DATA 670 9040 DATA ANALYTICS

Matthew DeMarco (mattdee@gmail.com)

Dr. Jon McKeeby (jon.mckeeby@faculty.umuc.edu)

Capstone Project: Final Results

December 2nd, 2018

Executive Summary

Social media is a pervasive force in today's world. Social media platforms are an important medium for businesses, governments and political figures to interact directly with consumers, citizens and constituents. With the power of these platforms users can influence the reputation of a brand, push a political agenda or lobby for a specific candidate. The reach of social media enables a small number of users to influence the greater whole as shown by the 2011 research in "*We're all connected: The power of the social media ecosystem*" (Richard Hanna, 2011). 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.

Social media can and has been weaponized to influence businesses, governments and political figures (Zeitzoff, 2018). Given this new reality it is particularly important for one to understand the underlying sentiment of a social media posting. With this understanding, one can adjust messaging to be better suited towards the target audience. This project will benefit various entities to allow them leverage machine learning lexicon for sentiment analysis.

The recent United States of America 2016 Presidential election was influenced by social media and distorted interactions on social media platforms (Alessandro Bessi, 2016). Additional research from Zeitzoff has shown that social media can influence real-world events. Understanding the intent as well as the underlying sentiment of a social media post can allow organizations to adjust messaging to better target their demographic.

The social media platforms for this project will be Twitter and Yelp. Twitter is an online service that allows users to create content with posts and interact with other users via 'tweets'. Twitter's current global user base was 328 million (Fiegerman, 2017). This large, international user base enables this project to analyze a massive live data set. Yelp is an online search service that is powered by user sourced feedback on local business as well as online reservation service. Yelp has also enabled businesses have the opportunity to respond to user feedback. Studies have shown that positive user reviews of businesses have increased sales by 5 to 9 percent (Luca, 2011). This project will analyze the publicly available Yelp academic dataset which contains data on users, reviews, businesses, and tips.

Table of Contents

Project Scope

The scope of this project is to collect and analyze the sentiment of a social media postings to understand the negative, neutral or positive tone of a post. Additionally, we will extract referenced entities in the social media posting into categories. The data delivery will be done using various data visualizations technologies.

Problem Description

The recent United States of America 2016 Presidential election was influenced by social media and distorted interactions on social media platforms (Alessandro Bessi, 2016). Additional research from Zeitzoff has shown that social media can influence real-world events. Research by Luca has shown that for every star a business receives on Yelp, sales increase by 5 to 9 percent (Luca, 2011). Understanding the intent as well as the underlying sentiment of a social media post can allow organizations to adjust messaging to better target their demographic.

Social media has become an extremely influential force in recent history. To understand the influence of a posting one must reference current trends, topics or 'memes' in order to fully appreciate the implied meaning. Businesses and political organization should be able to interpret the sentiment of a posting in order to judge whether said posting merits a response or rebuttal but with the volume of postings it is impossible for people to interpret and interact with this data. Sentiment analysis of postings allows for organizations to response to posts that are detrimental in order to circumvent negative feedback.

Internet media can and has been weaponized to influence businesses, governments and political figures (Zeitzoff, 2018). Given this new reality it is particularly important for one to

understand the underlying sentiment of a social media posting. With this understanding, one can adjust messaging to be better suited towards the target audience.

Business Understanding

Organizations are influenced by their consumers. Whether said organization is a traditional business serving customers or a political group, they must be aware if their messaging, both intentional or casual, impacts audience. With the raise of social media groups can and do interact with a wide audience that is targeted and intentional. This new medium needs to be managed like any interaction with the consumer.

Interactions with consumers of social media needs to be moderated. In order to quickly judge the sentiment of a posting, an organization needs to access, rapidly, whether a posting merits response. Modern systems allow for organizations to assess and respond by leveraging machine learning to judge the intent of a post. This enables quick interactions with the end user audience to continues a positive dialogue and/or stop negative feedback.

Organization

Organizations that have a public persona need to be aware of their public perception. Regardless of intent these groups need to manage how their messaging is affecting their intended and unintended audiences. A posting on social media can both be benign or malignant, only the interpretation of the end user defines the effect. Therefore, it is of special importance that social media interactions be managed.

The volume and velocity of social media makes management of interactions impossible unless machine learning is applied to this problem space. Leveraging modern compute power allows for organizations not only to be aware of user sentiments but interact when messaging or intent is trending negatively.

Stakeholders

Social perception of an organization or a topic, political or commercial, is critically important in today's modern world. Key stakeholders for management of audience interaction are the chief executive officer, chief marketing officer and sales executive.

The chief executive officer is typically incentivized on returning value back to stockholders. To do so, he/she must be aware public perception of the brand and its relative value in the market. Ignorance of market perception can negatively affect value as well as cause poor decision making on product direction. The same principles apply to political figures. Should a candidate for office not be aware of public perception, they can misalign decisions that are not representative of the public.

Another key stakeholder is the chief marketing officer. This individual must be keyed into the public understanding of the brand and its relative value. If the public feels the brand is not worth the cost of acquisition the messaging needs to change, or the brand has to adjust down market. The same principle applies for political marketing. If the messaging of a candidate does not resonate with the electorate, they need to change messaging and adapt.

Lastly, the sales executive needs to be acutely aware of public perception of the brand in order to apply an effective sales strategy. For example, should the public perception for database be a commodity then this executive needs to adjust messaging and pricing to adapt to the market. Again, as with previous examples, the same principles apply for political campaigns. Should the regional campaign manager not under how the party's or candidate's message effect the electorate then they will be met with disastrous results.

Define Business Area

The business area for this project will be both political topics, brand awareness and user feedback directed towards a business. The project will collect social media postings from Twitter and Yelp to analyze the sentiment of intended meaning of a post or user feedback. A social media posting will be classified as either neutral, negative or positive on the intended subject. The subject or topic of the posting will be extracted from postings with entity extraction techniques. For example, we will collect tweets then extract entities such as people, events or companies.

Business Objective

The business object is to understand how various topics and user feedback are trending on social media. Companies and political groups need to manage social perception as mismanagement can lead to disastrous results; either the loss of brand value or the loss of office.

Objective 1: Understand a political figure and business' online sentiment. Online interaction by political figures has taken a new-found importance with President Trump use of Twitter. Understanding how the messaging of the President of the United States of America is important to judge the relative impact of various topics. We will capture and measure tweets that mention President Trump for their implied sentiment. We will additionally capture and measure Yelp user sentiment feedback as they apply to various businesses.

Objective 2: Understand political topics and business category's sentiment. Using sentiment measures, we will see the impact of various political topic and business catergories. We will measure the sentiment measure over time and observe whether a topics trends negative, neutral or positive.

Objective 3: **Understand brand sentiment.** Using the same sentiment analysis, we will measure the online sentiment scores of various commercial brands such as Nike, NFL and Apple

as well as local businesses. The selection of these companies is done to represent companies with significant market presences across industry.

Business Success Criteria

The success factors for this project will be based on the ability to understand the sentiment of a Twitter and Yelp post, score said post on a spectrum of negative to degrees of positive as well as visualize trends over time. This work will make it easier for organizations to understand their social profile and adjust if needed.

Background

The recent United States of America 2016 Presidential election was influenced by social media and distorted interactions on social media platforms (Alessandro Bessi, 2016). Additional research from Zeitzoff has shown that social media can influence real-world events. Understanding the intent as well as the underlying sentiment of a social media post can allow organizations to adjust messaging to better target their demographic.

The social media platform for this project will be Twitter and Yelp. Twitter is an online service that allows users to create content with posts and interact with other users via tweets. Twitter's current global user base was 328 million (Fiegerman, 2017). This large, international user base enables this project to analyze a massive live data set. Yelp is an online search service that is powered by user sourced feedback on local business as well as online reservation service. Yelp has also enabled businesses have the opportunity to respond to user feedback. Studies have shown that positive user reviews of businesses have increased sales by 5 to 9 percent (Luca, 2011). This project will analyze the publicly available Yelp academic dataset which contains data on users, reviews, businesses, and tips.

Research

Analysis of social media has a long-storied history and has occurred since the first online communities (Lee, 2008). Understanding latent meaning has typically vexed researchers on this subject. The landmark work on defining a lexicon specifically for social media has been done by developers of the VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media micro-blog content and typically outperforms the general purpose Linguistic Inquiry and Word Count (LIWC, pronounced "Luke") (Gilbert, June 2014). This lexicon is a high performant utility for measuring sentiment online and typically outperforms human classifiers in correctly identifying sentiment.

Gaps in this Problem Resolution

The major problem in this approach by VADER is the user interacting with the data. For example, a negative score can go unnoticed if a topic is not of interest to the analyst. This is not a downfall of the lexicon or the model applied by the lexicon but merely a lack of observation by the person.

In order to overcome this shortcoming in the VADER approach, this project will implement an entity extraction library. Entity extraction is the process of analyzing speech to determine the entities like PERSON, COMPANIES, PRODUCTS, etc. The library of choice for this project will be the spaCy library. spaCy offers the fastest syntactic parser in the world and that its accuracy is within 1% of the best available natural language parsers (NLP). This accuracy and performance makes it an obvious choice for this project.

Proposed Project

Social media can and has been weaponized to influence businesses, governments and political figures (Zeitzoff, 2018). Given this new reality it is particularly important for one to understand the underlying sentiment of a social media posting. With this understanding, one can adjust messaging to be better suited towards the target audience. This project will benefit various entities to allow them leverage machine learning lexicons for sentiment analysis.

Key Performance Indicators

KPI 1: Monitored sentiment scores over time. This project will extract topics from the body of a posting and monitor the sentiment score over time. This is important to understand if a topic or brand perception changes over time. For example, an initial response of the public could negative but trend to neutral or positive over time.

KPI 2: Count of the number of tweets per public figure/topic/brand. This key performance indicator will measure the relative volume of a public figure/topic/brand over time. Understanding the volume of interest is a critical measure in understanding relative value/sentiment of interest.

KPI 3: Average sentiment score of tweets per public figure/topic/brand. Gathering the data and measuring the average sentiment for a public figure/topic/brand allows for the measured subject to adjust messaging over time. In order to do so, one must understand the current relative sentiment.

KPI 4: Measure the sentiment of user feedback in various businesses categories. This key performance indictor will be used to measure the positive, neutral, negative of various businesses categories as determined by user reviews in Yelp.

Project Insights of your Data Analysis

The project insights reaped from the data analysis will show one how perceived sentiment are affected over time. This is important as social media a powerful force in modern interactions in both political and business environments. The expected results of this work are to measure public sentiment on public figures, topics and brands.

For public figures, it is important to understand their online presence and to measure sentiment of postings directed at or posted by these figures.

Understanding the public sentiment of a topic can help shape messaging around said topic. We expect that using the VADER lexicon will allow rapid understanding of public perception.

Lastly, for brands, it is also important to understand how a brand is publicly perceived. For example, we would expect to overall measure of Nike to trend positive given their renewed marketing campaign.

Project Milestones

The major planned milestones for this project are as follows:

Milestone 1: Procuring access to Twitter streaming API.

Milestone 2: Procuring access to the Yelp dataset.

Milestone 3: Development of database schema for both Twitter and Yelp datasets.

Milestone 4: Development of a streaming extraction, transformation, and load (ETL) process to ingest the Twitter stream.

Milestone 5: Development of ETL process to ingest the Yelp JSON datafiles to RDBMS tables.

Milestone 6: Creation of visualization showing above key performance indicators.

Milestone 7: Analysis of work and trends observed during the period of study.

Completion History

Week 1	Reviewed requirements for project and data needs.
Week 2	Gained access to Twitter API for academic needs.
Week 3	Review Twitter API and Tweet data structure.
Week 4	Developed Twitter RDBMS data model, Python program, and data cleansing procedure.
Week 5	Additional of Yelp dataset. Presentation 1 to peer group. Peer group feedback suggestion of entity extraction added to project.
Week 6	Work on ETL process to extract data from Yelp JSON datafiles. Test of new visualization tool, Metabase.
Week 7	Continued work on ETL process to extract Yelp JSON data. The code on Yelp's github does not work. Found an archived dataset on Kaggle for the Yelp data in CSV.
Week 8	Successfully loaded Yelp dataset. Tested 5 data visualizations tools to find a performant solution. Presented project status update to class on November 1st.
Week 9	Developed visualization tools to fully explore the Twitter and Yelp datasets.
Week 10	Researched additional visualization tools that can be used to explore the datasets in new and novel ways.

