Variables

API generated Variables

Rank(Ordinal): The variable giving the current ranking of the article among top 25

Active_users(Int): Number of active users on the r/news subreddit at the sampling time.

Author(Str): The name of the Redditor Who posted the article

Num_Comments(Int): Number of Comments at the time of sampling

Created(DateTime): The date and time the article was posted on Reddit in UTC

Crossposts(Int): The number of times the article was shared across different subreddits and on other social media.

Domain(Nominal): Name of the news organization from the where the article was taken

Id(Nominal): A unique identifier of the news article

Score(Int): The total number of points for the article after taking the difference between upvotes and downvotes (Upvotes-Downvotes)

Title(Str): The title of the news post

Upvote_Ratio(Float): A decimal number indicating the % of upvotes for the article among all the votes.

Structural Variables

active_users_flag: A flag variable indicating the number of active users during the two sampling times. A value of 1 refers to the number of active users at night and 0 refers to number of active users in the morning at the time of sampling.

Images(Int): The number of images in the news posts

(Now its added under Image due to sparsity) AnimatedGifs (Int) : The number of animated gifs in the post

(Now its added under Image due to sparsity)Non_AnimatedGifs(Int): The number of non-animated gifs in the post

Videos(Int): The number of videos in the post

Video1Len(Time): The length of the 1st video in seconds

Video2Len(Time): The length of the 2nd video in seconds

Video3Len(Time): The length of the 3rd video in seconds

Video4Len(Time): The length of the 4th video in seconds

AudiLen(Time): Length of the audio tape in seconds if present

EmbeddedDocuments(Int): Number of documents embedded in the posts(Eg. Pdfs, Memos etc)

DocLen(Ordinal): Length of the news article (Word Count < 500 is short, Between 500-1000 words is medium and greater than 1000 words is denoted as Long)

SocialMedia Variables

Tweets(Int): Number of tweets embedded in the post

FB(Int): Number of Facebook posts embedded in the post

Instagram(Int): Number of instagram Pictures embedded in the post

Meta Variables

Topic1(Str): The Main topic of the article (Taken mainly from the news websites tags)

Topic2(Str): Additional Topic Tag

Topic3(Str): Additional Topic Tag

Region(Nominal): The region/state where the event is associated with

Source Type(Nominal): Type of news organization

OriginalDate(Date): The date on which the article was posted on the news website

OriginalTime(Time): The time at which the article was posted on the news website

Report(Int): If the article is a report then the number of pages of that report is mentioned.

Article Strength Variables

Repeated(Ordinal): Variable indicating the number of times a news post was present in the top 25. A value of 1 indicates that it was the first instance of the article and a value of 2 indicates the 2nd time and so on.The higher the value the longer the life of the article in the top 25.A value of 0 indicates that the article appeared only once in the top 25 list.

Linguistic Variables

LIWC features: http://liwc.wpengine.com/compare-dictionaries/

Featured Variables

Social_Media: Variable of combined values of the Tweets, FB and Instagram Variables

Video_Len:The sum of all the video length variables in seconds

Upvotes: No of upvotes on the article as calculated from the formula presented below

Downvotes: No of downvotes on the article as calculated from the formula presented below

Date: Date the article was posted to Reddit obtained after formatting the created variable.

Time: Time the article was posted to Reddit in UTC

Day_of_week: Indicates the day of the week.

