Temporal Information on Twitter

Si Shen

Twitter.com

---

hash-tags, re-tweets

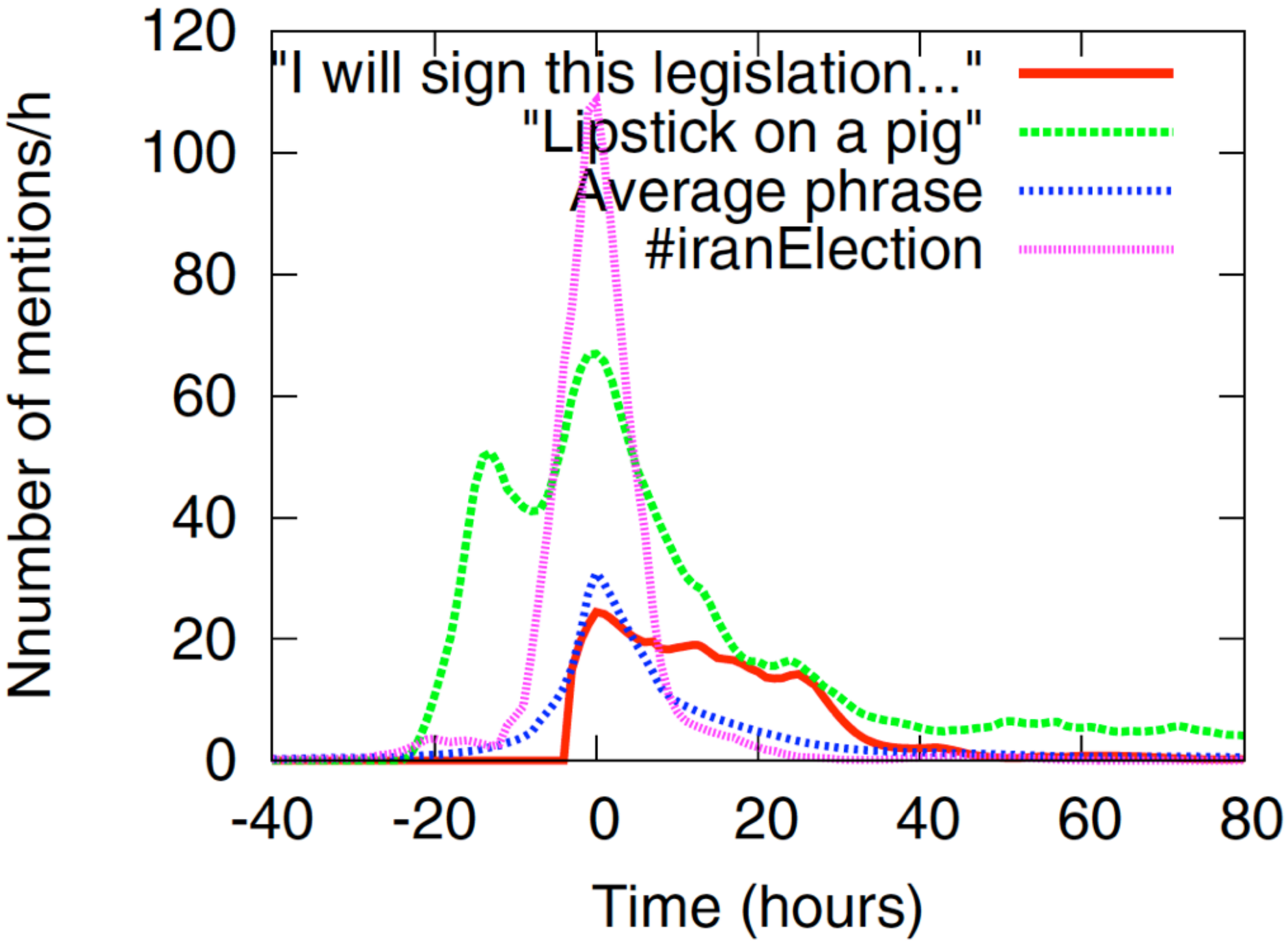
References

- Yang, J., & Leskovec, J. Patterns of temporal variation in online media. (WSDM '11)