Week 11	Refined database schema for more efficient joins during query runs making data visualization run faster.

Lessons Learned

Week 1	The requirements seemed confusing. Decided on Twitter data sentiment problem as real time analysis and deployment seemed applicable to real world business requirements.
Week 2	Twitter required a detailed explanation on the rationale for access.
Week 3	The Tweet data structure has additional information that is not of interest to the study and must by parsed out
Week 4	The RDBMS that can ingest the Twitter volume is not available on either IBM, AWS or Google cloud platforms. This forced the study to be deployed to an on-premise solution (my laptop).
Week 5	Addition of the Yelp dataset. The dataset is comprised of JSON files with millions of nested records. Introduction of the spaCy entity extraction engine and addition of ENTITY table to the database.
Week 6	Visualization tools do not handle the scale of millions of rows of data very efficiently. Researched other visualization tools for project.
Week 7	The extraction code posted on Yelp's GitHub does not work natively. Spent the week attempting to rewrite the code to use the data. Found archived dataset in CSV which enabled easier loading of data.

Week 8	The Yelp dataset contained a Microsoft Windows line terminator that I did not detect, and this caused the data to load incorrectly. I had to open the files in 'vi' in order to see the hidden character.
Week 9	Tableau does not scale to the either dataset therefore causing additional need of other visualization tools.
Week 10	I deployed Metabase and Zoomdata, both which run as a server on the database versus using a desktop bound application like Tableau. Zoomdata appears to need significant resources so I must scale down the Twitter data ingestion if I need to create a visualization it this tool.
Week 11	Development of the final presentation was challenging as I had to distill a semester's worth of work and analysis in 20 minutes.
Week 12	The operationalizing of machine learning has real impact to audiences outside of the data science community. I've presented my capstone project to sales management at my company and we are looking at ways to monetize the solution.

Data Set Description

Data Set Description

The first proposed dataset for this project will be the real-time feed from Twitter. The data is encoded in JavaScript Object Notation (JSON) and contains the following data in Table 1.

Data Entity	**Data Description**
created_at	Datetime stamp of tweet creation
id_str	Randomly generated identification number
text	Body of the tweet
user	Twitter username
place	If the tweet is geotagged this will contain latitude/longitude coordinates
entities	An array of hashtags, user mentions, URLs, cashtags, and native media
extended_entities	If the Tweet contains media it will be described here

Table 1. Twitter dataset

The text data entity will be used for the sentiment analysis to determine a score of the tweet.

To store the Twitter data, we will develop the following table to consume and query the information as shown in Figure 1.

Figure 1. RAW_TWEET Table.

Additionally, we will store extracted entities from the text of the tweet in a separate table called ENTITIES as show in Figure 2.

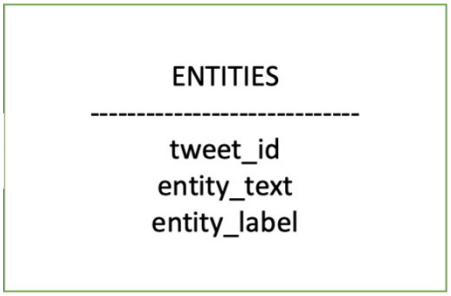
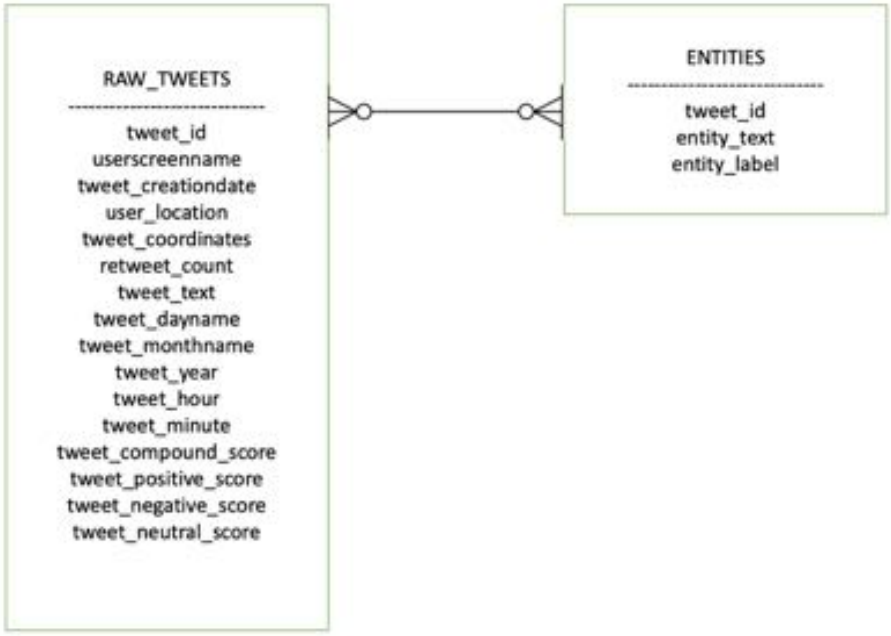


Figure 2. ENTITIES Table

These extracted entities will also be related in a ‘many to many’ relationship from this table to the RAW_TWEET table on the TWEETID attribute as show in Figure 3. This relation means that a single tweet can have many entities and an entity can belong to many tweets.



The second proposed dataset for the project is static information from Yelp. The Yelp dataset is stored statically in comma separated value (CSV) files. The files that are the subject of this project are as follows in Table 2.

Data Entity	Data Description
yelp_business.csv	Contains business data including location data, attributes, and categories.
yelp_review.csv	Contains business data including location data, attributes, and categories.
yelp_user.csv	Contains business data including location data, attributes, and categories.
yelp_checkin.csv	Checkins on a business.
yelp_tip.csv	Tips written by a user on a business.

Table 2. The Yelp dataset.

The table structure in the relational database management system will map tables to files. For example, we will create a BUSINESS table to load the yelp_business.csv into.

The table structure for Yelp is as follows in Figure 4.

Figure 4. Yelp RDBMS Table structure.

Data Definition/Data Profile

For the Twitter dataset, we have created three objects in the RDBMS; RAW_TWEETS, ENTITIES, and TWEETS. The RAW_TWEETS table stores the information extracted from a Python program to store the tweet text, user screen name, user location, tweet id, tweet compound score as yielded from VADER, tweet neutral score as yielded from VADER, tweet

positive score as yielded from VADER, tweet negative score as yielded from VADER, tweet creation date with time increments parsed out, tweet coordinates (if available) and retweet count.

The ENTITIES table stores the extracted entities from the tweet text using the spaCy natural language processing (NLP) library. This table is related to the RAW_TWEETS table via the TWEETID, which is parsed from the ID_STR in the Twitter API streaming JSON.

We also have created a database view, TWEETS, which defines positive tweets that have positive compound scores (<= '-0.0000000000000001') and negative tweets that have negative compound scores (>= '0.0000000000000001'). This view is merely for convenience to allow queries to the system to quickly assess the sentiment of a tweet content.

For the Yelp dataset we have created five tables ingest the data. The static dataset is composed of 1,185,348 tips by 1,518,169 users with aggregated check-ins over time for each of the 188,593 businesses. There are over 1.4 million business attributes such as hours, parking, availability, and ambience (Yelp, n.d.). The tables are YELP_BUSINESS, YELP_CHECKIN, YELP_REVIEW, YELP_TIP, and YELP_USER. These tables are a 'one to one' match for the CSV datafiles. For the YELP_REVIEW table, we will add columns to capture sentiment scores of the review in the TEXT attribute. Using the VADER lexicon we will measure negative, neutral, positive scores.

Data Preparation/Cleansing/Transformation

Data Preparation

In order to work with the Twitter dataset, we will need to create a Python program to prepare the data for consumption into the RDBMS. This requires the following:

1. Subscribe to the Twitter API
2. Parse the Twitter API JSON object into fields
3. Initialize the VADER sentiment analyzer
4. Run the VADER sentiment analyzer against the Tweet text
5. Calculate the Tweet text's compound, positive, neutral or negative score.
6. Cast all objects from the JSON object into strings for RDBMS consumption
7. Construct a SQL insert statement
8. Open a connection to the RDBMS
9. Insert the Tweet with sentiment scores

The Python libraries involved for this custom program are sys, os, tweepy, spaCy, and memsql. The sys and OS libraries are used by the interpreter to interact with functions and variables of the system as well as the operating system. The tweepy library conducts the majority of the work for the program.

Tweepy allows access to Twitter via OAuth module to set a token for the program. It also enables error handling for status codes that Twitter may pass back to the program. Lastly, Tweepy grabs the content from a Twitter post in the form of a JSON object.

The scaCy library is used in processing the entity extraction of the context from the Tweet text. scaCy is a high preforming Natural Language Processing (NLP) toolkit that has been proven to out perform other NLP toolkits. The main performance optiomization of spaCy is that

it is written in Cython, an optionally statically-typed language that compiles to C or C++, which is then loaded as a C extension module (Honnibal, 2015).

For the Yelp dataset the data was stored in CSV format. This format is comma separated with delimiters marking the content in each column plus the end of each record. This format required minimal effort to convert to a format that was easily ingested by the native RDBMS toolset.

Data Cleansing

The tools of choice for this project to clean the data to the required format for loading are as follows:

1. Python
2. csvkit library
3. VADER sentiment analysis library
4. spaCy NLP toolkit library

Python is a general-purpose programing language that is extremely easy for the novice to learn or the experienced to exploit. The central power of Python is that it has a wide array of libraries to extend its native features. These libraries are easily installed via the Python PIP package management suite.

To create the data schema for the Yelp dataset we leveraged the csvkit library found in Python. The csvkit library is a full featured toolkit for working with both CSV files and a multitude of RDBM. To create tables, one merely needs to pass the RDBMS of choice plus the CSV file and the output will be a data definition language SQL string. This make creating tables extremely trivial and the need for cleaning data relatively non-existent.

The creation of the Twitter database schema as driven off the attributes we needed to capture from the Twitter API as well as entity extraction. The data in the Twitter API is present as a JSON array which needs to be cleansed to enable easy database consumption. To consume the Twitter JSON, we parse data payloads into Python variables, the main one for this project being the TEXT variable. Once the TEXT variable is captured, we pass that variable to the spaCy NLP library to extract the entities referenced in the content. The TEXT variable is passed to the VADER sentiment analyzer to understand the latent sentiment of the tweet. The last process in the data cleaning pipeline, is to convert all variables to a string format.

Data Transformation

The data in the Twitter dataset is presented by the Twitter API in JSON format. This data needs to be transformed to a format that can be easily ingest by the RDBMS. In order to do this, we parse out the JSON array to grab the context from the Twitter STATUS payload. The data that is captured from the STATUS payload is ID_STR, USER.SCREEN_NAME, CREATED_AT, USER.LOCATION, COORDINATES, RETWEET_COUNT, and TEXT.

When the TEXT data is captured, we pass the context of the TEXT to the VADER sentiment analyzer to determine the sentiment score. The sentiment score is classified in four areas of compound, positive, neutral, and negative scores. These attributes plus ID_STR, USER.SCREEN_NAME, CREATED_AT, USER.LOCATION, COORDINATES, RETWEET_COUNT, and TEXT are all transformed into strings for convenience to load into the RDBMS.

The data for the Yelp dataset is a static dataset stored in CSV format. This data can be loaded in most RDBMS with native features. For speed and ease reasons we decided to leverage

native features of the RDBMS of choice, MemSQL. MemSQL is a 'NewSQL' database that is an 'in-memory first' database uses memory as the default storage medium for data storage. Additional MemSQL has includes a pipeline feature that enables the database to native ingest data from Kafka, Hadoop, Amazon s3, and filesystems.

With the MemSQL filesystem pipeline feature, we were able to map the Yelp dataset CSV files to table columns in the RDBMS. Using this feature, we placed the files into individual directories to ingest the data. The average time to load using this feature was, on average, four seconds to load, an average, of 2,024,508.4 rows into five tables.

