SOCIAL MEDIA
SENTIMENT ANALYSIS & ENTITY EXTRACTION

MATT DEMARCO

DATA 670 - DATA ANALYTICS CAPSTONE

PRESENTATION 3 - FINAL REPORT

PROFESSOR DR. JON MCKEEBY

11.20.2018

PROJECT BACKGROUND & IMPORTANCE

- Social media can and has been weaponized to influence businesses, governments and political figures (Zeitzoff, 2018).
- Positive user reviews on social media have been shown to increase sales by 5 to 9 percent (Luca, 2011).
- This project will benefit various entities to allow them leverage machine learning natural language processing for sentiment analysis and entity extraction.

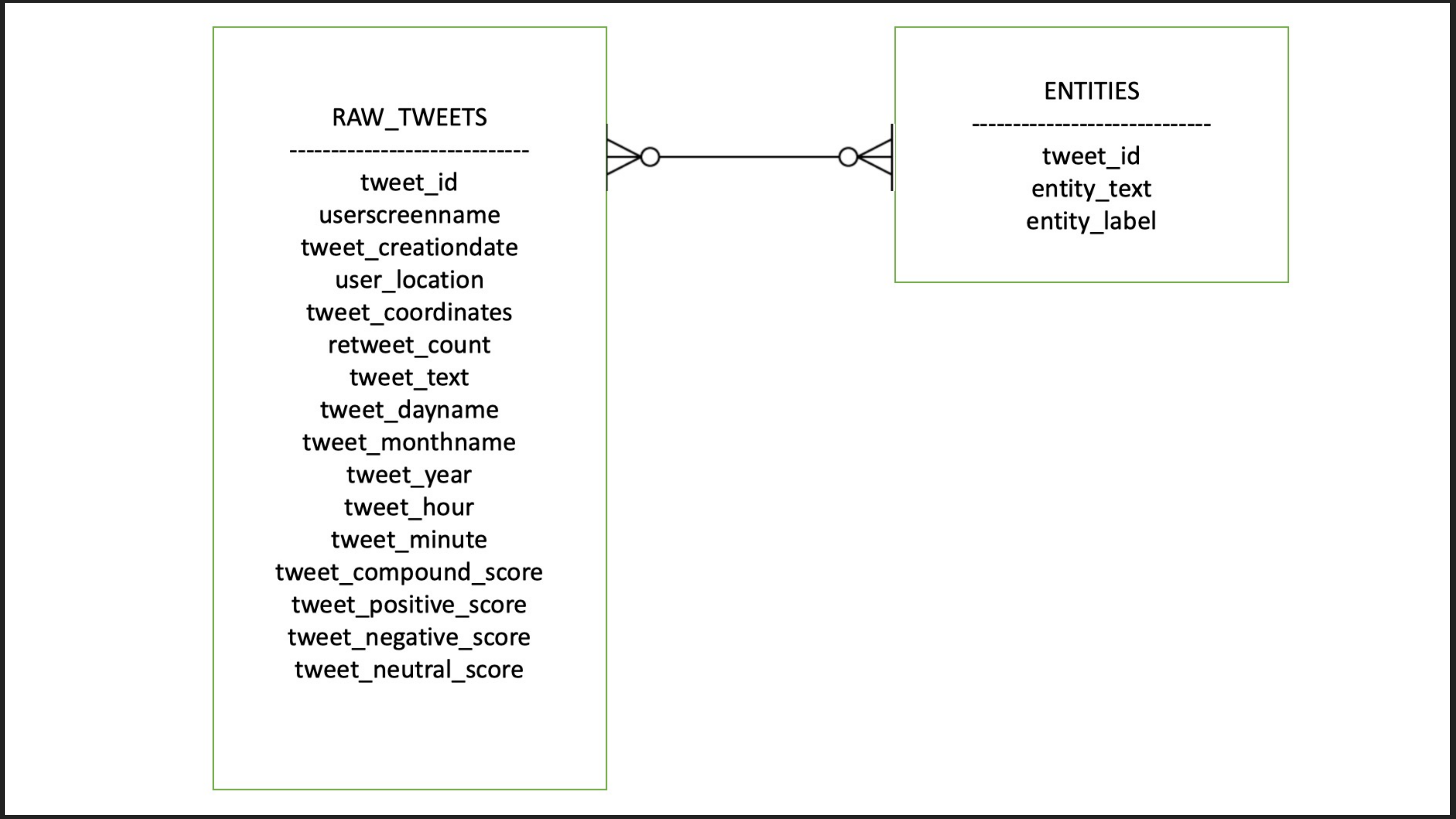
PROJECT SCOPE

- The scope of this project is to collect and analyze the sentiment of a social media posting to understand the negative, neutral or positive tone of the post.
- Data delivery will be done via dashboarding tools.

TWITTER DATA SET

Data Entity	Data Description
created_at	Datetime stamp of tweet creation
id_str	Randomly generated identification number
text	Body of the tweet (subject of analysis)
user	Twitter username
place	Geotagged location of tweet
entities	An array of hashtags, user mentions, etc…
extended_entities	Media tags, if needed