- Asur, S., & Huberman, B. A. Predicting the Future with Social Media. (WI-IAT '10)

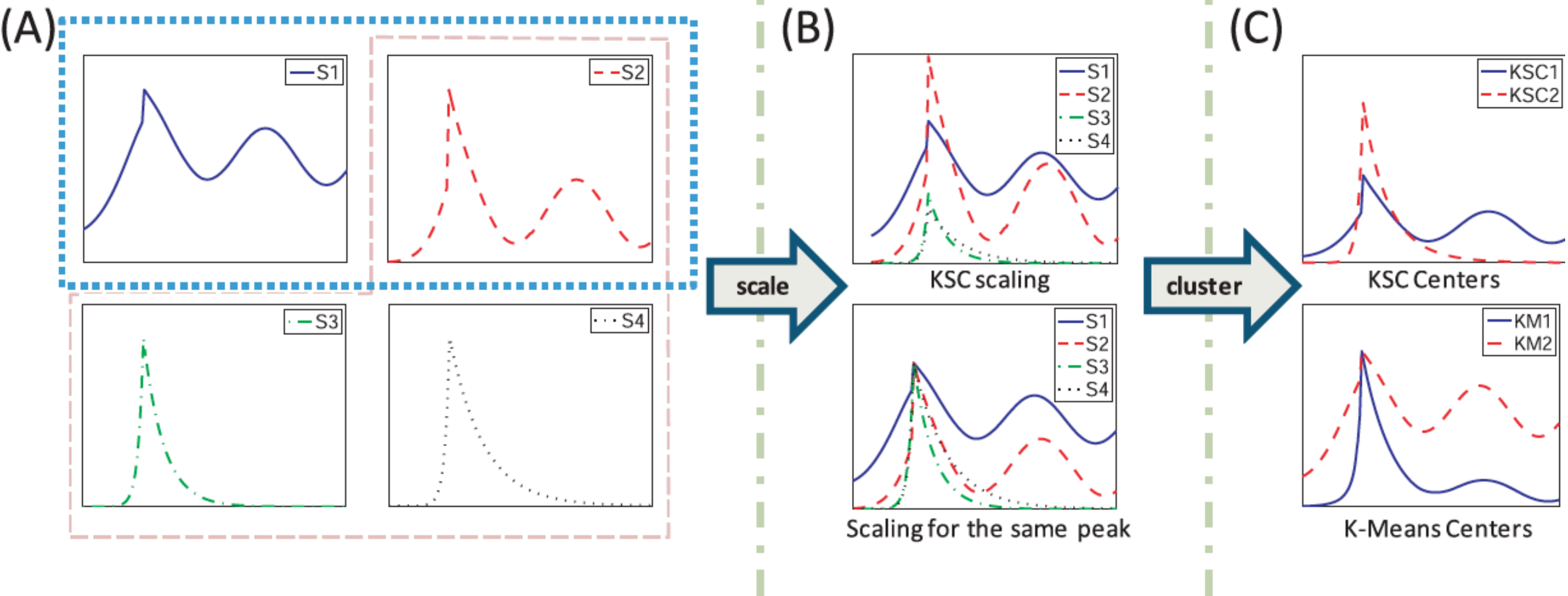
Patterns of Temporal Variation in Online Media

- Problem Definition
- K-Spectral Centroid Clustering
- Experiments



Problem Definition

To find patterns of temporal dynamics invariant to scaling and shifting



The Work-flow

K-means vs K-SC

Distance Metric

- y(q): the result of shifting time series y by q time units

- Closed-form expression to compute the optimal alpha for fixed q

$$\hat{d}(x,y) = \min_{\alpha,q} \frac{||x - \alpha y_{(q)}||}{||x||}$$

- Initialize q’ such that the time series peak at the same time then search for optimal q

$$\mu_k^* = \arg\min_{\mu} \frac{\mu^T M \mu}{||\mu||^2}$$

Cluster Center

Since the new centroid minimizes the spectral norm of M, we call it the Spectral Centroid, and the complete algorithm the K-Spectral Centroid clustering.