Data Analysis

Data analysis is key in making practical use of the data that is captured in this project. To fully exploit the possible data in this project we will leverage the following tools for data analysis:

1. Zoomdata
2. Metabase
3. Tableau

Zoomdata is a business intelligence tool that enables live ingest from data sources. This capability makes this tool a natural fit for analysis of the sentiment analysis of the live feed captured from Twitter. The volume of data that we are capturing from Twitter makes other tools only capture a 'point in time' for analysis. Zoomdata enables continuous capture and 'drill down' capabilities to explore the data. This tool will allow us to explore the data and relative sentiment of tweets live as it happens.

Metabase is an additional business intelligence tool selected for this project. Metabase is a business intelligence tool that enables users to ask questions. Native connections to databases are available in Metabase. This tool allows users to dynamically interact with data sources and save results into dashboards.

Lastly for data analysis we will be leveraging Tableau for data visualization. Tableau has a multitude of features available for creating visualizations. The features we will be using for this project are scatter plot, histograms, and side-by-side circles. A scatter plot will be used to compare at least two measures. Histograms will show the distribution of the data. The side-by-side circle feature will enable comparison of multiple measures.

Data Visualization

The main technique employed for the analysis of both the Twitter and Yelp datasets is descriptive statistics. Descriptive statistics are used to summarize features about a collection of data. For the analysis of these datasets, this approach has proven most useful across a wide array of tools.

Data Visualization 1

In order to get a base understanding of the data we will create groupings of the data. For the Twitter dataset we have grouped the data on the extracted tweet subject by on count as shown in Figure 5 with the removal of NULL values from the dataset.

The visualization for this data is a pie/donut chart. This chart is especially useful when reviewing metrics broken down by a dimension. In this case the metric is a count of the volume of each tweet subject.

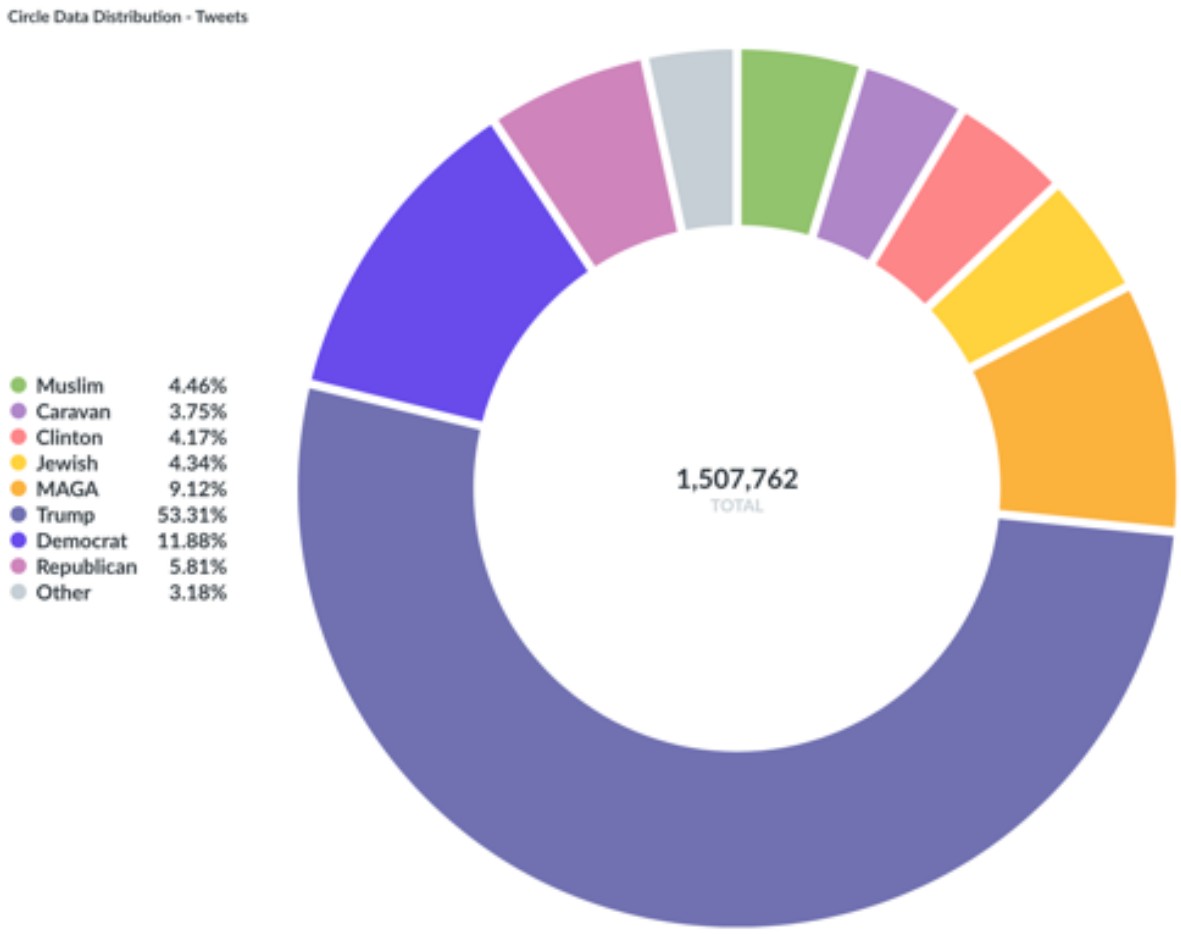


Figure 5. Tweets by subject.

As we can see from Figure 5, the majority of tweets contain information on or about President Donald Trump at 53.31 % of a total of 1,507,762 tweets. Next at 11.88% is the extracted subject of Democrats followed by 9.12% for the slogan of MAGA, which the acronym for 'Make American Great Again'.

It was unexpected to have such a large number of tweets associated with the President of the United States. This finding has motivated us to perform additional analysis on the sentiment of such a poplar topic on this social media platform.

For the Yelp dataset we have grouped the data on the business' locations that participate in the Yelp platform as shown in Figure 6. With this data we will again use the pie/donut chart. The metric used in this analysis is a count of the volume of each business by state.

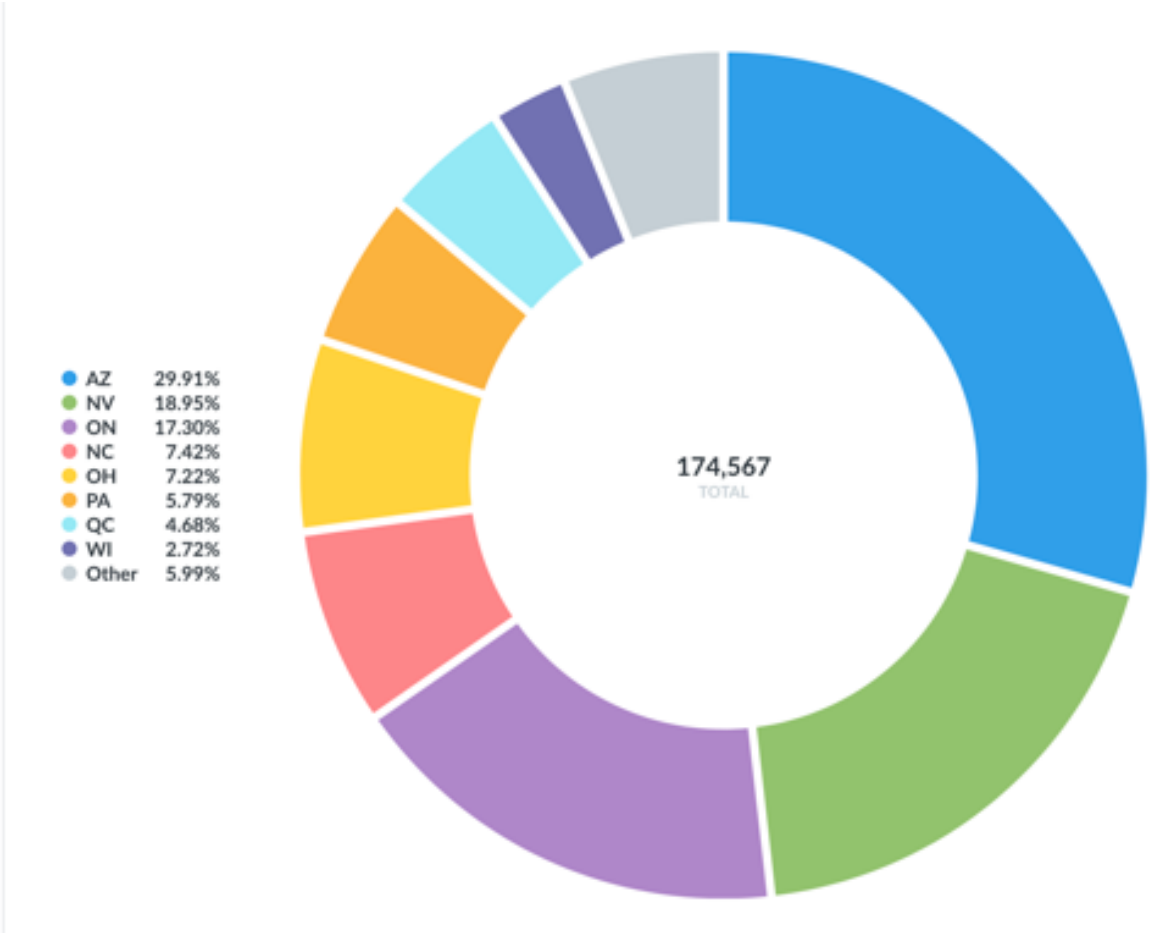


Figure 6. Yelp businesses by state.

As we can see in Figure 6, a majority of businesses that participate in the Yelp platform are found the United States and Canada. The majority of businesses being in Arizona at 29.91%, Nevada at 18.95% and Ontario at 17.30% of the 174,567 total businesses.

An unexpected result of the analysis is the fact that businesses are located in the state of 'Other'. In order to explain this, we examined the data further as shown in Figure 7 using an area chart visualization technique. The area chart technique allows us to further review all the data points in the dataset.

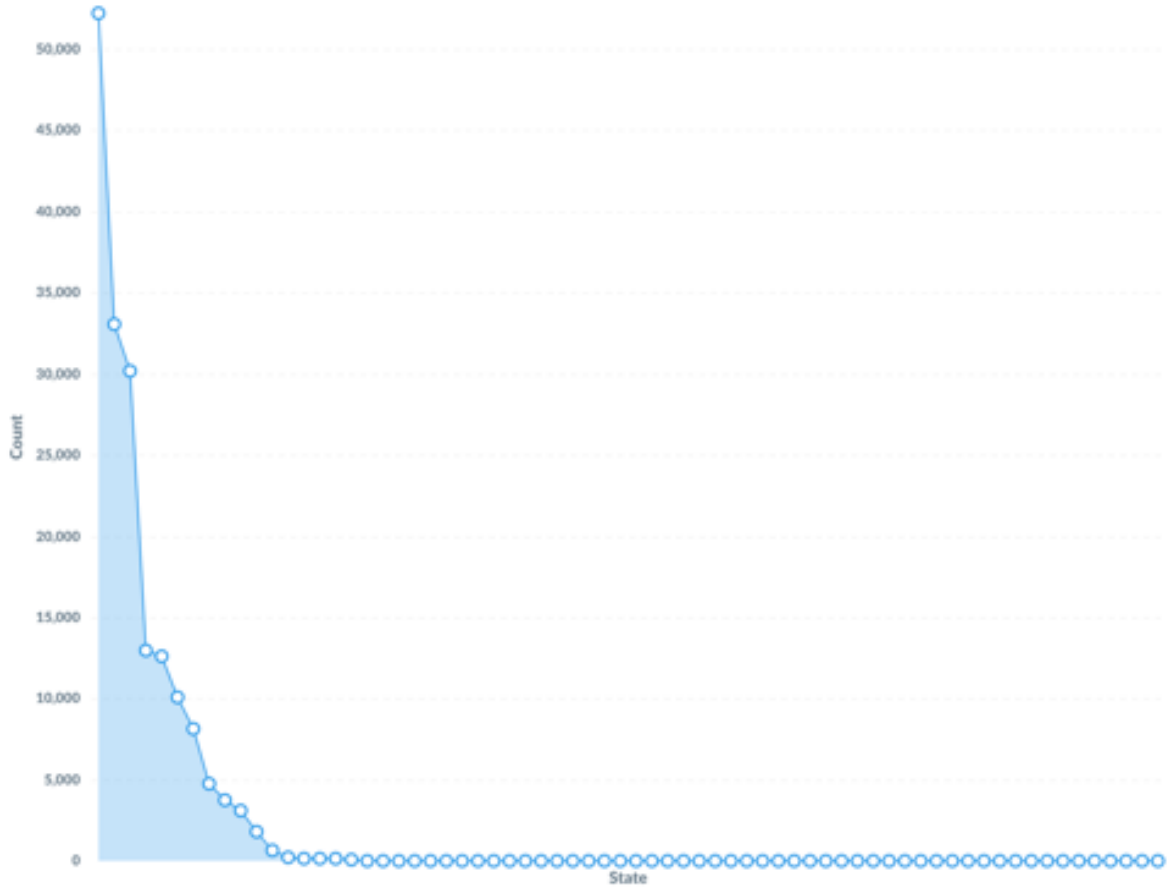


Figure 7. Area chart of states.

