Tweetin' in the Rain: Exploring societal-scale effects of weather on mood

Aniko Hannak† Eric Anderson† Lisa Feldman Barrett† Sune Lehmann‡ Alan Mislove† Mirek Riedewald†

†Northeastern University ‡Technical University of Denmark

Motivation

Significant interest in using sentiment of postings on OSNs
- Predicting the stock market [ICWSM'10]
- Forecasting movie success [WI'10]
- Traditional polls replaced by Twitter data [ICWSM'10]

Most methods for sentiment analysis use scored word lists
- Affective Norms of English Words (ANEW)
- Wilson, Wiebe and Hoffmann (WWH)
- Hu and Liu (HL)

Can we construct a more accurate, tailored word list?

Previous psychological studies on patterns of sentiment
- Daily, weekly, seasonal
- Geographic
- Climate-related

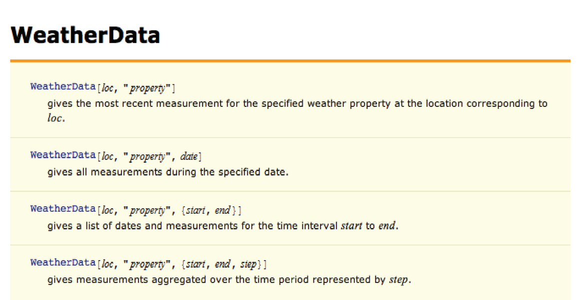
To what extent do these translate to population-wide patterns?
To what extent is aggregate sentiment itself predictable?

Data

Use data collected from Twitter
- User profiles, tweets and their timestamps
- January - August 2009
- ~40 M users and ~1.3 B tweets

Map users to their geographical location
- Self-reported location in users' profiles
- Interpret location using Google Maps API
- Only use locations in 20 largest US metropolitan areas

Correlate tweets with weather at corresponding location
- Mathematica's WeatherData package
- Uses National Weather Service data
- For each of the 20 metropolitan areas, every hour:
  - *Cloud cover, Humidity, Precipitation, Temperature ,Wind speed*



Measuring Sentiment

Limitations of existing sentiment inference techniques:

Word list-based techniques:
- Contain few words (low 1,000s)
- Not Twitter-specific, often ignore:
  - Abbreviations (e.g. OMG, LOL)
  - Nelogisms (e.g. truthiness)
  - Hashtags (e.g. #fail)
- Expensive to create (manual process)

Natural language processing:
- Techniques don't scale to Twitter corpus
- Unique use of language on Twitter
  - Many misspellings
  - Incorrect grammar
  - Missing punctuation

Creating EMOT word list
- Consider subset of tweets containing emoticons :) :-) :(:-(
- For each token, measure relative fraction of co-occurance with emoticons

Advantages of the EMOT list
- Automatically created
- Captures Twitter-specific syntax
- Much larger than existing lists
- Easily extends to other languages

Happy Bday OBAMA!!! :)
It's Obama's bday . . . :-(
Obama is the best!!! :-)
Oh no, it's raining again!!! :(
Happy Bday @Anna !!! :) http://youtube...

bday = 2 x 🙂 / (2x 🙂 + 1x 🙁) = 0.67

obama = 2 x 🙂 / (2x 🙂 + 1x 🙁) = 0.67

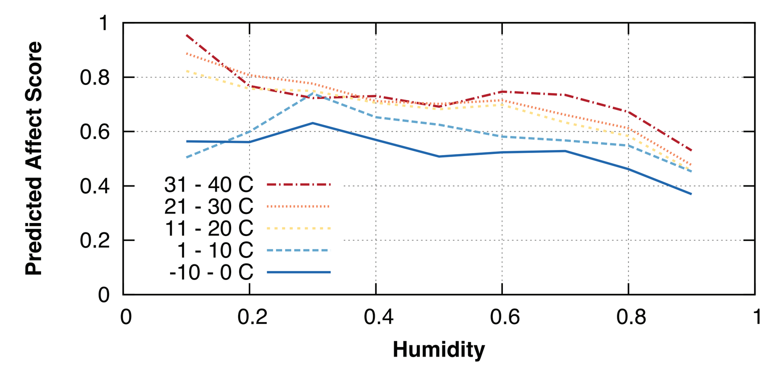
!!! = 3 x 🙂 / (3x 🙂 + 1x 🙁) = 0.75

Results

Observe significant correlation
- All variables: 0.79 ROC area
- Most useful variables: W, T

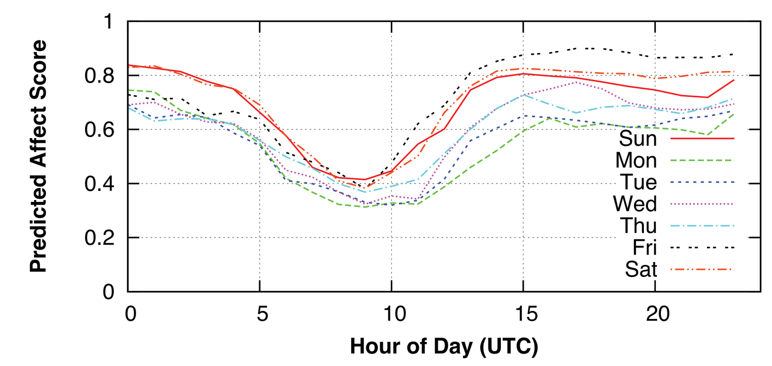
Can predict sentiment with significant accuracy