for $j = 1$ to K **do** {Refinement step}

$M \leftarrow \sum_{i \in C_j} (I - \frac{x_i x_i^T}{||x_i||^2})$

$\mu_j \leftarrow$ The smallest eigenvector of M

Cluster Quality

- KM-NS: K-means with peak alignment only
- KM-P: K-means with peak alignment and scaling

Method	F (lower is better)	$\sum \hat{d}(\mu_i, \mu_j)^2$ (higher is better)
KM-NS	122.12	2.12
KM-P	76.25	3.94
K-SC	**64.75**	**4.53**

Incremental K-SC

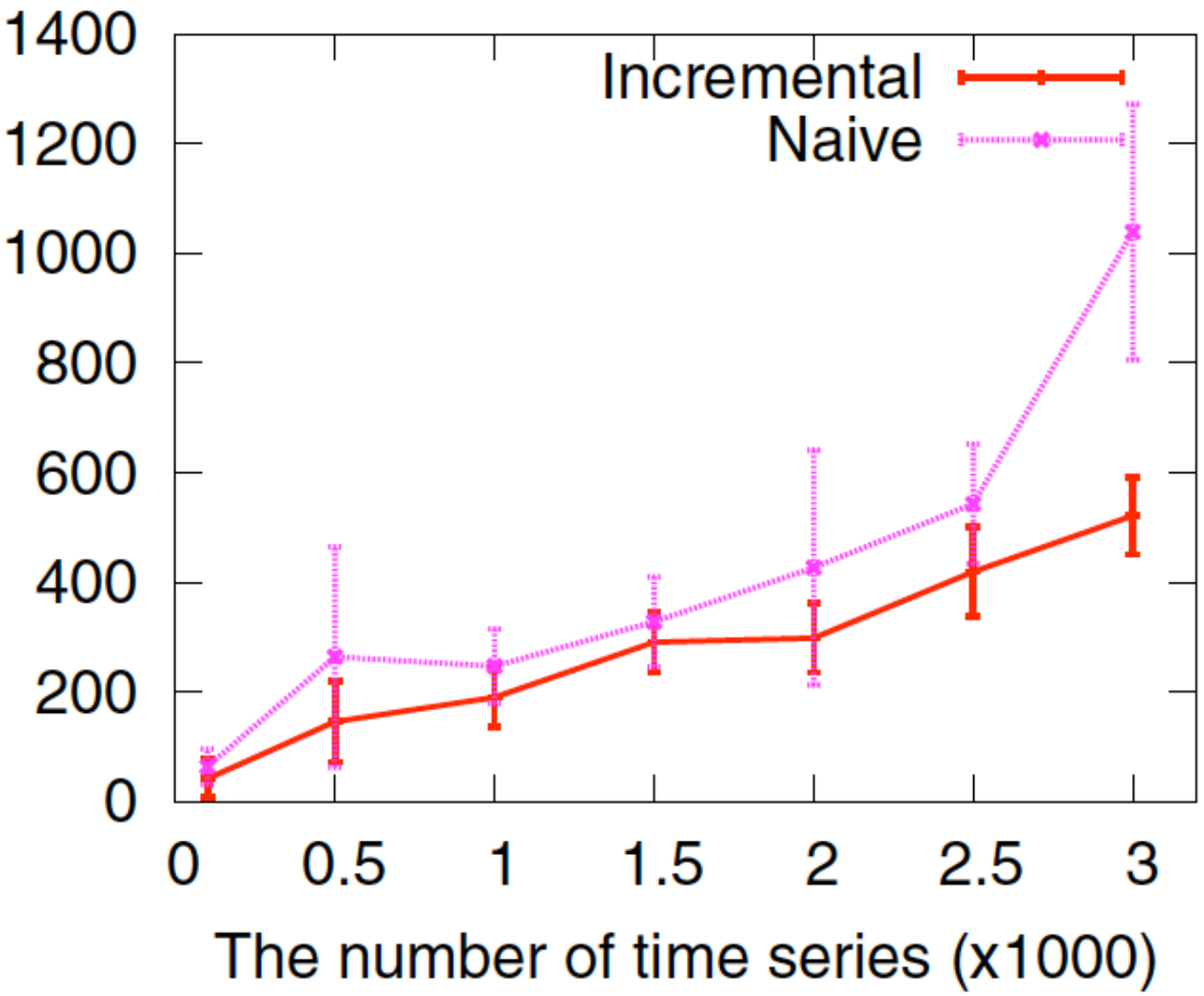
- The algorithm may be very slow if the initial centers are poorly chosen

