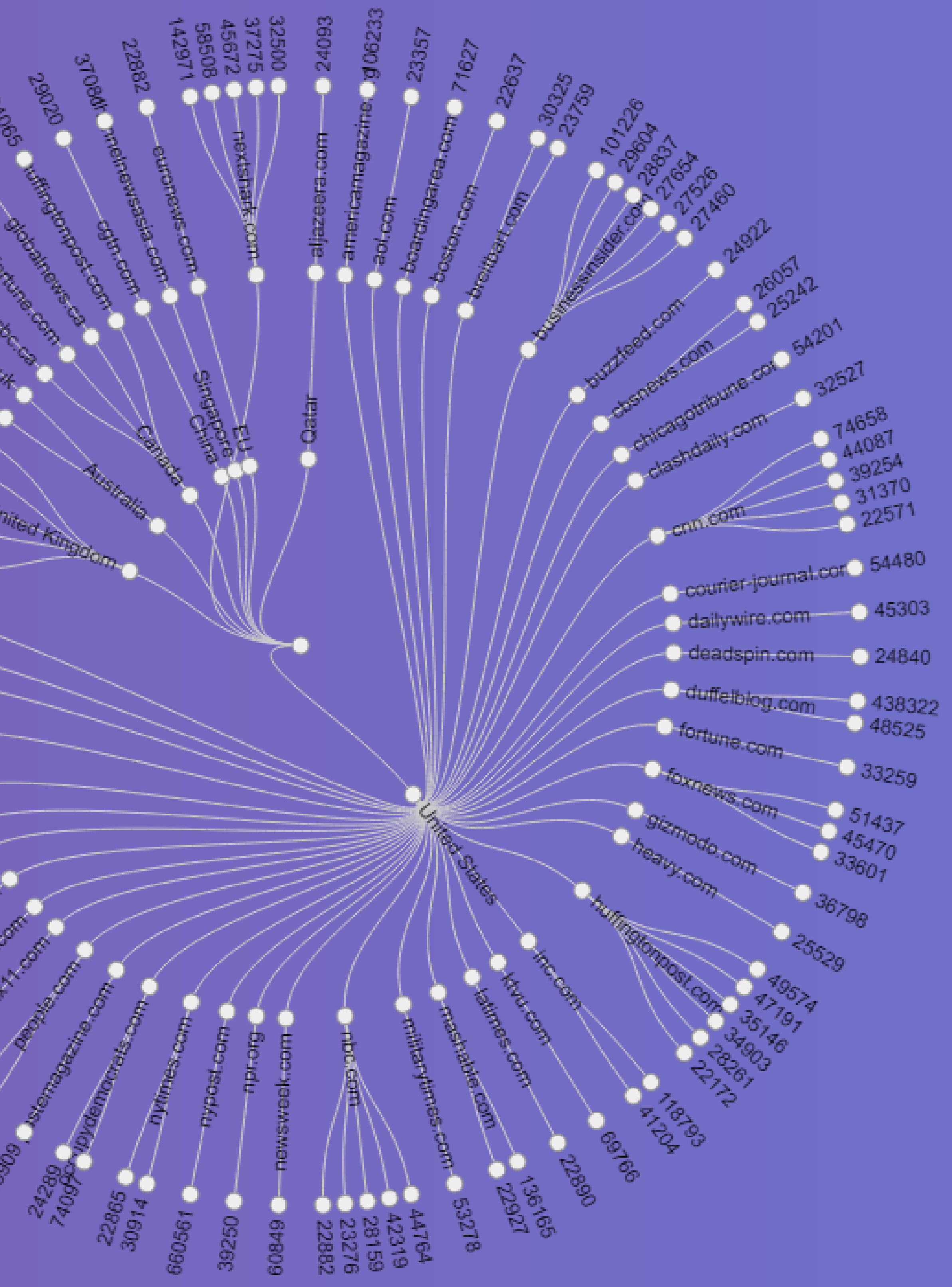
EPIC Data Challenge

QR Bunch

Prepared By: Raihanah Nabilah Fatinah, Pan Jia Qi, Sewen Thy, Sowmya Uppili Raghavan



Section 1: Introduction and Research Question

INTRODUCTION

SITUATION

United Airlines is one of the major airlines in the United States.

PROBLEM

United Airlines has underwent several high-profile cases in crisis reporting where news publications and syndication networks have reported negative feedback about the Airlines' service or incidents that took place during the customer journeys.

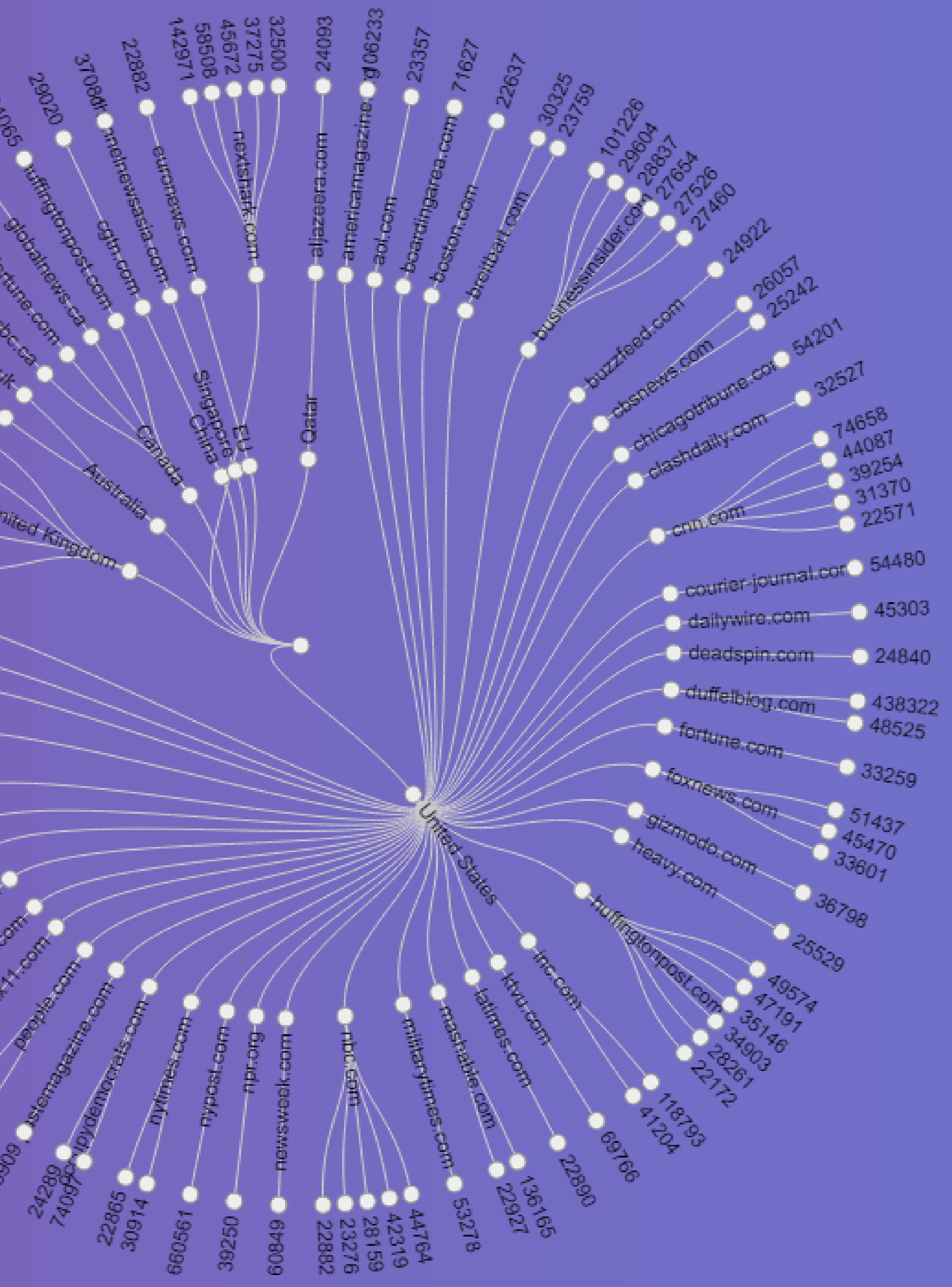
What are the key factors that lead to negative sentiment in news reporting about United Airlines during crisis events?

OUR KEY PROBLEM STATEMENT

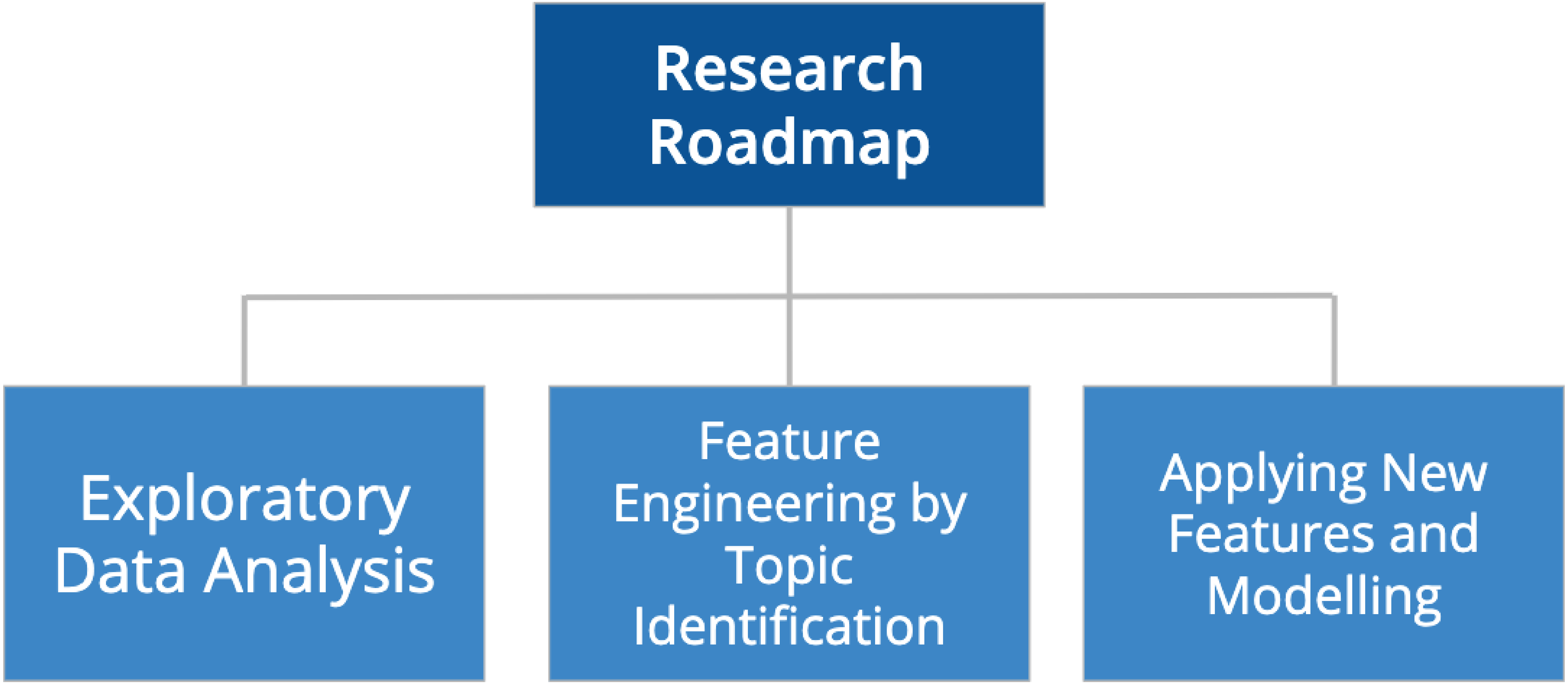
Pattern Detection

QUESTION B: WHO IS SHARING THE CONTENT?

