# General Assembly

## DSI7-SF

Project team:

- Bill Yu
- Kun Guo
- Evelyn Li
- Manu Kalia

## NEWS API ARTICLES CLASSIFICATION DATA ANALYSIS:

## Disaster or Not-disaster?

# PROBLEM STATEMENT

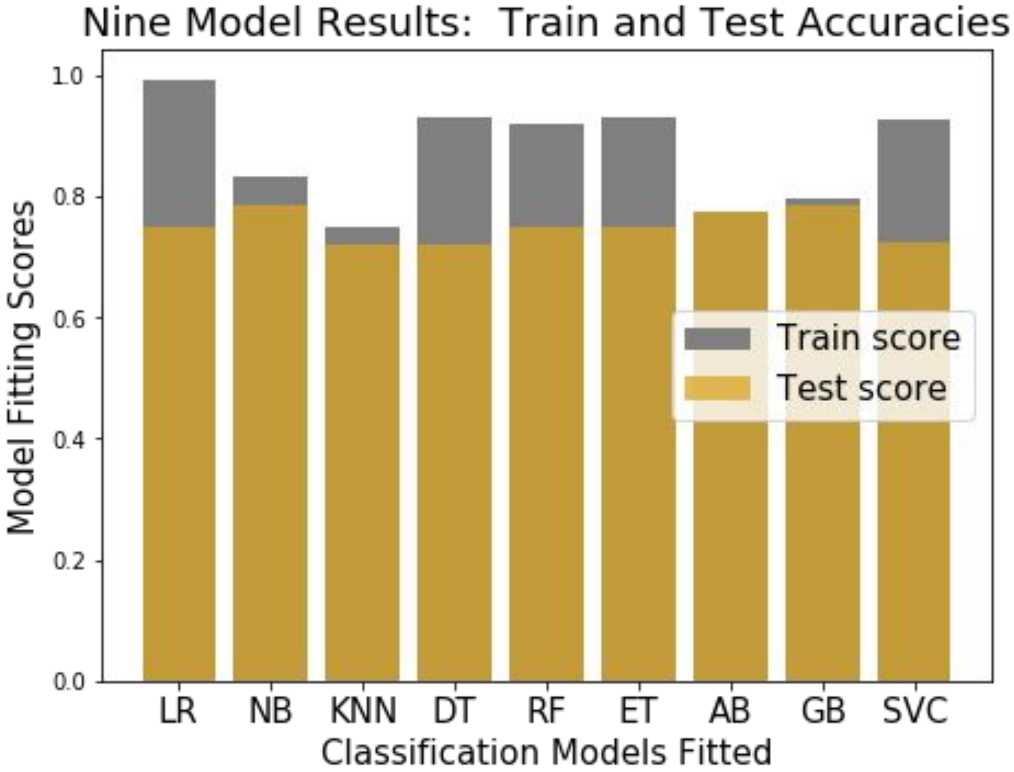
***Natural Language Processing*** has a number of very high-utility application areas: classification, machine translation, sentiment analysis, chat bots/ cust service, marketing message, targeting, etc.

***The question here is***: can a classification model successfully classify news article posts into one of two categories… DISASTER or NOT-DISASTER? If so, which estimator performs best in terms of accuracy and compute resources?

# THE ANSWER

Yes!

Naive- Bayes achieved 78% accuracy, followed closely by Logistic Regression at 77%



# EDA & VISUALIZATIONS

Missing Values

- *Since there are only a few rows with missing values (~60), which is significant fewer than the total (11,100), we filled them with “ ”.*

Decide which columns to include in the NLP process:

- *contents, descriptions, and titles all contain essential words*

Check unbalanced class:

- *Make sure there are equal numbers of 1s and 0s (5551, 5551)*

Prepare for vectorization

- *Check if there are any special characters such as “,”, numbers, “\n”, “_”... etc. included in the Words column*