- Discrete Haar Wavelet Transform

  - Cluster the coarse-grained representations of the time series

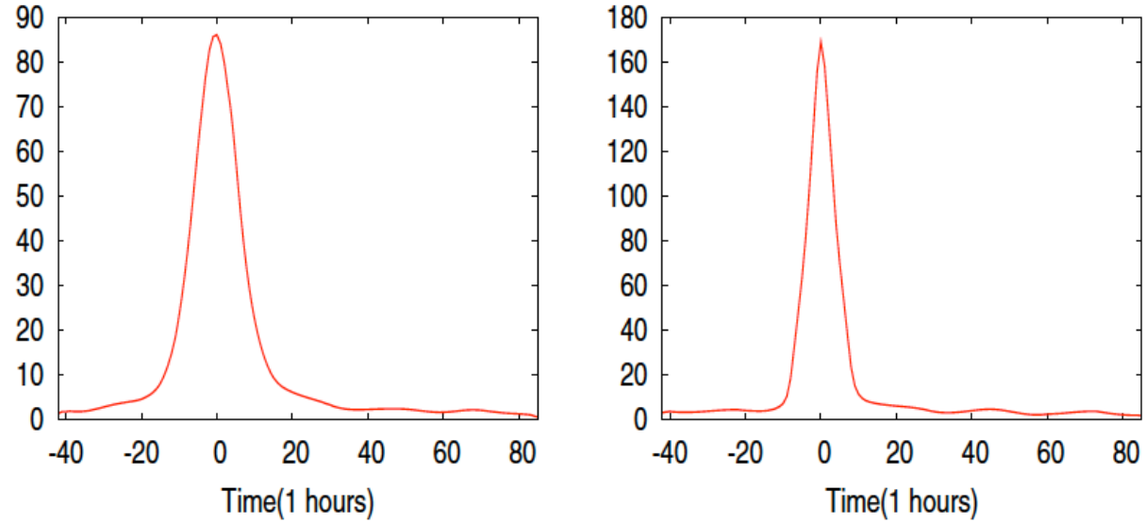
  - Move to the next level of resolution of the time series, initialization based on previous run