- Are there any meaningful segments of influencers or publishers that you can identify? What are the common characteristics in each segment?
- Does the sentiment towards articles differ between the segments you develop?
- How do you identify top publishers and influencers? Which publishers and influencers should United Airlines pay attention to and why?



Section 2: Research Methodology

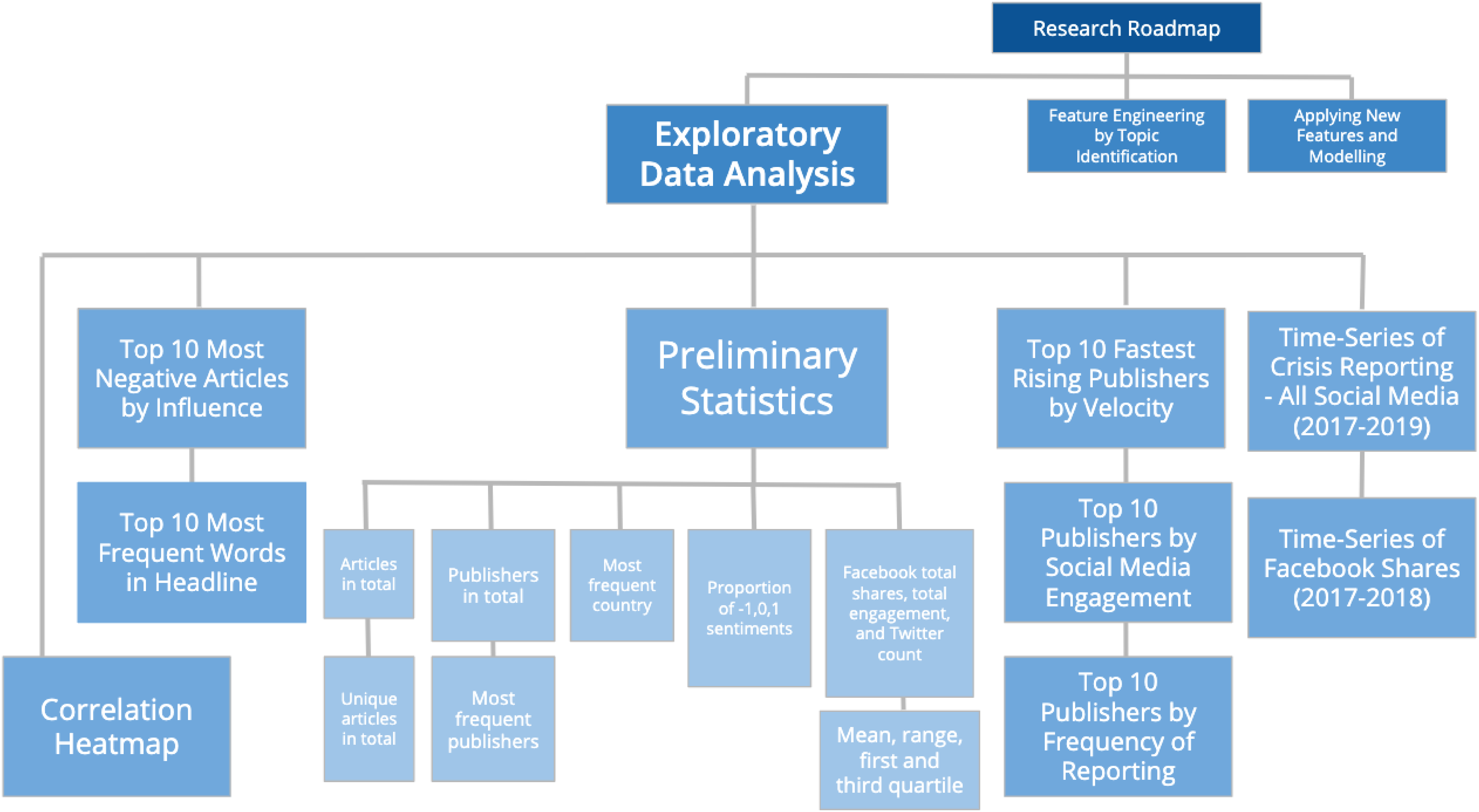


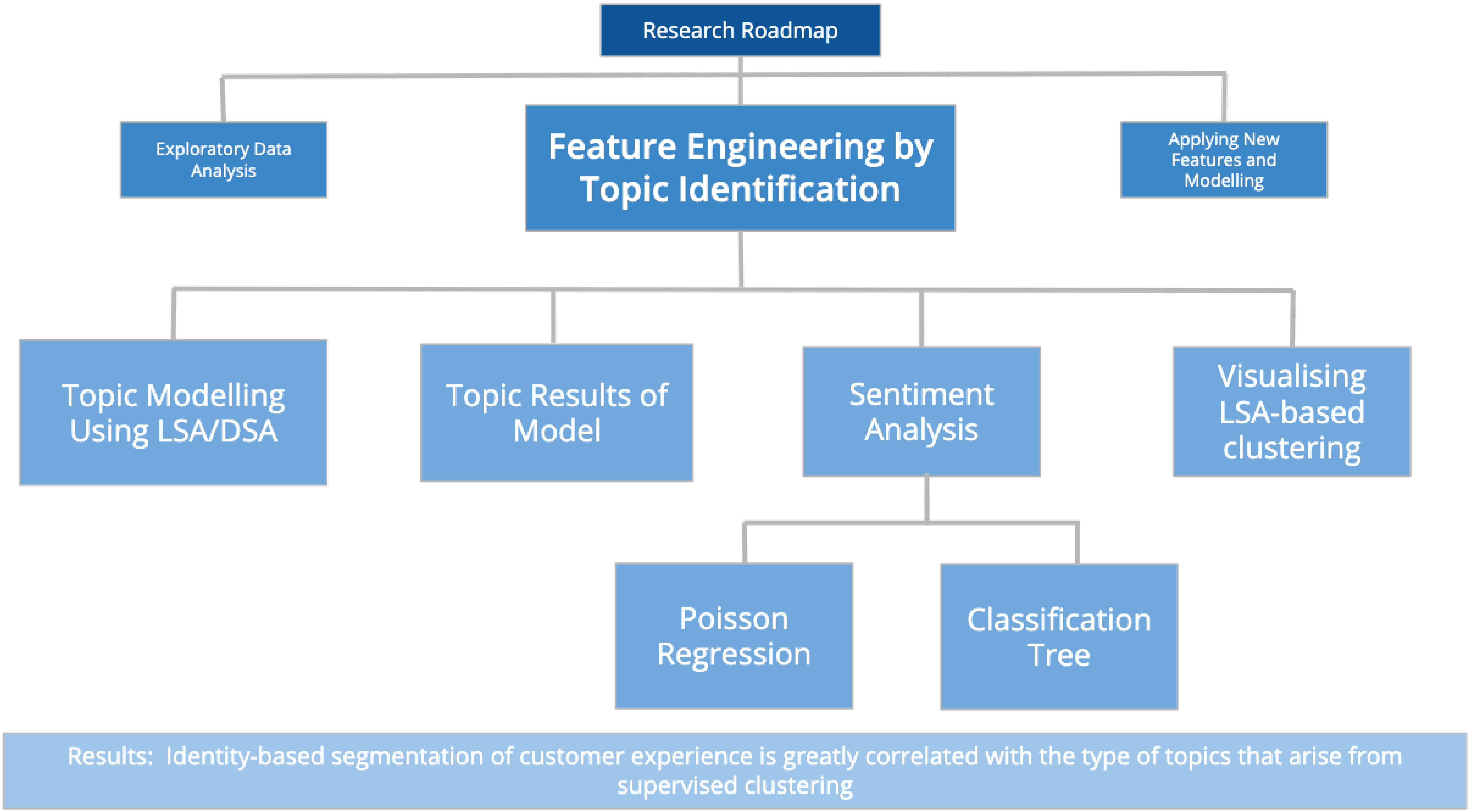
Research Roadmap

Exploratory Data Analysis

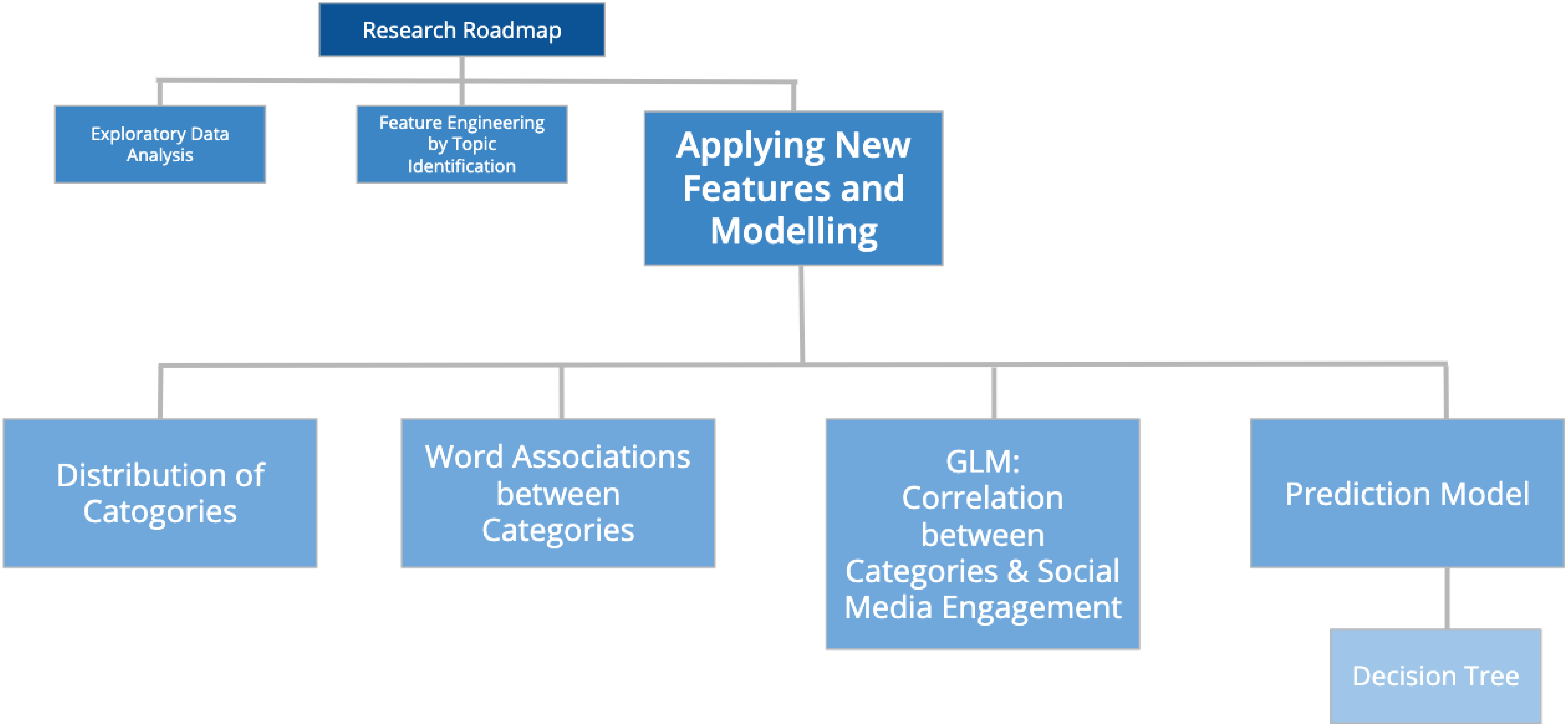
Feature Engineering by Topic Identification