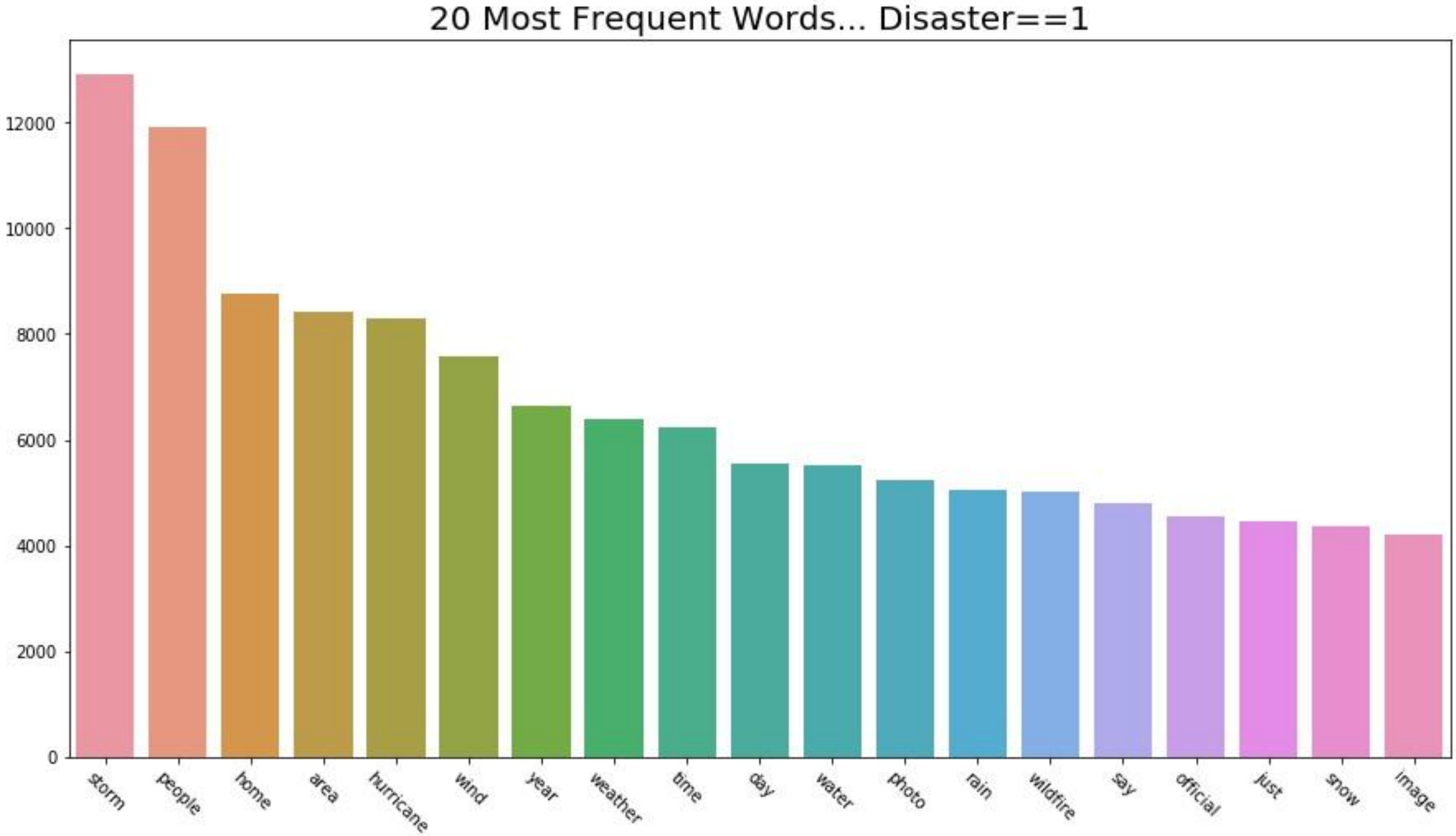
# EDA & VISUALIZATIONS

Count vectorized word frequencies

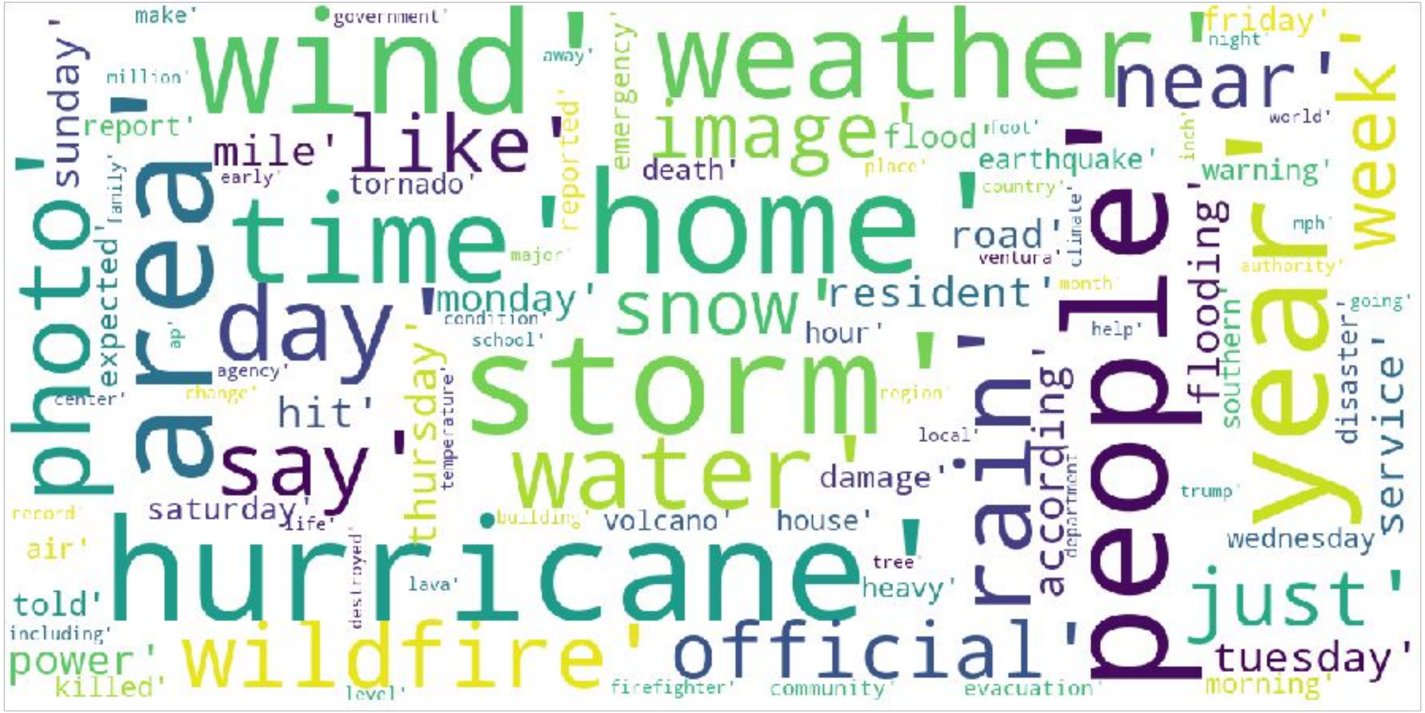
- Sorted the Top frequent words appeared in yes_disaster=1, in descending order
- Sorted the Top frequent words appeared in yes_disaster=0, in descending order
- With stopwords/without stopwords
- Top 20, 50, 100, 200 words

