TLP: Twitter Language Processing

Matthew Yates

Overview

- Motivation
- Goal
- Process
- Data
  - Text - 92 GB of tweets
  - Labels - Dow Jones, Rasmussen
    Presidential Approval, Gold Price
- Results
- Conclusions

Goal

See what if any knowledge can be gleaned from Twitter data

Motivation

Wisdom of Crowds
Text Processing
Machine Learning
+
___
Usable Information

It's an uphill battle

Goal: Answer Three Questions

Will the price of an ounce Gold in US Dollars rise?

Will President Obama's Rasmussen daily approval rating increase?

Will the Dow Jones Industrial average close at a higher price?

Features Compared

word n grams
- n = 1, 2, 3, and 4

character n grams
- n = 1, 2, 3, and 4

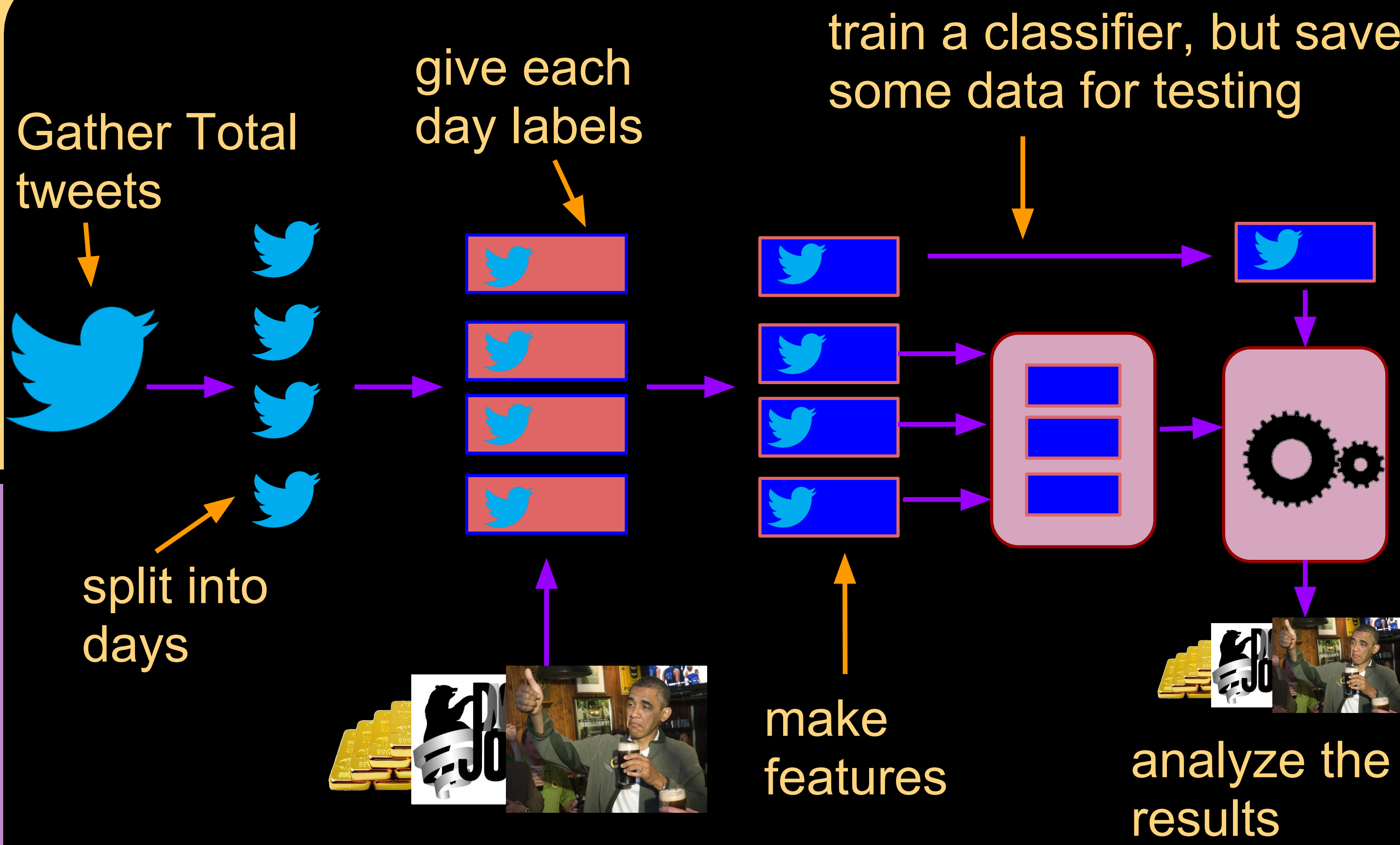
usernames, hashtags, and urls

word n grams with actual usernames, hashtags, and urls
- n = 1, 2, 3, and 4

word n grams with pseudo usernames, hashtags, and urls
- n = 1, 2, 3, and 4

word n grams with usernames appended
- n = 1 and 2

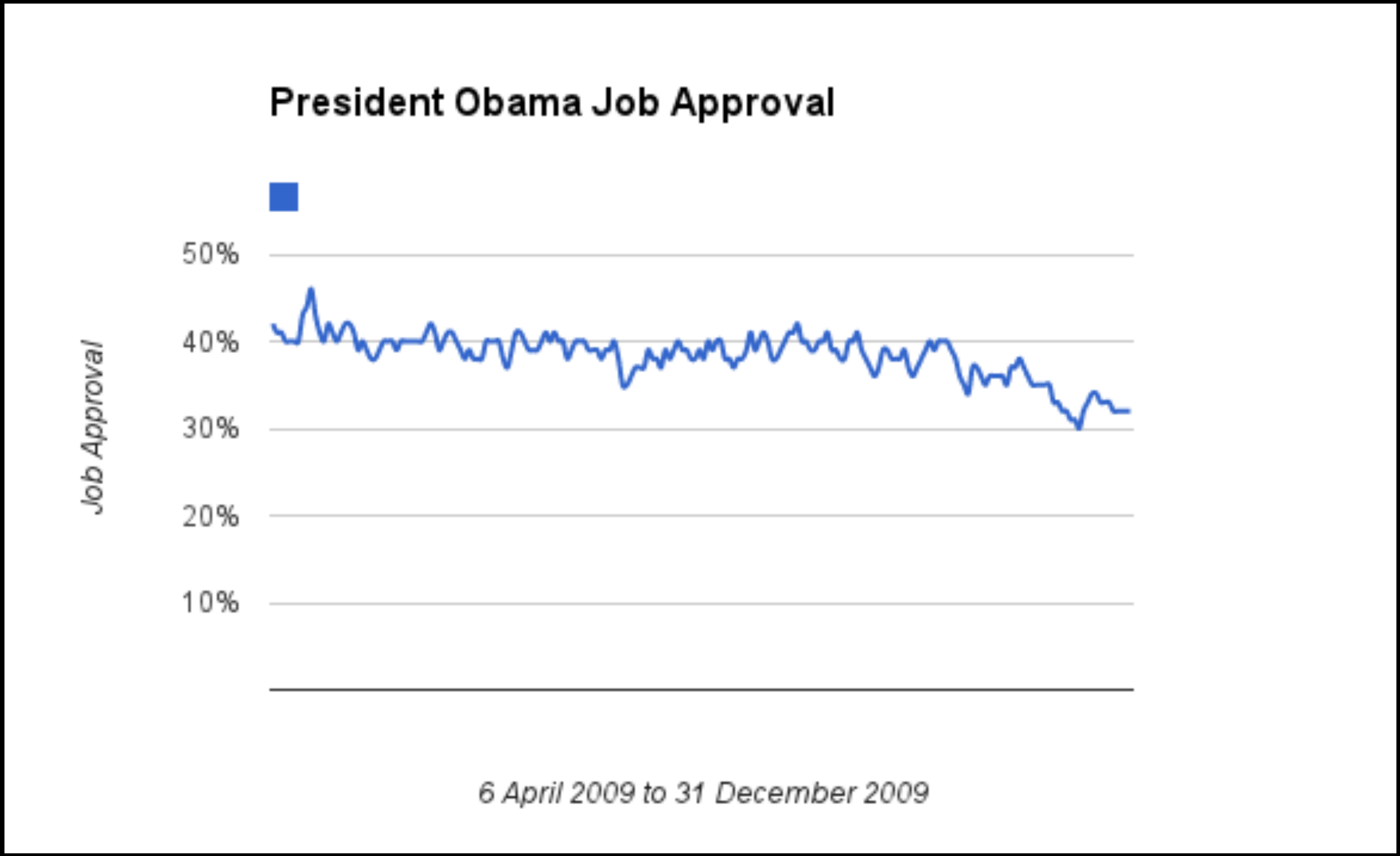
Project Diagram



Gold Prices

June 2009 - December 2009

Obama Rasmussen Job Approval



April 2009 - December 2009

Dow Jones Industrial Average

June 2009 - December 2009

Data Labels: Processing

Decision is made considering an entire days worth of tweets

Binary: Will value be greater tomorrow than today?

Days for which no data was collected (closed markets or no polling for holidays) used last good value for baseline