Applying New Features and Modelling


Research Roadmap
Exploratory Data Analysis
Feature Engineering by Topic Identification
Applying New Features and Modelling
Top 10 Most Negative Articles by Influence
Top 10 Most Frequent Words in Headline
Correlation Heatmap
Preliminary Statistics
Articles in total
Unique articles in total
Publishers in total
Most frequent publishers
Most frequent country
Proportion of -1,0,1 sentiments
Facebook total shares, total engagement, and Twitter count
Mean, range, first and third quartile
Top 10 Fastest Rising Publishers by Velocity
Top 10 Publishers by Social Media Engagement
Top 10 Publishers by Frequency of Reporting
Time-Series of Crisis Reporting - All Social Media (2017-2019)
Time-Series of Facebook Shares (2017-2018)


Research Roadmap
Exploratory Data Analysis
Feature Engineering by Topic Identification
Applying New Features and Modelling
Topic Modelling Using LSA/DSA
Topic Results of Model
Sentiment Analysis
Visualising LSA-based clustering
Poisson Regression
Classification Tree

Results: Identity-based segmentation of customer experience is greatly correlated with the type of topics that arise from supervised clustering



Research Roadmap

Exploratory Data Analysis

Feature Engineering by Topic Identification

Applying New Features and Modelling

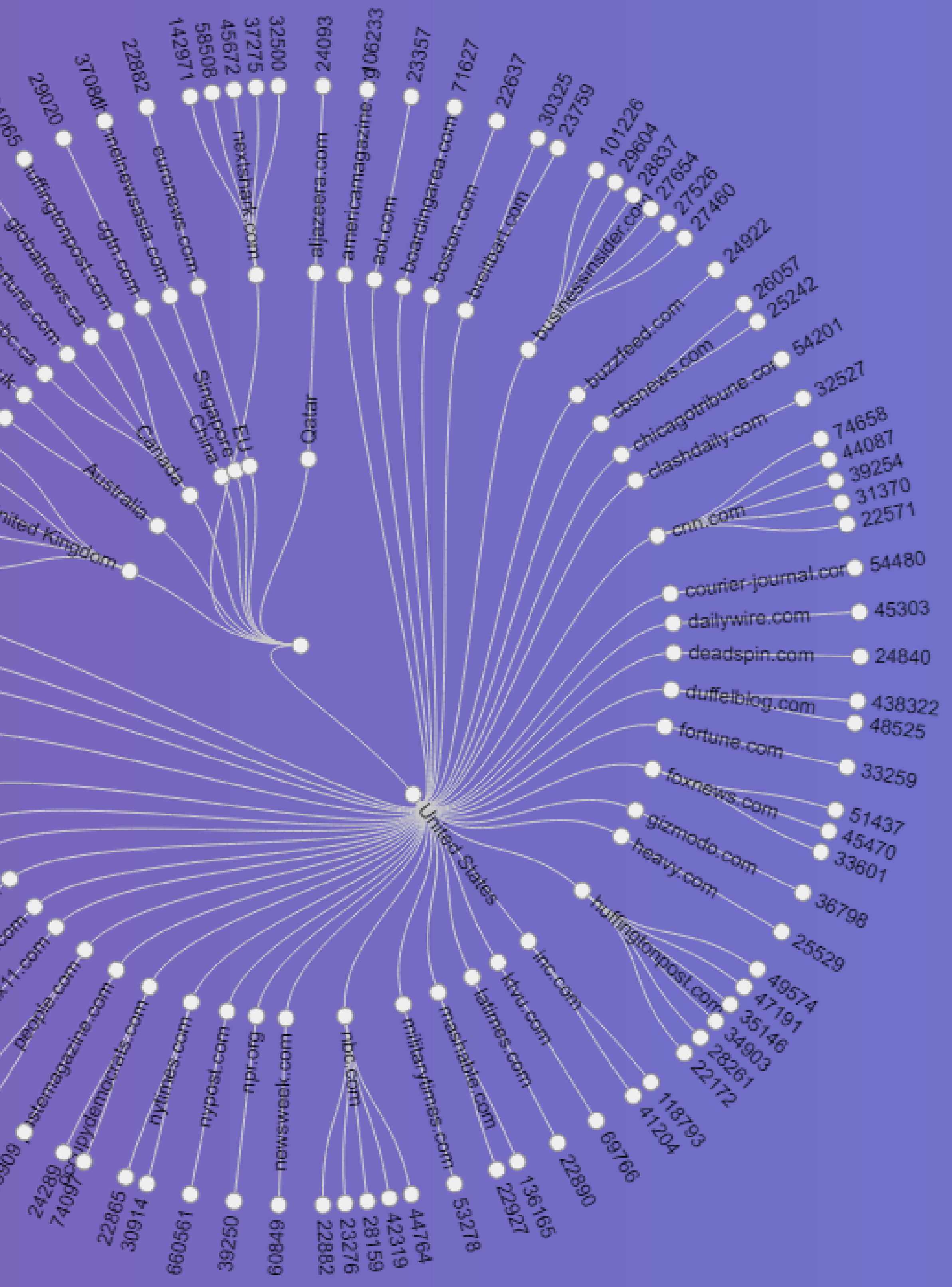
Distribution of Catogories

Word Associations between Categories

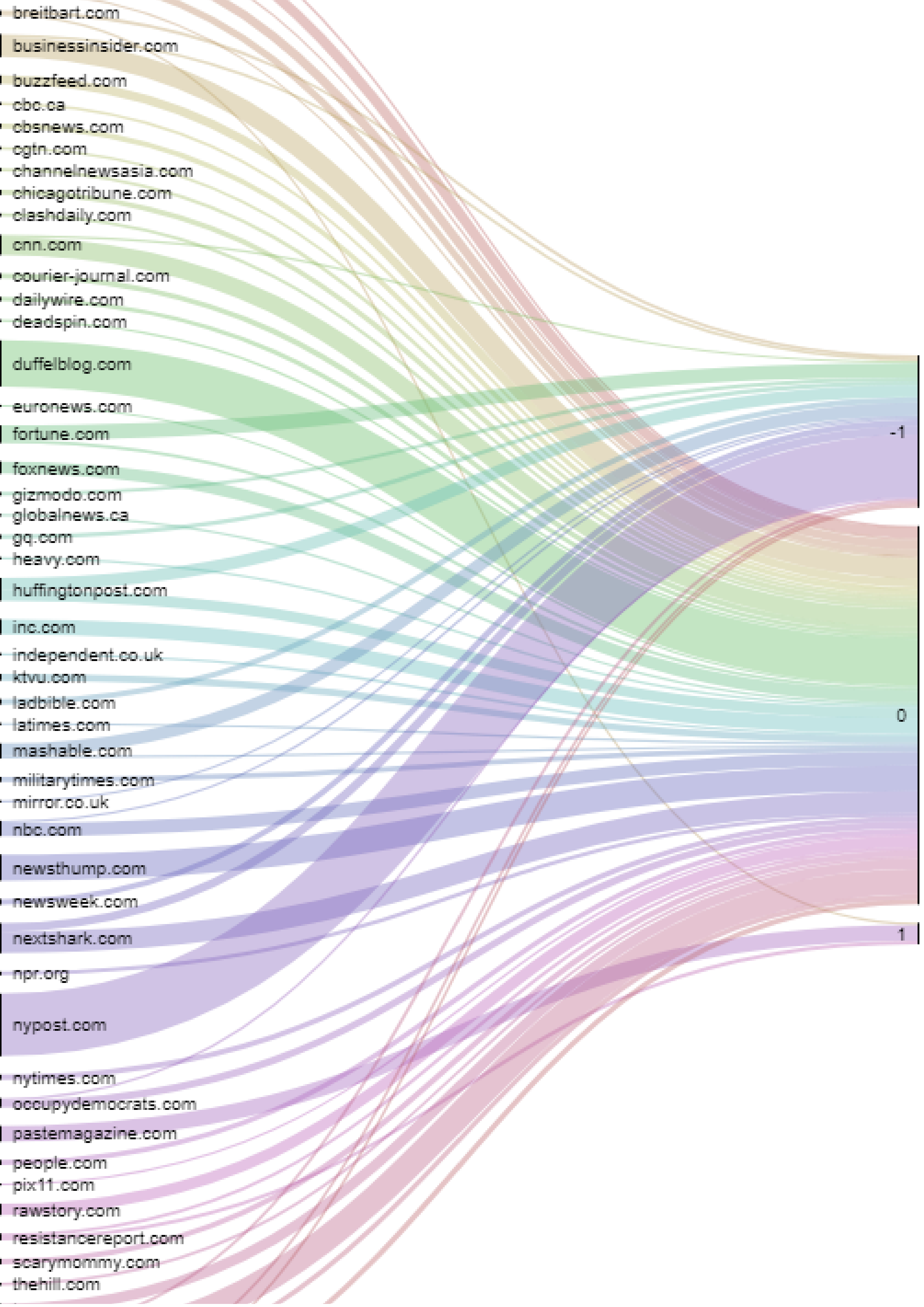
GLM: Correlation between Categories & Social Media Engagement

Prediction Model

Decision Tree



Section 2.1: Exploratory Data Analysis



Insight #1:

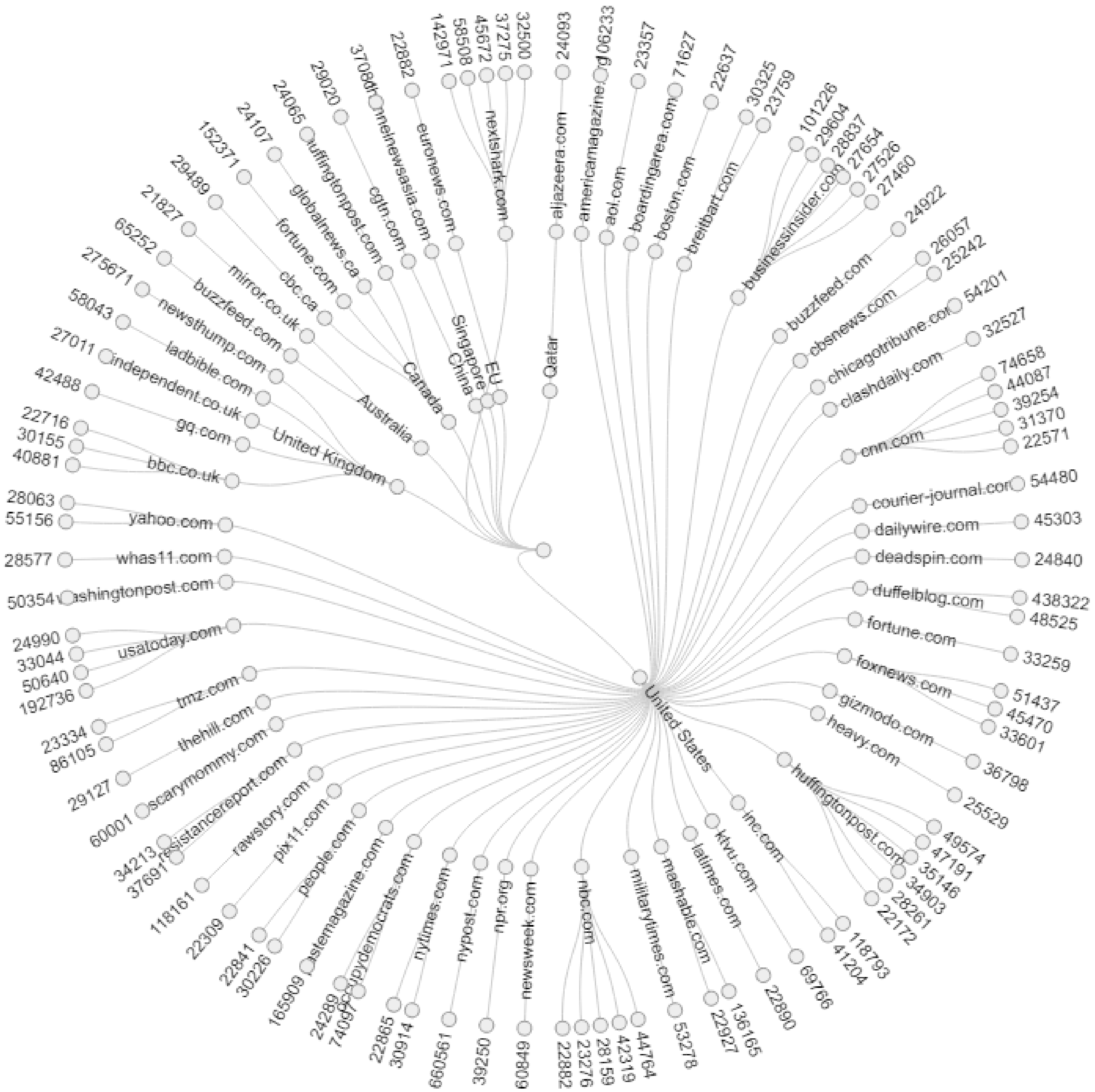
The most shared content by publishers are mostly negative or neutral. The least shared content is negative.