Using the area chart, we were able to determine that the 'Other' category shown in Figure 6 was a arbitrary grouping done by the visualization tool. Therefore, we quickly determined that the data was loaded correctly with no incorrect values.

Data Visualization 2

As the Twitter data is streamed into the database, it is scored on a spectrum of negative, neutral, positive and compound score for the intended sentiment if the tweet using the VADER lexicon. The compound score is computed by summing the scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive) (Gilbert, June 2014). We find this to be the most useful metric to get the sentiment of the entire text.

Using the compound score, we create the visualization shown in Figure 8 called a Packed Bubble Chart. The Packed Bubble Chart is used to display data in a cluster of circles. This is done using the extracted tweet subject and the compound score in Tableau.

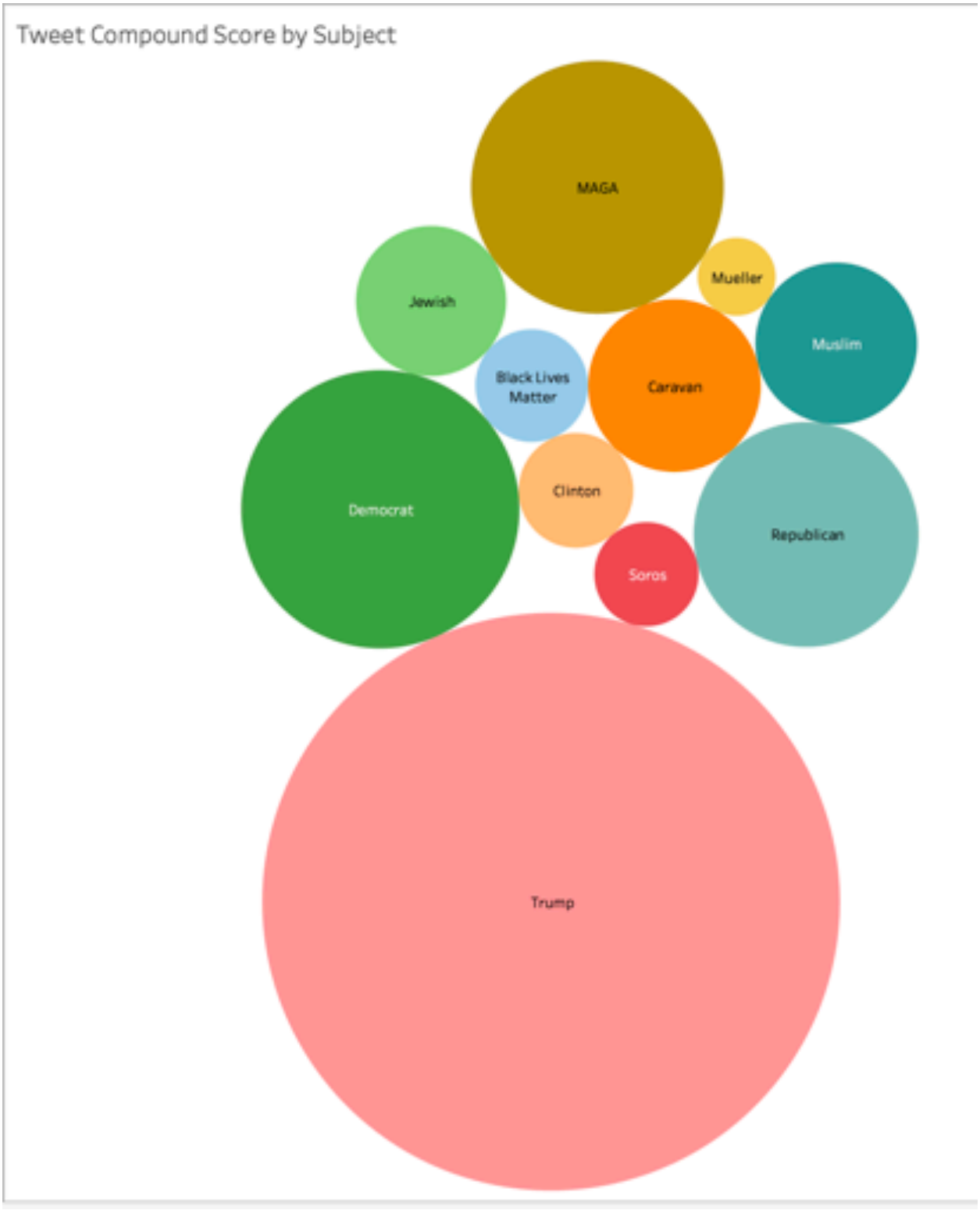


Figure 8. Packed bubble chart

As we can see from Figure 8, the same general trend holds true that was discovered in the first data visualization that is that President Trump has a vast majority tweets, but they are scored in with a high positive compound score. We can also observe that the extracted terms of Democrat and MAGA (Make America Great Again) trend towards a high positive compound scores.

The Yelp user review data is scored using the same VADER lexicon. We again leverage the compound score because it is a standardized threshold for classifying sentences as either positive, neutral, or negative (Gilbert, June 2014). Using this score, we look at sentiment as grouped by business category as show in Figure 9.

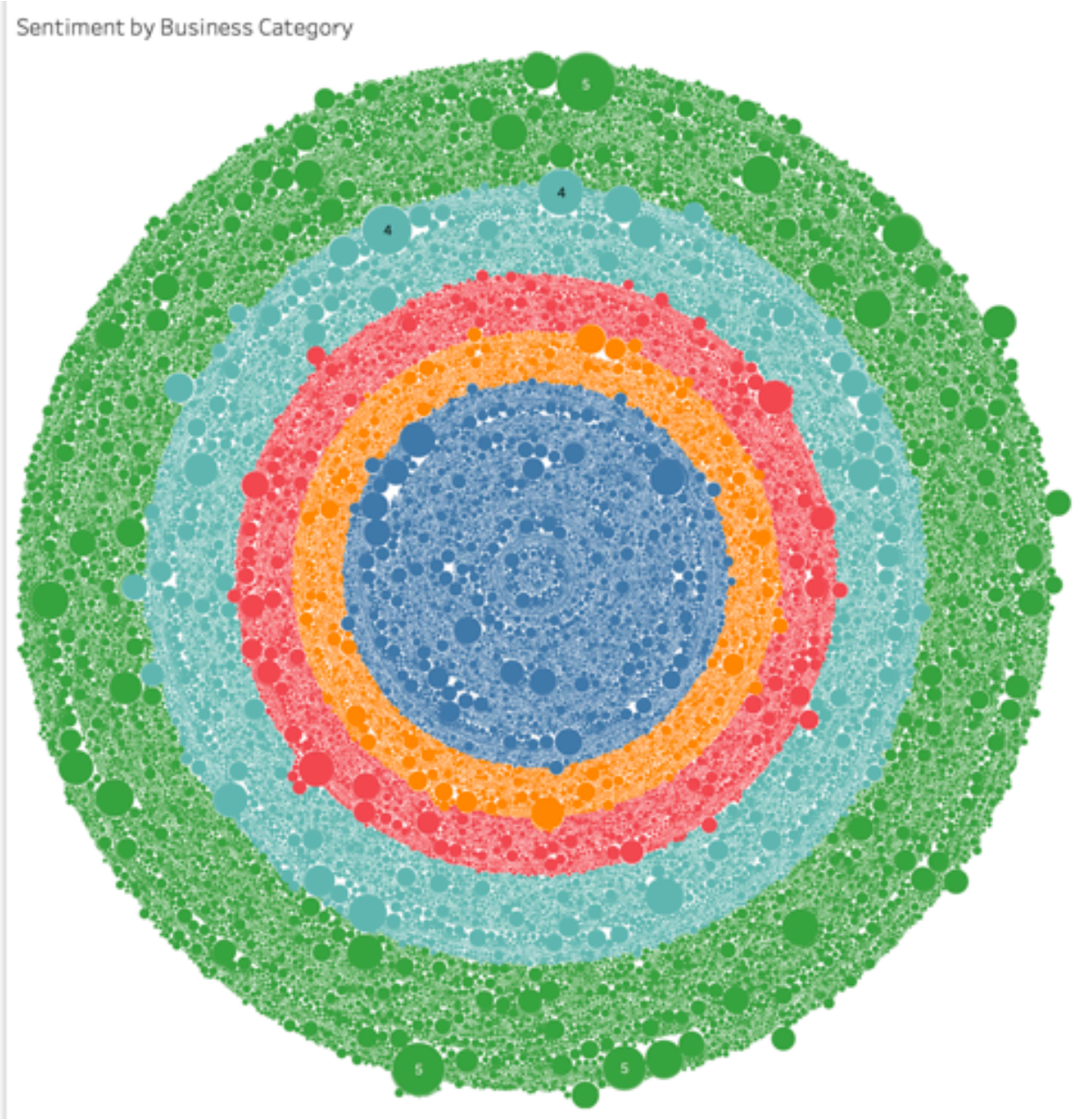


Figure 9. Sentiment by business category

As shown in Figure 9 we can see the breakdown of sentiment by business category. The packed bubble chart, in this instance, has the lowest ranked businesses in the center moving towards the highest ranked businesses on the outer ring.

The packed bubble chart is a good general-purpose choice for visualizing data on datasets that have dimensions and measures.

Data Visualization 3

Understanding the sentiment score of tweets is the main objective of data analysis on the Twitter data stream. We leverage the VADER lexicon to process sentiment in real time during ingest. Tweets are scored on a spectrum from negative, neutral, positive and a compound score. As shown in Figure 10 we can see the extracted tweet topics and their relative scores.

Figure 10. Sentiment score by subject

To view all the scores of the extracted subjects we leverage an area chart with stacked values. The ability to stack the scores allows us to review all scores across the multiple extracted subjects.

As we can see in Figure 10, some topics skew sharply negative. The movement called 'Black Lives Matter' tends to score negative sentiment as Muslim, Soros, and Mueller. These scores are generated by VADER and the context in the tweet and not any implied overarching social context. From these results we can observe these topics are either negatively discussed or mentioned.

For the Yelp dataset we will apply the same sentiment analysis techniques towards user business reviews by state. The sentiment scores are on a range of negative, neutral, positive and with a compound score. As we can see in Figure 11, scores varied wildly state by state.

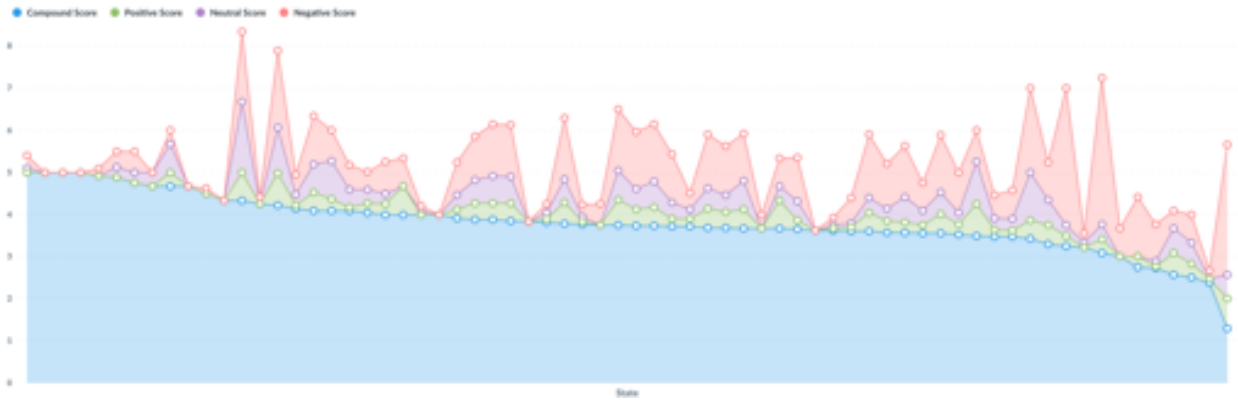


Figure 11. Sentiment score by state

As can be seen from Figure 11 most states scored a positive score (pink color) and the compound scored trended relatively positive. To see if these trends hold, we opted to do additional analysis. In Figure 12, we group the sentiment scores by categories.