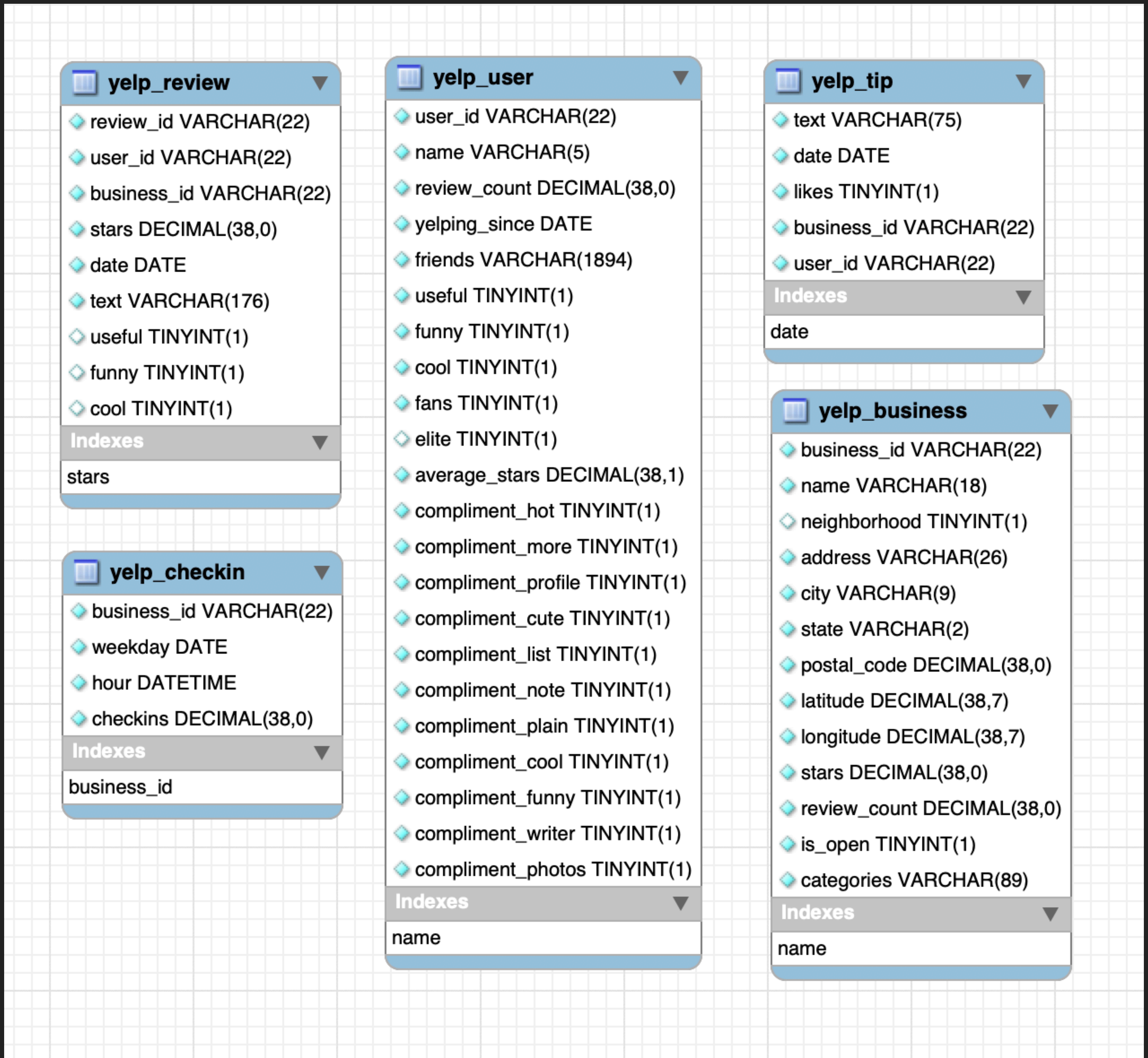
TWITTER SCHEMA



YELP DATA SET

Data Entity	Data Description
business.csv	Contains business data including location data, attributes, and categories.
review.csv	Contains full review text data including the user_id that wrote the review and the business_id the review is written for. (Subject of analysis)
user.csv	User data including the user's friend mapping and all the metadata associated with the user.
checkin.csv	Checkins on a business
tip.csv	Tips written by a user on a business