  - Repeat this procedure until reaching the full resolution of the time series

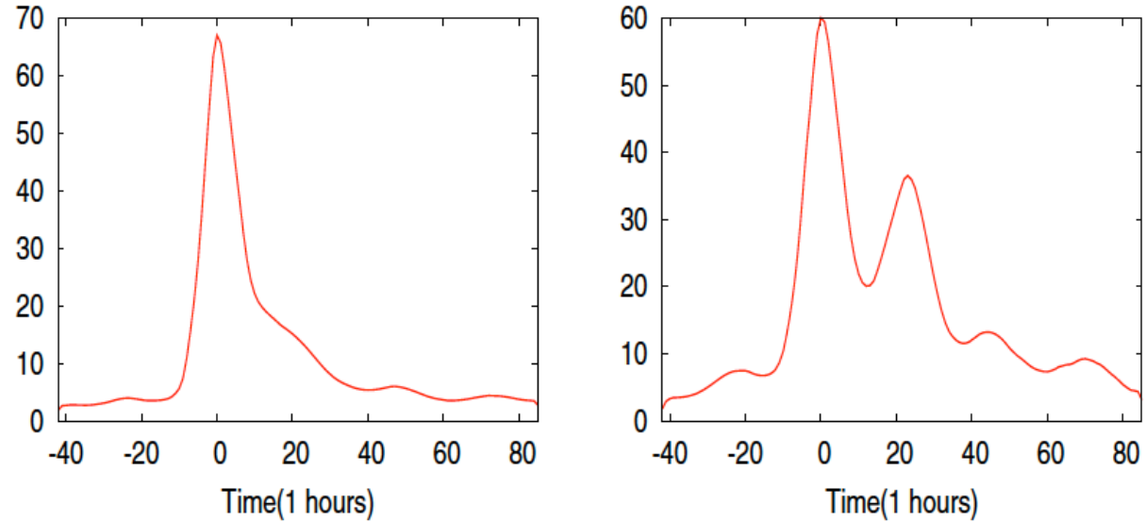


Experiments

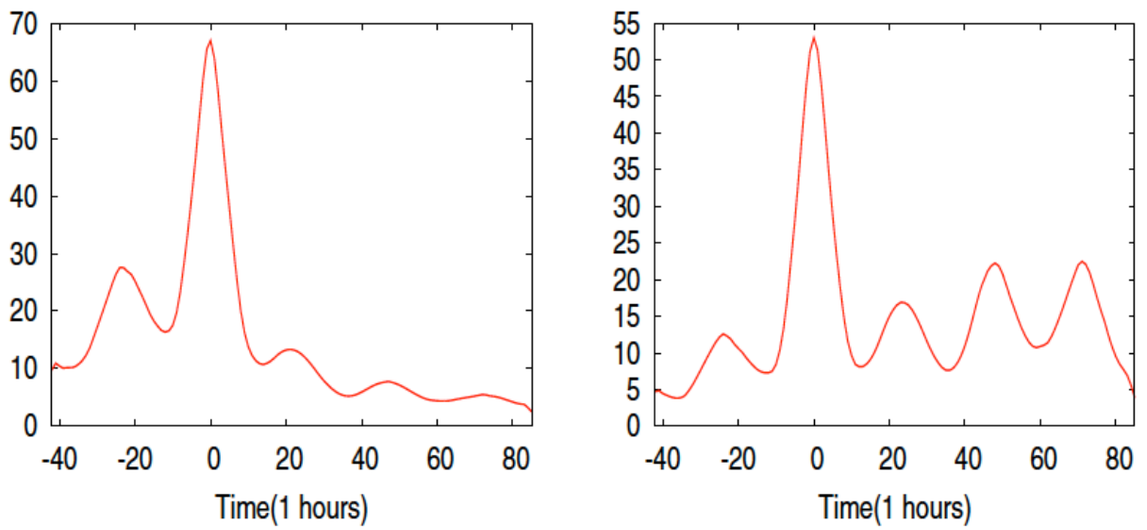
- 1,000 most frequently mentioned hash-tags and 100,000 related most active users from a period of 8 month
- 1 hour time unit and time series of 128 hours around the peak volume, peak located at 1/3 of the 128 hours
- K = 6



(a) Cluster T1 (b) Cluster T2



(c) Cluster T3 (d) Cluster T4



(e) Cluster T5 (f) Cluster T6

Number of features	50	100	200	300
Temporal features	69.53%	78.30%	88.23%	95.35%
Volume features	66.31%	71.84%	81.39%	92.36%
TF-IDF features	64.17%	70.12%	79.54%	89.93%