Word-overlap rates for groups of top words by frequency:

- Words that appeared in both yes_disaster=1 and yes_disaster=0…
  52.6% (top 20 wds), 36.7% (top 50), 45.5% (top 100), 56.8% (top 200)


20 Most Frequent Words... Disaster==1
12000
10000
8000
6000
4000
2000
0
storm
people
home
area
hurricane
wind
year
weather
time
day
water
photo
rain
wildfire
say
official
just
snow
image

# Most Frequent Words Classified as 'yes_disaster' = 1



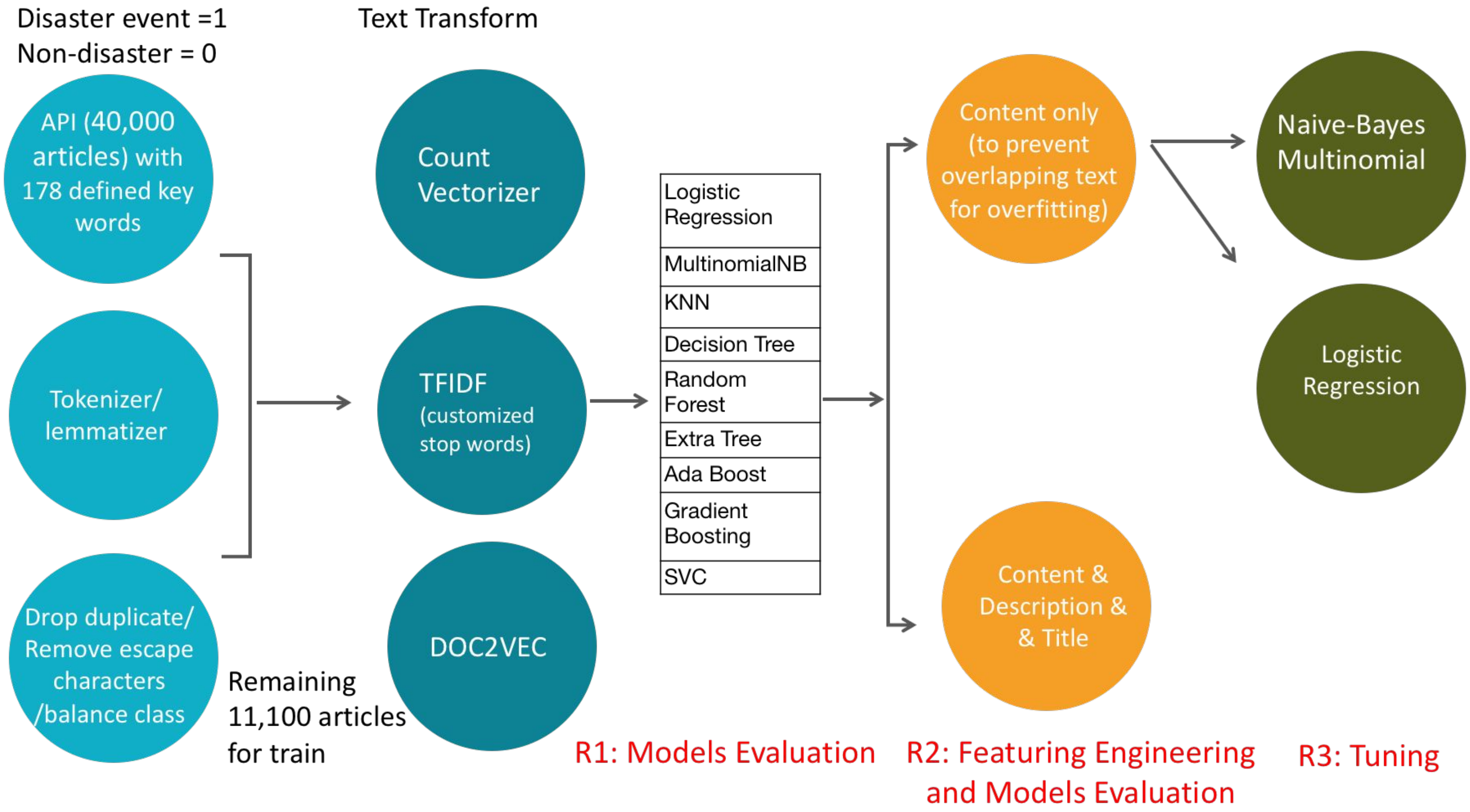
# CUSTOM STOP WORDS

Employed english stop words, with additional location words added to prevent the models from training on locations instead of disaster related words.  Training the classification models on location words would not generalize well to future disasters, but rather to any future news that occurs in locations at which disasters occured in the training set (e.g. news occurrances that occur in Paradise, CA the site of one of 2018's wildfire disasters.

**2,490** total stopwords:

- 318 starting stopwords list
- 193 UN countries
- 1,000 US cities
- 51 US states + D.C.
- 928 Tsunami, river-flood, storm-surge, and earthquake prone regions

# ROADMAP: MODELING STRATEGY



Disaster event =1
Non-disaster = 0

API (40,000 articles) with 178 defined key words

Tokenizer/ lemmatizer

Drop duplicate/ Remove escape characters /balance class

Remaining 11,100 articles for train

Text Transform

Count Vectorizer

TFIDF (customized stop words)

DOC2VEC

Logistic Regression
MultinomialNB
KNN
Decision Tree
Random Forest
Extra Tree
Ada Boost
Gradient Boosting
SVC

R1: Models Evaluation

Content only (to prevent overlapping text for overfitting)

