EPIC DATA CHALLENGE 2019

COMPETITION BRIEF

Detect Digital News Patterns

BACKGROUND

When a digital crisis occurs, every second counts. Brands must act swiftly - from drafting an official statement to managing the grievances expressed by their customers and wider community.

For example: On April 9th, 2017, Dr. David Dao refused to give up his seat on the over-booked United Airlines Flight 3411. He was then forcibly removed from the plane, whilst receiving severe injuries in the process. The video was widely circulated on social media, and the incident was reported in major news outlets – the story went viral and led to a huge PR crisis for United Airlines.

In a hypothetical scenario, it is critical for both United and Edelman to understand the digital crisis landscape - the kinds of content that are shared, the individuals and publications doing the sharing, and how this process of information spread occurs. These insights will go on to drive inform crisis response and subsequent reputation management strategies.

PROBLEM STATEMENT

Phase 1 - Pattern Detection

(Teams will choose **only one topic** between A and B; **C is compulsory**)

A. What content is being shared?	B. Who is sharing the content?
• Are there any meaningful themes of the articles? What are the common characteristics in each group? • Within news articles, what are issues that United needs to focus on with their media monitoring? And why? • Can you identify any recurring issues that United Airlines face?	• 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?

C. How is content being shared?
• Are there any syndication patterns across publishers, e.g. how articles of same issue are being shared by different publishers? • Are there any general propagation patterns, e.g. when the engagements start to go up, where and when an article gets picked up by social media, etc.? **If choose Topic A in Phase 1, answer** • Does the syndication / propagation pattern change if it is a recurring issue for United Airlines? **If choose Topic B in Phase 1, answer** • Does the syndication / propagation pattern vary among different publishers and influencers segments that you develop?

Phase 2 - Recommendations

(Compulsory)

- Based on your findings, what recommendations would you give for digital crisis monitoring?
- What further analytical questions would you develop to understand digital crises and what additional data sources or analytical approach would you suggest for improvement of the analysis?

DATA

You will be working on three datasets provided in JSON format in this Data Challenge. External public datasets are allowed in the challenge with compliance to Terms & Conditions.

1. **Articles Dataset**

   File Name: articles.json

   Description: This dataset includes articles mentioned United Airlines published between January 1st, 2017 and April 30th, 2019 (both dates inclusive). In this dataset, you can find data points about article profile and the engagements they gathered from social media.

   Data Dictionary: Fields with * are collections of data points

Field	Name	Definition
article_id	Article IDs	An alphanumeric string that uniquely identifies a piece of content.
authors	Authors	A list of the author(s) of the content, when available.
contents	Article Content	The full content of the article.
excerpt	Excerpt	A short extract of the content, when available.
fb_data*	Facebook Data	This is a collection of different data points that relate to Facebook.
has_video	Video Flag	A flag to show whether we detected video in the content.
headline	Headline	The heading of the content.
image_link	Image Link	If we have detected an important image for the content, this is the link to it.
keywords	Keywords	A list of keywords sometimes provided by the publisher. These are provided verbatim, and so may not be consistent from publisher to publisher.
li_data*	LinkedIn Data	This is a collection of different data points that relate to LinkedIn.
link	Link	The URL of the content.
max_velocity	Maximum Velocity	The maximum social velocity of a piece of content so far.

pi_data*	Pinterest Data	This is a collection of different data points that relate to Pinterest.
publication_timestamp	Publication Timestamp	The date and time that the content was published. It is a UNIX timestamp in milliseconds.
sentiment	Article Sentiment	This indicates the sentiment analysis result from the data provider. -1 means negative sentiment, 0 means neutral and 1 means positive.
source*	Source	This is a collection of different data points that relate to the publisher.
topics	Topics (sometimes referred to as Categories)	The theme associated with the content's source, based on Topics.
tw_data*	Twitter Data	This is a collection of different data points that relate to Twitter.
velocity	Current Velocity	The current social velocity of a piece of content.
fb_data collection		
total_engagement_count	Facebook Total Engagement Count	The total number of engagements (reactions, shares, comments) a piece of content has had on Facebook.
likes	Facebook Likes	The number of likes of posts relating to a piece of content.
shares	Facebook Shares	The number of times posts relating to a piece of content were shared on Facebook.
comments	Facebook Comments	The number of comments made on posts relating to a piece of content.
li_data collection		
li_count	LinkedIn Count	The number of engagements a piece of content has had on LinkedIn. See Known issues for updates on the availability of LinkedIn data.
pi_data collection		
pi_count	Pinterest Count	The number of pins a piece of content has had on Pinterest.
source collection		

