Social Media Data Analysis

Advisor : Amarnath Gupta

Amir Shirkhani, Mohammed Tawashi, Hanu Pathuri, Naveed Mohammed, Vamsi Namuduri

Introduction

The project focuses on studying how Twitter images impact the narrative of hashtags.

Hypothesis: Media files (Images/videos) are being used to influence the narratives of hashtags.

Objective: The project aims at finding specific patterns in tweets where media files (images) are used to change the narrative of the corresponding hashtag and co-occurred hashtags. To study the importance of images in the Tweets by modeling aggregated text of hashtags with and without images.

Targeted Audience:

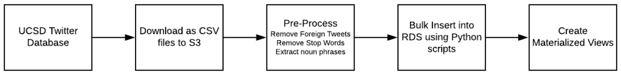
- ❖ Election organizer as part of a campaign can study and detect endorsing and opposing trends and act by countermeasures using similar techniques
- ❖ Social Media platform and specially Twitter itself can detect patterns and potentially restrict the behavior.
- ❖ Journalists can report to general public on how a potential small group of influencers can sway a narrative and push various agendas.

Data Sources

UCSD Twitter database is the main data source for the project. Below are the main entities that are downloaded from this database.

- ➢ Users
- ➢ Tweets
- ➢ Hashtags
- ➢ Media URLs

The data collected is processed and inserted into RDS Postgres instance. Materialized Views are created on top of the data.

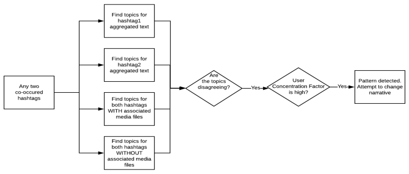

UCSD Twitter Database
Download as CSV files to S3
Pre-Process
Remove Foreign Tweets
Remove Stop Words
Extract noun phrases
Bulk Insert into RDS using Python scripts
Create Materialized Views

Methodology

Hashtag Co-occurrence: This is pair of hashtags that were used together in a Tweet.

User Concentration Factor: Is the number of tweets per User.

Pattern Recognition:


Any two co-occured hashtags
Find topics for hashtag1 aggregated text
Find topics for hashtag2 aggregated text
Find topics for both hashtags WITH associated media files
Find topics for both hashtags WITHOUT associated media files
Are the topics disagreeing?
Yes
User Concentration Factor is high?
Yes
Pattern detected. Attempt to change narrative

Community recognition based on user mentions:

NFM Model with NetworkX sub-graph traversal parameter K-Core = [2, 6]
(K is the in-degree + out-degree)

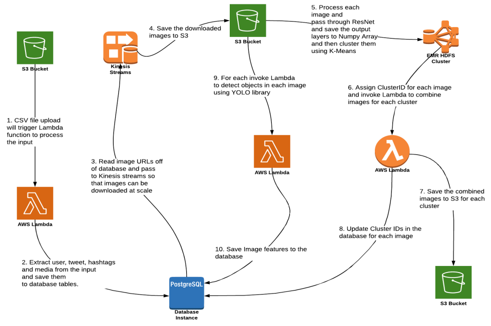
No of Topics = 10

No of words per topic = 20.

At K core = 6, the model converges and provides optimal communities based on the topics.