Content & Description & & Title

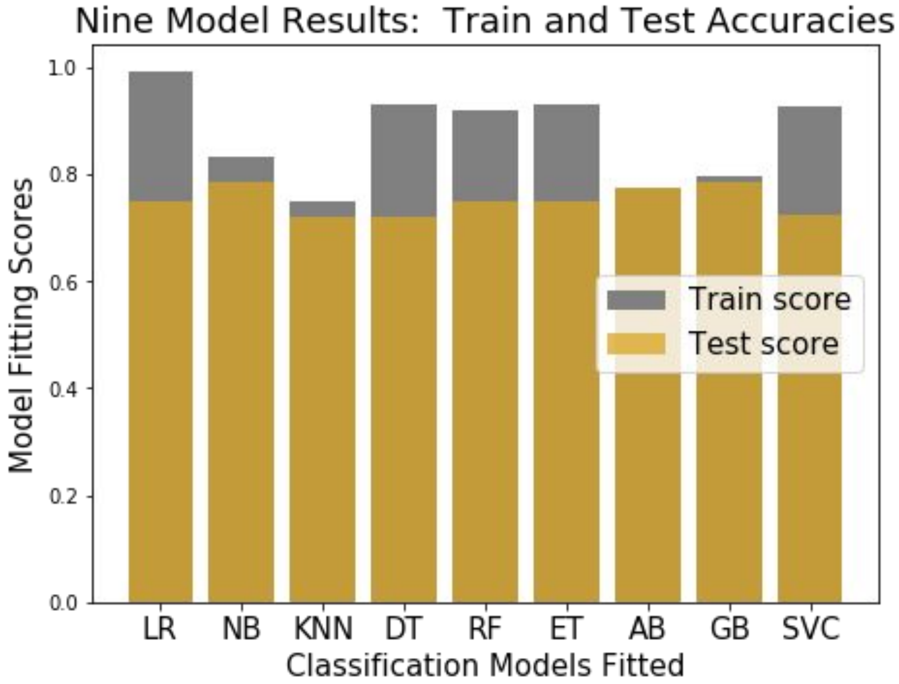
R2: Featuring Engineering and Models Evaluation