Type_of_Day: Indicates whether its a weekday or weekend

Date_EST: Date the article was uploaded to reddit standardized to EST using the UTC Timestamp obtained from the API

Time_EST: Time the article was uploaded to reddit standardized to EST using the UTC Timestamp obtained from the API

Life_Span: Indicates the duration of the article in the top 25 at the time of sampling

Short: The article appeared only once in the dataset
Medium: The article appeared for the second time in the dataset
Long: The article appeared more than two times in the dataset

Rank_Difference: Rank change during subsequent sampling of the article

Time_Difference: Time elapsed from the original time of posting to the time it was uploaded
On Reddit

Rank_Sign: Indicates if the article moved up the ranking or moved down the ranking in Subsequent sampling

Time_Of_Day: The time the article was sampled from Reddit

Expansions of Variable tags

Source Type Expansions

INO: International News Organization
NNO: National News Organization
TNO: Technology News Organization
LNO: Local News Organization
IFNO: International Financial/Business News Organization(Its now under **BNO** due to sparsity)
BNO: Business/Financial News Organization
UNO: University News Organization
MNO: Magazine News Organization
GNO:Government News Organization
SNO:Sports News Organization

General Notes

The original time and date has been standardized to EST from different timezones as mentioned on those sites. If it is a US local news website and the time timezone is not mentioned then for now it is assumed to be EST.
*The times are now converted to EST and all the times in the dataset represent EST, but still timezone conversion cannot be used to calculate the difference between publish time and upload time as the publish time information is not consistent across the urls.

The tweets, FB,Instagram variables will be combined into one variable called **Social Media** the animated, non-animated gifs will be removed and instead if there are gifs they will considered as Images.

Formula to convert unix timestamp to excel date :
=(A1/86400)+DATE(1970,1,1)

More info at the below link

https://exceljet.net/formula/convert-unix-time-stamp-to-excel-date

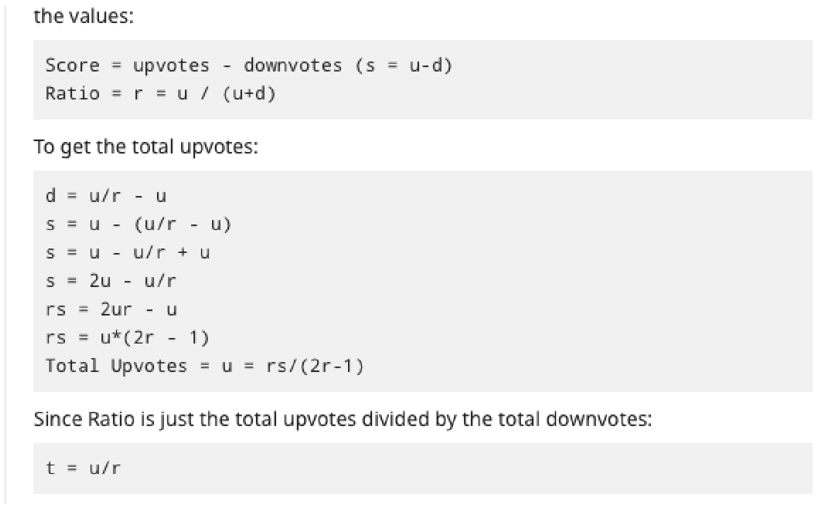
Formula to split date from timestamp
=INT(A2) (A2 is the cell you need to split by)
Formula to split time from timestamp
=A2-C2 (A2 is the cell you split by, and C2 is the date cell)

More info on splitting the time stamp can be found at below link
https://www.extendoffice.com/documents/excel/2758-excel-split-date-and-time.html

Life Span Variable Logic: Value 0 in the Repeated column is called Short, values 1 and 2 are called medium as it represents that the article appeared again in the dataset. Any value 3 and above is denoted as Long lifespan.

https://exceljet.net/formula/get-days-hours-and-minutes-between-dates

Derivation for Calculating the Upvotes from Score and upvote ratio values provided by PRAW



the values:

```
Score = upvotes - downvotes (s = u-d)
Ratio = r = u / (u+d)
```

To get the total upvotes:

```
d = u/r - u
s = u - (u/r - u)
s = u - u/r + u
s = 2u - u/r
rs = 2ur - u
rs = u*(2r - 1)
Total Upvotes = u = rs/(2r-1)
```

Since Ratio is just the total upvotes divided by the total downvotes:

```
t = u/r
```