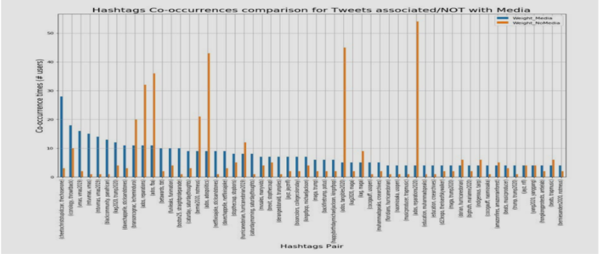
Data Pipeline

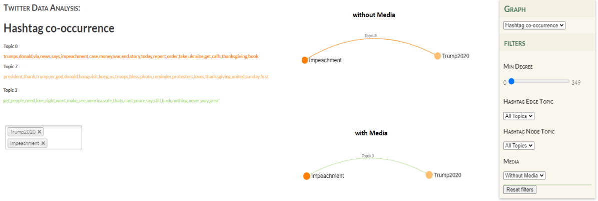
Solution Architecture


S3 Bucket
1. CSV file upload will trigger Lambda function to process the input
AWS Lambda
2. Extract user, tweet, hashtags and media from the input and save them to database tables.
PostgreSQL
Database Instance
3. Read image URLs off of database and pass to Kinesis streams so that images can be downloaded at scale
Kinesis Streams
4. Save the downloaded images to S3
S3 Bucket
5. Process each image and pass through ResNet and save the output layers to Numpy Array and then cluster them using K-Means
EMR HDFS Cluster
6. Assign ClusterID for each image and invoke Lambda to combine images for each cluster
AWS Lambda
7. Save the combined images to S3 for each cluster
S3 Bucket
8. Update Cluster IDs in the database for each image
9. For each invoke Lambda to detect objects in each image using YOLO library
AWS Lambda
10. Save Image features to the database

Results

Hashtag Co-occurrence with and without media





Twitter Data Analysis:

Hashtag co-occurrence

Topic 8

trumps,donald,via,news,says,impeachment,case,money,war,end,story,today,report,order,fake,ukraine,get,calls,thanksgiving,book

Topic 7

president,thank,trump,mr,god,donald,hong,visit,kong,us,troops,bless,photo,reminder,protesters,loves,thanksgiving,united,sunday,first

Topic 3

get,people,need,love,right,want,make,see,america,vote,thats,cant,youre,say,still,back,nothing,never,way,great

Trump2020 ✕
Impeachment ✕

without Media

with Media

Graph: Hashtag co-occurrence

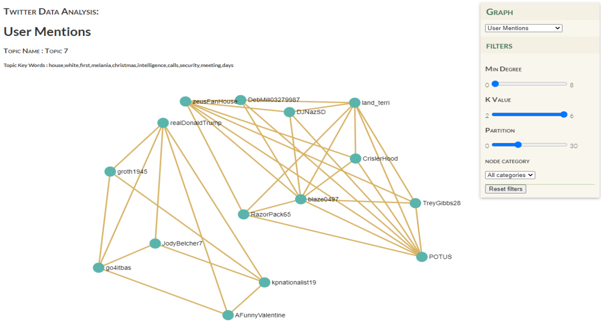
Filters

Min Degree: 0 – 349

Hashtag Edge Topic: All Topics

Hashtag Node Topic: All Topics

Media: Without Media



Twitter Data Analysis:

User Mentions

Topic Name : Topic 7

Topic Key Words : house,white,first,melania,christmas,intelligence,calls,security,meeting,days

Graph: User Mentions

Filters

Min Degree: 0 – 8

K Value: 2 – 6

Partition: 0 – 30

Node category: All categories

Conclusion

Conclusions that came out of the data analysis from Twitter.

- ➢ Text extraction from media files associated with tweets are, in most cases, supplementing the narrative of the tweet's text.
- ➢ Media files are being used as a powerful tool to contaminate the original narrative of single/co-occurred hashtag pairs.
- ➢ Object detection from images did not result in any additional insights.

Future Work and Acknowledgements

Future Work:

The model we developed can be extended on a few different ideas:

- ○ Tweets related to Trump dominated the dataset. We can do the analysis by filtering those tweets.
- ○ Include videos and GIFs/memes in the analysis.
- ○ Expand the dataset to include wide range of topics instead of just political tweets.

Acknowledgements:

We would like to thank:

- ✓ Our advisor Prof. Amarnath Gupta
- ✓ All the DSE professors and teaching assistants.
- ✓ Staff at Supercomputer Center (SDSC).