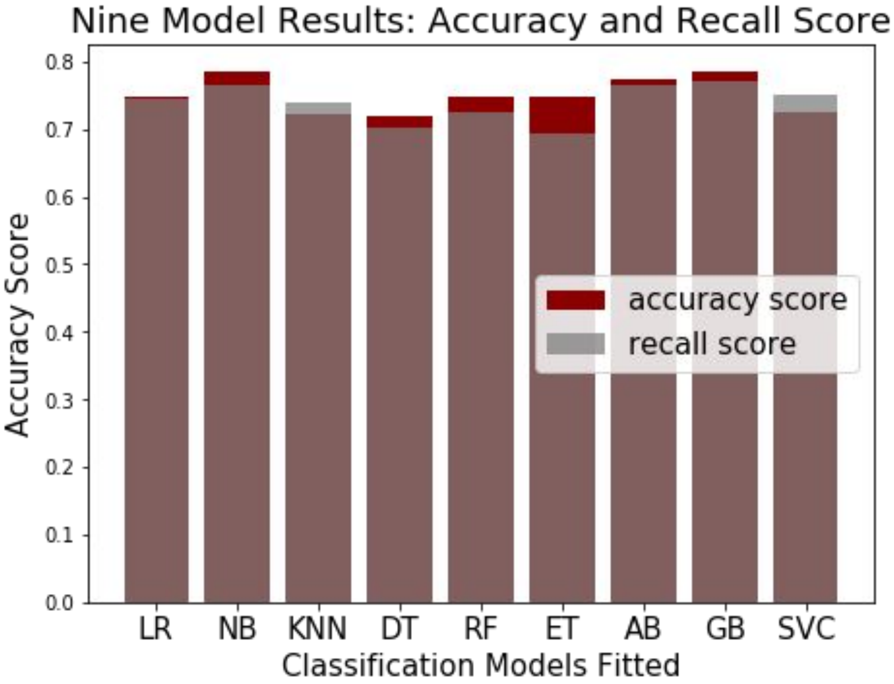
Naive-Bayes Multinomial

Logistic Regression

R3: Tuning

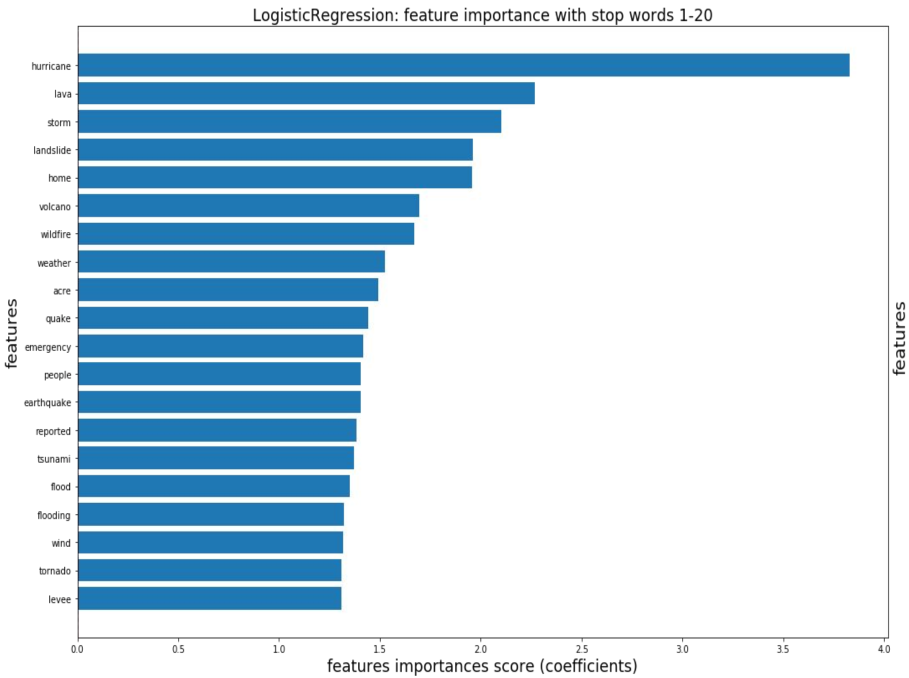
# TEST-TRAIN SCORES

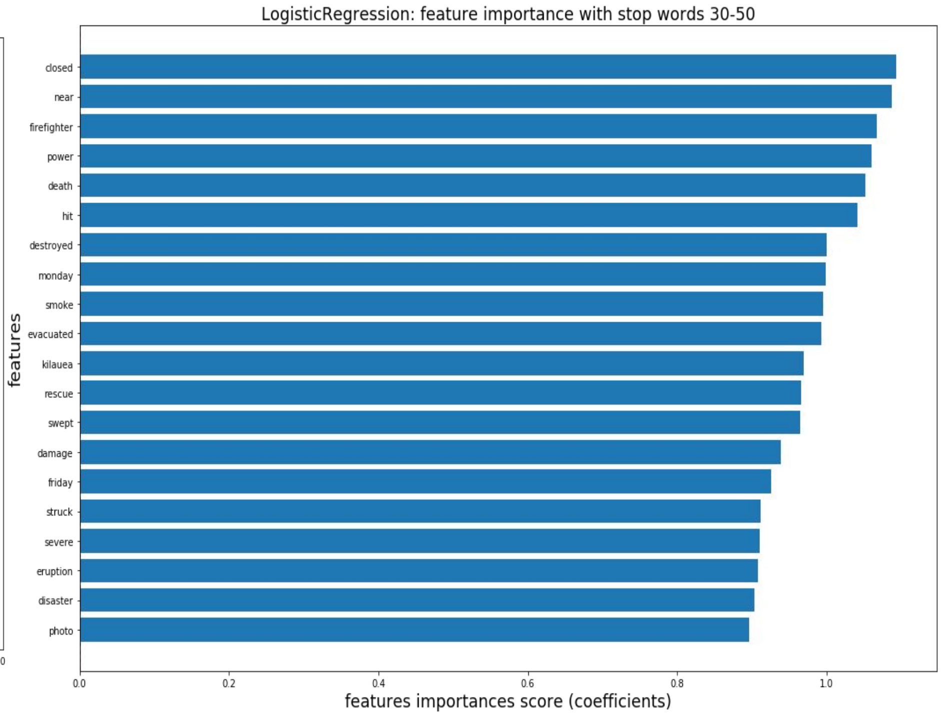




With a consideration of computing power, speed, and interpretability, Logistic Regression and Naive Bayes could classify a large volume of news without scarifying computation resources and accuracy.

# Feature Importance





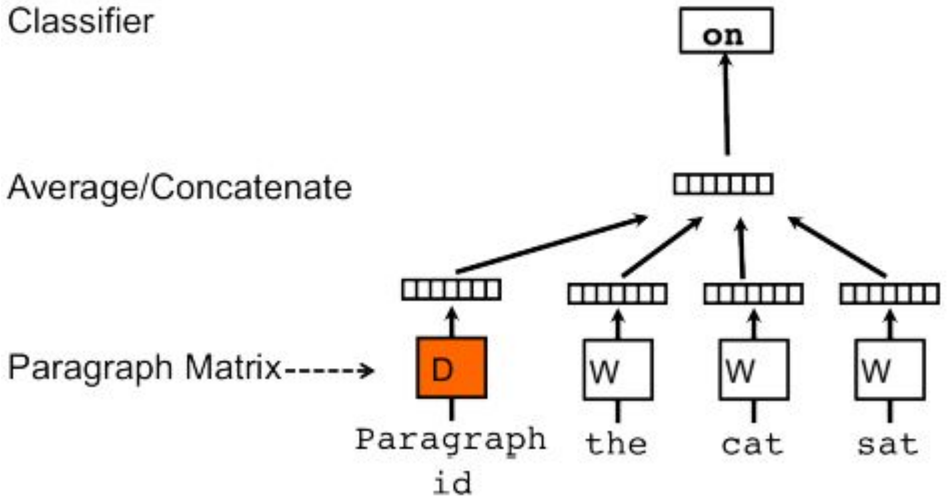
In general, The top 20 features words in Logistic Regression and NB highlight the status of the natural disaster such as hurricane, earthquake, and storm. Theses words will contribute the most to the prediction; Meanwhile, the next 30 - 50 words seems to suggest the consequences after the disaster. We perceived such words as firefighter, rescue, evacuate, and damage. The occurrence of these words meet our intuition on how to define a disaster.

