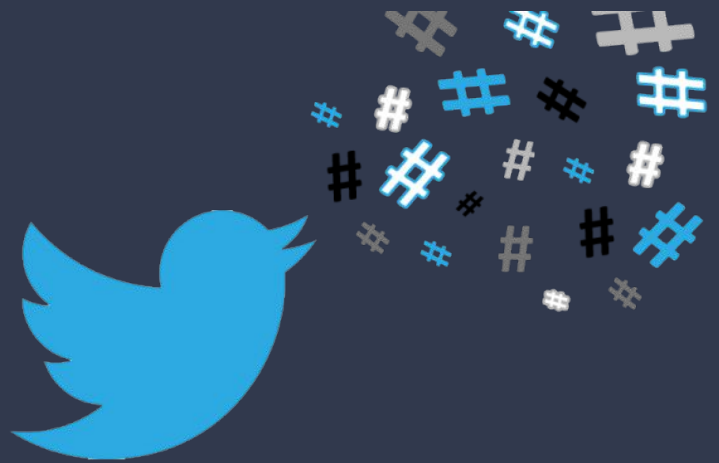
#How Is Twitter Feeling About…

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

kaggle

Sentiment140 dataset with 1.6 million tweets

Twitter API

tweepy™

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

Feature Engineering

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

Text Preprocessing

- Convert all characters to lowercase
- Remove punctuations
- Remove extensive list of stop words
- Lemmatization with POS tags

Text Vectorization

TFIDF Vectorizer

Example: Word like "we" appears often in sentences, hence we are issuing less weights on it.

EXAMPLE

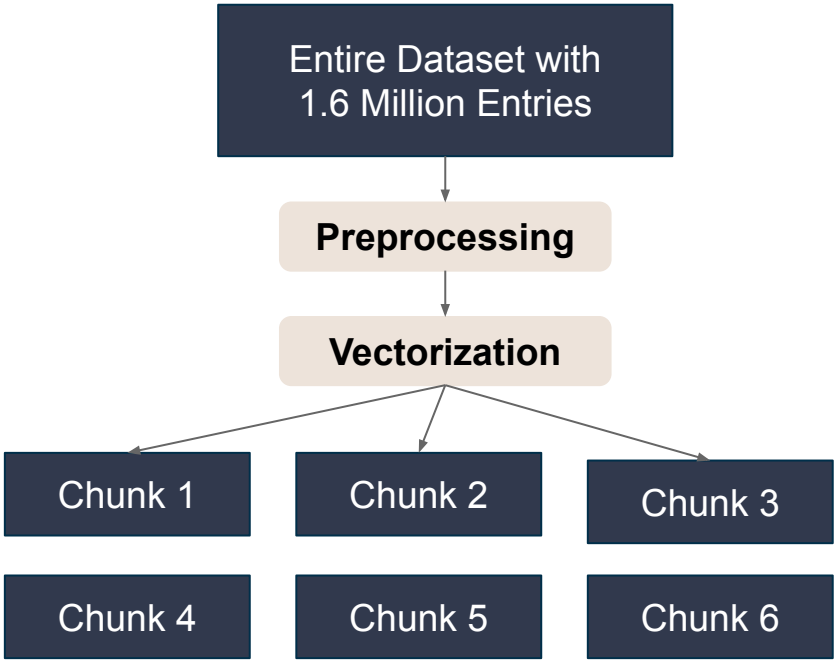
Raw Data

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	58.43%
SVM	62.18%

TYPICAL ML APPROACH

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

WEB APP DEMO

#How is Twitter Feeling About...

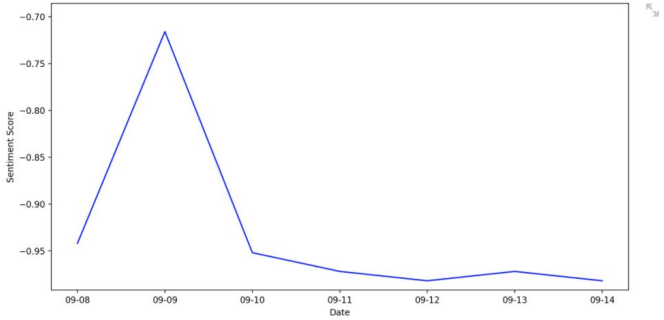
Which Twitter topic would you like to explore?

Raptors

Topic sentiment in the past 7 days:

Sentiment Score on a scale of -1 to 1

(with -1 being the most negative and 1 being the most positive)



Top 10 Emojis Used:

	Emoji	Count
43	📰	104
130	🇨🇦	97
15	🏀	81
65	❤️	74
70	😂	57
35	😔	33
60	🙏	32
116	🙏🏻	29
26	🙏	27
18	🧯	27

Associated Hashtags Word Cloud: