Exploration of Similarity Calculation to Enhance Transductive Learning

Presented by Lin Gong

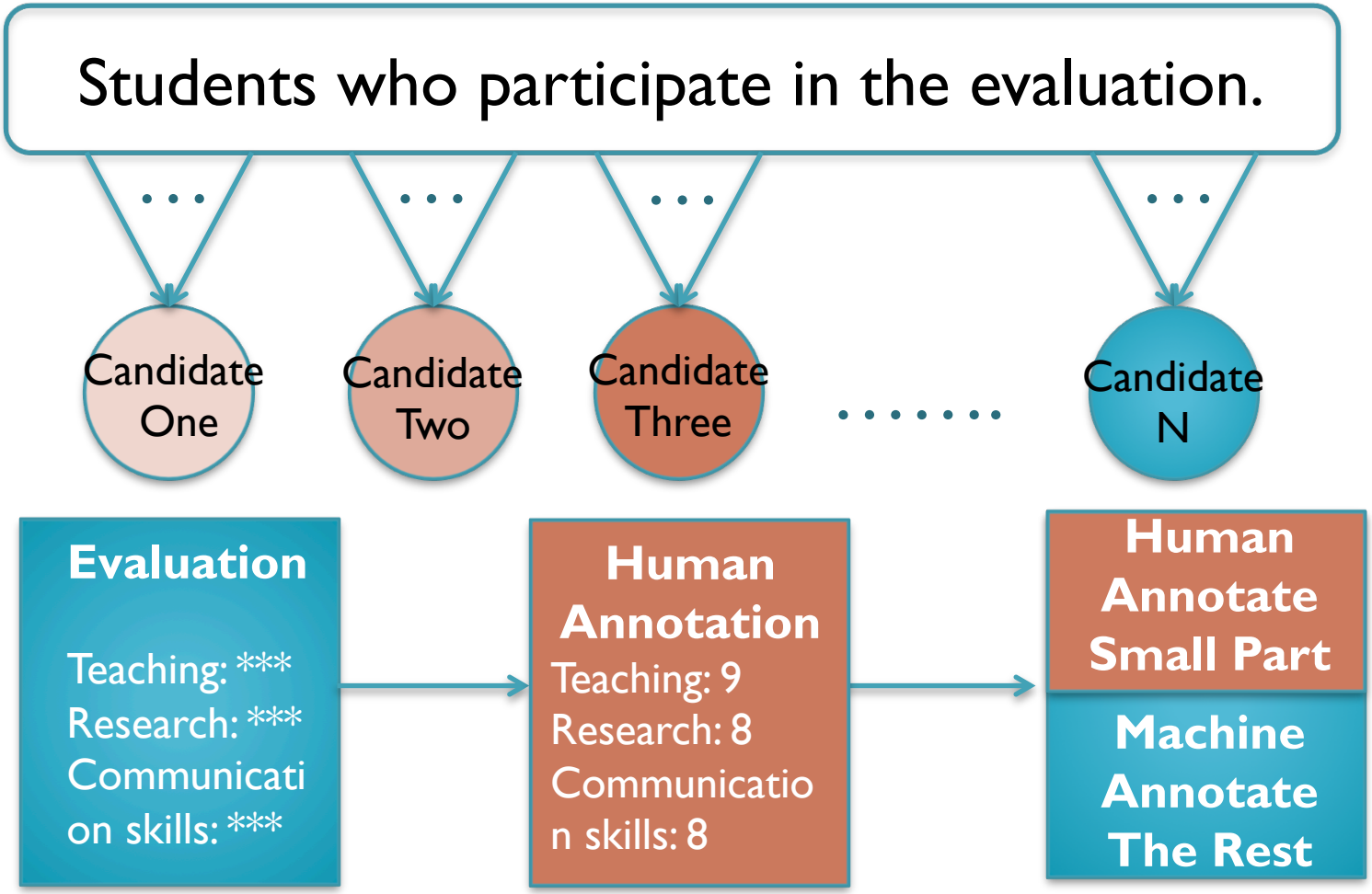
May 2nd, 2015

Roadmap

- Introduction
  - What? Why?
- Related Work
- Methodology
  - Metric Learning
  - POS Tagging
- Experimental Results
- Further Work

