Weather Detection Using Twitter

By Alan Chen and Eoin Nugent

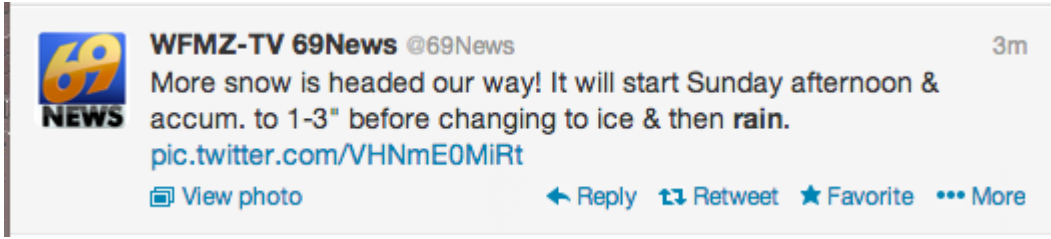
Motivation

What can a machine learn about its surroundings given simple human interactions?

Are the manpower of labels or the learning of the machine the limiting factor?

“Predicting the weather”

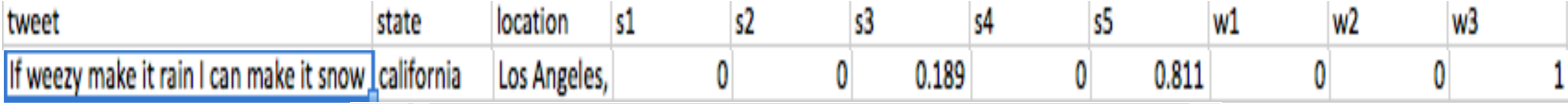
The goal of our project is to classify tweets based on temperament, time period, and type of weather.



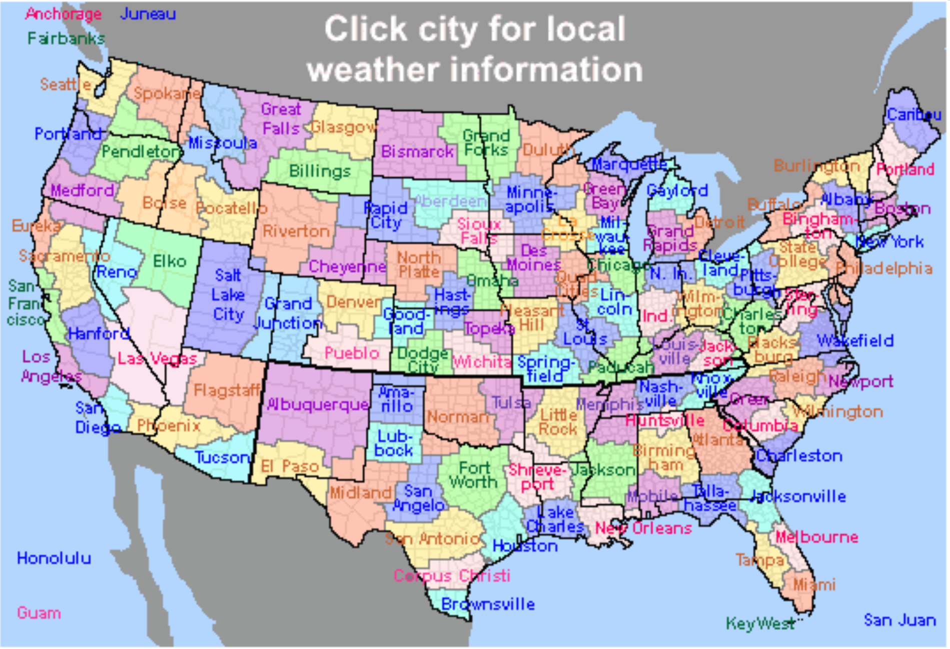
WFMZ-TV 69News @69News 3m

More snow is headed our way! It will start Sunday afternoon & accum. to 1-3" before changing to ice & then **rain**. pic.twitter.com/VHNmE0MiRt

The Data



tweet	state	location	s1	s2	s3	s4	s5	w1	w2	w3
If weezy make it rain I can make it snow	california	Los Angeles,	0	0	0.189	0	0.811	0	0	1



Getting Features

Word count

Weight by term frequency-inverse document frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Uses training corpus as vocabulary of words

Classify

Fit features to tweet labels

First used Support Vector Machine

Accuracy: Sentiment = 30%

Time = 78%

Weather = 45%

Prediction Difficulties

Weather can have multiple valid labels

Fallibility of human raters

More Classifiers

Decision Tree

Random Forest

AdaBoost

Gaussian Naive Bayes

Next Steps...

Feature Improvement:

- n-grams
- “social connectedness”

Applications

Practically plug n’ play with other labels/categories

Streaming: understanding human language in real-time

Crowd-sourced knowledge maps

Q&A