country	Country	The human-readable country name for the publisher.
country_code	Country Code	The ISO 3166 country code that represents the country of the publisher.
link	Link	The URL of the content.
publisher	Publisher	The name, usually the domain, of the publisher of the content.
tw_data collection		
tw_count	Twitter Count	The number of Influencer Shares a piece of content has had on Twitter.

2. **Twitter Influencers Dataset**

   File Name: twitter_influencers.json

   Description: This dataset includes Twitter users who shared articles mentioned United Airlines in the Articles Dataset. You can find data points about the Twitter users and the corresponding tweets and articles they shared. Only Twitter verified users are included in this dataset.

   Data Dictionary: Fields with * are collections of data points

Field	Name	Definition
followers_count	Facebook or Twitter Followers Count	The number of people following a given Twitter account.
following_count	Twitter Following Count	The number of other Twitter accounts followed by a given Twitter account.
likes_count	Twitter Likes Count	The number of posts that this user has liked.
max_retweet_value	Maximum Retweets	The number of retweets on this user's most influenced article.
social_referrals*	Social Referral	This is a collection of data points that relate to the articles shared by this Twitter user.
statuses_count	Twitter Posts Count	The number of Tweets the user has made.
twitter_handle	Twitter Handle	The account name of a Twitter user.
twitter_id	Twitter User ID	A numeric string that uniquely identifies a Twitter user.
social_referrals collection		
article_id	Article ID	An alphanumeric string that uniquely identifies a piece of content.
article_url	Article URL	The URL of the shared content.

created_at	Created At	The date and time that a Twitter post was published. A UNIX timestamp in milliseconds.
description	Description	A short description of the article shared in Twitter post.
favorite_count	Twitter Likes Count	The number of likes that the specified post has received.
rt_count	Twitter Retweets Count	The number of retweets that the specified post has received.
twitter_id	Twitter User ID	A numeric string that uniquely identifies a Twitter user.
twitter_url	Twitter URL	The URL of the Twitter post.

3. **Tweets Dataset**

   File Name: tweets.json

   Description: This dataset includes tweets in which articles mentioned United Airlines in the Articles Dataset were shared. You can find not only the Twitter users and their tweets covered in Twitter Influencers dataset, but also the replies and retweets on those tweets. You can find data points about both the users and the tweets in this dataset.

   Data Dictionary: This dataset is retrieved from Twitter API. Data dictionaries are available on Twitter, Tweet Object and User Object are potentially most useful for you.

4. **Notes**

   - For Twitter user's information, there are data points overlapping between Twitter Influencers Dataset and Tweets Dataset. In the case of discrepancy, you might want to use the information in Tweets Dataset given it was retrieved more recently.
   - Not all articles from Article Dataset can be found in Twitter Influencer Dataset because some of them are not shared in social media; Not all Tweets from Twitter Influencer Dataset can be found in Tweets Dataset because some of them are deleted by the users by the time we collected them.
   - Recommendations should be substantiated by your data insights.
   - If you have any questions regarding the dataset, you can email us at EPICDataChallenge@edelman.com. We will get back to you within 24 hours. For commonly asked questions, we will consolidate and share to all the teams for clarity through email update.

TIPS & SUGGESTIONS

- Teams are **NOT** expected to focus on analyzing the Flight 3411 incident specifically. Take United Airlines as a hypothetical client, use data to help them understand the overall digital news landscape.
- Clearly mention the problem statements that you are tackling in your deliverables.
- There are many data points shared. Take time to understand thoroughly the relationships among them and selectively use relevant data for your analysis.
- Learning is as important as the outcome. Do share with us your learnings, key takeaways, and any further suggestions to solve the problem.

HOSTED BY

Edelman Predictive Intelligence Centre

PARTNERS

DataKind

DSR

GENERAL ASSEMBLY

SGINNOVATE