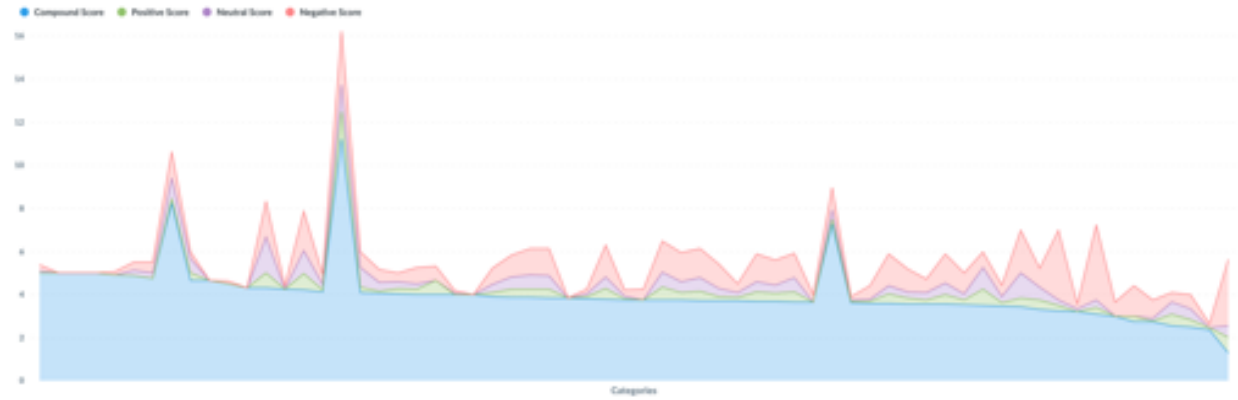


Figure 12. Sentiment score by category

Reviewing Figure 12, we can see that the relatively positive scores are exceptionally high in some business categories. With the visualization tool we are using, Metabase, we can 'drill down' into data points. Using this ability, we can see that the most positive reviewed businesses are in the Nightlife; Bars; Pubs categories as shown in Figure 13.

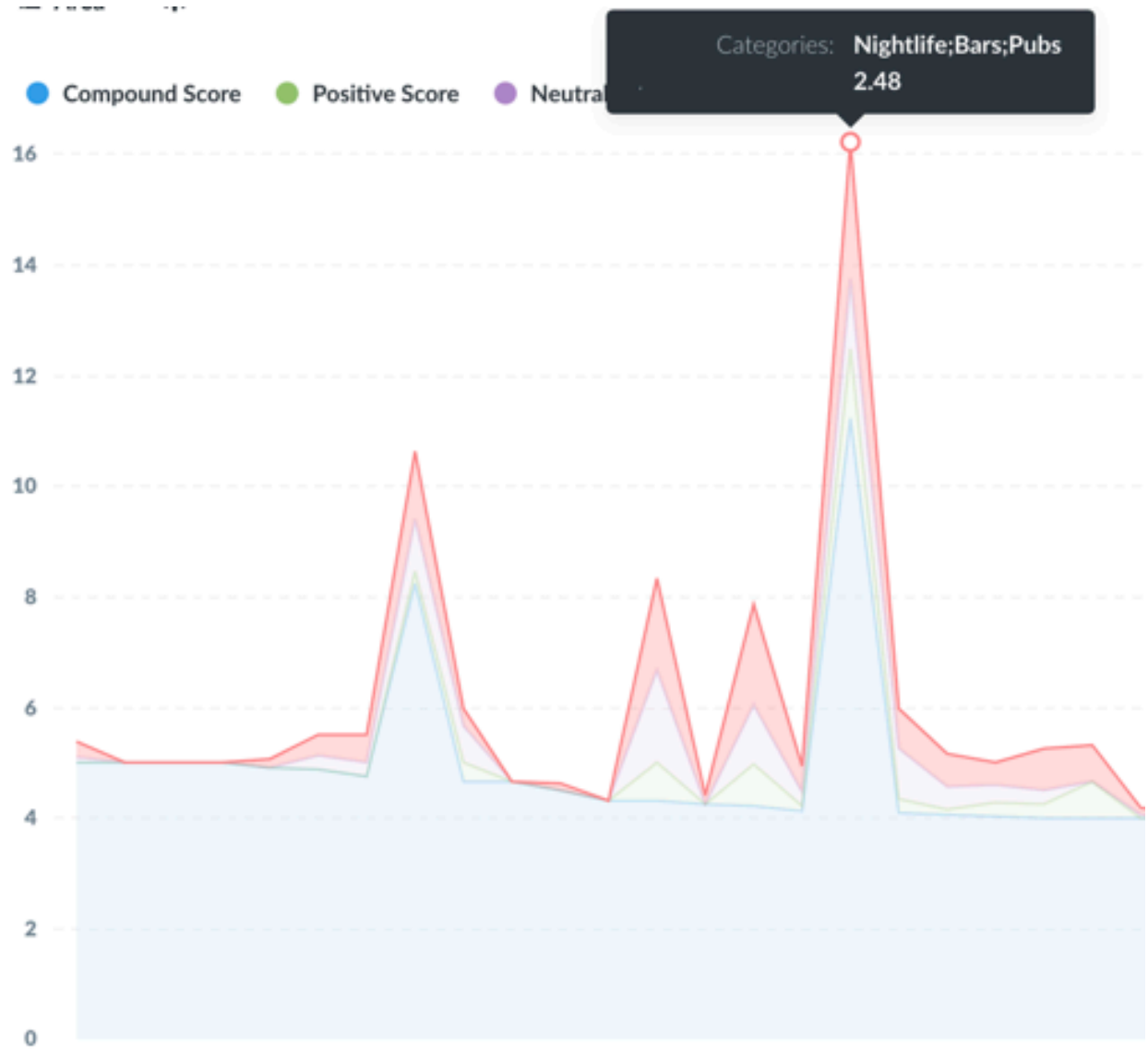


Figure 13. Drill down on positive scores

Proposed Visualizations

The Twitter and Yelp datasets both contain geographic information. This information can be used as a geospatial visualization to show on map where the most popular topics or categories are originating from as well as locations with the highest compound score.

One problem with the geospatial information in the Twitter dataset is that the location information is self-reported by the Twitter user. This could prove problematic as we would have to geocode the data to something that is useful as can be seen in Figure 14. Geocoding is the process of getting the geographical coordinates from a provided location. For example, for the user reported location of London, we would have to geocode that to latitude/longitude coordinates of 51.5074° N, 0.1278° W.

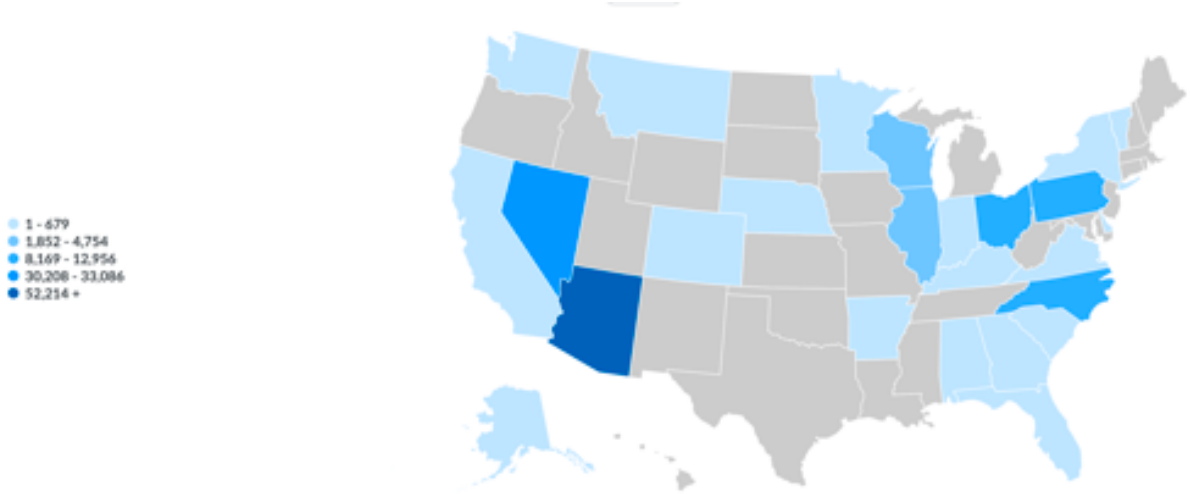


Figure 14. Self-reported Twitter user locations

Another proposed visualization for the Twitter and Yelp datasets is a Sankey diagram. A Sankey diagram is a type of flow diagram. For the purpose of data visualization, the flow to be visualized would be the dataset of interest. With the data flowing from the left side of the visualization to the right into groups.

For Twitter dataset, we would do this based the Twitter volume flowing to the negative, neutral and positive score categories. For the Yelp dataset we would use the volume of Yelp reviews flowing to the negative, neutral and positive score categories.

The main benefit of Sankey diagrams is that they are quickly understood. Sankey diagrams put emphasis on the flow of the diagram as shown in Figure 15.

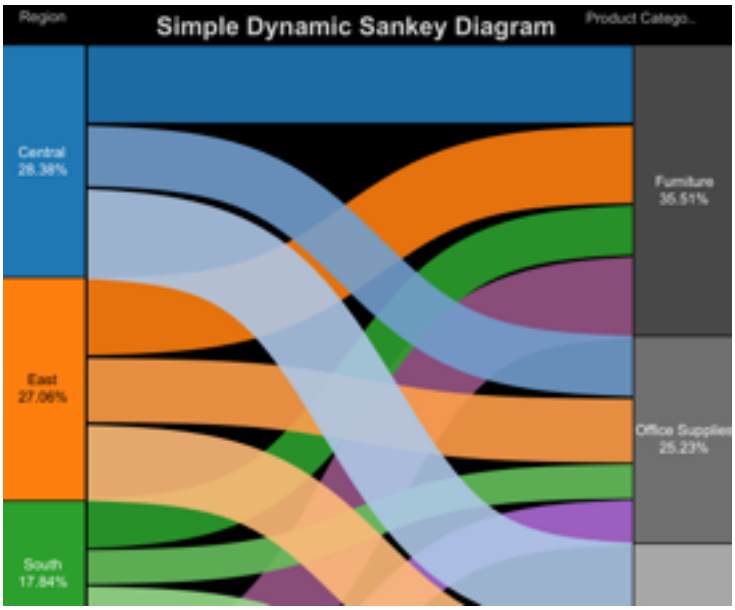


Figure 15. Sankey example

Predictive Models

The main objective of this project is the application of natural language processing tools to understand an unstructured data corpus. Natural language processing enables the understanding of data, particularly unstructured natural language, by a computer. For this project we use two natural language processing libraries for sentiment analysis and entity extraction. We will also discuss our analytical technique for gleaning an understanding of trends on the datasets.

Sentiment Analysis

For sentiment analysis we use the 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), which is the main datasets of this project. The need for a domain specific lexicon that applies to social media versus traditional lexicons drove the choice for VADER.

In the natural language processing domain one can chose from many 'gold standard' lexicons such as Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW) or the General Inquirer (GI) that have been validated time and time again. LIWC and GI are polarity-based lexicons that organize words into one of two categories, positive or negative, related to sentiment analysis. ANEW is a valence-based lexicon which determines the binary state of language, positive or negative, but also includes the in-sentiment intensity. Words in the ANEW lexicon have an associated sentiment valence ranging from one to nine with five being the midpoint. This valence-based lexicon is what is required in determining the sentiment of a micro-blog. Unfortunately, these 'gold standard' lexicons do not

account of understanding the deeper lexical properties from in most micro-blog content therefore the need to use the VADER rule-based sentiment analysis lexicon.

