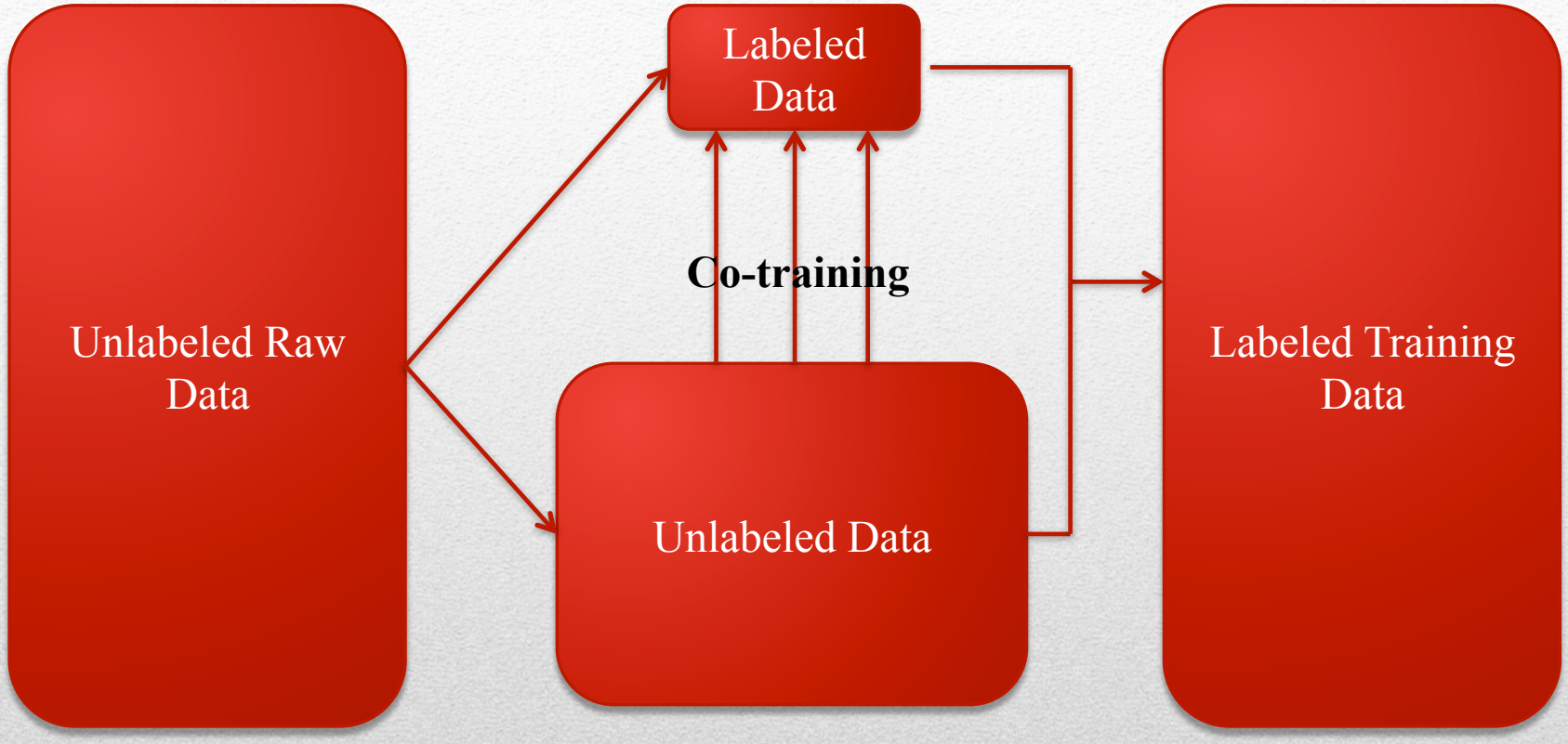
# wPod

Weibo Public Opinion (Polarity) Detection

Haotian He & Sanae Sato

- Microblog is always updating, and new issues and terms come out. There are not many valuable handy labeled data.
- Co-training is a semi-supervised learning technique that provides a good way for such a case.
- Co-training employs two classifiers, with two sets of features separately, in a loop to label all the unlabeled examples. Each classifier takes turns to select the most confidently predicted examples and add these into the training set. Both classifiers then re-learn on the enlarged training set so that they take into account the newly added data.

