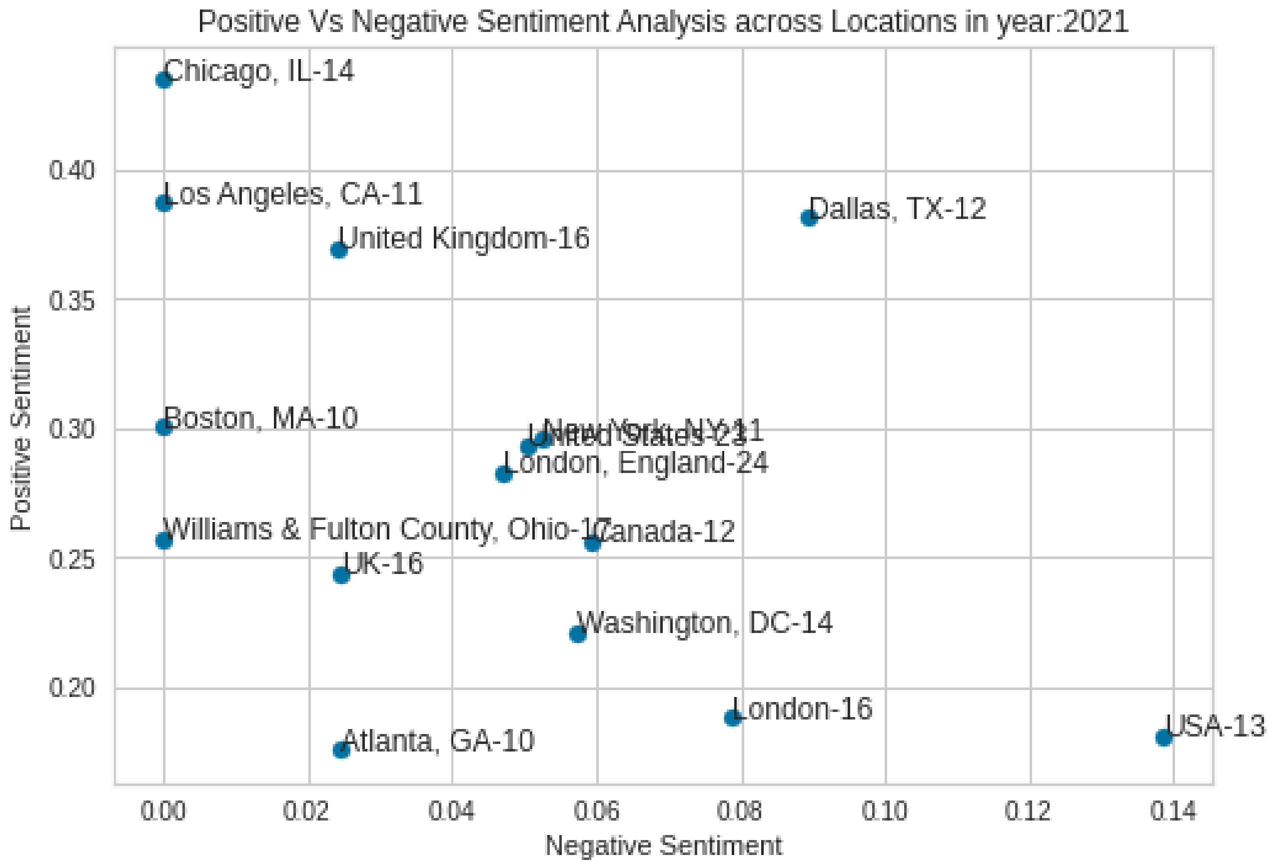

- Overall Sentiment mostly positive towards SMB across time
- Slight movement towards more positive and less negative in time

# Sentiment across Locations and Time




Carnegie Mellon University

# Sentiment across Locations and Time



# Agenda

1. Motivation
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. Technical Challenges
5. Sentiment Analysis
6. Exploratory Data Analysis
7. Results
8. **Conclusion**
9. Next Steps

# Conclusion

- As discussed earlier we can see there was an **improved positive sentiment during covid** towards Small Businesses
- We're able to see **different locations** having **different** sentiments/**opinions** on small businesses
- **Unexpected Results:** Egypt and Nigeria have shown immense support to Small Businesses
- **Speculation :** for the latest data we can see that UK has good place for SMB

# Agenda

1. Motivation
2. Where do we get the Data?
3. Data Cleaning and Transformation
4. Technical Challenges
5. Sentiment Analysis
6. Exploratory Data Analysis
7. Results
8. Conclusion
9. **Next Steps**

# Next Steps

- **More Data:** We plan to extract even more data and perform better location clustering to get even more insights about the data
- **Improved Location Mapping :** Current flaw in processing location needs fix
- **Improved Clustering :** As of now, we have tried LDA (n-grams) and initial NMF clustering techniques. We plan to dig deeper in the clustering techniques to get a better overview about the topics and capture more relevance from the data
- **External Data Sources :** We plan to merge the current tweet data with the data related to the Covid waves and see further trends explained by them
- **Publish Research :** Conversations on this space are new (try **INFORMS** Confs.)

# Thank you!

Happy to connect :)

Davidson Siga
dsiga@cmu.edu
linkedin.com/in/davidsonsiga
MISM 16, CMU

Rajanikant Tenguria
trajanikant@cmu.edu
linkedin.com/in/algorrt
MISM 12, CMU