# WORD VECTORIZATION WITH DOC2VEC

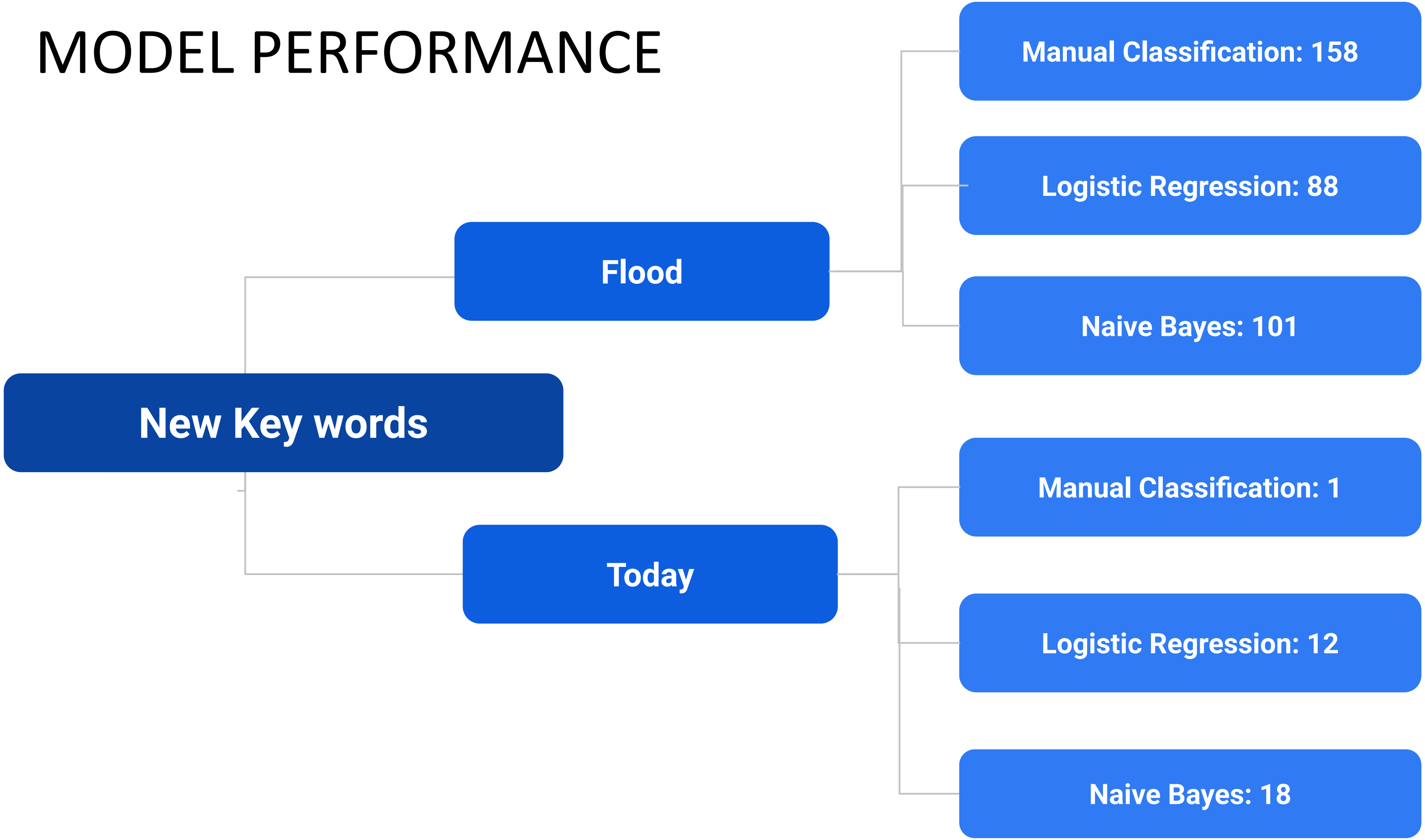
- 9 classification models
- No significant improvement on accuracy score comparing to CountVectorizer or TFIDF
- Loss interpretability after transformation

Key Takeaway:

- Documentation structure does not matter in this cause due to similar content structure and words
- What matter the most is the keywords



