# Cannot Predict Comment Volume of a News Article before (a few) Users Read It

**Lihong He, Chen Shen**
Department of Computer & Information Sciences
Temple University

**Arjun Mukherjee**
Department of Computer Science
University of Houston

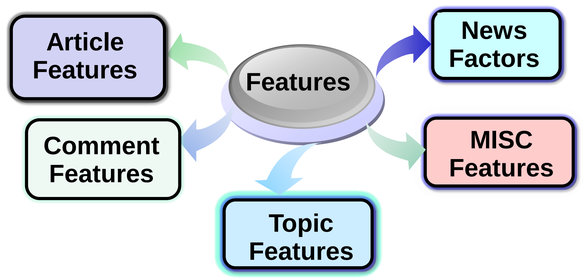
**Slobodan Vucetic, Eduard Dragut**
Department of Computer & Information Sciences
Temple University

## Motivation

- What?
  - Predict comment volume of a news article
- Why?
  - Disagreement between previous works and communication community.
  - Previous Works
    - Features related to article alone work
  - Communication Community
    - Comment features should be considered
- How?
  - Build different feature groups for comparison

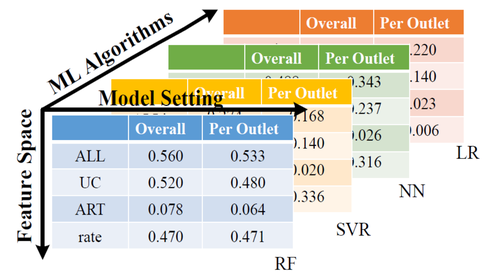
## Features

- Five Feature Groups
  - Understand the importance of different feature group in the prediction task




## Methodology

- Model Setting
  - Global Model
  - Local Model
- ML Algorithm
  - RF, SVR, NN
  - LR
- Feature Space
  - **ALL** Features
  - Comment Features (**UC**)
  - Article Features (**ART**)
  - **Rate** = $^{\alpha}/_{t_\alpha - t_1}$




## Dataset

- ~20K articles from 6 outlets
  - **Washington Post**: 6,470
  - **Daily Mail**: 6,046
  - **Wall Street Journal**: 2,516
  - **Fox News**: 1,739
  - **The Guardian**: 1,697
  - **New York Times**: 965
- 465 comments per article in average

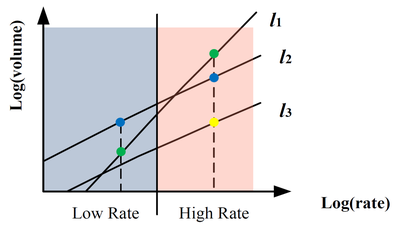
## Experimental Results

- $R^2$ on global setting
  - Comment features are important
  - Linear ML algorithm cannot solve problem
  - rate: dominant single feature

| | RF | SVR | NN | LR |
|---|---|---|---|---|
| ALL | 0.560 | 0.472 | 0.499 | 0.413 |
| Comment Features | 0.520 | 0.479 | 0.502 | 0.400 |
| Article Features | 0.078 | 0.021 | 0.016 | 0.020 |
| rate | 0.470 | 0.465 | 0.459 | 0.370 |

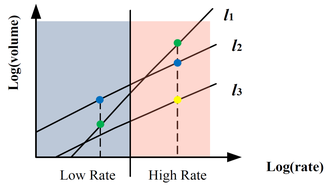
## Rate Model

- Linear fit lines
  - Cross each other: $l_1$, $l_2$
  - In parallel : $l_2$, $l_3$
- rate → volume
  - Consider $l_1$ and $l_2$
    - Low rate area: larger volume in $l_2$
    - High rate area: larger volume in $l_1$
  - Consider $l_2$ and $l_3$
    - Always larger volume in $l_2$



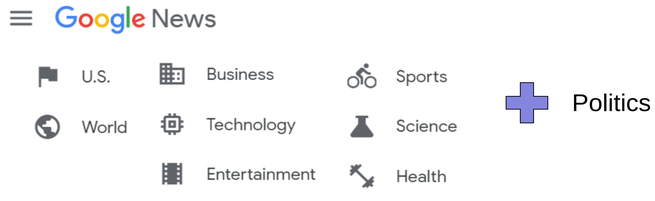

## Rate Model Study

- Study across Outlets
  - $l_1$: Fox News
  - $l_2$: New York Times
  - $l_3$: Daily Mail



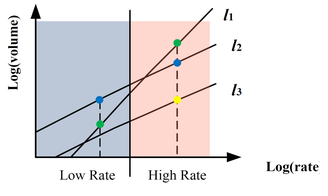

## Data Split by Categories

- Split articles into 9 categories




## Rate Model Study

- Study across Categories
  - $l_1$: Politics
  - $l_2$: US
  - $l_3$: Business




## Acknowledgement

This work is supported by U.S. NSF Award **#1838145, #1838147, and #2026513**

## References

[Tsagkias09] Tsagkias M, Weerkamp W, De Rijke M. **Predicting the volume of comments on online news stories**. In CIKM 2009.

[Balali17] Balali A, Asadpour M, Faili H. **A supervised method to predict the popularity of news articles**. In Computación y Sistemas, 2017.

## References

[Backstrom13] Backstrom L, Kleinberg J, Lee L, Danescu-Niculescu-Mizil C. **Characterizing and curating conversation threads: expansion, focus, volume, re-entry**. In WSDM, 2013.

[Cheng14] Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. **Can cascades be predicted?**. In WWW, 2014..