YELP SCHEMA

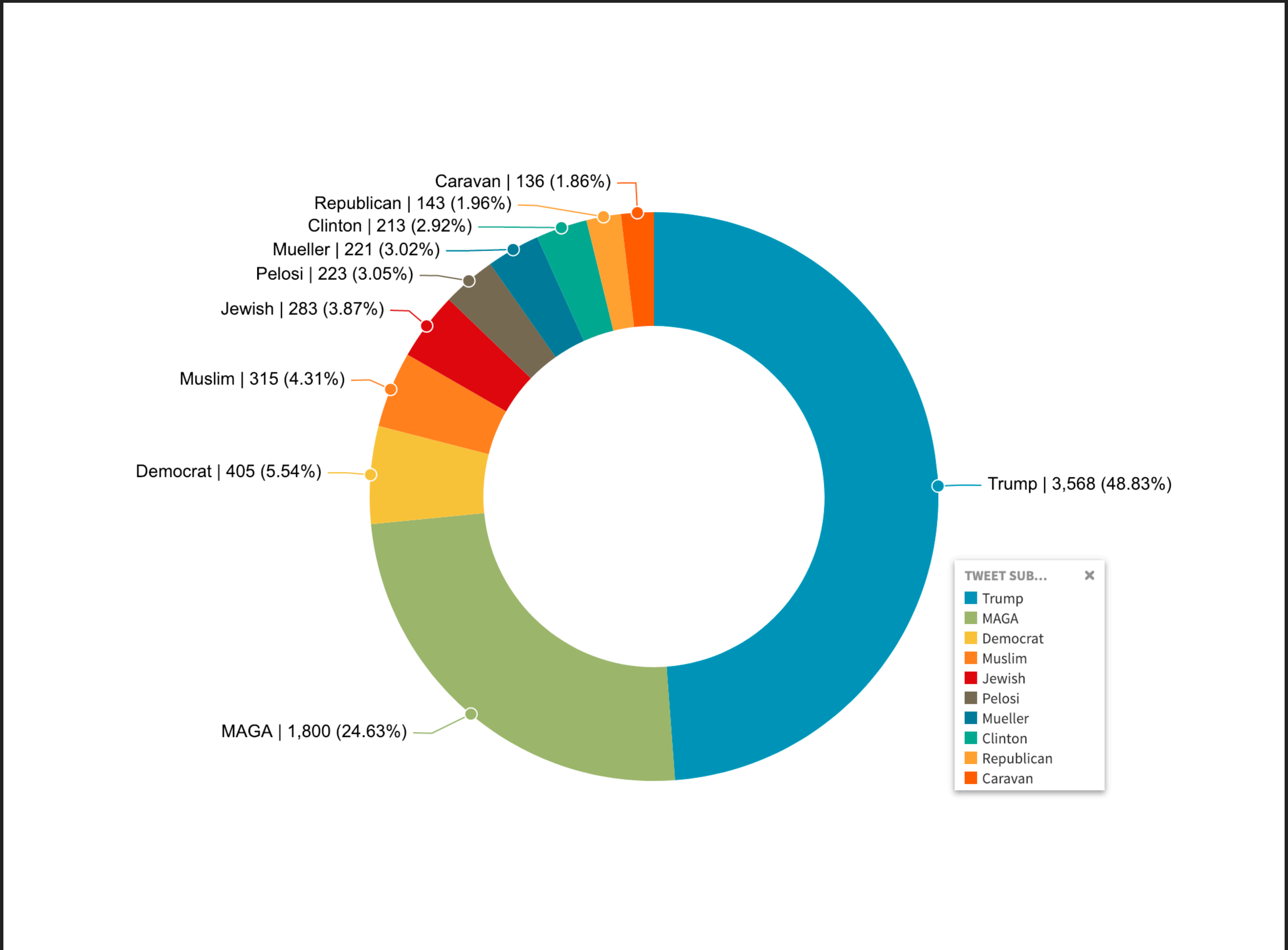

yelp_review
review_id VARCHAR(22)
user_id VARCHAR(22)
business_id VARCHAR(22)
stars DECIMAL(38,0)
date DATE
text VARCHAR(176)
useful TINYINT(1)
funny TINYINT(1)
cool TINYINT(1)
Indexes
stars
yelp_checkin
business_id VARCHAR(22)
weekday DATE
hour DATETIME
checkins DECIMAL(38,0)
Indexes
business_id
yelp_user
user_id VARCHAR(22)
name VARCHAR(5)
review_count DECIMAL(38,0)
yelping_since DATE
friends VARCHAR(1894)
useful TINYINT(1)
funny TINYINT(1)
cool TINYINT(1)
fans TINYINT(1)
elite TINYINT(1)
average_stars DECIMAL(38,1)
compliment_hot TINYINT(1)
compliment_more TINYINT(1)
compliment_profile TINYINT(1)
compliment_cute TINYINT(1)
compliment_list TINYINT(1)
compliment_note TINYINT(1)
compliment_plain TINYINT(1)
compliment_cool TINYINT(1)
compliment_funny TINYINT(1)
compliment_writer TINYINT(1)
compliment_photos TINYINT(1)
Indexes
name
yelp_tip
text VARCHAR(75)
date DATE
likes TINYINT(1)
business_id VARCHAR(22)
user_id VARCHAR(22)
Indexes
date
yelp_business
business_id VARCHAR(22)
name VARCHAR(18)
neighborhood TINYINT(1)
address VARCHAR(26)
city VARCHAR(9)
state VARCHAR(2)
postal_code DECIMAL(38,0)
latitude DECIMAL(38,7)
longitude DECIMAL(38,7)
stars DECIMAL(38,0)
review_count DECIMAL(38,0)
is_open TINYINT(1)
categories VARCHAR(89)
Indexes
name

DATA TECHNIQUES

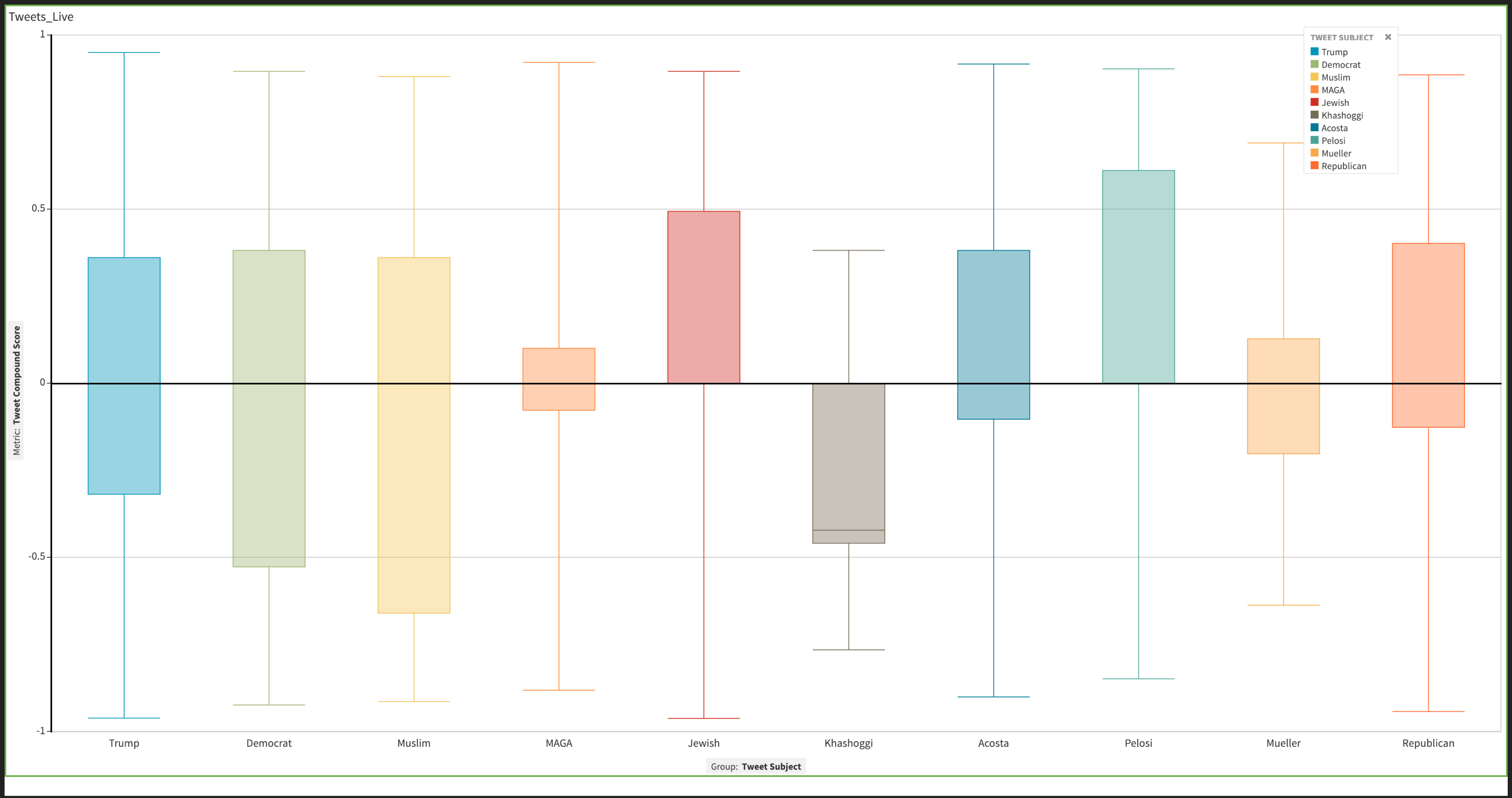
- Python for data ingestion, augmentation and/or cleansing
- VADER Analyzer for sentiment analysis
  - Tuned for microblog content
- spaCy for entity extraction
  - High performance NLP engine