# MODEL PERFORMANCE



New Key words

Flood

Manual Classification: 158

Logistic Regression: 88

Naive Bayes: 101

Today

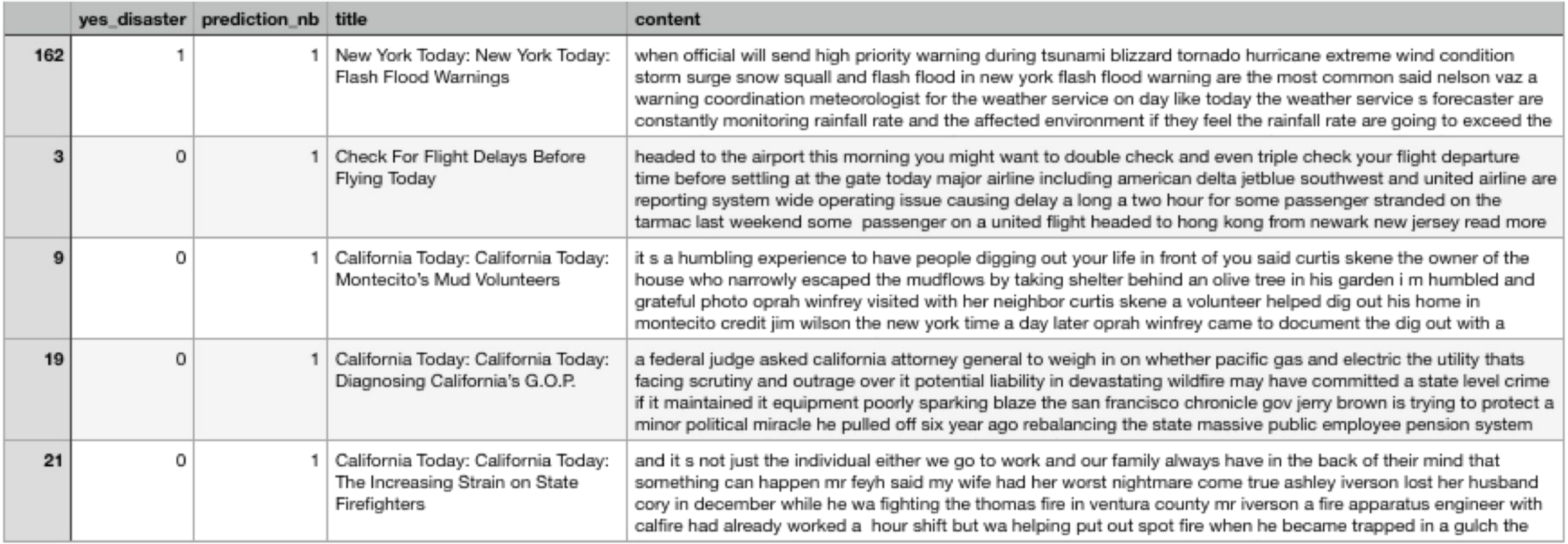
Manual Classification: 1

Logistic Regression: 12

Naive Bayes: 18

# PERFORMANCE ON UNSEEN DATA

Downloaded 250 articles (no keyword) from yesterday.  Below are five of the 17 articles identified by the model as yes_disaster = 1.  All but one are false positives, but all contained words highly relevant to disasters.



| | yes_disaster | prediction_nb | title | content |
|---|---|---|---|---|
| 162 | 1 | 1 | New York Today: New York Today: Flash Flood Warnings | when official will send high priority warning during tsunami blizzard tornado hurricane extreme wind condition storm surge snow squall and flash flood in new york flash flood warning are the most common said nelson vaz a warning coordination meteorologist for the weather service on day like today the weather service s forecaster are constantly monitoring rainfall rate and the affected environment if they feel the rainfall rate are going to exceed the |
| 3 | 0 | 1 | Check For Flight Delays Before Flying Today | headed to the airport this morning you might want to double check and even triple check your flight departure time before settling at the gate today major airline including american delta jetblue southwest and united airline are reporting system wide operating issue causing delay a long a two hour for some passenger stranded on the tarmac last weekend some  passenger on a united flight headed to hong kong from newark new jersey read more |
| 9 | 0 | 1 | California Today: California Today: Montecito's Mud Volunteers | it s a humbling experience to have people digging out your life in front of you said curtis skene the owner of the house who narrowly escaped the mudflows by taking shelter behind an olive tree in his garden i m humbled and grateful photo oprah winfrey visited with her neighbor curtis skene a volunteer helped dig out his home in montecito credit jim wilson the new york time a day later oprah winfrey came to document the dig out with a |
| 19 | 0 | 1 | California Today: California Today: Diagnosing California's G.O.P. | a federal judge asked california attorney general to weigh in on whether pacific gas and electric the utility thats facing scrutiny and outrage over it potential liability in devastating wildfire may have committed a state level crime if it maintained it equipment poorly sparking blaze the san francisco chronicle gov jerry brown is trying to protect a minor political miracle he pulled off six year ago rebalancing the state massive public employee pension system |
| 21 | 0 | 1 | California Today: California Today: The Increasing Strain on State Firefighters | and it s not just the individual either we go to work and our family always have in the back of their mind that something can happen mr feyh said my wife had her worst nightmare come true ashley iverson lost her husband cory in december while he wa fighting the thomas fire in ventura county mr iverson a fire apparatus engineer with calfire had already worked a  hour shift but wa helping put out spot fire when he became trapped in a gulch the |

# KEY TAKEAWAYS FROM MODEL PERFORMANCE

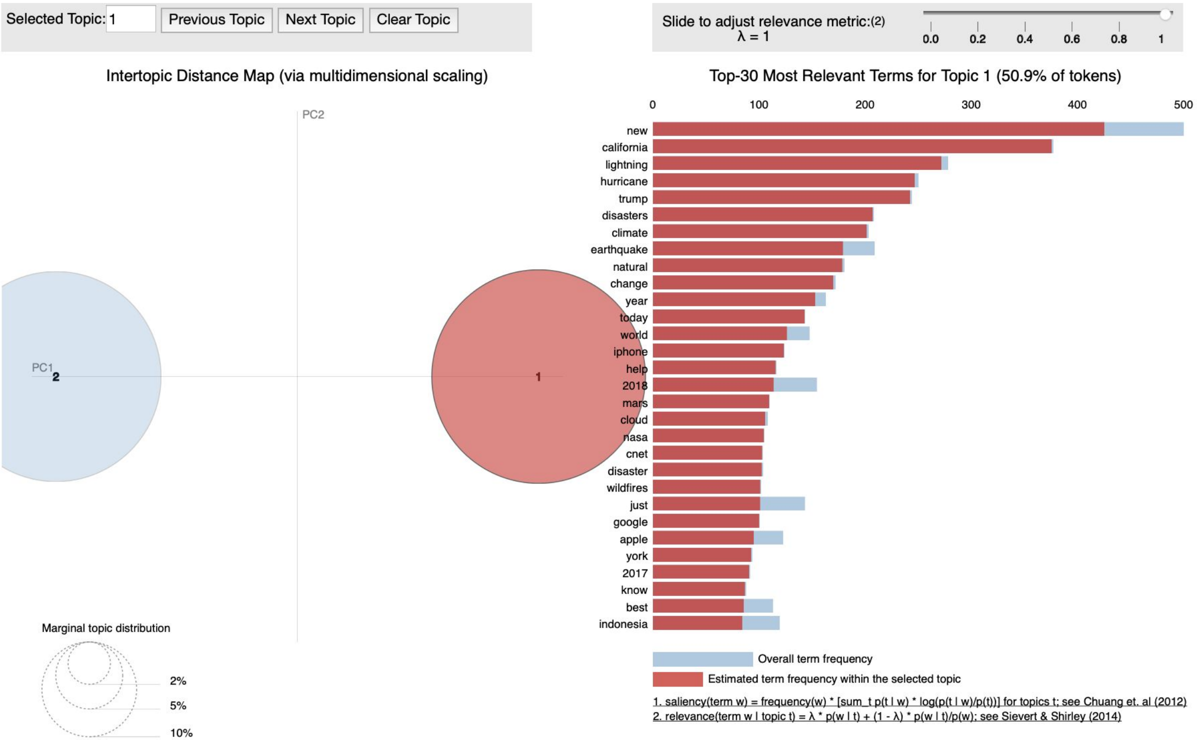
1. Model does a fair job identifying relevant news articles