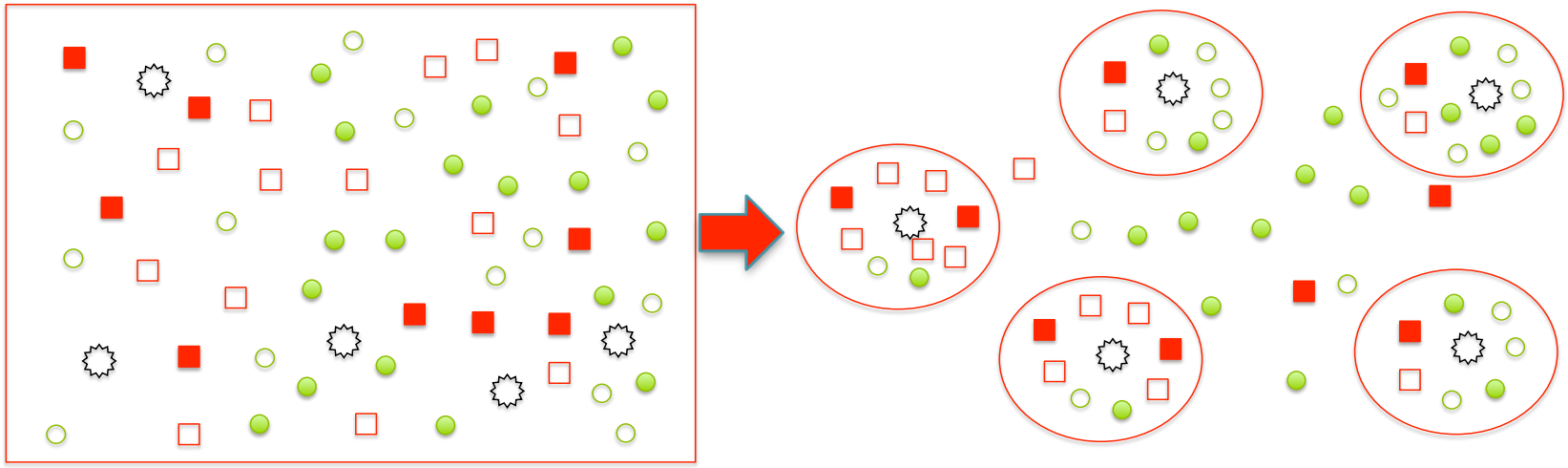
What?Why? Transductive learning?

- Problem: Recruit new faculty members.



What? Why? Tranductive Learning?

- Use both labeled and unlabeled data
  - ■ vs ● : different classes
  - ■ vs □ : labeled and unlabeled samples
  - ✹ : target

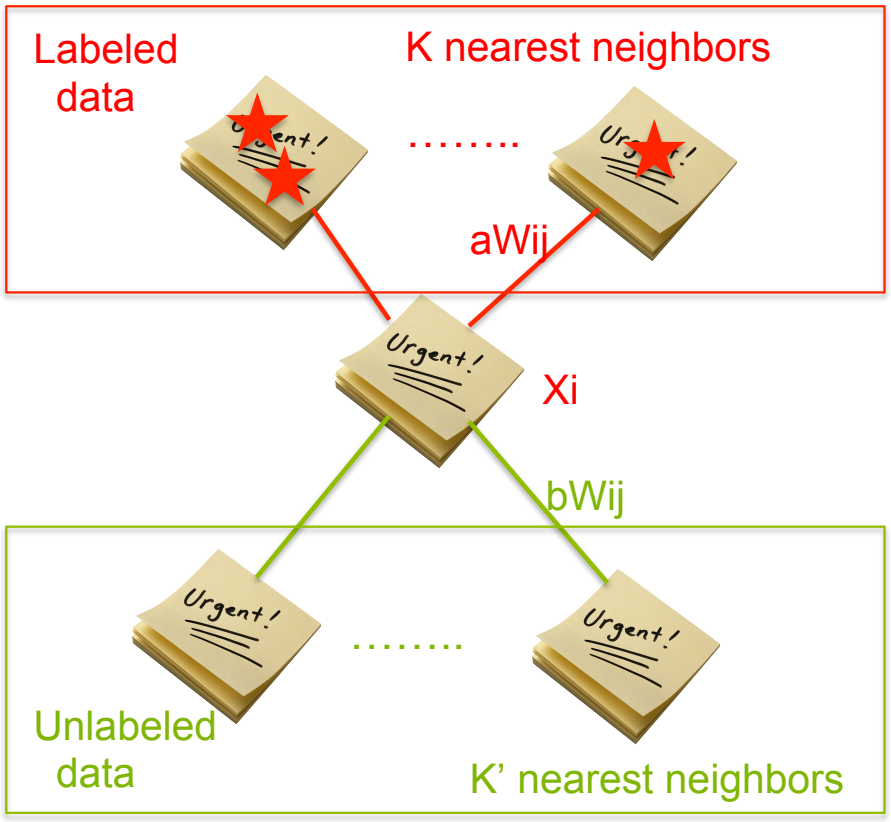


Related Work

- Different transductive learning alogrithms:
  - Spectral methods, random walks, etc.

- Similarity Calculation:
  - Cosine similarity
  - PSP: positive sentences percentage
  - Define a new similarity matrix.

Methodology-Transductive Learning



- Take the average of the neighbors.
- Iterate until it converges.
- Similarity plays an important role.

Methodology-Metric Learning

- Information Theoretic Metric Learning(ITML)

$$d_A(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T A(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

- Measure distance by relative entropy

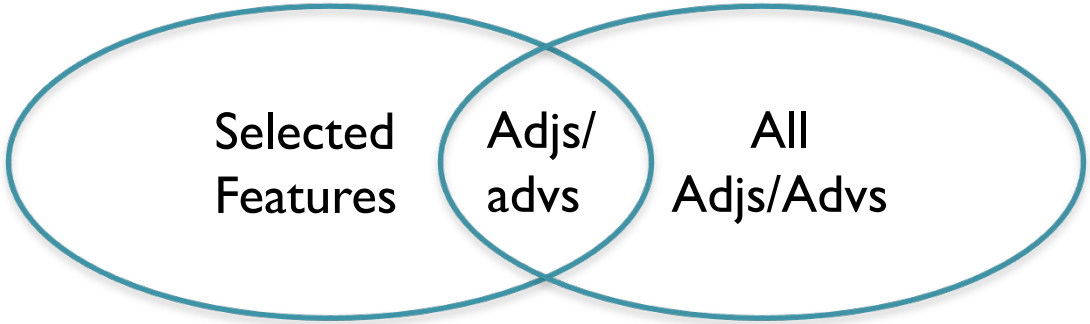
$$KL(p(x; A_0) \parallel p(x; A_0)) = \int p(x; A_0) \log \frac{p(x; A_0)}{p(x; A)} dx.$$

- Add similar/dissimilar pairs as constraints

$$\begin{aligned} \min_{A} \quad & KL(p(x; A_0) \parallel p(x; A)) \\ \text{subject to:} \quad & d_A(x_i, x_j) \leq u \quad (i, j) \in S, \\ & d_A(x_i, x_j) \geq l \quad (i, j) \in D. \end{aligned}$$

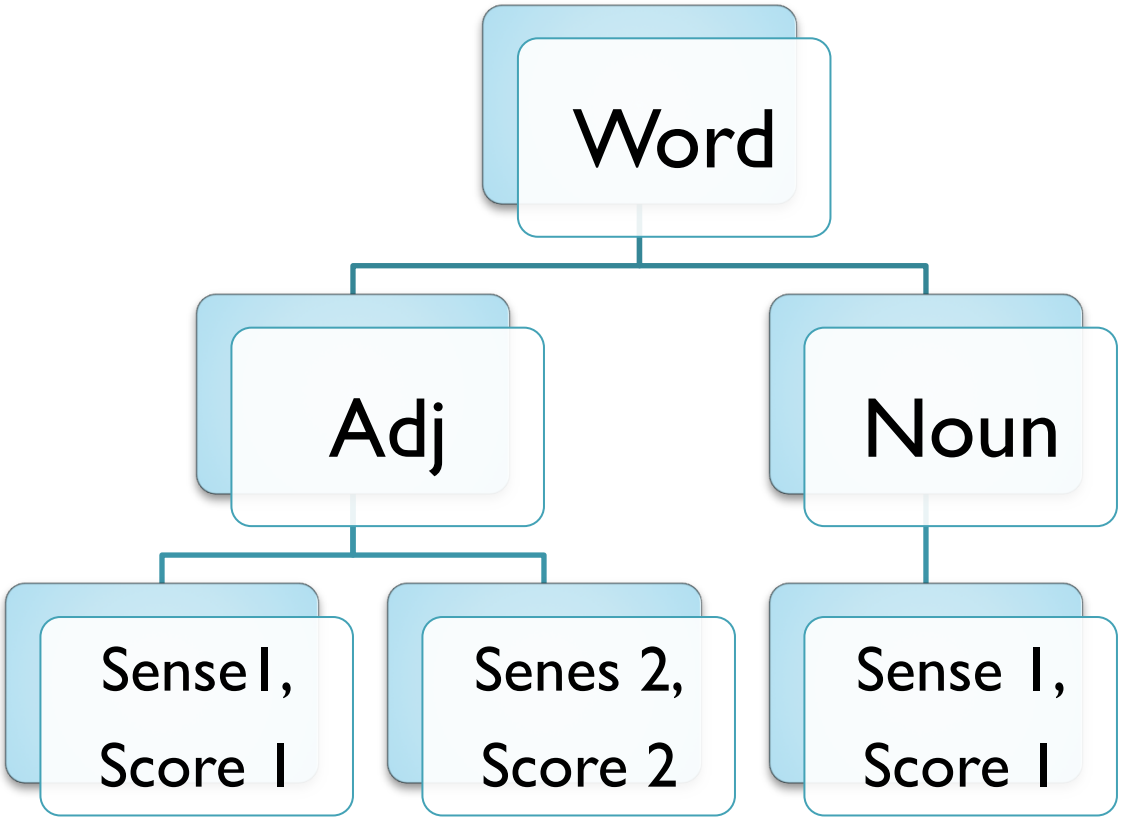
Methodology-POS Tagging

- Extract adjs/advs



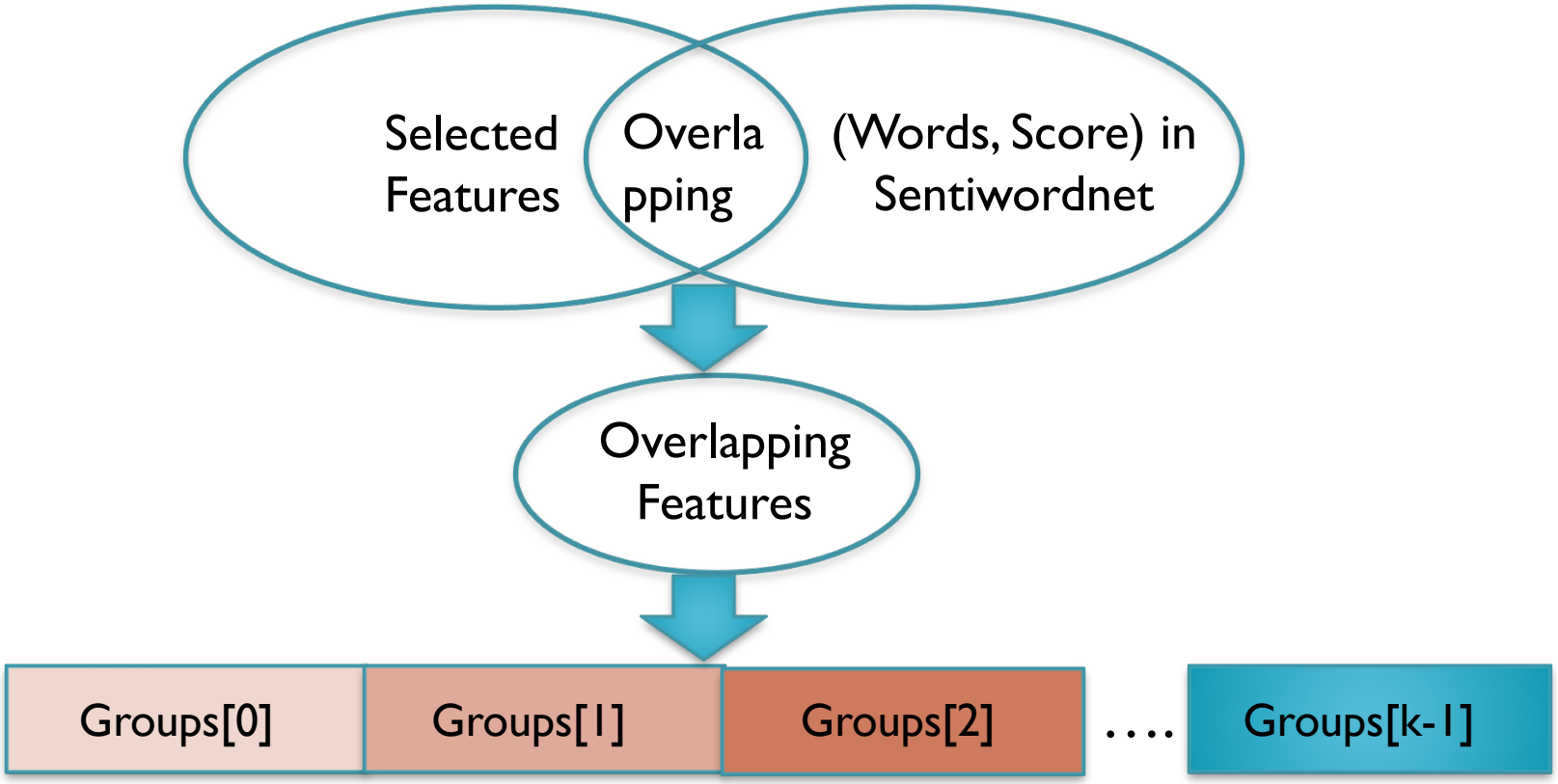
Methodology-POS Tagging

- Extract words shown in sentiwordnet
  - Word in Sentiwordnet:



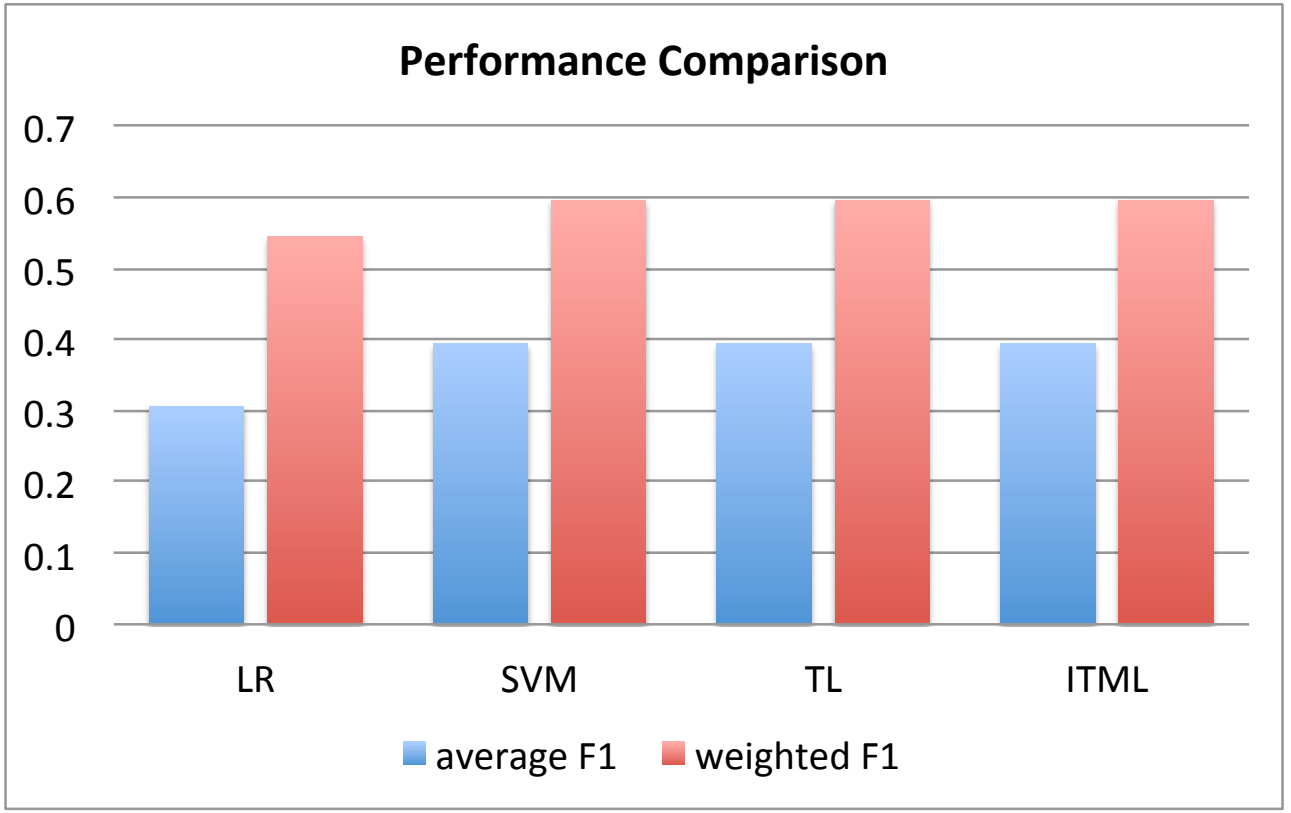
Methodology-POS Tagging

- Group them into k groups



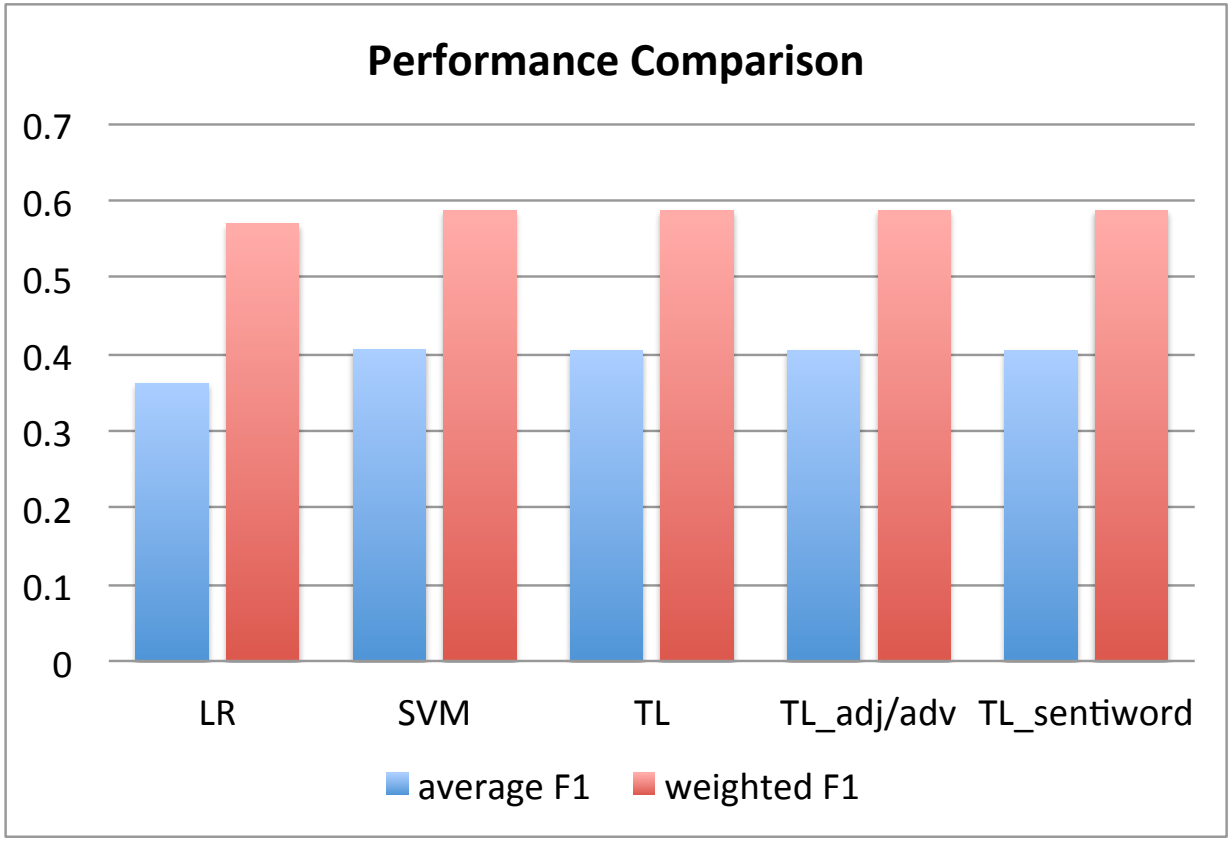
Experimental Results

- ITML: 5-class classification
  - Setup:2692 Amazon reviews from tablets.
  - 1844 features selected from CHI.



Experimental Results

- POS Tagging
  - Setup: 17073 Amazon reviews from tablets.
  - 5608 features selected from CHI.



Experimental Analysis

- Metric Learning
  - ITML
    - Computation Limitation.
- POS Tagging
  - Adj/advs:
    - Overlapping are sparse.
  - Sentiwordnet:
    - Incomplete words list.
    - Score is not representative.

Further Work

- Optimize the computation of ITML.
- Find out consistent patterns among reviews and ratings.
- Develop new ways of representing users' sentiment.

Thanks!

Q&A