Saturday's tweets will be used to decide if the Dow closed higher on Monday than

Data Labels: Stats

Increase: 73

No Increase: 131

Global Baseline: 64.22%

Increase: 68

No Increase: 136

Global Baseline: 66.67%

Increase: 114

No Increase: 90

Global Baseline: 55.88%

Text Data - USNA Twitter Corpus

204 days of Tweets over 7 months

From 11st June 2009 to December 31st 2009

94,534,846 tweets

11.45 GB

1,578,010,381 words

463,406 tweets a day

7,735,345 words a day

Each day's tweets will be treated as one entity

Roughly 1% of all tweets during time period

What's in a Tweet?

140 unicode characters

Can be in 1 of 33 languages

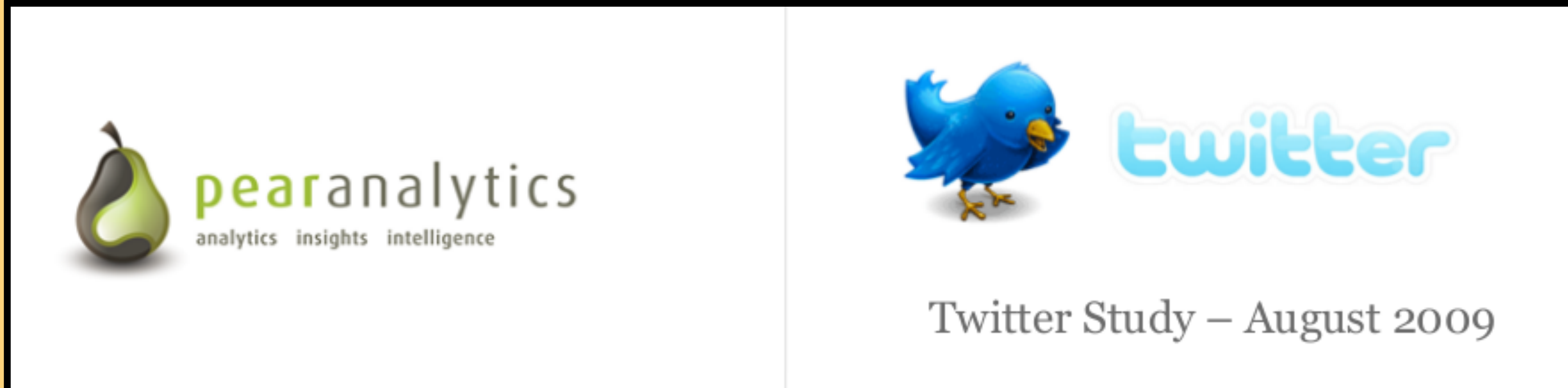
Worldwide, 24 hours a day, 7 days a week