2. Needs improvement on classifying borderline articles that relate to disasters, but are not specifically about unfolding disaster events

3. Testing dataset is small, 250 news due to time constraints, but in real life the model should perform better

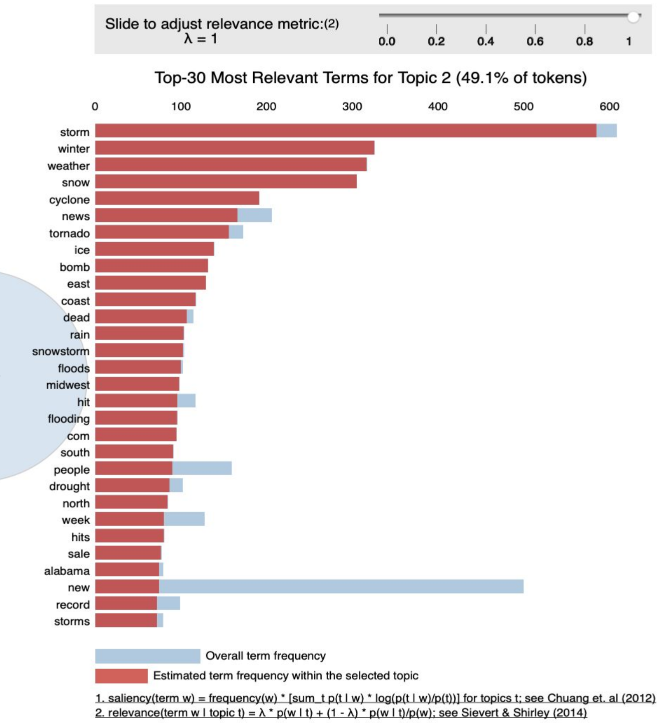
# UNSUPERVISED LEARNING WITH LDA

## TOPIC ONE



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

## TOPIC TWO



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

# RECOMMENDATION 1: Use Keywords for incoming news/tweets filtering to display on a webpage

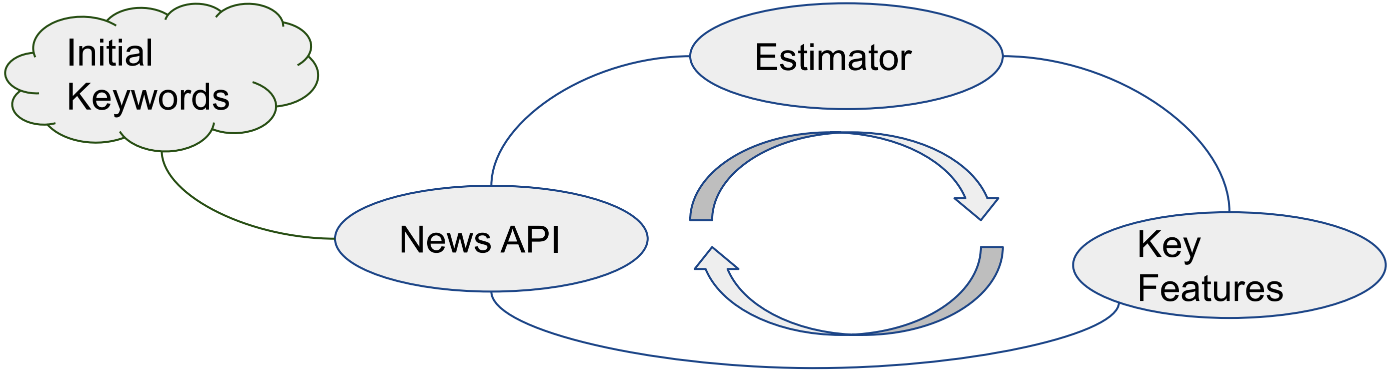
1. Keywords to use: top coefficient words from the final logistic regression model

2. Set up a scheduled script (as in CRON) to ingest all news posts (and/or tweets) 1-4 times per day. Use this downloaded set of articles as a new 'test' dataset for our trained and fitted model.

3. Display all the news items that are returned as positive for the disaster class. There will be some false positives that are posts containing disaster words, but are not necessarily about a disaster in progress.

# RECOMMENDATION 2: Rerun the model & update the key features list

Periodically (2-3 times per year) use the top 50 or so key features to download more articles for training the classifier(s). The update cycle will collect news articles, and check for duplicates to articles in the existing training dataset. Re-run the Naive-Bayes and Logistic Regression models to keep the keywords set updated.



# RECOMMENDATIONS:  Continuously Grow the Dataset

1. Collect large new sets of articles and characterize each one as "disaster" or not-disaster" by human inspection.  This is a time and labor-intensive task, so it is important to minimize the labor expense whenever possible...
   a. Free lance services
   b. Local student contests
   b. Internships
   c. Disaster assistance volunteer days

2. Flag all news articles, and store for future incorporation, from days on a known disaster (and perhaps for x days afterwards).

# RECOMMENDATIONS 4: Future Improvements

Future Enhancements to the Model...

a. Secondarily classify the disaster types (wildfires, storms)

b. Identify a "Disaster Condition" using the number of articles classified as 'yes_disaster' during a certain time-period (perhaps one day or several hours).