TWITTER DATA VISUALIZATIONS

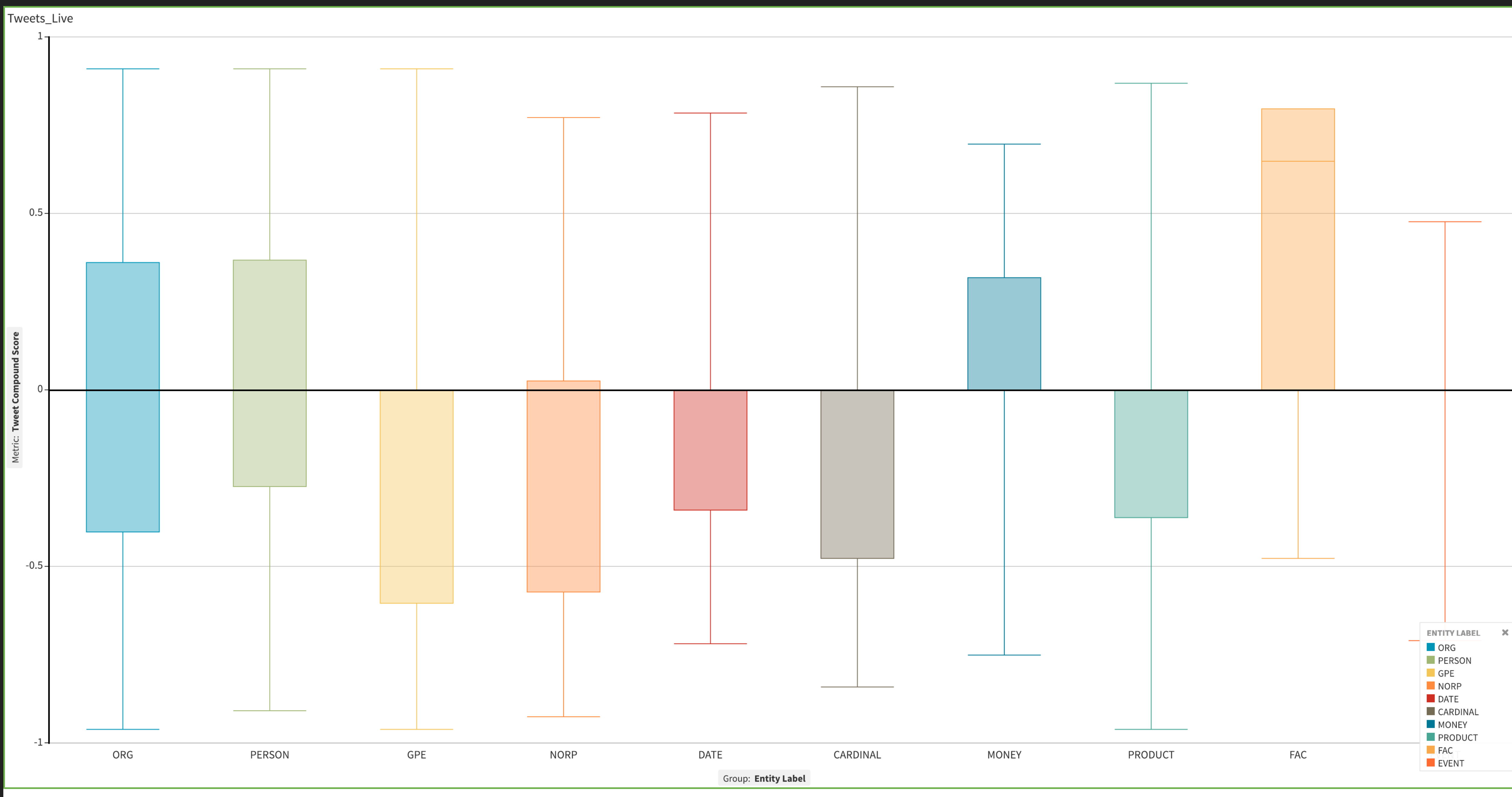
TWEET VOLUME BY SUBJECT



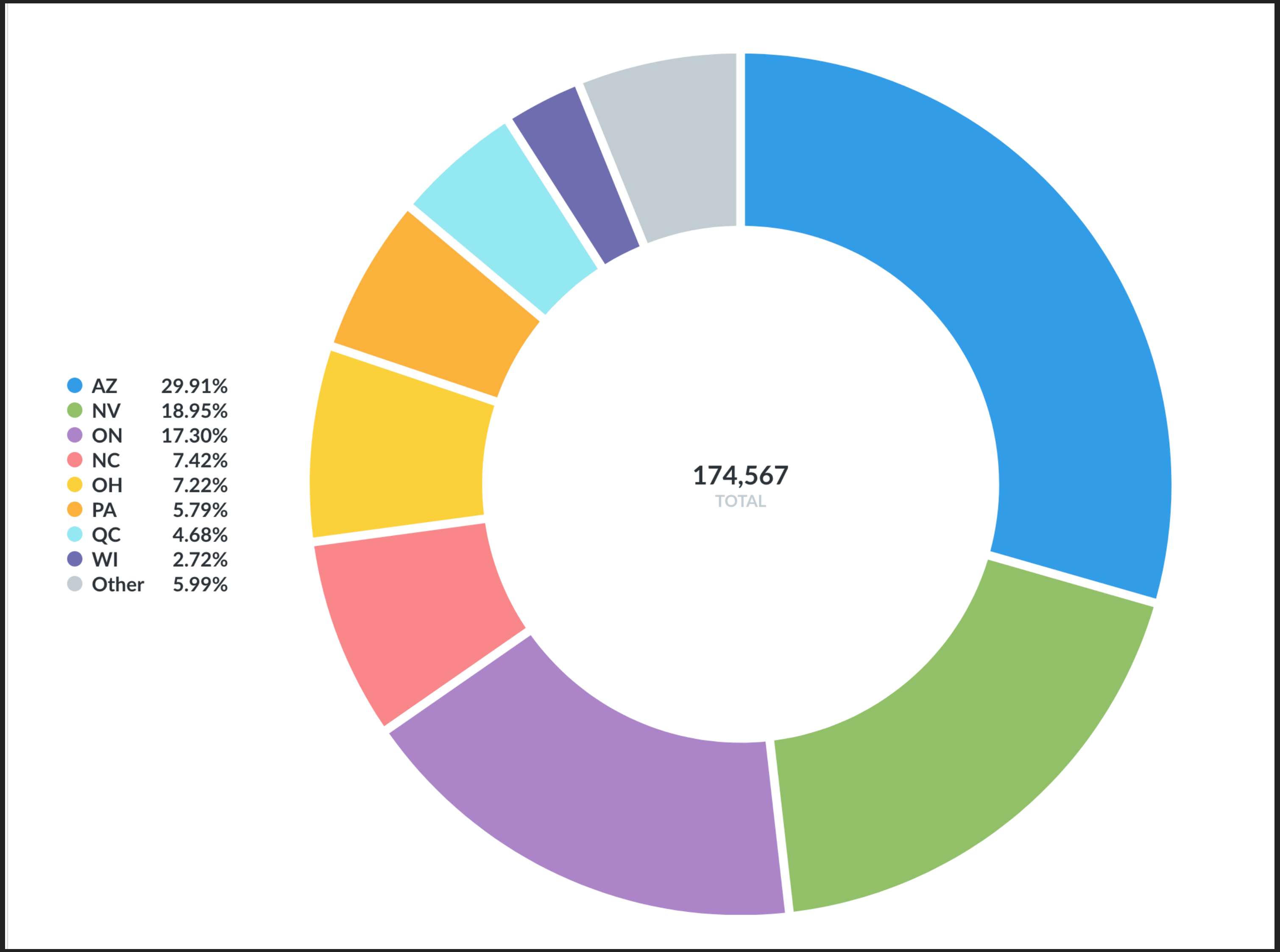
TWEET SENTIMENT SCORE BY SUBJECT

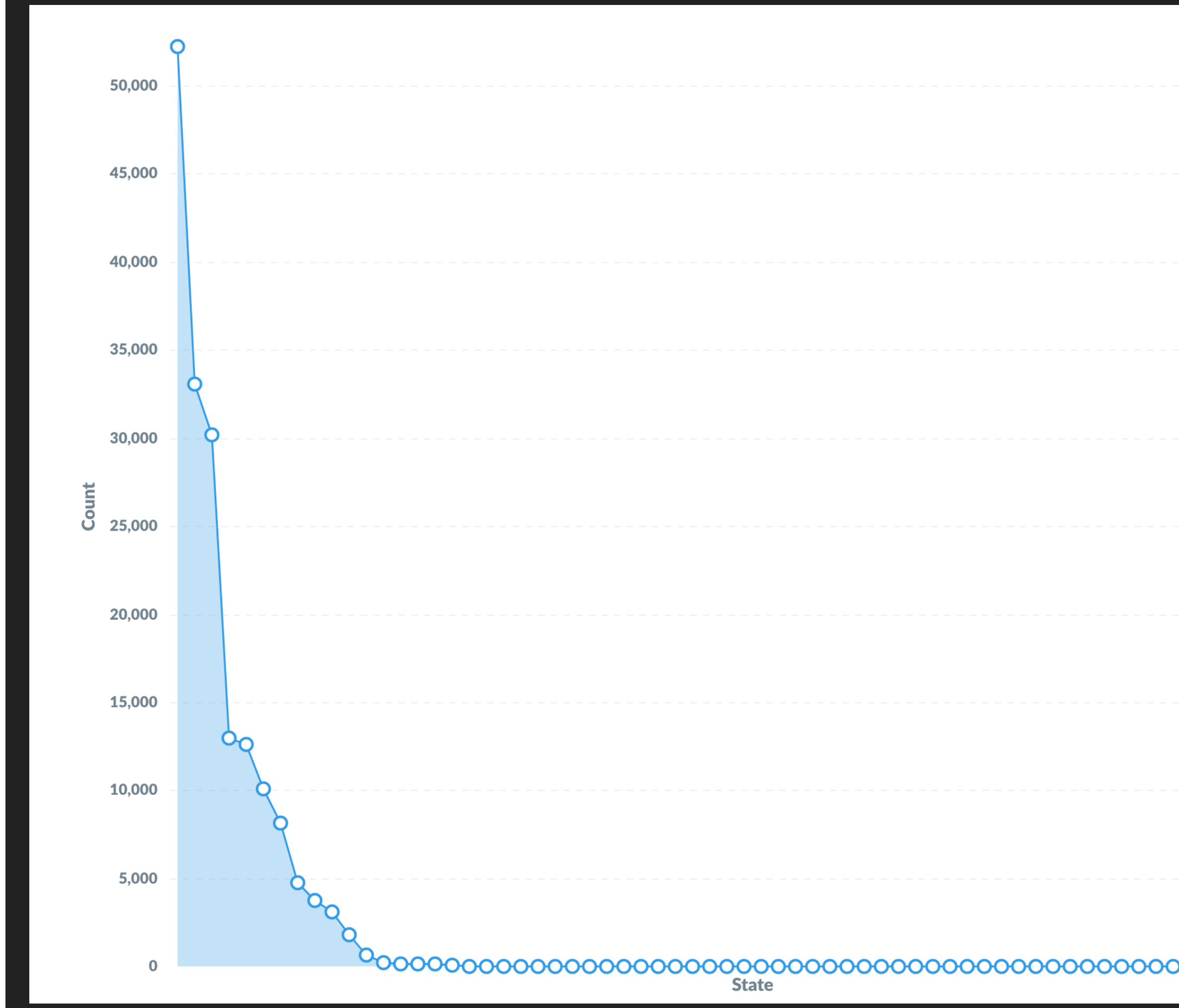


TWEET SENTIMENT SCORE BY ENTITY

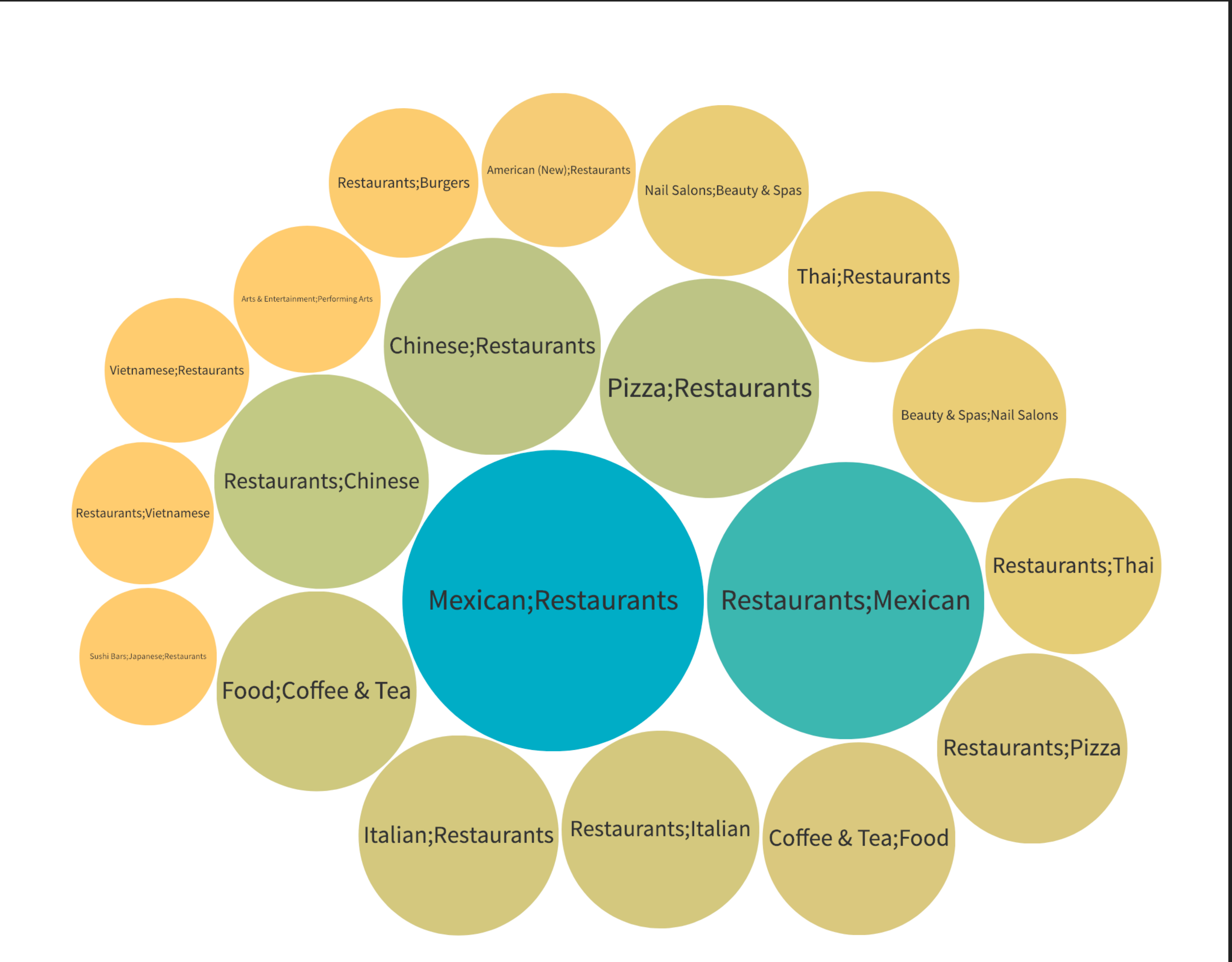


YELP BUSINESSES BY STATE

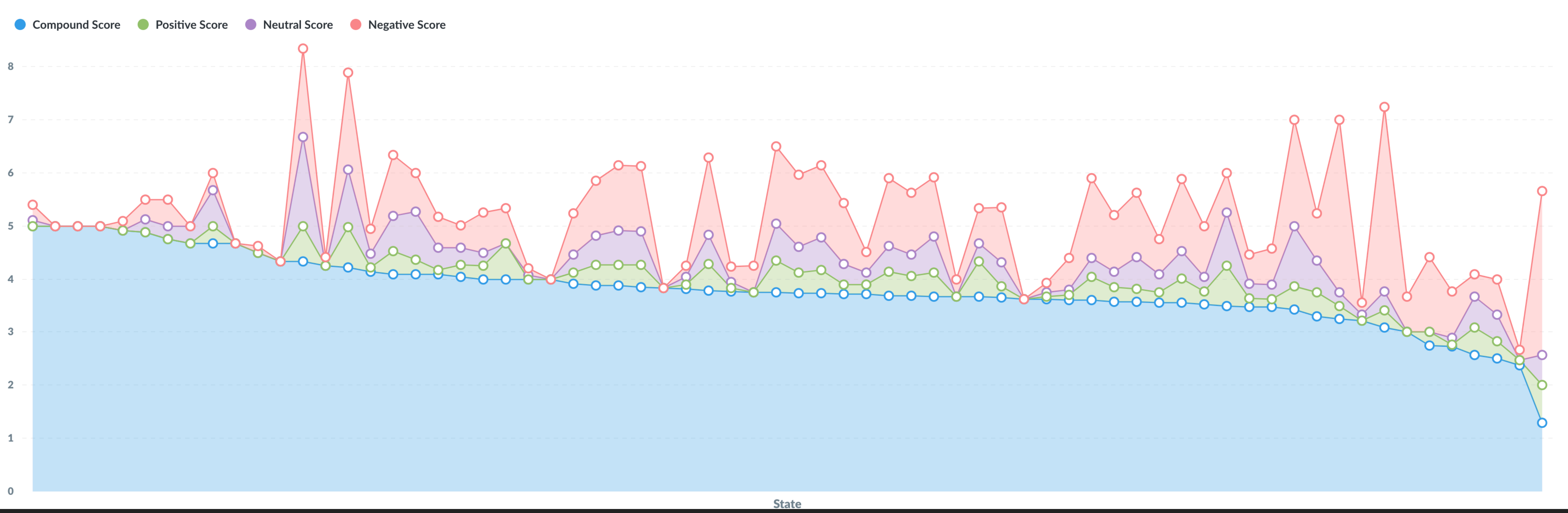




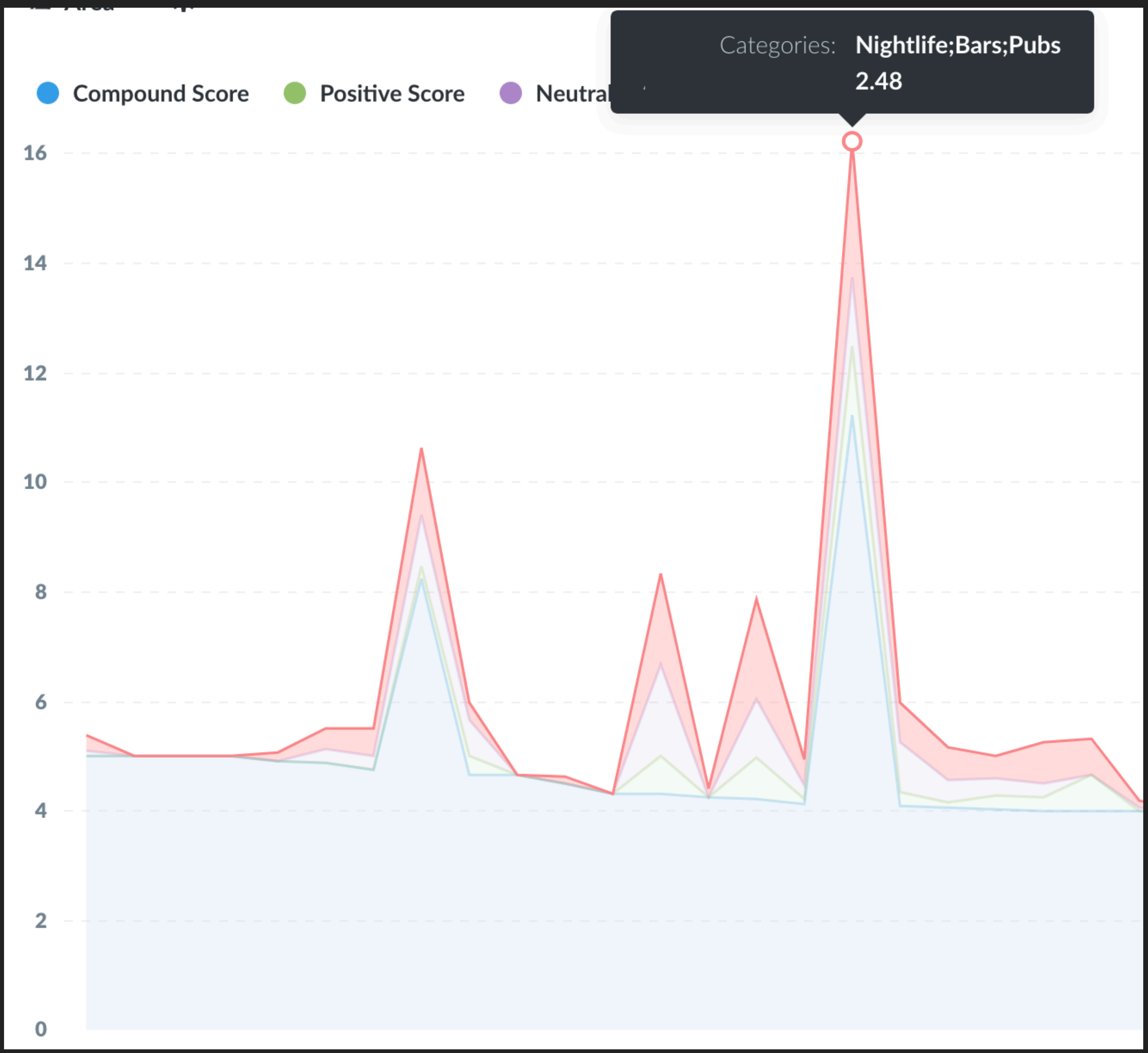