VADER is specifically attuned for micro-blog content that is found on social media and outperforms traditional machine learning classifiers and traditional lexicons. In development of VADER, the researchers compared VADER against machine learning classifiers such as Naïve-Bayes, Maximum Entropy, and Support Vector Machine. As we can see from Figure 16, VADER outscores traditional machine learning algorithms for extracting sentiment from micro-blog content on various social media platforms.

	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

Figure 16. VADER accuracy scores.

Additionally, VADER scores just as high or in some cases higher than human classifiers on determining the sentiment of micro-blog content such as Twitter. As we can see from Figure 17, VADER outperforms humans at classifying the sentiment of a tweet.

	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

Figure 17. VADER accuracy scores versus human.

VADER is designed to outperform previous natural language lexicons, machine learning classifiers and humans at determining sentiment of micro-blog content. VADER is some performant it exceeds in scoring sentiment on streaming data. The data that we consume from Twitter is a streaming dataset and VADER correctly scores sentiment as the data is streaming from Twitter into our database. The chose for VADER as our sentiment analysis is obvious when we see how well accurate the lexicon performs for domain specific content from micro-blog sites.

Entity Extraction

Entity extraction is the ability to locate and classify named entities in unstructured text. For entity extraction our main objective was to consume data from both Twitter and Yelp while determining entities referenced in the content all in a highly performant manner. The data we consume from Twitter is a streaming dataset therefore the requirement of do entity extraction in

a way that does not impact the ingestion of this data. For the Yelp dataset, entity extraction was not as needed as the data is classified by the data model into businesses.

Given the requirement for the streaming dataset we chose the entity extraction library of spaCy. spaCy is designed to be extremely fast and has models trained using convolutional neural networks. spaCy outperforms other natural language processing entity extraction solutions. As we can see from Figure 18, spaCy has an accuracy of 92.6% on correct entity identification while processing thousands of words a second.

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

Figure 18. spaCy speed and accuracy.

spaCy leverages a convolutional neural network to train its models for entity extraction. For this project we use the *en_core_web_sm* model for entity extraction. This model is an English multi-task convolutional neural network that assigns word vectors, parts of speech tags, and named entity recognition. As we can see from Figure 19, spaCy has incredible syntax accuracy and named entity recognition.

SYNTAX ACCURACY		NER ACCURACY	
UAS	91.72	NER F	85.30
LAS	89.80	NER P	84.97
POS	97.04	NER R	85.63

Figure 19. spaCy accuracy scores on syntax and named entity recognition.

When parsing unstructured data to determine entities objects will be classified into the categories in Figure 20.

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.

Figure 20. Entity classifications.

The application of this named entity recognition can be seen when applied to the streaming Twitter dataset. Figure 21 shows us that a majority (57.02%) of tweets contain some reference of either an organization (33.56%) or a person (23.46%)

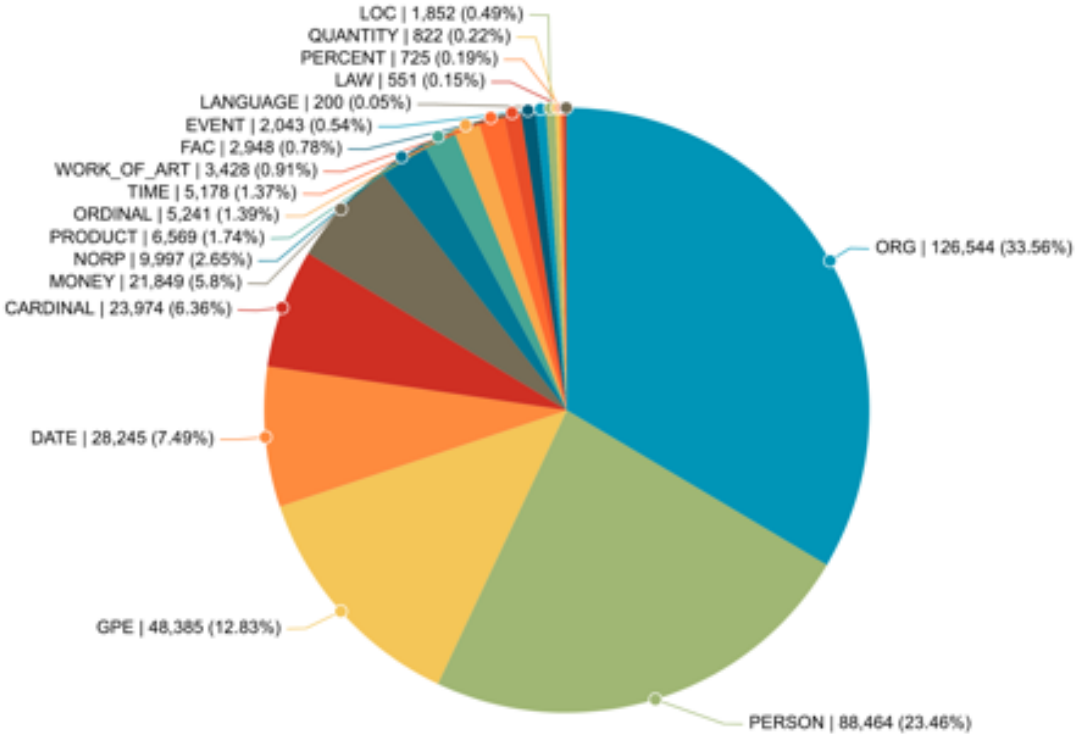


Figure 21. Entity extraction from Twitter dataset.

Given the speed and accuracy of spaCy, we believe the choice to use this library for entity extraction is self-evident.

Analytical Technique

In order to get a deeper understanding of the Twitter and Yelp datasets we apply descriptive statistics. Descriptive statistics enable understanding data in a dataset by using mean, median and mode.

Mean is the sum of all the scores divided by the number of scores. Mean calculations can be strongly influenced by the appearance of extreme measures that sway the calculation. For example, five customers have a bill of $10 but the sixth has a bill of $100. These extreme outliers can cause a normal data distribution to skew. In order to combat this skew, we can use the median measure.

Median is defined as the value in the middle position in a dataset when the data is sorted in ascending or descending order. Using this allow the data to distributed in exactly two halves. Median is often referred as a positional average. As stated previously, median is not distorted by outliers.

Lastly, we can leverage mode to get an understanding of a dataset. Mode is defined as a value that occurs most often in a dataset. In some datasets, one cannot determine a mode as a measure only appears once. On the other extreme some datasets can have more than one mode if the dataset has two or values of equal frequency that appear more often than any other data. Mode is typically not a useful measure in gaining insight into a dataset unless there is a central tendency in the dataset distribution (Manikandan, 2011).

The application of these techniques enables us to understand the data at a deeper level. As we can see Figure 22, we develop a box plot on the Twitter dataset. The box plot in Figure 22 shows us the min, median and max with quartiles of compound sentiment scores from extracted entities from Twitter.

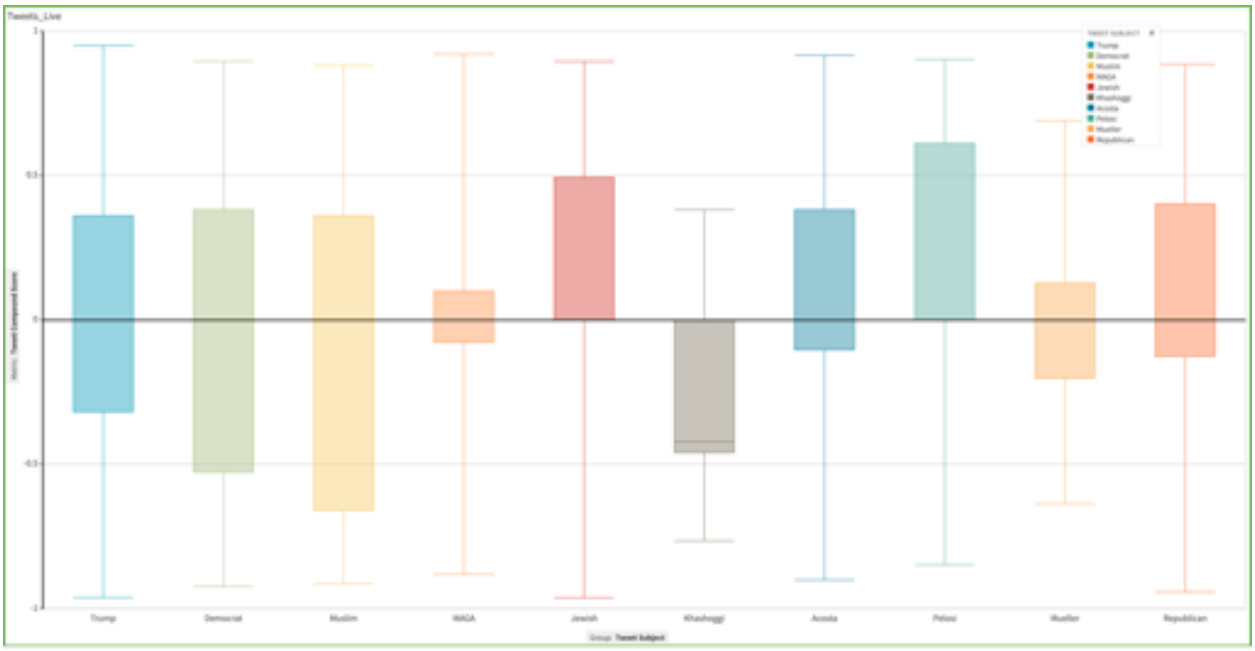


Figure 22. Box plot on Twitter dataset.

Sentiment Analysis, Entity Extraction & Analytical Technique Review

The main objective of this project is to use natural language processing tools to get an understanding of an unstructured data corpus. The datasets we used were a streaming dataset from Twitter and a static export from Yelp. We used VADER for sentiment analysis on the micro-blog content from both Twitter and Yelp. spaCy was used for entity extraction for Twitter but was not applicable to Yelp. Lastly, for an understanding of the data we applied descriptive statistics.

VADER has been shown to outperform traditional natural language processing lexicons and human classifiers. The creators of VADER also compared it to machine learning classification techniques, but VADER far exceeded what these algorithms can do in the natural

language processing domain. The reason VADER outperforms other choices in this domain is due to the fact it is purpose build for streaming, micro-blog content. It also has a modern list of lexical features along with their associated sentiment intensity measure (Gilbert, June 2014) making it the natural choice for this project.

spaCy is purpose built to be incredibly fast and accurate for large scale entity extraction tasks (Industrial-Strength Natural Language Processing, n.d.). spaCy is trained using convolutional neural network models that assigns word vector, parts of speech, named entities. We have shown that spaCy performs so well and is so accurate that it makes it choice to meet the requirements of this project.

Lastly, we applied descriptive statistics techniques to get a deep understanding of the data sets. Using median scores, we are able to understand the data distribution and create compelling data visualizations to quickly convey understanding.

Final Results

The results of the project validated the theory that machine learning natural language processing can successfully score the sentiment of a social media posting and extract the referenced entities in said post. The solutions we developed enabled us to scale up to in order to handle volume, velocity and variety of the ingested data, sboth streaming and batch. Additionally, the data visualizations technologies deployed decoupled the solution from the constraints of desktop compute resources.

Analysis Justification

Social media is a pervasive force in today's world. These platforms have become a vital medium for businesses, governments, and individuals to interact directly with their target audiences. Users of these platforms can engage directly with businesses, governments, and other users. These posts can, and have, influence others that read the posting therefore it is critically important to understand the sentiment of a post as well as all reference entities. It is not practical for humans to continuously monitor these sites therefore we have applied machine learning natural language processing for this task.

Leveraging machine learning natural language processing allows us compute the sentiment score of a post as well as extract the entities of the post. We have shown that the VADER lexicon can correctly score the sentiment of a posting. The most useful measure for a understanding the sentiment of a post was the VADER compound score. This is a single unidimensional measure of sentiment for a given posting (Gilbert, June 2014). With this we were able to track sentiment of postings about given extracted entities.

For the entity extraction, we used the high-performance entity extraction library spaCy. With spaCy, we were able to extract all referenced entities in a social media posting. This

enabled the ability to search across multiple dimensions when analyzing the data. We were able to use this capability to understand sentient scores as it applied to the various extracted entities both in real-time and after batch loading.