Due to small message sizes, links use url shorteners

Users can use hash tags to identify some category of tweet

- ie #murica

About Twitter

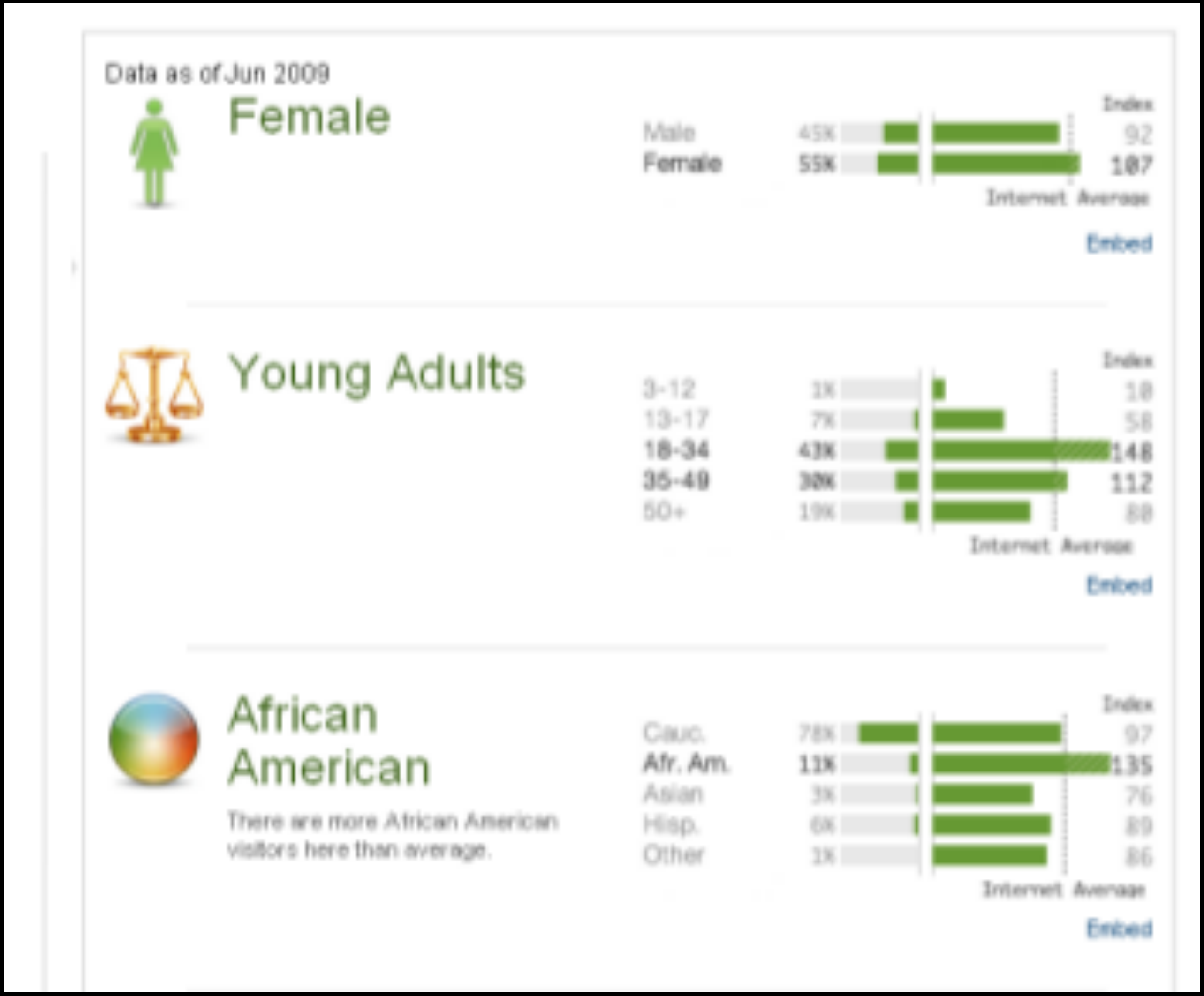


Ryan Kelly, ed. (August 12, 2009). "Twitter Study – August 2009" (PDF). *Twitter Study Reveals Interesting Results About Usage*. San Antonio, Texas: Pear Analytics. Archived from the original on 2011-07-15