Methodology:

Using an **alluvial diagram**, we mapped the **top 400** publishers to their respective sentiment categories. The width of the connection was determined by the extent of the Facebook total engagement count, an inhouse metric that measures the impact of likes, shares and comments to a continuous numeric variable.

Analysis:

We found that while the bulk of the sentiment analysis is neutral (0), a significantly larger proportion of articles are negative compared to positive. This confirms our initial assumption that crisis events are often shaped by negative sentiment. This is understandable, as crisis events reflect customer dissatisfaction with the United Airlines brand and customer experience.



Insight #2:

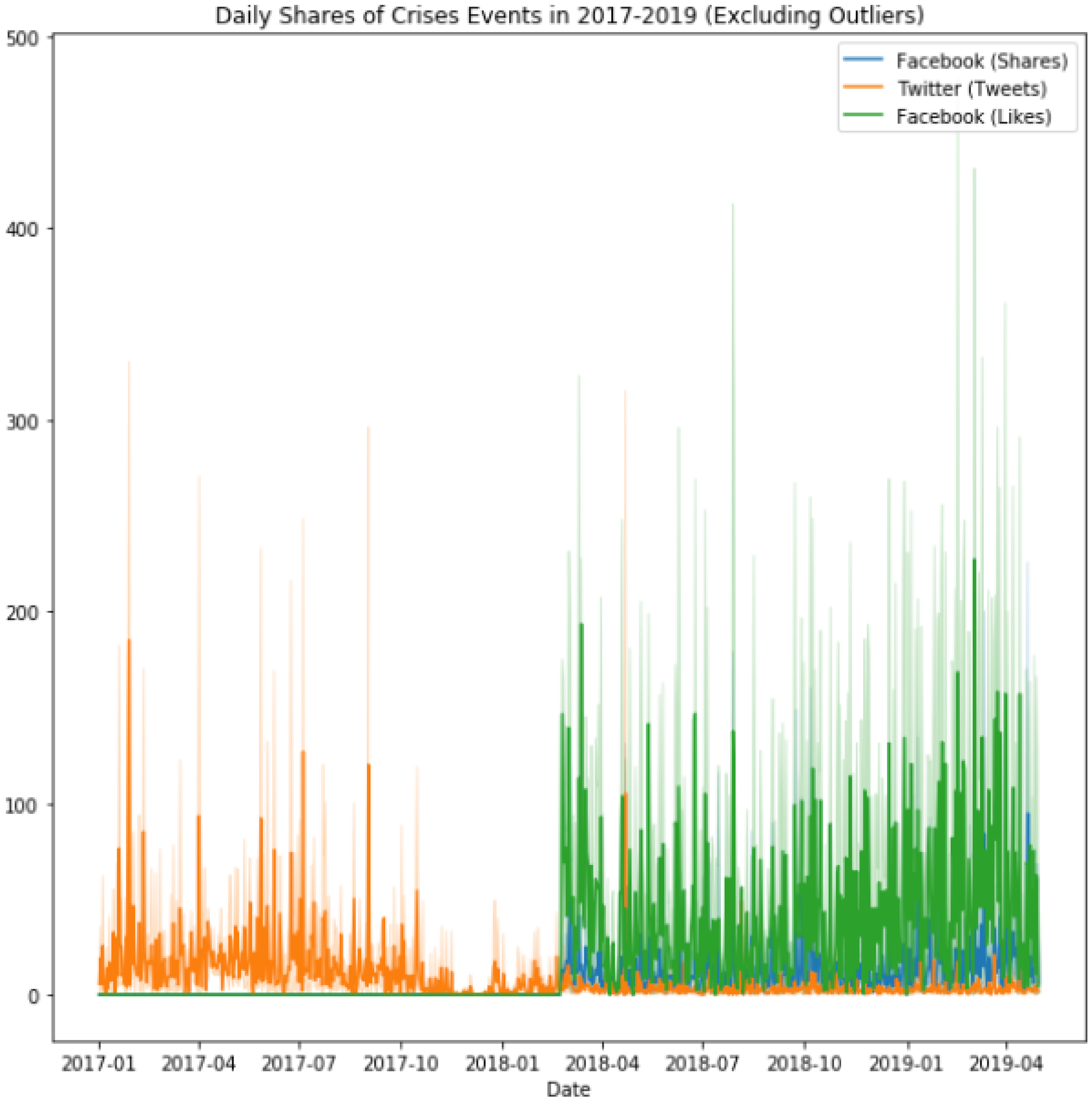
The most shared content by publishers are mostly negative or neutral. The least shared content is negative.

Methodology:

We constructed a circular dendrogram, which maps from the root node (country origin of the publisher) to the publishers segmented by date, and the overall total Facebook engagement. Publishers with high frequency shares on the same date are grouped together.

Analysis:

Crisis reporting are correlated with multiple shares by the same clusters of publishers on the same date. This indicates a sharp spike in crisis events



Insight #2.5:

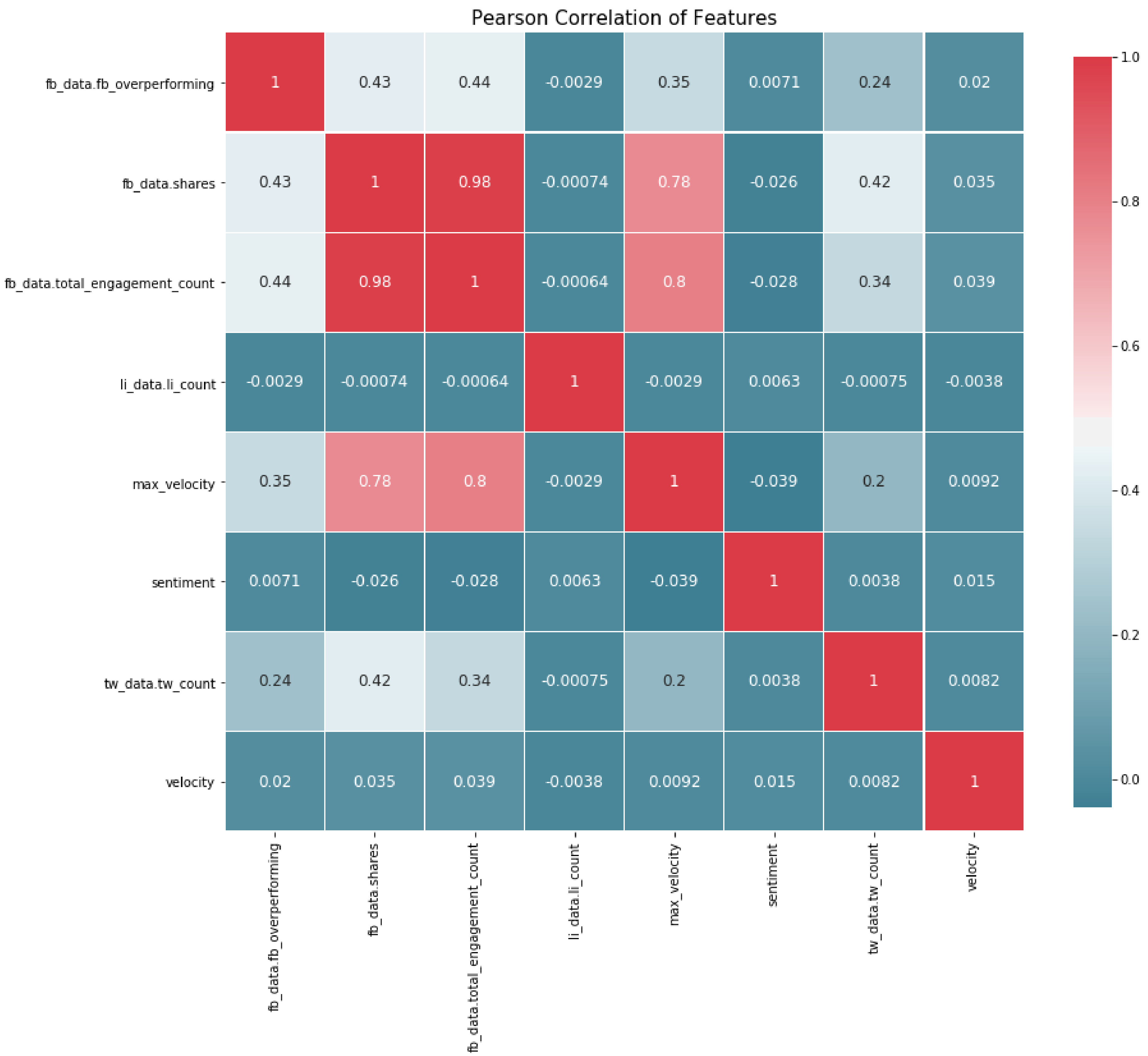
There are sharp spikes in social media engagement during crisis reporting events, that last on average 1-3 days after the initial reporting event.

Methodology:

We constructed a time-series analysis of Facebook total engagement count, Twitter Shares and Facebook shares over a two year time period between 2017 and 2018.

Analysis:

We found that there are huge spikes in social media engagements, such as tweets or Facebook shares and likes, during crisis reporting events, lasting usually 1-3 days. This is understandable, as social media are platforms frequently used by a large population of people and built on the basis of sharing. This shows us the “wildfire” nature of social media, whereby news or information about a crisis can spread and persist through this medium.