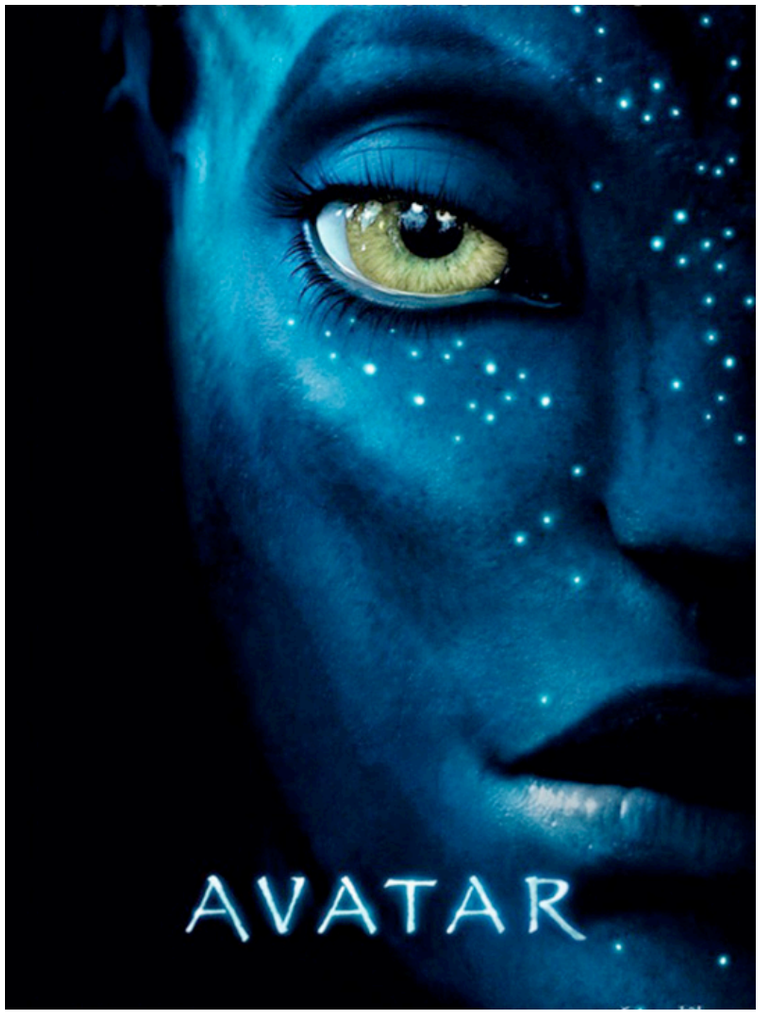
Classification Accuracy

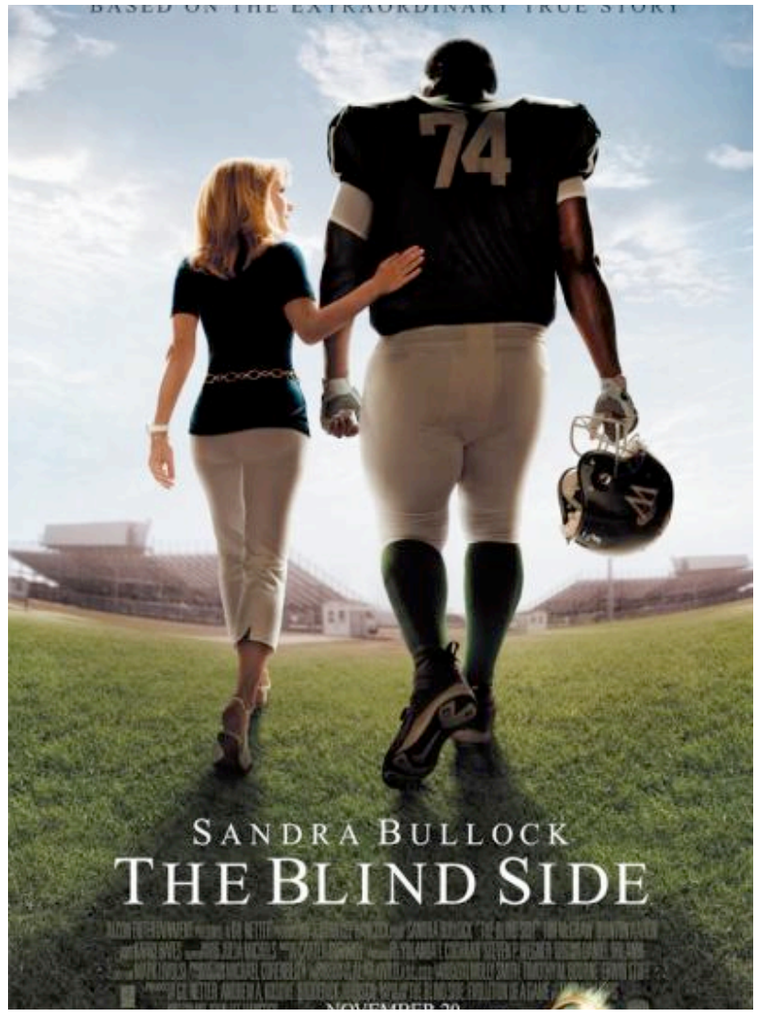
Timing is important

References

- ~~Yang, J., & Leskovec, J. Patterns of temporal variation in online media. (WSDM '11)~~

- Asur, S., & Huberman, B. A. Predicting the Future with Social Media. (WI-IAT '10)