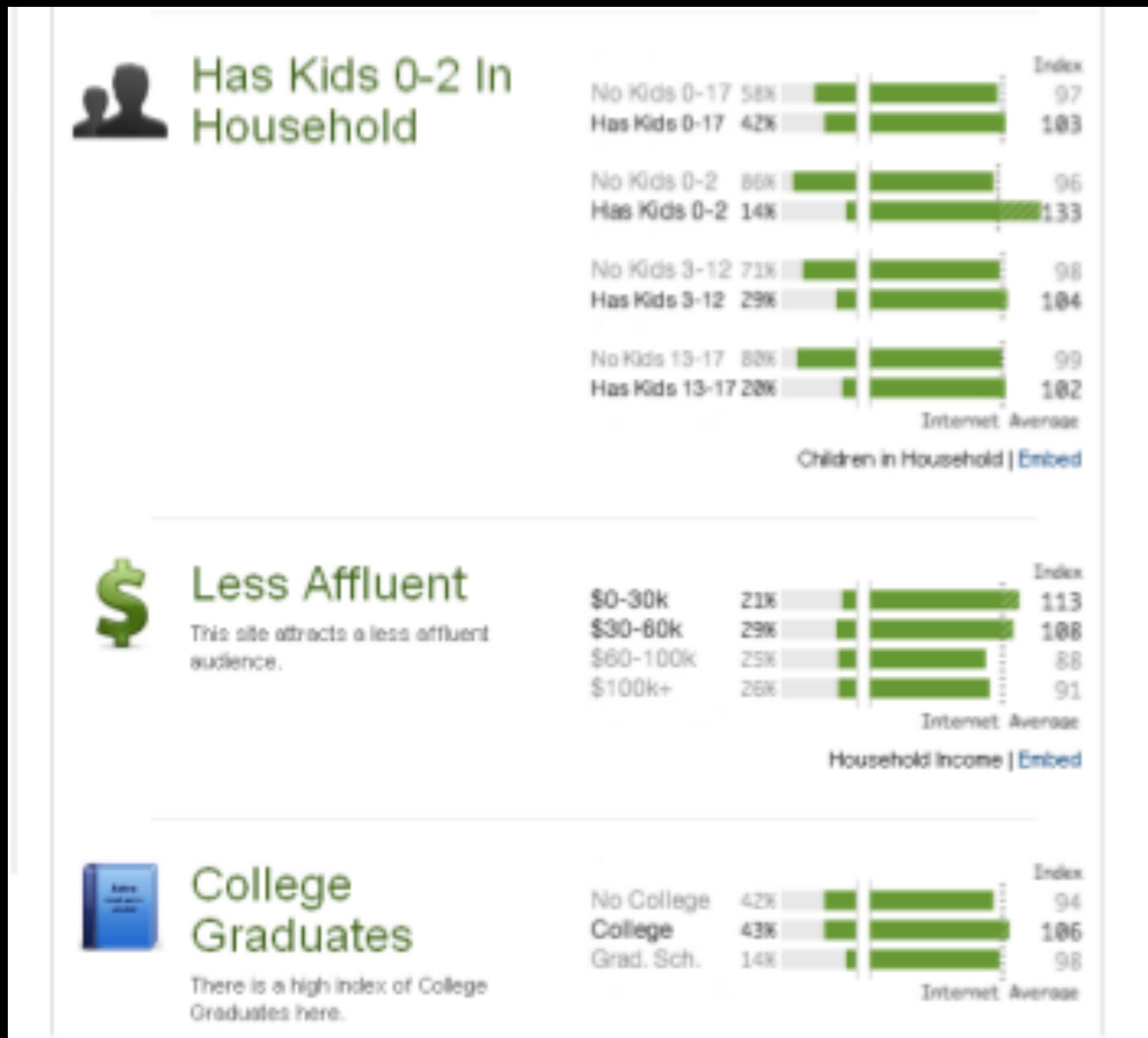
About Twitter: The Users

- 27 million people per month in the U.S.
- 55% are female
- 43% are between 18 and 34
- 78% Caucasian, but African American users are 35% above Internet average
- Average household income is between $30 and $60k
- 1% of the addicts contribute 35% of the visits
- 72% are passers-by, while only 27% are regular users

About Twitter: The Users (cont.)



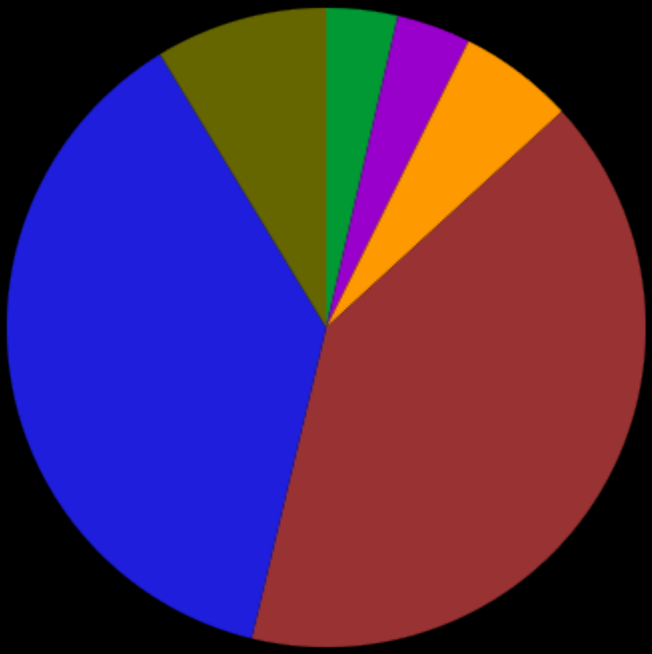
About Twitter: The Users (cont.)



About Twitter: The Tweets

Content of Tweets according to Pear Analytics

- News (3.6%)
- Spam (3.8%)
- Self-promotion (5.9%)
- Pointless Babble (40.1%)
- Conversational (37.6%)
- Pass-along value (8.7%)



About Twitter: The Tweets (cont.)

News

Any sort of main stream news that you might find on your national news stations such as CNN, Fox or others. This did not include tech news or social media news that you might find on TechCrunch or Mashable.

Spam

These are the tweets such as "See how I got 3,000 followers in one day" type of tweets.

Self-Promotion

These are typical corporate tweets about products, services, or "Twitter only" promos.

Pointless Babble

These are the "I am eating a sandwich now" tweets.

Conversational

These are tweets that go back and forth between folks, almost in ar instant message fashion, as well as tweets that try to engage followers in conversation, such as questions or polls.

Pass-Along Value

These are any tweets with an "RT" in it.

Now, if there were any tweets that could fit **into** more than one category (which was rare), if it started with "@", we deemed it as conversational, even if it was a news item or self-promotion.

Legal Considerations

The text has not been anonymized

Due to Twitters then privacy policy regarding research data, the Twitter data can not leave USNA

Hardware

Used Data and Lore from Naval Academy's CS department

both have 12 core, with hyperthreading

both have 216 GB

Processing was still slow...