Pearson Correlation of Features

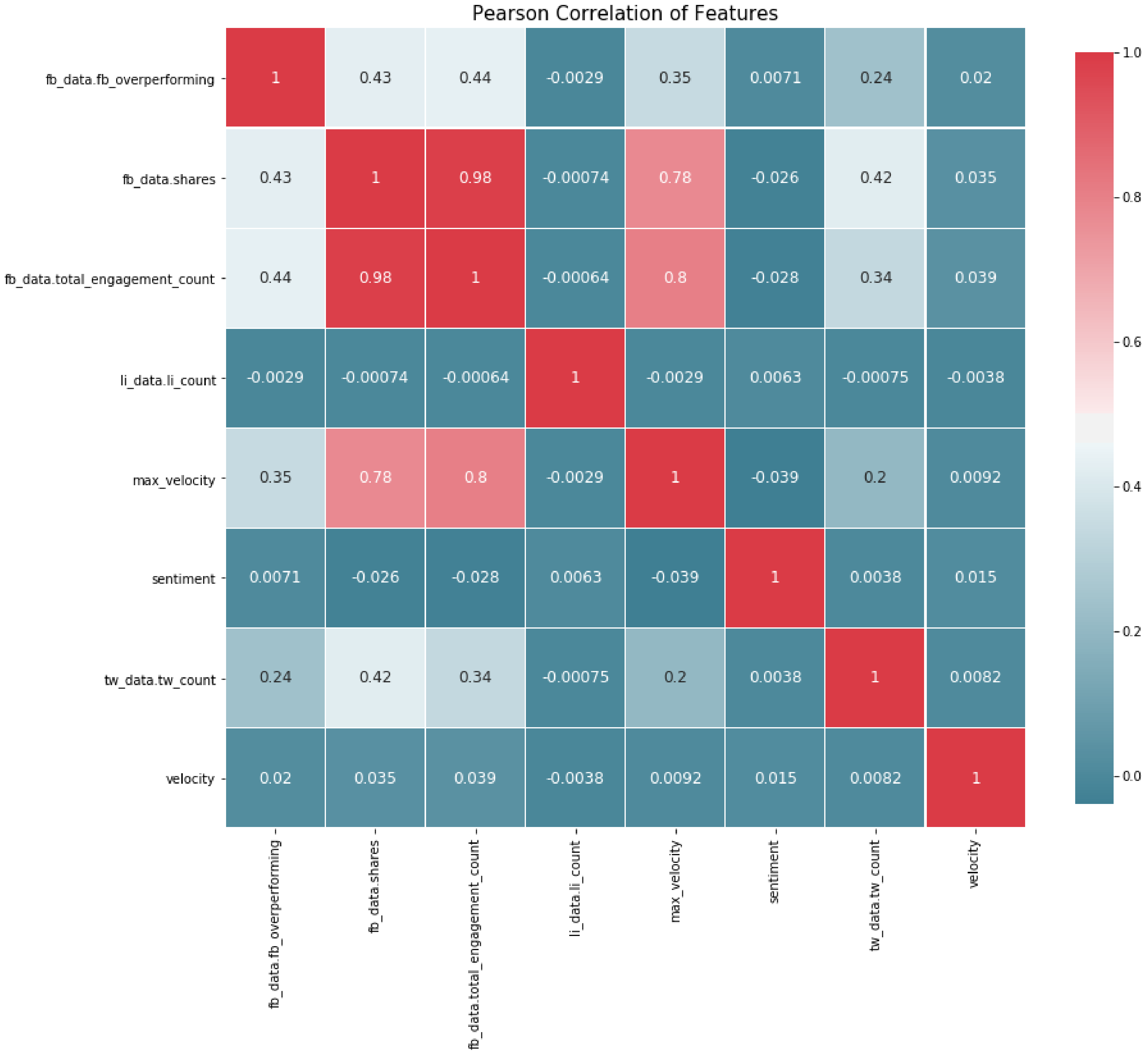
	fb_data.fb_overperforming	fb_data.shares	fb_data.total_engagement_count	li_data.li_count	max_velocity	sentiment	tw_data.tw_count	velocity
fb_data.fb_overperforming	1	0.43	0.44	-0.0029	0.35	0.0071	0.24	0.02
fb_data.shares	0.43	1	0.98	-0.00074	0.78	-0.026	0.42	0.035
fb_data.total_engagement_count	0.44	0.98	1	-0.00064	0.8	-0.028	0.34	0.039
li_data.li_count	-0.0029	-0.00074	-0.00064	1	-0.0029	0.0063	-0.00075	-0.0038
max_velocity	0.35	0.78	0.8	-0.0029	1	-0.039	0.2	0.0092
sentiment	0.0071	-0.026	-0.028	0.0063	-0.039	1	0.0038	0.015
tw_data.tw_count	0.24	0.42	0.34	-0.00075	0.2	0.0038	1	0.0082
velocity	0.02	0.035	0.039	-0.0038	0.0092	0.015	0.0082	1

Insight #3:

Publishers with high-performing articles in Facebook tend to enjoy higher shares in Twitter as well, and experience cross-platform popularity.

Methodology:

We visualized the correlation matrix between the numeric variables within the dataset to suss out if there were high correlations between specific possible predictors within the model. The more red the value appears, the higher the Pearson Correlation Coefficient.



	fb_data.fb_overperforming	fb_data.shares	fb_data.total_engagement_count	li_data.li_count	max_velocity	sentiment	tw_data.tw_count	velocity
fb_data.fb_overperforming	1	0.43	0.44	-0.0029	0.35	0.0071	0.24	0.02
fb_data.shares	0.43	1	0.98	-0.00074	0.78	-0.026	0.42	0.035
fb_data.total_engagement_count	0.44	0.98	1	-0.00064	0.8	-0.028	0.34	0.039
li_data.li_count	-0.0029	-0.00074	-0.00064	1	-0.0029	0.0063	-0.00075	-0.0038
max_velocity	0.35	0.78	0.8	-0.0029	1	-0.039	0.2	0.0092
sentiment	0.0071	-0.026	-0.028	0.0063	-0.039	1	0.0038	0.015
tw_data.tw_count	0.24	0.42	0.34	-0.00075	0.2	0.0038	1	0.0082
velocity	0.02	0.035	0.039	-0.0038	0.0092	0.015	0.0082	1

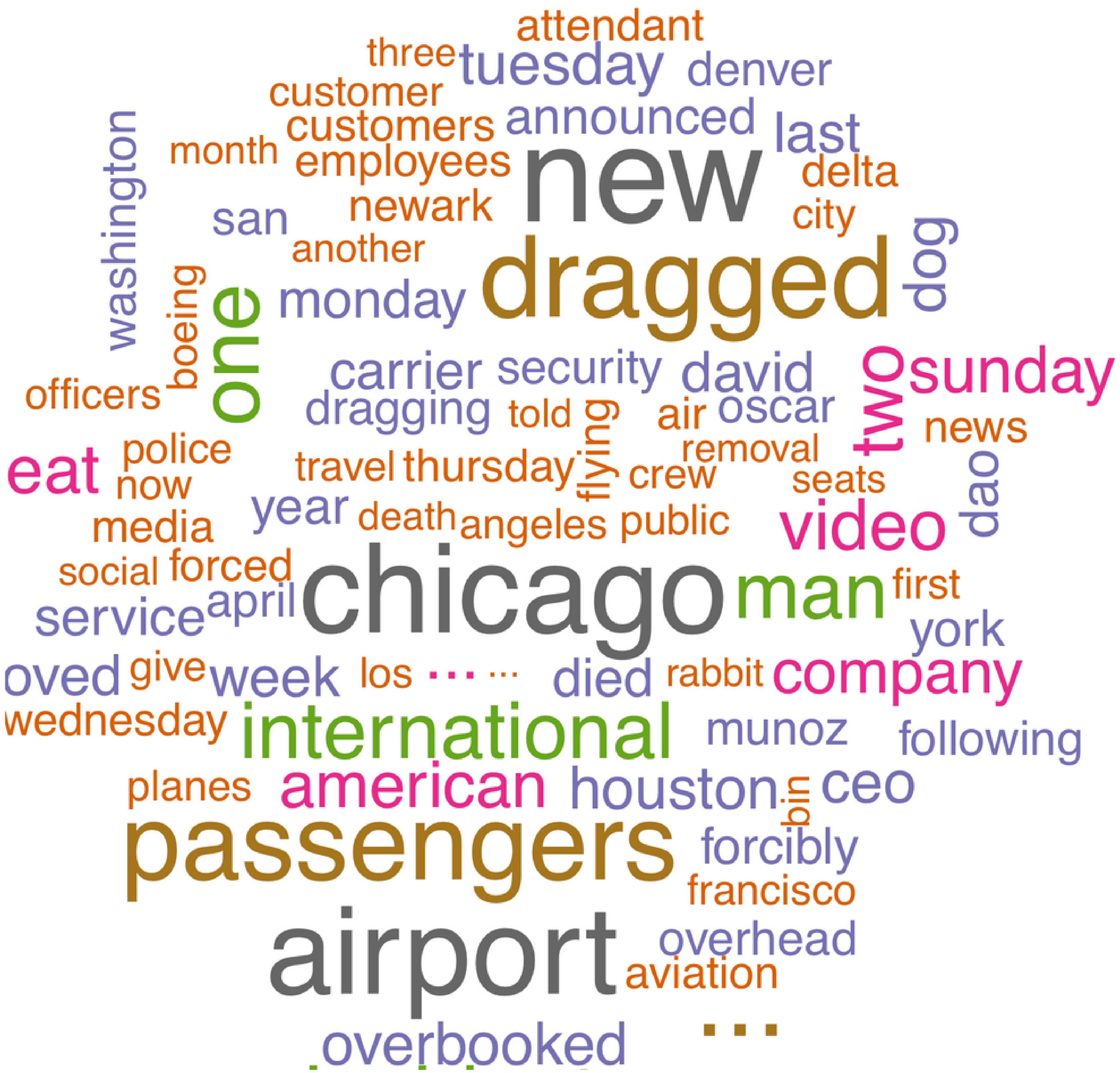
Insight #3 (cont):

Publishers with high-performing articles in Facebook tend to enjoy higher shares in Twitter as well, and experience cross-platform popularity.

Analysis:

We found various interesting correlations between some of the data given, such as the strong correlation between Facebook shares and tweets. The most interesting result of the correlation matrix is the generally weak correlation between sentiment and the other categories.

Since there is a weak correlation between sentiment and other categories, we decided to feature engineer sentiment metrics we can use to examine the relationship between 'sentiment' and the social media influence of an article.



Insight #4:

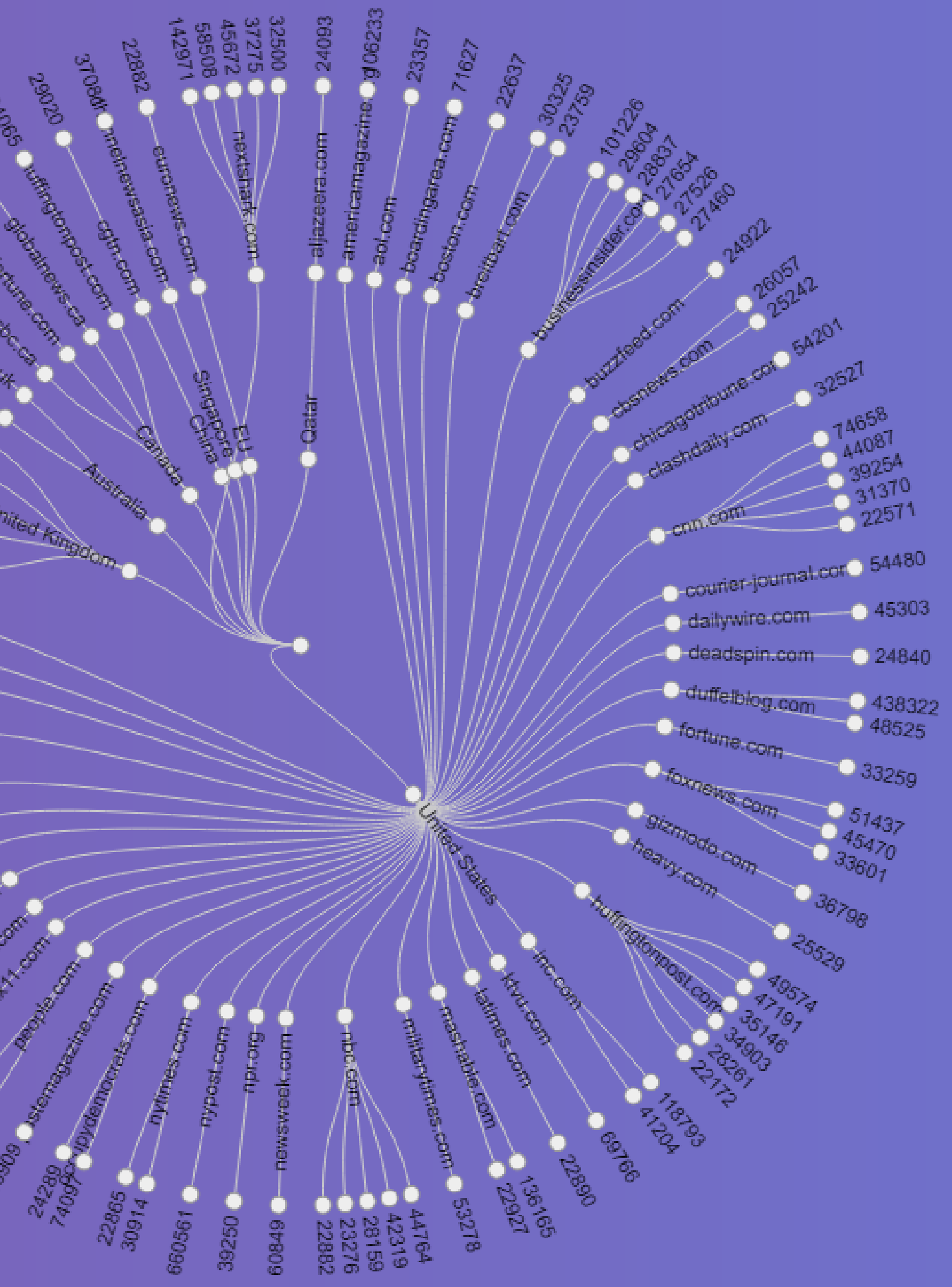
The most frequent words appear to relate to the dragged passenger incidents and the death of pets.

Methodology:

This word cloud is created using the excerpts by cleaning out the stop words and mining for meaningful words by removing frequent industry words e.g. flights, planes, United, etc with the size corresponds to its frequency of appearance.

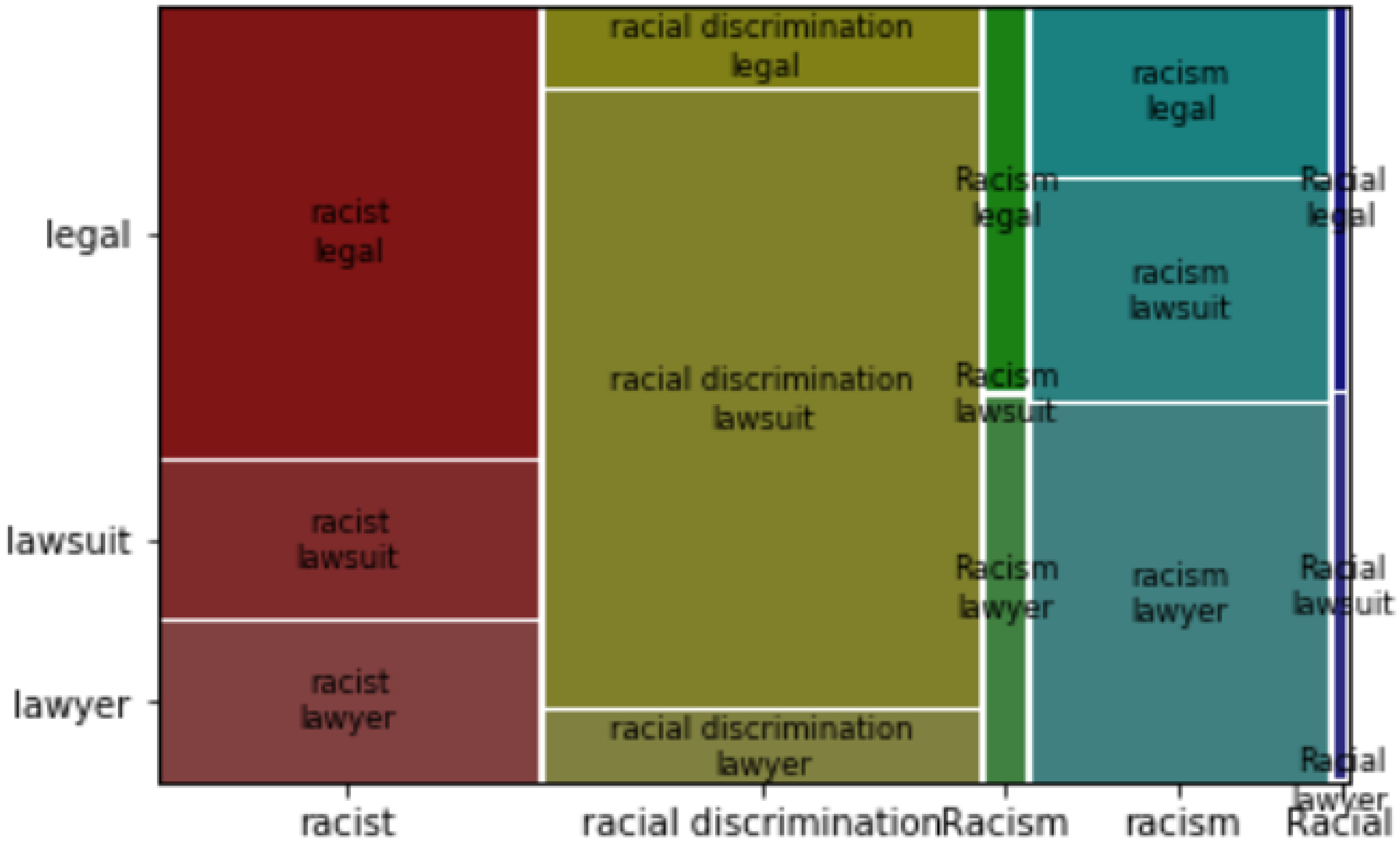
Analysis:

We found that the most frequent mentions of United Airlines are with the incident of the dragged passenger on the overbooked flight. This is due to the nature of the incident dealing with force and, thus, reflecting a level of customer discomfort as it could happen to them. Furthermore, the pets related incidents appeared as a strong second as it is also a personal matter that negatively affects client satisfaction with the airlines.



Section 2.2: Feature Engineering by Topic Identification

Section 2.3: Applying New Features and Modelling

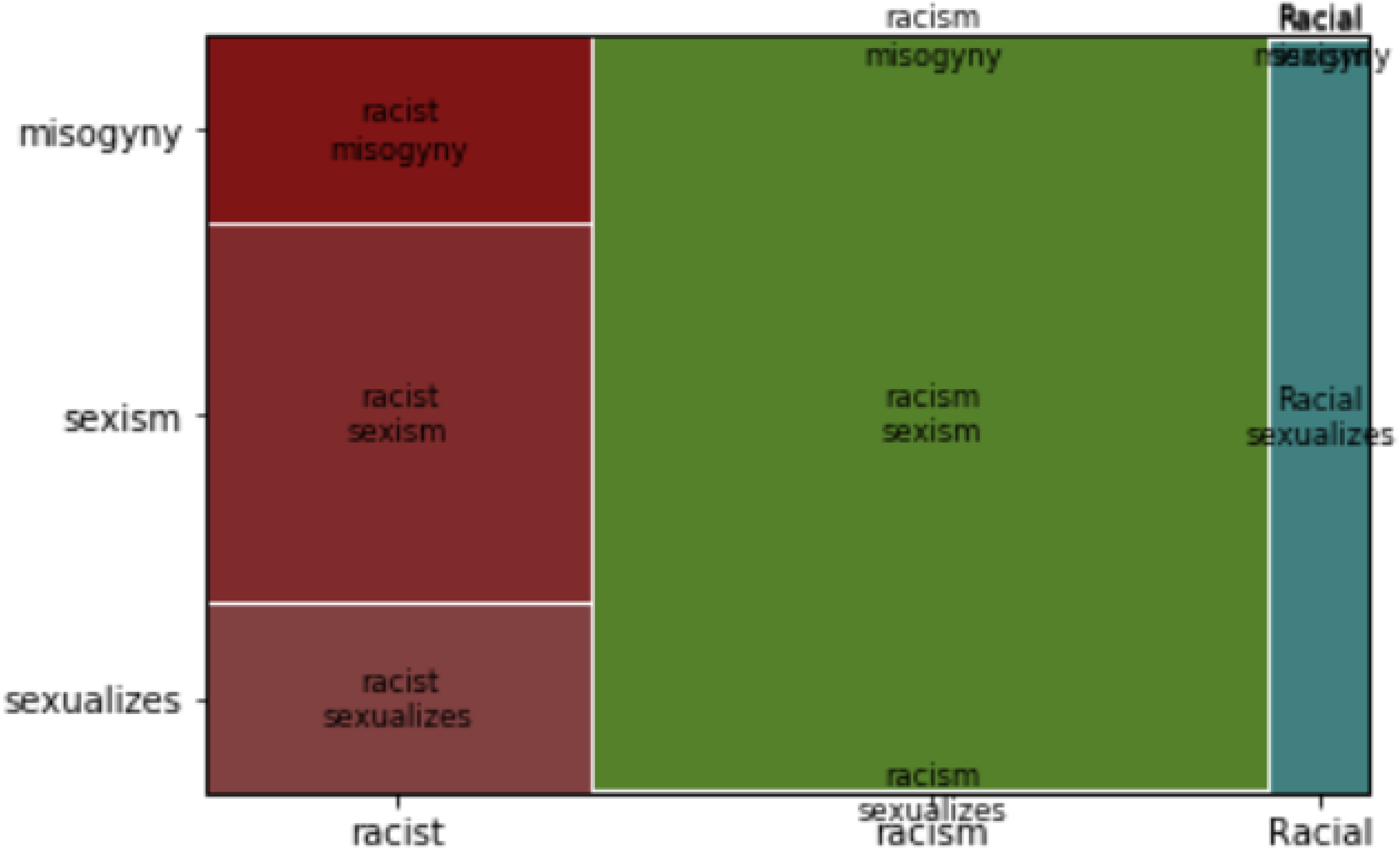


Feature #1:

Sentiment Analysis by Identity Categories

When segmenting publishers by mentions of identity categories, mentions of racism is correlated with both reports of sexism and the reports of lawsuits by publishers.