This capability to quickly understand the referenced entities and sentiment of a social media posting provides a unique tool to protect the reputation of for businesses, governments, and individuals. Reputation management is vital to protect the economic impact from negative posts (Fombrun, 1995).

Findings

This project explored the application of machine learning natural language processing for entity extraction and sentiment analysis for social media postings with streaming and batch processed datasets from Twitter and Yelp, respectively. With the streaming Twitter dataset, we were able to ingest over 31,000,000 tweets while extracted referenced entities and scoring the sentiment of the posting as seen in Figure 23.

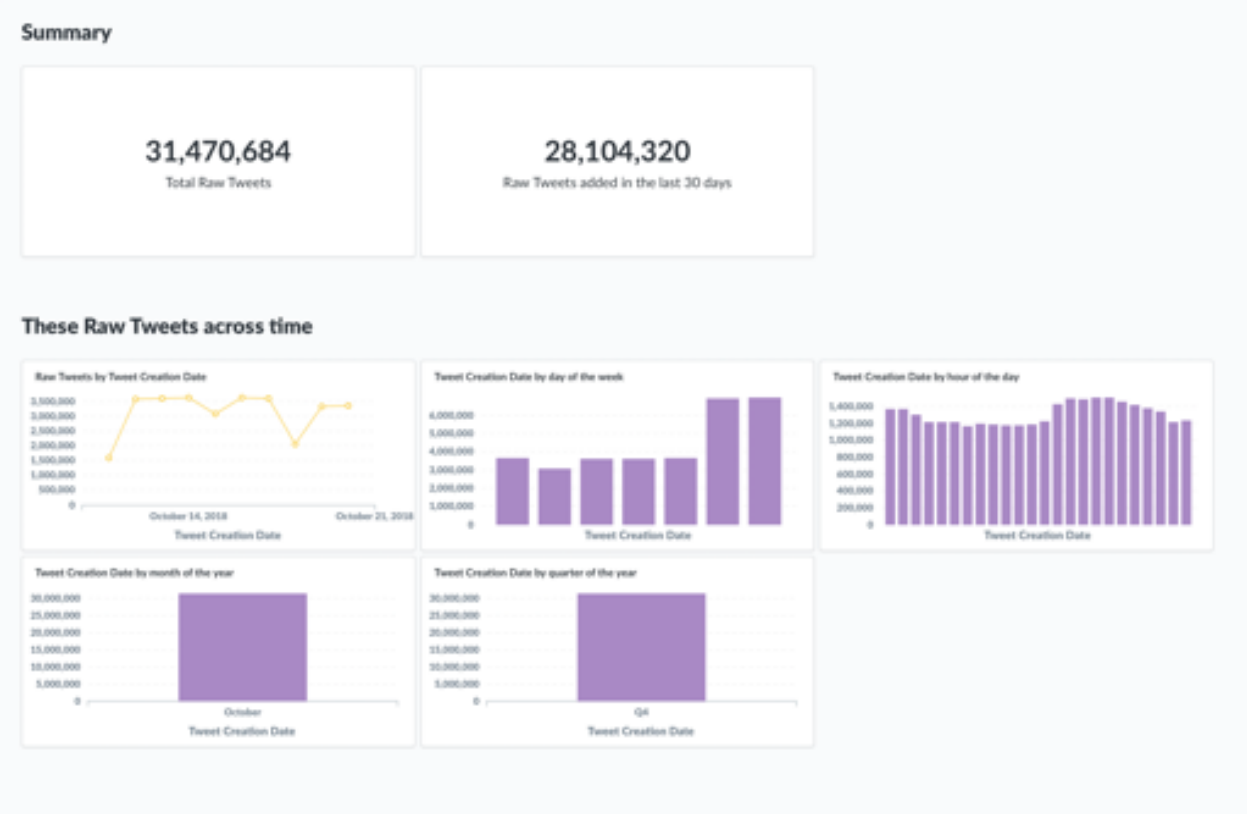


Figure 23. Total ingested Twitter dataset.

The Yelp dataset was a static dataset that contained 174,567 businesses with 17,746,271 user reviews across various geographies (Figure 24). We were able to score the sentiment as the data was loaded into the database. These sentiment scores mapped nicely to the 'star' rating of the review.

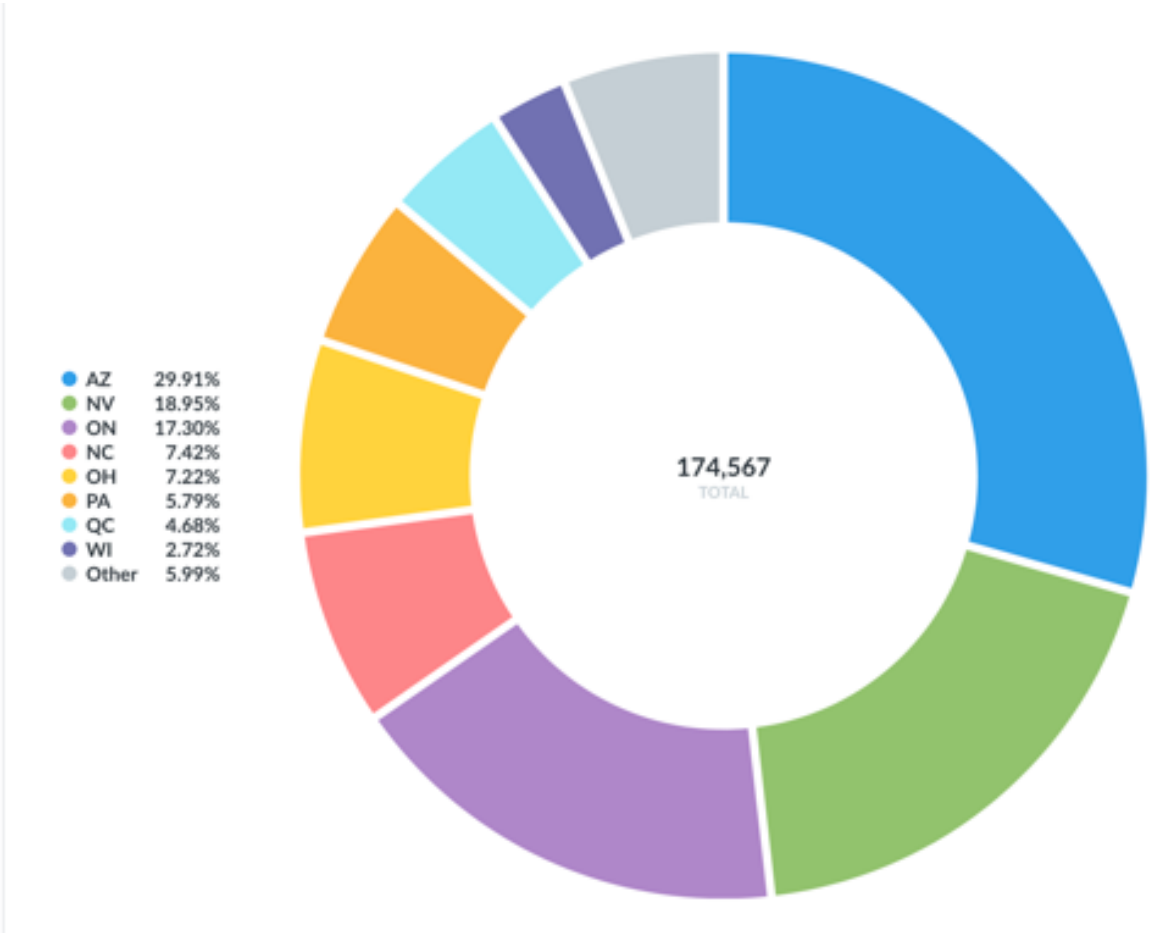


Figure 24. Yelp businesses.

The real-time, streaming Twitter workload forced us to explore technologies that would be able to handle the streaming rate. The technology solutions used in this project allowed us to not only ingest a traditionally difficult dataset but also visualize it effectively. Using Python libraries that were purpose built with performance in mind enabled the development of a highly scalable and performant architecture. The spaCy entity extraction was able to quickly identify, and extract referenced entities from a social media posting. VADER lexicon correctly scored the sentiment of the social media posting. This augmented data flowed into a NewSQL, in-memory first, distributed database with ANSI SQL support called MemSQL.

This ANSI SQL support was critical as it allowed us to use both desktop and server-based data visualizations. We found that the desktop-based tools could not handle the scale and speed of the Twitter dataset. This forced us to use server-based data visualizations tools, notably Zoomdata and Metabase. Zoomdata was an exceptional tool for visualizing the streaming Twitter dataset. Metabase allowed data exploration through a simple to use interface.

We found that this technology stack was extremely capable of handling both datasets for effectively extracting entities and scoring sentiment. Using a database with ANSI SQL support allowed maximum flexibility for working with most data visualization tools. We would recommend this technology stack for similarly based workloads.

Review of Success

This project had defined the business success criteria as ability to understand the sentiment of a Twitter and Yelp post, score said post on a spectrum of negative to degrees of positive as well as visualize trends over time. Additionally, we defined the following key performance indicators:

KPI 1: Monitored sentiment scores over time. This project will extract topics from the body of a posting and monitor the sentiment score over time. This is important to understand if a topic or brand perception changes over time. For example, an initial response of the public could negative but trend to neutral or positive over time.

KPI 2: Count of the number of tweets per public figure/topic/brand. This key performance indicator will measure the relative volume of a public figure/topic/brand over time. Understanding the volume of interest is a critical measure in understanding relative value/sentiment of interest.

KPI 3: Average sentiment score of tweets per public figure/topic/brand. Gathering the data and measuring the average sentiment for a public figure/topic/brand allows for the measured subject to adjust messaging over time. In order to do so, one must understand the current relative sentiment.

KPI 4: Measure the sentiment of user feedback in various businesses categories. This key performance indictor will be used to measure the positive, neutral, negative of various businesses categories as determined by user reviews in Yelp.

We have successfully developed an operational machine learning natural language processing technology stack that enables the sentiment understanding of a social media posting from both Twitter and Yelp. This solution is exceedingly flexible and can be adapted to additional datasets, either streaming or static and can support the addition of data visualization tools.

We have successfully met all the key performance indicators for this project. We have successfully monitored sentiment scores over time to see possible fluctuation of scores as shown in Figure 25.