Data & Lore

Efficiency tricks

Threading any computations could be threaded

Serialize, compress, and write to disk all large objects

Try to split operations into separate steps to allow for less memory overhead

Cron jobs to run processes at night

Classifier

Maximum Entropy Classifier from Stanford's Java NLP Library

Normalized by the Vector length

- L2 normalization

Really complicated math, but really easy to use

Maximum Entropy Classifier

$$\tilde{p}(x, y) \equiv \frac{1}{N} \times \text{number of times that } (x, y) \text{ occurs in the sample}$$

$$\widetilde{p}(f) \equiv \sum_{x,y} \widetilde{p}(x, y) f(x, y)$$

$$p(f) \equiv \sum_{x,y} \widetilde{p}(x) \boxed{p(y \mid x)} f(x, y)$$

$$p(f) = \widetilde{p}(f)$$

$$\sum_{x,y} \widetilde{p}(x) p(y \mid x) f(x, y) = \sum_{x,y} \widetilde{p}(x, y) f(x, y)$$

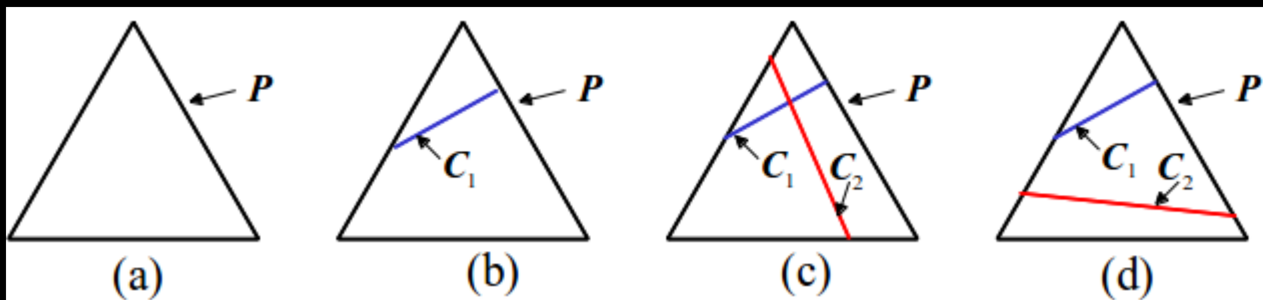
Maximum Entropy Classifier (cont.)

Suppose that we are given n feature functions f_i, which determine statistics we feel are important in modeling the process. We would like our model to accord with these statistics

That is, we would like p to lie in the subset C of P defined by

$$C \equiv \{ p \in P \mid p(f_i) = \tilde{p}(f_i) \quad \text{for } i \in \{1,2,\ldots,n\} \} \qquad (4)$$

Maximum Entropy Classifier (cont.)



C would be ideal but normally not possible

$$p(y \mid x) \geq 0 \quad \text{for all } x, y.$$

$$\sum_y p(y \mid x) = 1 \quad \text{for all } x.$$

- This and the previous condition guarantee that p is a conditional probability distribution

$$\sum_{x,y} \tilde{p}(x) p(y \mid x) f(x, y) = \sum_{x,y} \tilde{p}(x, y) f(x, y)$$

for $i \in \{1, 2, \ldots, n\}$.

- In other words, $p \in \boldsymbol{C}$, and so satisfies the active constraints $\boldsymbol{C}$

Maximum Entropy Classifier (cont.)

$$H(p \equiv) - \sum_{x,y} \tilde{p}(x) p(y \mid x) \log p(y \mid x) \qquad (5)$$

$$\begin{aligned} p^* &= \arg\max_{p \in C} H(p) \\ &= \arg\max_{p \in C} \left(-\sum_{x,y} \tilde{p}(x) p(y \mid x) \log p(y \mid x) \right) \end{aligned}$$

$$\begin{aligned} \xi(p, \Lambda, \gamma) \equiv &- \sum_{x,y} \tilde{p}(x) p(y \mid x) \log p(y \mid x) \\ &+ \sum_i \lambda_i \left(\sum_{x,y} \tilde{p}(x, y) f_i(x, y) - \tilde{p}(x) p(y \mid x) f_i(x, y) \right) \qquad (8) \\ &+ \gamma \left(\sum_y p(y \mid x) - 1 \right) \end{aligned}$$

Maximum Entropy Classifier (cont.)