Accusations of racist treatment and sexist treatment are indicators of poor customer experience for United Airlines, and the prevalence of lawsuits also signal as a metric towards the long term costs of the



Feature #1 (cont):

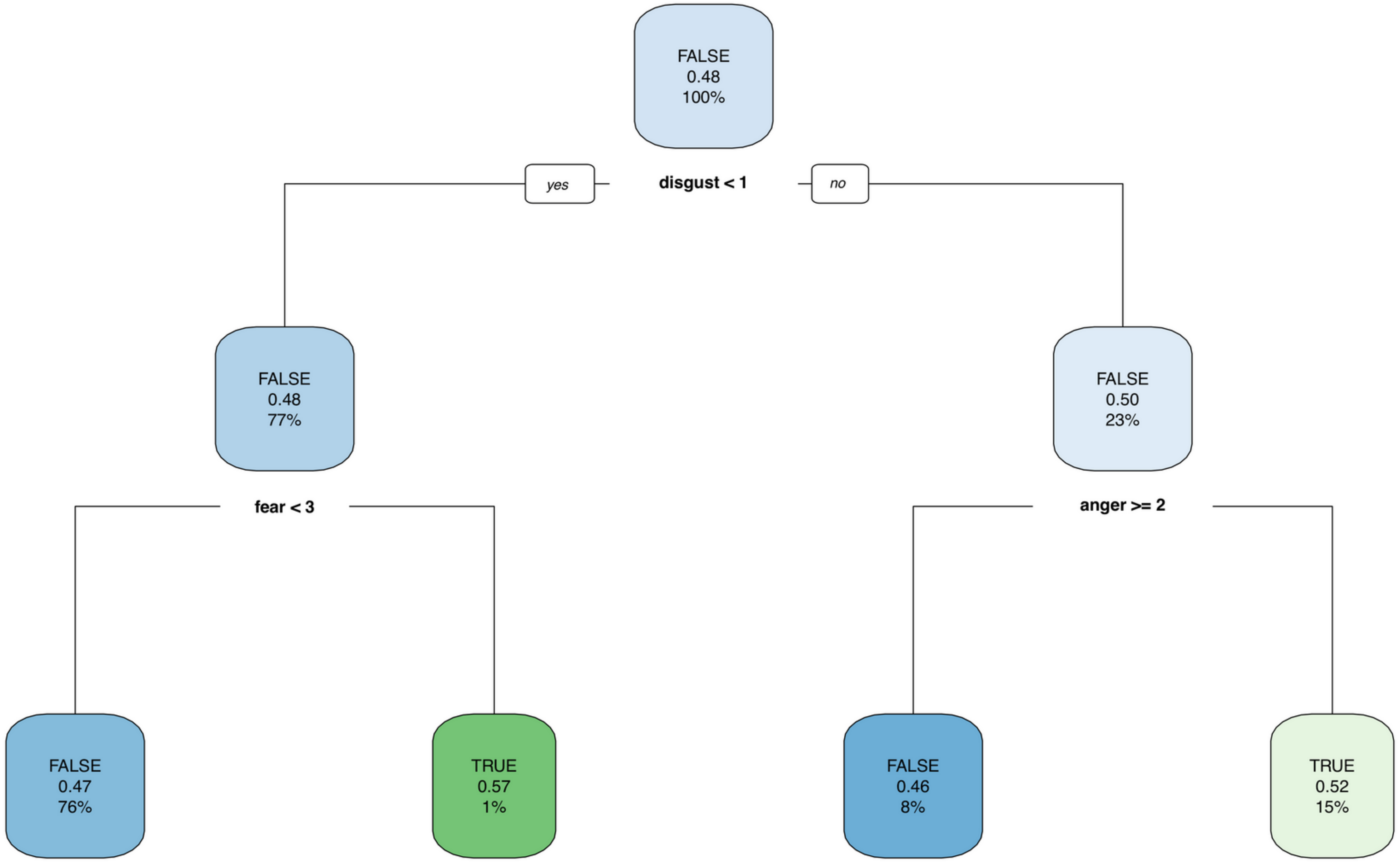
Sentiment Analysis by Identity Categories

Methodology:

Using the Mosaic analysis tool in Python, we group the prevalence that two categories (x,y) occur together in the same article. The ones with a higher frequency by ratio have a higher proportion within the mosaic tile.

Analysis:

We found that the occurrence of racism and sexism in articles appear very frequently, indicating a high usage of these terms across the publishers. Accusations of racist treatment and sexist treatment are indicators of poor customer experience for United Airlines, and the prevalence of lawsuits also signal as a metric towards the long term costs for the airlines and the negative publicity such lawsuits will garner. The frequent usage of such terms will hence paint United Airlines quite negatively in the news and create a perception of poor customer experience in the masses.



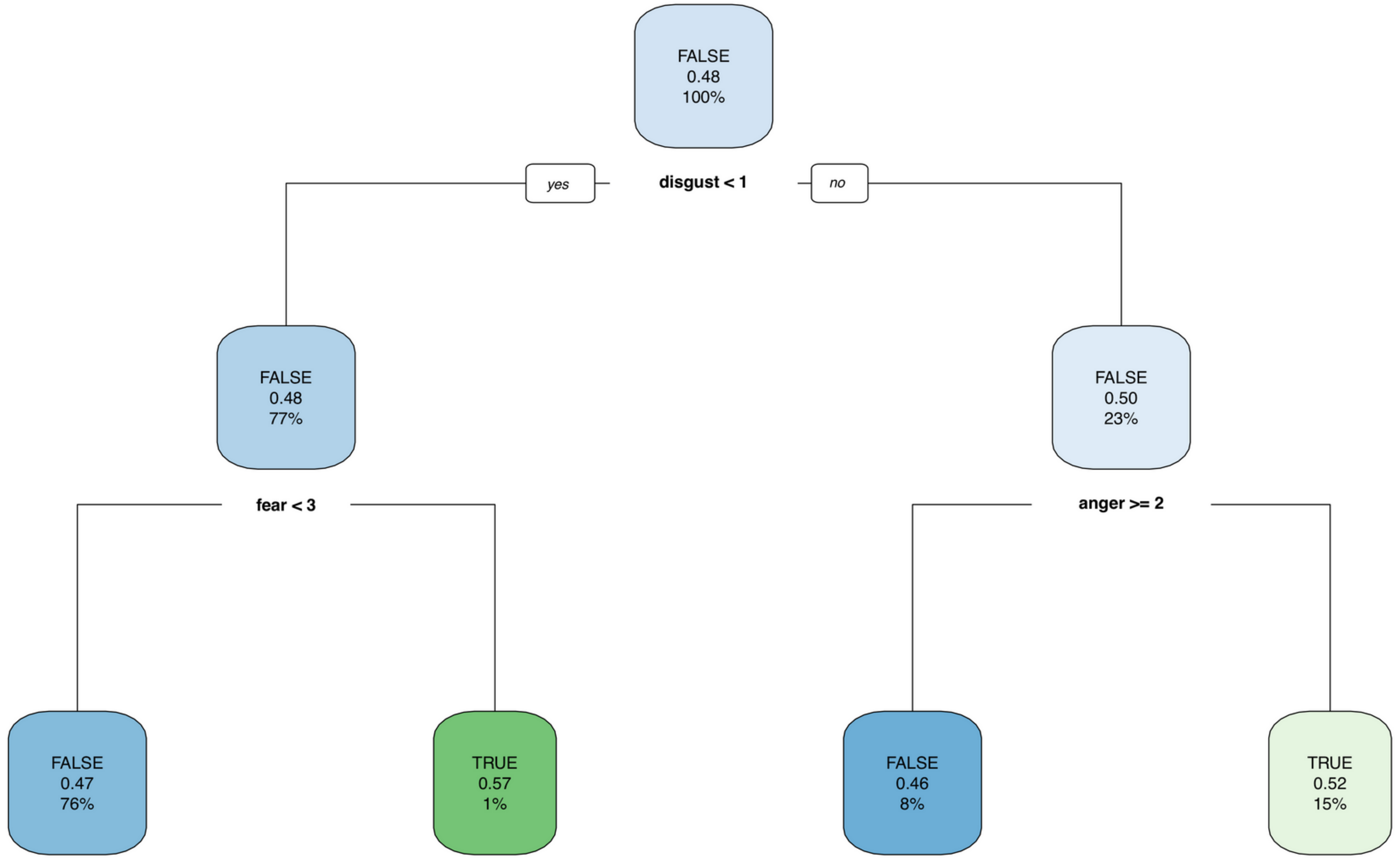
Feature #2:

Classifying Sentiment Analysis by Emotion

Disgust and fear seem to be the sentiments that influence the maximum velocity of the articles whilst anger is almost negligible.

Methodology:

This decision tree model is a classification model to determine whether an article will meet our minimum max_velocity of 50(the third quartile of max_velocity values). It is made using the articles' excerpts sentiment analysis cross-validated using four k-fold to minimize our error margin. The sentiments are generated using the NRC Word-Emotion Association Lexicon (aka EmoLex) (Saif, n.d.) which associates the words with their emotional meaning.



Feature #2 (cont):

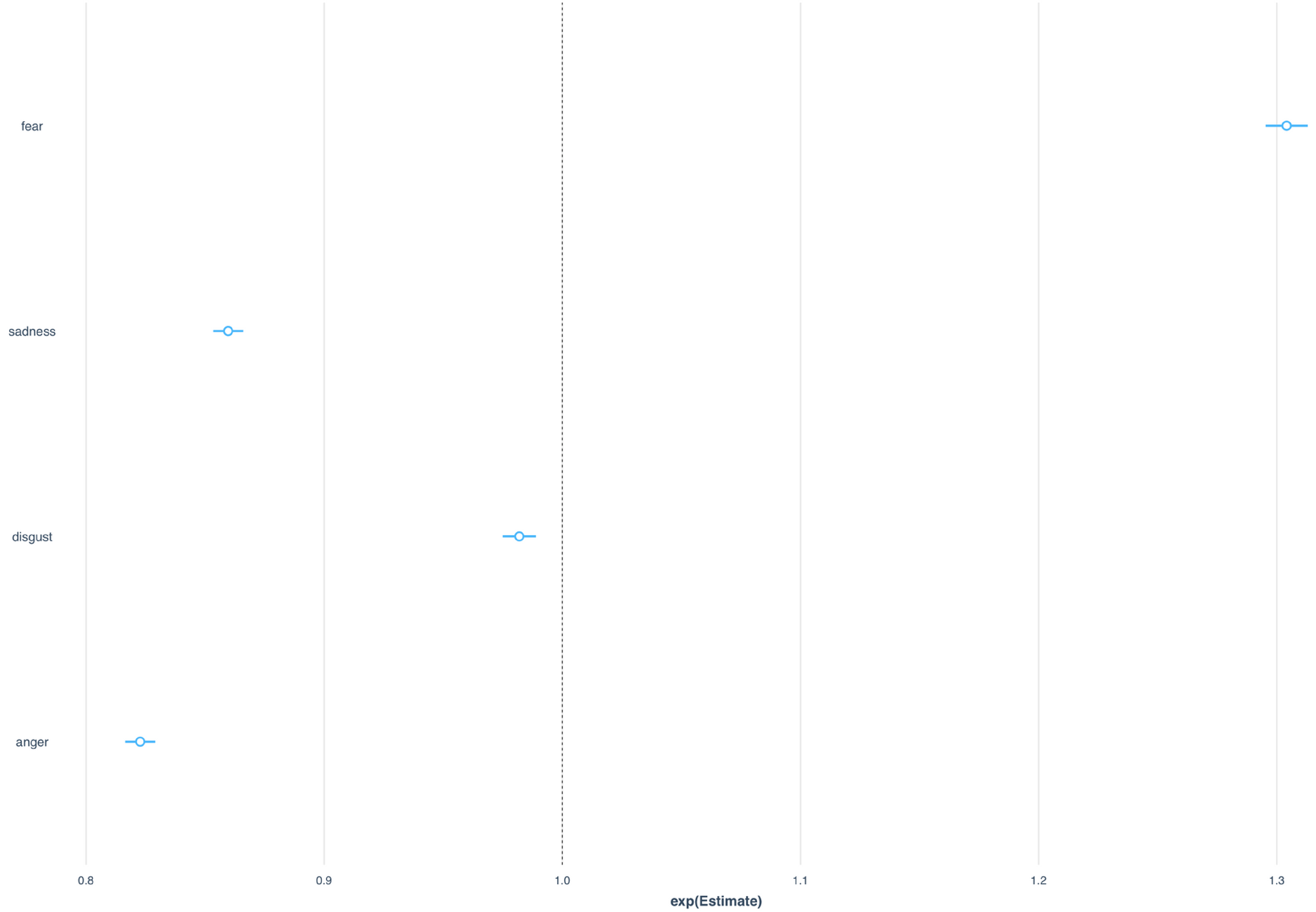
Classifying Sentiment Analysis by Emotion

Analysis:

We found that disgust is the main segmenting factor that determines whether the news will meet the minimum max_velocity or not as most of the articles that spread appeared on the higher end of the disgust spectrum. Although disgust may not be as high, if the article were to score higher on the fear spectrum, it will spread too. However, the fact that the mentioned conditional probability only accounts for 1% of the articles means that this should not be focused on. We will explore the different impacts between the two dominant features: fear and disgust in the next model.