Variable classes	Area Under ROC Curve
G, S	0.6585
W, S	0.7427
T, S	0.7450
W, G	0.7561
T, W	0.7724
G, T	0.7753
W, G, T, S	0.7857

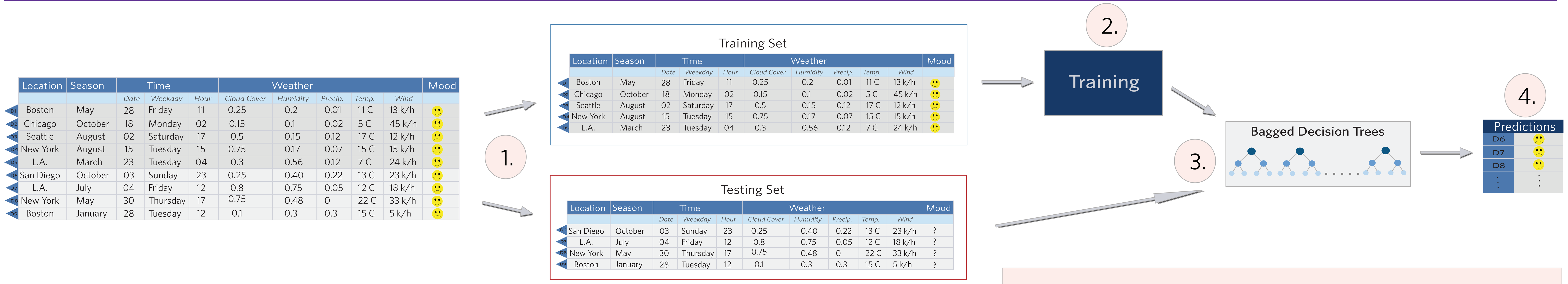


Partial dependence plots
- Can ask predictor for relationships between variables
- Closely match intuition

As humidity increases, affect scores increase (with more pronounced effect at higher temperatures)



Predicting Sentiment



	Location	Season	Time: Date	Time: Weekday	Time: Hour	Weather: Cloud Cover	Weather: Humidity	Weather: Precip.	Weather: Temp.	Weather: Wind	Mood
D1	Boston	May	28	Friday	11	0.25	0.2	0.01	11 C	13 k/h	🙂
D2	Chicago	October	18	Monday	02	0.15	0.1	0.02	5 C	45 k/h	🙁
D3	Seattle	August	02	Saturday	17	0.5	0.15	0.12	17 C	12 k/h	🙁
D4	New York	August	15	Tuesday	15	0.75	0.17	0.07	15 C	15 k/h	🙂
D5	L.A.	March	23	Tuesday	04	0.3	0.56	0.12	7 C	24 k/h	🙂
D6	San Diego	October	03	Sunday	23	0.25	0.40	0.22	13 C	23 k/h	🙂
D7	L.A.	July	04	Friday	12	0.8	0.75	0.05	12 C	18 k/h	🙁
D8	New York	May	30	Thursday	17	0.75	0.48	0	22 C	33 k/h	🙂
D9	Boston	January	28	Tuesday	12	0.1	0.3	0.3	15 C	5 k/h	🙁

Training Set

	Location	Season	Time: Date	Time: Weekday	Time: Hour	Weather: Cloud Cover	Weather: Humidity	Weather: Precip.	Weather: Temp.	Weather: Wind	Mood
D1	Boston	May	28	Friday	11	0.25	0.2	0.01	11 C	13 k/h	🙂
D2	Chicago	October	18	Monday	02	0.15	0.1	0.02	5 C	45 k/h	🙁
D3	Seattle	August	02	Saturday	17	0.5	0.15	0.12	17 C	12 k/h	🙁
D4	New York	August	15	Tuesday	15	0.75	0.17	0.07	15 C	15 k/h	🙂
D5	L.A.	March	23	Tuesday	04	0.3	0.56	0.12	7 C	24 k/h	🙂

Testing Set

	Location	Season	Time: Date	Time: Weekday	Time: Hour	Weather: Cloud Cover	Weather: Humidity	Weather: Precip.	Weather: Temp.	Weather: Wind	Mood
D6	San Diego	October	03	Sunday	23	0.25	0.40	0.22	13 C	23 k/h	?
D7	L.A.	July	04	Friday	12	0.8	0.75	0.05	12 C	18 k/h	?
D8	New York	May	30	Thursday	17	0.75	0.48	0	22 C	33 k/h	?
D9	Boston	January	28	Tuesday	12	0.1	0.3	0.3	15 C	5 k/h	?

How to capture non-linear correlations?
How to see the effect of combined variables?

Treat sentiment prediction as machine learning problem
- Goal: Predict sentiment from the other variables
- Ability to predict implies correlation

Methodology: Bagged decision trees
- Can handle all attribute types (even missing values)
- Easy to analyze effect of variables from tree structure
- Works well with fairly little tuning

Input variables
- Geography (G): the metropolitan area
- Season (S): the month-of-year
- Time (T): day-of-month, day-of-week and hour-of-day
- Weather (W): five weather variables and historic data

1. Split input data into training and testing sets
   - Training: 67%, Testing: 33%
   - Sentiment score simplified to happy/sad (1/0)
   - Data points aggregated into hour-long city buckets
2. Train our machine learning algorithm
   - Build bagged decision trees (1,000)
3. Run testing set on decision trees
   - Obtain predicted mood scores for testing set
4. Compare predictions with actual data
   - Measure accuracy with Area under the ROC curve