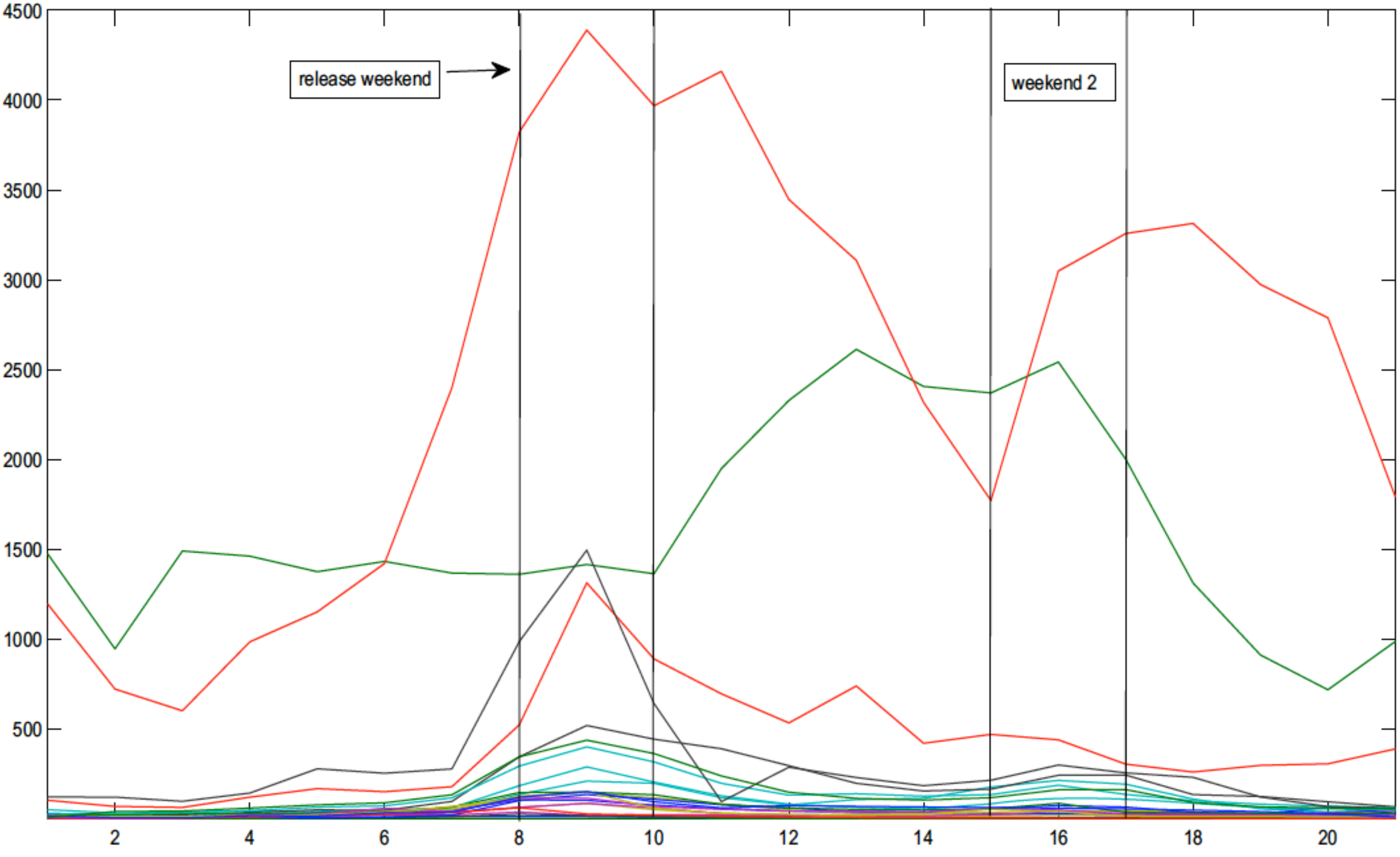


Predicting Box-office Revenue from Twitter

Data

2.89 million tweets referring to 24 different movies period of 3 months (Nov-Feb) from 1.2 million users

Movie	Release Date
Armored	2009-12-04
Avatar	2009-12-18
The Blind Side	2009-11-20
The Book of Eli	2010-01-15
Daybreakers	2010-01-08
Dear John	2010-02-05
Did You Hear About The Morgans	2009-12-18
Edge Of Darkness	2010-01-29
Extraordinary Measures	2010-01-22
From Paris With Love	2010-02-05
The Imaginarium of Dr Parnassus	2010-01-08
Invictus	2009-12-11
Leap Year	2010-01-08
Legion	2010-01-22
Twilight : New Moon	2009-11-20
Pirate Radio	2009-11-13
Princess And The Frog	2009-12-11
Sherlock Holmes	2009-12-25
Spy Next Door	2010-01-15
The Crazies	2010-02-26
Tooth Fairy	2010-01-22
Transylmania	2009-12-04
When In Rome	2010-01-29
Youth In Revolt	2010-01-08



Critical Period

Regression Variables

- Promotional Material: Urls and Retweets
- Rate of Tweet Mentions
- Sentiment Ratio

Features	Week 0	Week 1	Week 2