Feature #3:

Correlating Sentiment Emotion and Article Velocity



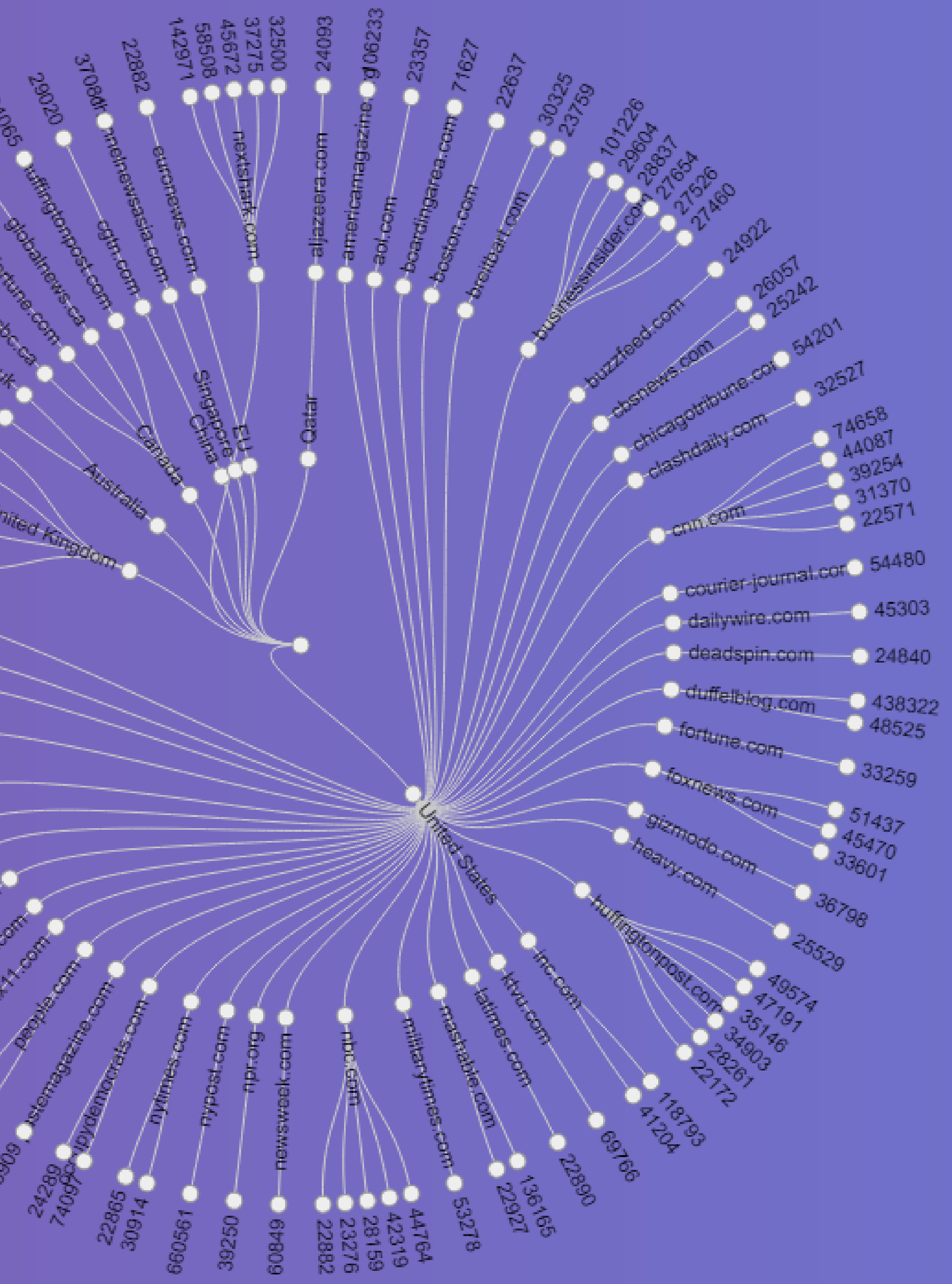
Feature #3:

Correlating Sentiment Emotion and Article Velocity

Methodology:
This Poisson model is a generalised linear model (GLM) that estimates the percentage increase in categorical count data by a subset of predictors. Essentially, we examined whether the existence of the emotion of fear, sadness, emotion and disgust correlate to the extent of velocity. A unit unit increase in velocity corresponds to an exp(Estimate) increase in the percentage of the existence of the specified emotional category. For example, a one unit increase in velocity results in a 266% increase in prevalence of the emotion fear.

It is made using a matrix of sentiment mined from the excerpts of the articles and against the max_velocity of the articles. Since we could not do a ballpark estimation with the previous classification, this model aims to be more precise in our prediction of the max_velocity of the article based on our generated sentiment.

Analysis:
The fastest-shared articles have a higher prevalence of fear compared to the emotions. In contrast to the decision tree model which states that disgust and fear perform relatively the same, we find that disgust has a lower estimated prevalence than fear.



Section 3: Recommendations

Recommendations

We recommend the following metrics to be utilised when monitoring digital media crisis patterns:

Monitoring combined social media engagement:
We found that totalling up the total Facebook/Twitter/LinkedIn shares was necessary in evaluating articles by influence. Since articles that perform well on Facebook also tend to perform well in Twitter, creating a custom social media engagement metric is necessary.

Monitoring the velocity of social media engagement:
We found that the most useful metric in classifying the training dataset was the speed through which the shares perform. Using a velocity threshold of the **third quartile** provided meaningful results in demonstrating the different emotions

Monitoring Sentiment of Articles by Emotion:
We found that crisis events have a higher prevalence of negative emotions. By using the NRC Word Emotion Lexicon algorithm, and pruning it by **correlating the sentiment** to the existing categorical variable of negative sentiment (-1), one can create a decision trees, and eventually scale up to an ensemble Random Forest model that can evaluate the

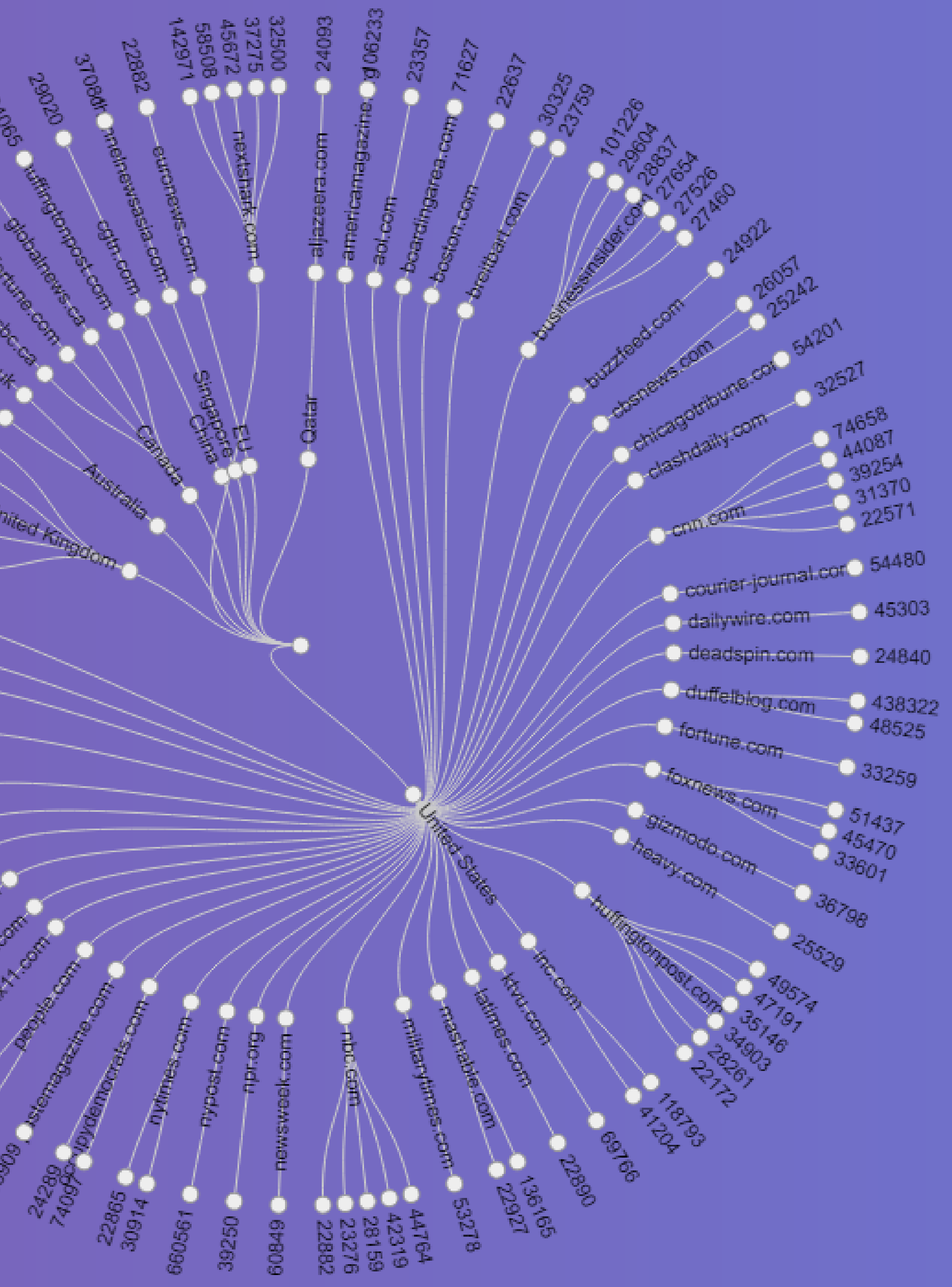
Recommendations

We recommend the following metrics to be utilised when monitoring digital media crisis patterns:

Monitoring Sentiment of Articles by Identity-Category:
A large portion of crisis-related events surround customer personas that perceive a sub-optimal customer experience related to their identity, either race/gender/sexuality. Predicting whether an event has an identity category can also predict **future** costs of the crisis monitoring, i.e. the possibility of lawsuits.

Further Analytical Questions

1. **What is the geospatial variation of the high-frequency Facebook shares?**
    a. **Necessary Data Sources:** GeoJSON data or SHAPE files that correspond to the unique Twitter shares or Facebook shares. By doing so, one can map and trace the geographic clusters of which individuals have shared negative sentiment.
2. **Who are the high-performing Twitter influencers?**
    a. Constructing a network graph that links the source.publishers (as the root) to the shares, and weights it by the amount of times each individual consistently shares from the publisher. This shows the ‘loyal’ sharers of each publisher, and therefore how many people effectively spread the crisis reporting.



Thank you so much!

References

Saif, M. (n.d.). NRC Word-Emotion Association Lexicon (aka EmoLex). Retrieved from https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm#targetText=NRC%20Word-Emotion%20Association%20Lexicon,were%20manually%20done%20by%20crowdsourcing

GitHub Link

https://github.com/sewen770/EPIC-Data-Challenge