$$\Psi(\Lambda) \equiv \xi(p^*, \Lambda, \gamma^*) \equiv -\sum_{x,y} \tilde{p}(x)\tilde{p}(y \mid x) \log p(y \mid x)$$

$$= -\sum_{x,y} \left[\tilde{p}(x) \cdot \tilde{p}(y \mid x) \cdot \log\left(\frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(x.y) \right)\right)\right]$$

$$= -\sum_{x,y} \left[\tilde{p}(x) \cdot \tilde{p}(y \mid x) \cdot \left(-\log Z(x) + \sum_i \lambda_i f_i(x.y) \right)\right]$$

$$= -\sum_{x,y} [-\tilde{p}(x) \cdot \tilde{p}(y \mid x) \cdot \log Z(x)] - \sum_{x.y} \left[\tilde{p}(x) \cdot \tilde{p}(y \mid x) \cdot \sum_i \lambda_i f_i(x.y) \right]$$

$$= -\sum_{x} [-\tilde{p}(x) \cdot \log Z(x)] - \sum_{x.y} \left[\tilde{p}(x,y) \cdot \sum_i \lambda_i f_i(x.y) \right]$$

$$= \sum_{x} [\tilde{p}(x) \cdot \log Z(x)] - \sum_{i} \left[\lambda_i \sum_{x,y} \tilde{p}(x,y) \cdot f_i(x.y) \right]$$

$$= \sum_{x} [\tilde{p}(x) \cdot \log Z(x)] - \sum_{i} [\lambda_i \tilde{p}(f_i)]$$

Maximum Entropy Classifier (cont.)

Computing the parameters

Algorithm 1 *Improved Iterative Scaling*

Input : *Feature functions* $f_1, f_2, \cdots f_n$; *empirical distribution* $\tilde{p}(x,y)$

Output : *Optimal parameter values* Λ_i^*; *optimal model* p^*

1. *Start with* $\lambda_i = 0$ *for all* $i \in \{1,2,\cdots,n\}$
2. *Do for each* $i \in \{1,2,\cdots,n\}$:

   a. *Let* $\Delta\lambda_i$ *be the solution to*

   $$\sum_{x,y} \tilde{p}(x) p(y \mid x) f_i(x,y) \exp \boxed{\Delta\lambda_i f^{\#}(x,y)} = \tilde{p}(f_i) \qquad (18)$$

   where $f^{\#}(x,y) \equiv \sum_{i=1}^{n} f_i(x,y)$ (19)

   b. *Update the value of* λ_i *according to* : $\lambda_i \leftarrow \lambda_i + \Delta\lambda_i$
3. *Go to step 2 if not all the* λ_i *have converged*

$$\sum_i \Delta\lambda_i f_i(x,y)$$

Development and Evaluation

- It would be cheating to keep tinkering with features until getting really good numbers
- Leads to overfitting
- Use a dev set, and eval set
- Dev was achieved with 4 fold cross fold validation over 183 days
- Eval set was 10% of total data
  - 21 days, picked at random

Best Results for Gold

character 4 grams, followed by word 2 grams

Cross Folds

64.444%

71.111%

64.444%

62.222%

Average:

65.5525%

Increase: 73

No Increase: 131

Baseline: 64.22%

With a score of 76.190% tied the evaluation set baseline

Best Results for Obama

word 4 grams, followed by word 3 grams

Cross Folds
60.000%
73.333%
71.111%
64.444%

Average:
67.22%

Increase: 68
No Increase: 136
Baseline: 66.67%

Tied the evaluation set baseline at 66.67%

Best Results for DJIA

word 1 grams and 2 grams, followed by word 2 grams

Cross Folds

62.222%

46.667%

60.000%

57.778%

Average:

56.6625%

Increase: 114

No Increase: 90

Baseline: 55.88%

Only achieved 57.143% on the evaluation set with a baseline at 66.67%

Conclusions

Appending the usernames to the bigrams did not help with classification

- maybe due to because the small amount of sampling

Only boring old fashion features work

Questions?