YELP SENTIMENT BY BUSINESS CATEGORY



YELP SENTIMENT BY STATE



YELP SENTIMENT BY STATE - ZOOMED



MODEL SELECTION

- Sentiment Analysis
  - Rule-based sentiment analysis lexicon VADER (Valence Aware Dictionary for sEntiment Reasoning)
  - VADER is specifically developed for micro-blog content from social media (Gilbert, June 2014)
  - 'Gold Standard' lexicons (LIWC, ANEW, & GI) are NOT developed for the deeper lexical properties from in most micro-blog content

MODEL SELECTION

- Sentiment Analysis
  - VADER outperforms traditional machine learning classifiers

	3-Class Classification Accuracy (F1 scores) Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	**0.96**	0.61	**0.63**	**0.55**
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	**0.75**	0.49	0.44
ME (movie)	0.56	**0.75**	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

MODEL SELECTION

- Sentiment Analysis
    - VADER outperforms human classifiers

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)				
Ind. Humans	**0.888**	0.95	0.76	0.84
VADER	0.881	**0.99**	**0.94**	**0.96**
Hu-Liu04	0.756	0.94	0.66	0.77
SCN	0.568	0.81	0.75	0.75
GI	0.580	0.84	0.58	0.69
SWN	0.488	0.75	0.62	0.67
LIWC	0.622	0.94	0.48	0.63
ANEW	0.492	0.83	0.48	0.60
WSD	0.438	0.70	0.49	0.56