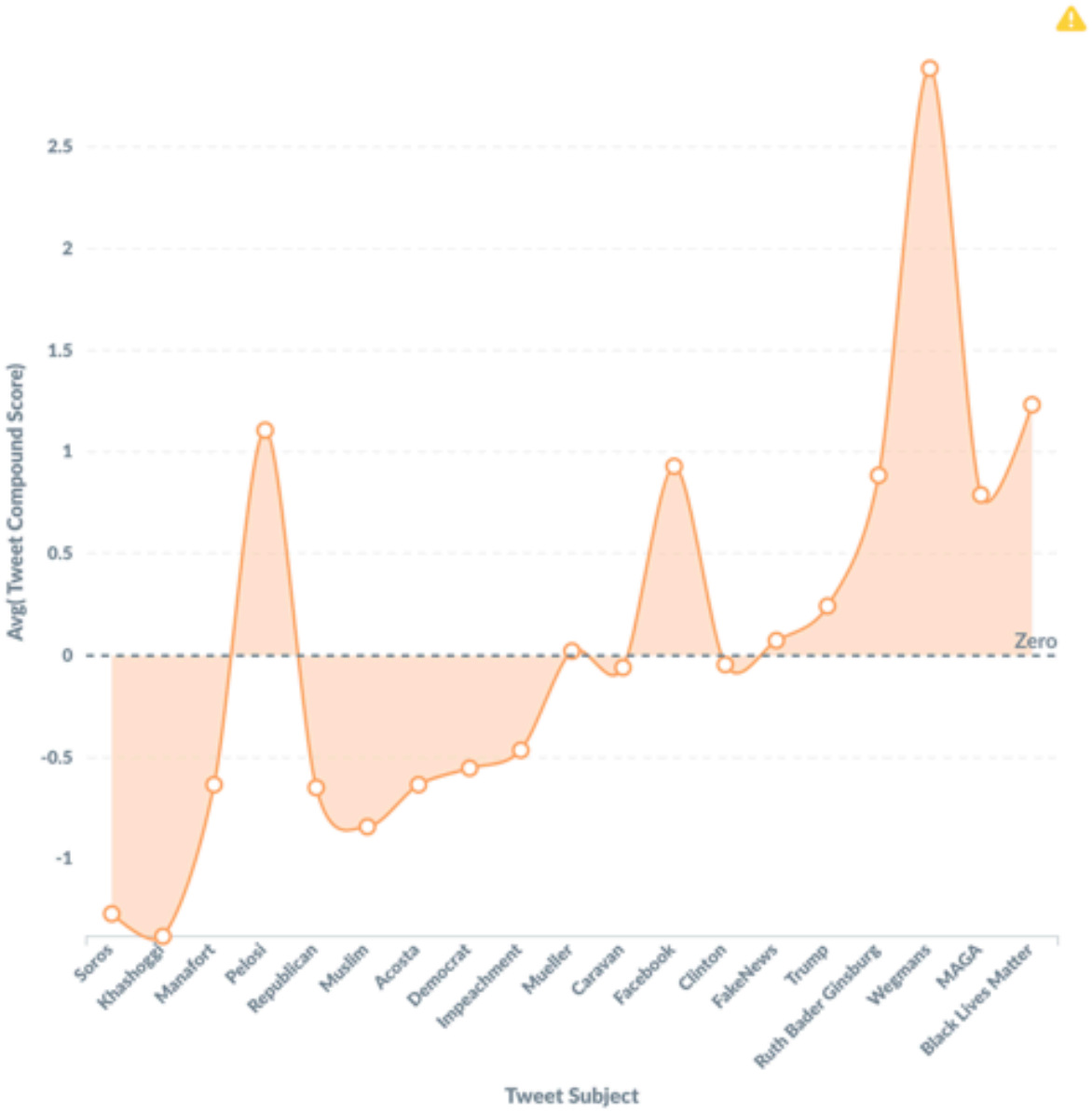


Figure 25. Sentiment scores monitored over time.

For key performance indicators two, we have successfully counted the number of tweets per extracted public figure/topic/brand. This metric allows us to quickly understand the trending topics posted to Twitter as shown in Figure 26.

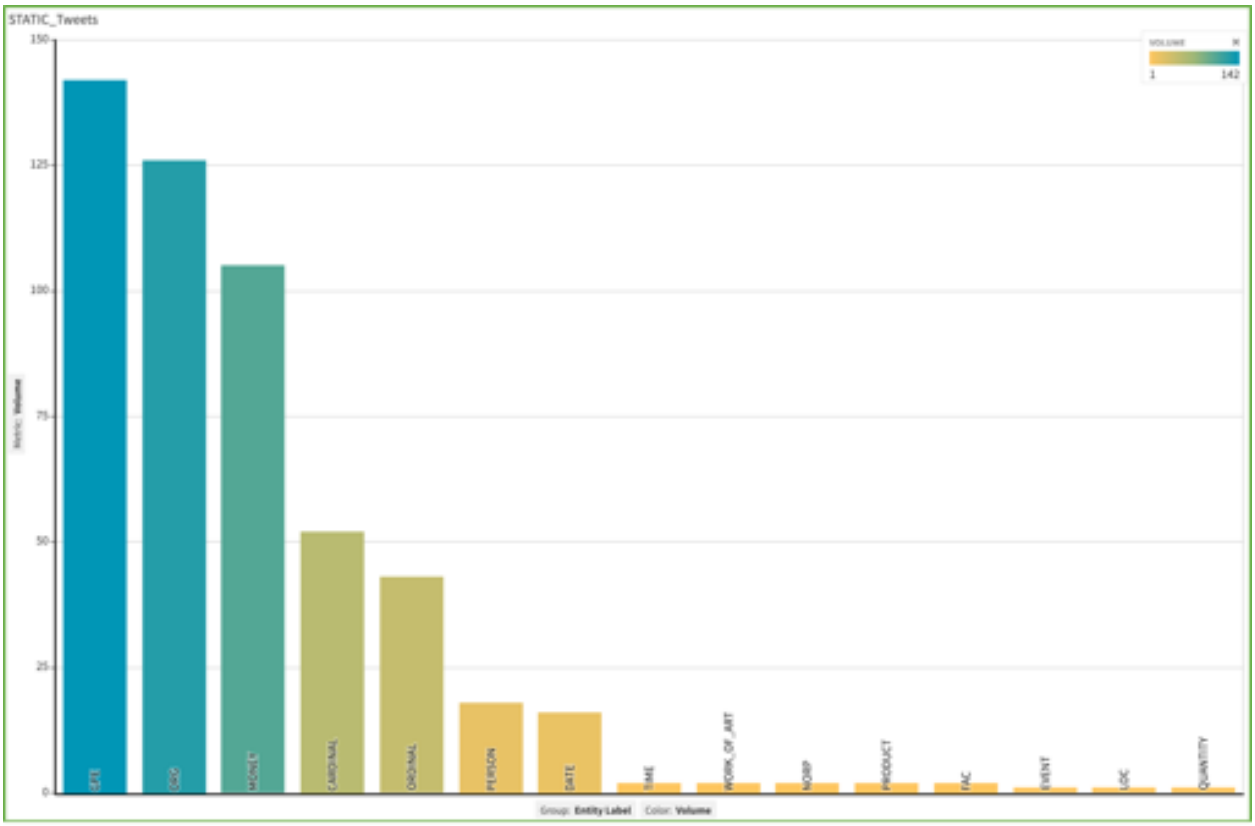


Figure 26. KPI 2 showing volume of tweets per category.

Key performance indicator three was successfully obtained by the project for monitoring the average sentiment score per public figure/topic/brand. This key performance indicator allows us to understand how an extracted entity is perceived on the social media platform as shown in Figure 27.

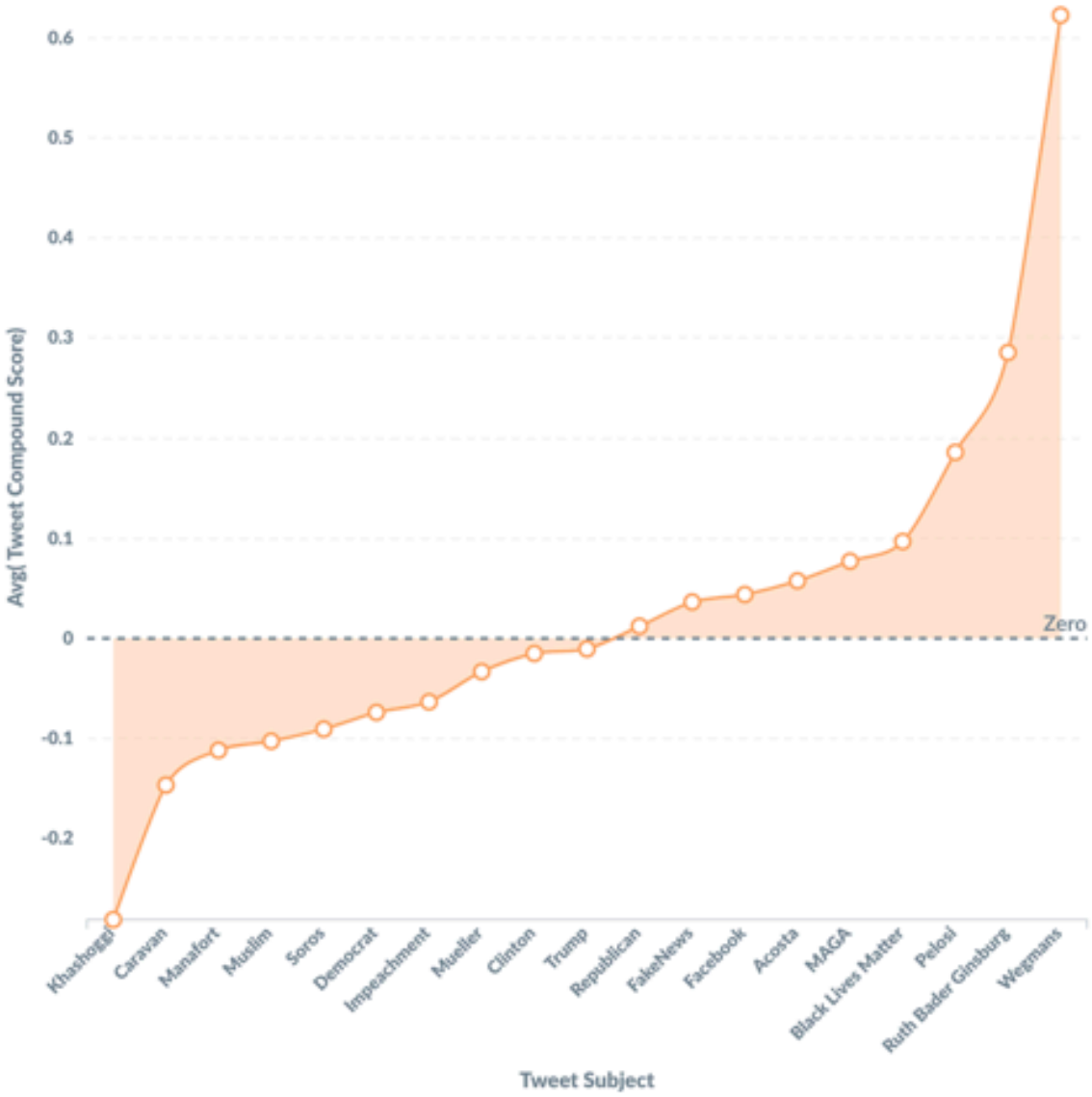


Figure 27. Average sentiment score of tweets per public figure/topic/brand

Lastly, for key performance indicator four we were able to successfully measure the sentiment score of user feedback of various business categories. This enables us to understand which businesses are the most popular with users on the social media platform as shown in Figure 28.

Figure 28. Measure the sentiment of user feedback in various businesses categories.

Recommendations for Future Analysis

This project was extremely successful at meeting both the business success criteria and key performance indicators but there is additional work that should be done. Continued work should focus on retraining the entity extraction model, enabling geocoding of user provided location information and development of a natural language grammar solution.

While the spaCy entity extraction engine performed exceedingly well, there needs to be retraining of the model. The convolutional neural network model does well to identify and extract most entities but does create miscategorized labels. For example, the model will correctly categorize the name "Jim Acosta" into the PERSON category but when only the surname of "Acosta" is in the corpus, it incorrectly categorizes it to the ORG category. Additional work should be done in order to train the model to identify surnames into the PERSON category.

With the Twitter dataset, users can provide a location of where the tweet originates from. This data is typically in the form of CITY, STATE which is not in the format to correctly map this data with geospatial tools. In order to map user locations, we need to geocode the provided information into a latitude/longitude coordinate pair. This would take the data point of NEW MARKET, MARYLAND into the coordinate pair of 39.38279000000006, -77.27231999999998. This capability would allow us to quickly understand where tweets originate from and visually map that data.

Lastly, additional work should be done to develop a natural language grammar solution. This net new capability would create a gist for a social media post that would allow a human analyst to understand, at an executive level, what a social media post is about. These summaries enable easy comprehension of a posting without the need of a detailed reading.

References

Alessandro Bessi, E. F. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*.

Fombrun, C. J. (1995). *Reputation: Realizing Value from the Corporate Image.* Harvard Business School Press.

Generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc. (n.d.). Retrieved from Matplotlib: https://matplotlib.org

Gilbert, C. H. (June 2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI*.

Honnibal, M. (2015, February 19). *spaCy.io*. Retrieved from Blog: https://explosion.ai/blog/how-spacy-works

Industrial-Strength Natural Language Processing. (n.d.). Retrieved from spacy.io: https://spacy.io

Lee, B. P. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*.

Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School Working Paper, No. 12-016*.

Manikandan, S. (2011). Measures of central tendency: Median and mode. *J Pharmacol Pharmacother*, 2-5.

Matplotlib. (n.d.). Retrieved from https://matplotlib.org.

Richard Hanna, A. R. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*, 265-273. Retrieved from http://www.sciencedirect.com.

Yelp. (n.d.). *Yelp Open Dataset* . Retrieved from https://yelp.com: https://www.yelp.com/dataset

Zeitzoff, T. (2018). Does Social Media Influence Conflict? Evidence from the 2012 Gaza Conflict. *Journal of Conflict Resolution*, 22-27.