

# #How Is Twitter Feeling About...

Jenna Fu



# OVERVIEW:

**Build an app, where user is able to enter a Twitter topic, and gauge the public opinions and sentiments of tweets related to that particular topic.**



# DATA PROCESSING ROAD MAP

Data Collection

Feature Engineering

Text Preprocessing

Text Vectorization

kaggle

Sentiment140 dataset with  
1.6 million tweets



Twitter API



An open source Python package  
that helps you access the Twitter  
API with Python

- Split date column into datetime features
- Retrieve numerical features such as count of hashtags, mentions and urls from the text column

- Convert all characters to lowercase
  - Remove punctuations
  - Remove extensive list of stop words
  - Lemmatization with POS tags
- Example:**  
**corpora => corpus**  
**better => good**

TFIDF Vectorizer

**Example:** Word like “corpora” appears seldomly in sentences, hence we are issuing more weights on it, than common terms like “better”.

# EXAMPLE

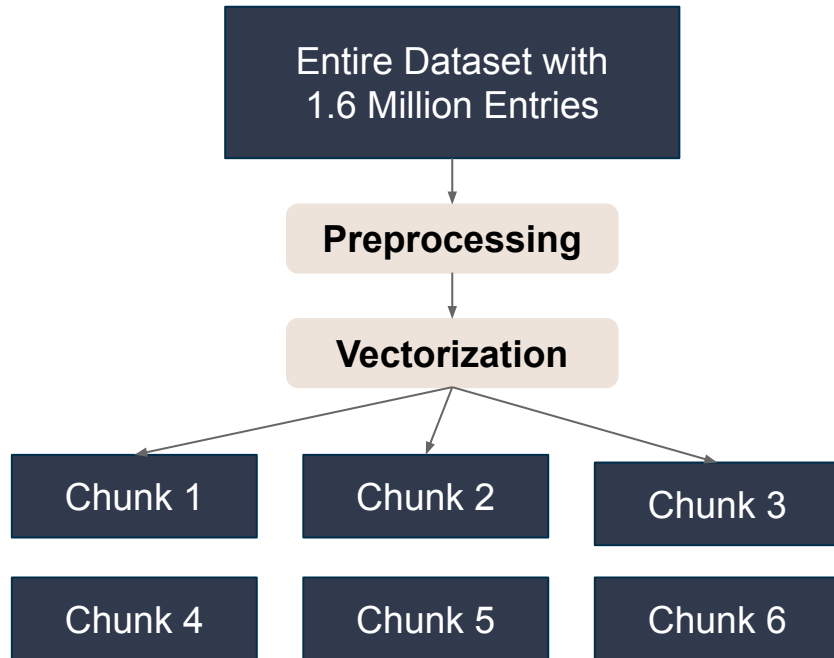
## Raw Data

tweet_id	date	flag	user	text
1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all...

## Preprocessed Data

date_hour	date_day	date_month	date_weekday	hashtag_counts	user_counts	url_counts	aa	aaa	aaaa	...	zoe	zombie	zombies	zomg	zone
22	6	4	1	0	1	1	0	0	0	...	0	0	0	0	0
22	6	4	1	0	0	0	0	0	0	...	0	0	0	0	0
22	6	4	1	0	1	0	0	0	0	...	0	0	0	0	0
22	6	4	1	0	0	0	0	0	0	...	0	0	0	0	0
22	6	4	1	0	1	0	0	0	0	...	0	0	0	0	0

# DATA MODELLING: TWO APPROACHES



## INCREMENTAL LEARNING

Training the model on the entire datasets by chunks with **SGDClassifier**.

- Linear classifiers (SVM, logistic regression, etc.) with stochastic gradient descent training.

## TYPICAL ML APPROACH

Training the model on a single chunk selected randomly, with **Logistic Classifiers, SVM and Decision Tree**.

# MODEL ASSESSMENT/ACCURACY

## INCREMENTAL LEARNING

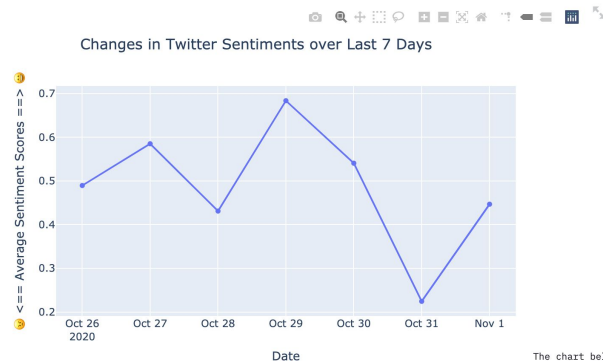
Model	Accuracy
Logistic	49.43%
SVM	55.18%

## TYPICAL ML APPROACH

Model	Accuracy
Logistic	58.43%
SVM	76.27%
Decision Tree	72.91%
KNN	73.76%

# WEB APP DEMO

The chart below returns the average sentiment scores of the last 7 days, from a scale of 0 to 1. The scores are calculated by the averaging the prediction scores (0 or 1) on a single day.



The chart below finds and ranks the top 10 most frequently appearing emojis in the tweets we have scrapped from Twitter.



The word cloud below represents the words or phrases from tweets associated with positive and negative sentiments accordingly.