MODEL SELECTION

- Entity Extraction
  - Entity extraction library spaCy
  - spaCy is designed to be extremely fast and has models trained using convolutional neural networks
  - Python library written Cython

MODEL SELECTION

- Entity Extraction
  - Entity extraction library spaCy for high accuracy and speed

SYSTEM	YEAR	LANGUAGE	ACCURACY	SPEED (WPS)
spaCy v2.x	2017	Python / Cython	**92.6**	*n/a* ?
spaCy v1.x	2015	Python / Cython	91.8	13,963
ClearNLP	2015	Java	91.7	10,271
CoreNLP	2015	Java	89.6	8,602
MATE	2015	Java	92.5	550
Turbo	2015	C++	92.4	349

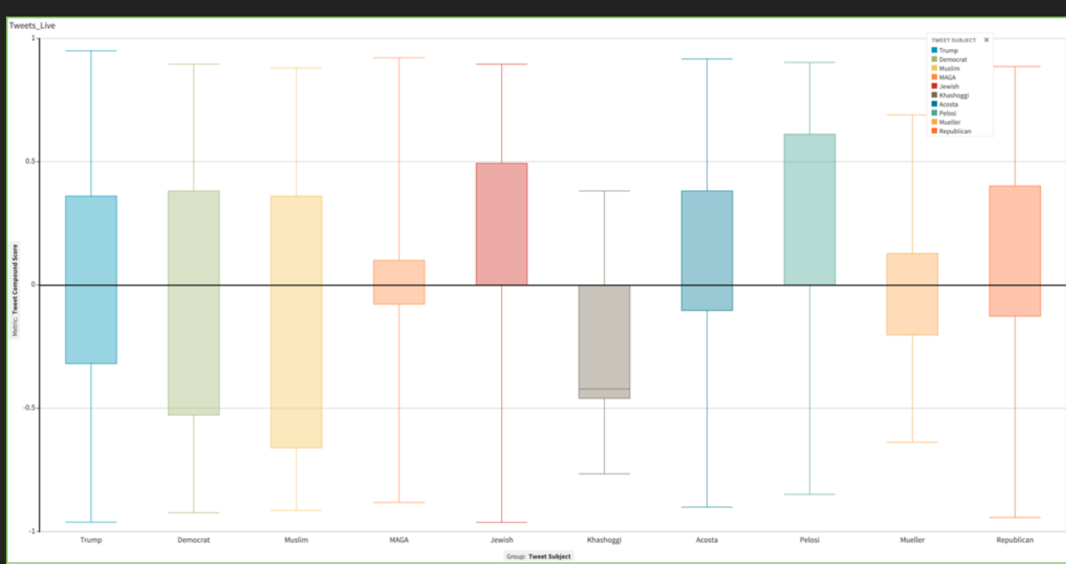
MODEL SELECTION

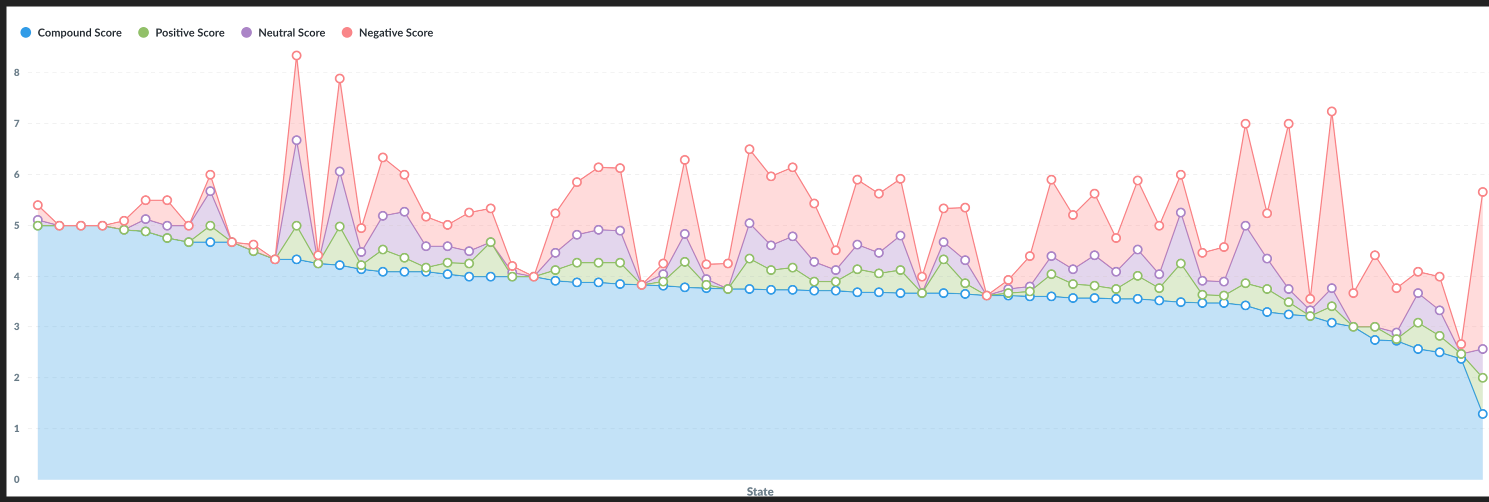
- Entity Extraction
  - Entity categories

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

ANALYTICAL TECHNIQUE

- Use of descriptive statistics
  - Min, median or mode
- Calculate the median score of sentiment from microblog content





CONCLUSIONS

- Real time sentiment analysis scoring and entity extraction is possible
- Compound sentiment score is the most balanced metric for determining sentiment
- Robust visualizations technologies are needed to handle both the speed and scale of a streaming dataset

FUTURE WORK

- Refine the entity extraction handling
    - Many false positives
- Enable geocoding of userlocation to map data
- Develop a content gist creation tool

QUESTIONS?