url	39.5	25.5	22.5
retweet	12.1	12.1	11.66

Features	Correlation	R^2
url	0.64	0.39
retweet	0.5	0.20

Urls and Retweets

Opening Weekend

- Tweet-rate: the number of tweets referring to a particular movie per hour
- thcnt: number of theaters the movies were released in
- R square of earlier work is 0.788
- Avatar (1212.8/hr): 77M
- New Moon (1365.8/hr): 142M

Features	Adjusted R^2
Avg Tweet-rate	0.80
Tweet-rate timeseries	0.93
Tweet-rate timeseries + thcnt	**0.973**
HSX timeseries + thcnt	0.965

$$Rev(mov) = \beta_0 + \sum \beta_i * Tweet\text{-}rate_i(mov) + \beta_{th} * thcnt$$

Hollywood Stock Exchange

Second Weekend

- PNRatio = # of positive tweets / # of negative tweets

- The Blind Side (5.02 to 9.65): 34M to 40.1M

- New Moon (6.29 to 5): 142M to 42M

Predictor	Adjusted R^2
Avg Tweet-rate	0.79
Avg Tweet-rate + thcnt	0.83
Avg Tweet-rate + PNratio	0.92
Tweet-rate timeseries	0.84
Tweet-rate timeseries + thcnt	0.863
Tweet-rate timeseries + PNratio	**0.94**

Thank You