# Algorithm




# Co-training Architecture

# Dataset

- Microblogs of 2565 users from the signed-up day of each user till March 15, 2014, more than 3 million in total.
- Randomly picked up 330 microblogs to manually label the polarity as the initial training dataset.
- Co-trained for loops to reach a training dataset as 65,000 microblogs.
- Microblogs for test dataset are extracted from March 3 to 5 during the NPC (The House) and CPPCC (Senate) annual joint conferences in 2014, and divided to 12 different categories according to the key words.
- Manually labeled two test datasets for evaluation.

| Category | Key Words | Count |
|---|---|---|
| Report of the Government (pre) | Premier, State Council, Report | 1762 |
| Report of the Supreme Court（spc) | Supreme Court, Report | 315 |
| Report of the Supreme Procuratorate (spp) | Supreme Procuratorate, Report | 158 |
| Education Equality (edu) | Education Equality | 123 |
| Second Child (sch) | Second Child | 210 |
| Environment, Air pollution (env) | Environment, pollution, PM2.5 | 4439 |
| Anti-corruption (cor) | Anti-corruption, … | 1062 |
| Medical System Reform (med) | Medical Reform | 456 |
| Public Funding Usage (pfu) | Public Funding Usage | 168 |
| State-Owned Enterprise Reform (ser) | State-Owned Enterprise Reform | 922 |
| Real Estate (hou) | Real Estate, Price, … | 7444 |
| Food Safety (fst) | Food Safety | 2333 |

# Dataset

# Features and Classifier

- Features:
  - Set 1: Unigram + Bigram + Polarity Words
  - Set 2: Trigram + Emoticon
- Classifier:
  - Naïve Bayes
  - Maximum Entropy (better performance / final choice)

| | Accuracy |
|---|---|
| Education Equality | 0.8048780487804879 |
| Second Child Policy | 0.8333333333333334 |

# Evaluation

| | Accuracy |
|---|---|
| Education Equality | 0.8048780487804879 |
| Second Child Policy | 0.8333333333333334 |

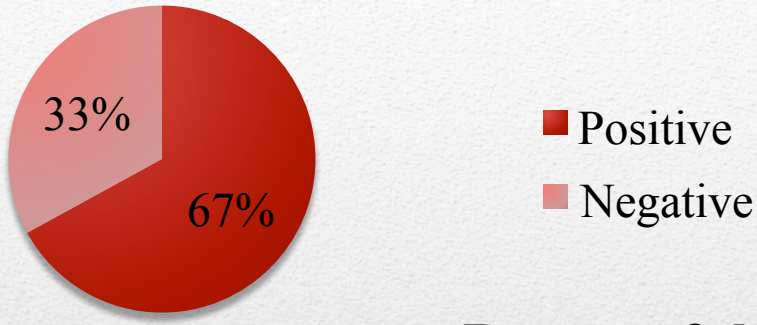
**Top 3 predicted negative microblogs for Second Child Policy:**

| Microblog text | Translation |
|---|---|
| 2013年两会也是马旭这只计生狗大言不惭的讲到单独二胎不会一下子放开。 | Ma Xu, this dog official, shamelessly said the second child policy would not be released in 2013 NPC and CPPCC. |
| 计生委早死。老百姓就不会死。计生委不死。老百姓就会断子绝孙 | If Family Planning Office is shut down, people would survive. If it is not, people would die without sons. |
| 反人类废除计划生育2013:请记住这计划生育利益集团代言人的丑陋嘴脸。 | Repeal the antihuman family plan 2013: Please remember the ugly face of the Family Plan interest group. |

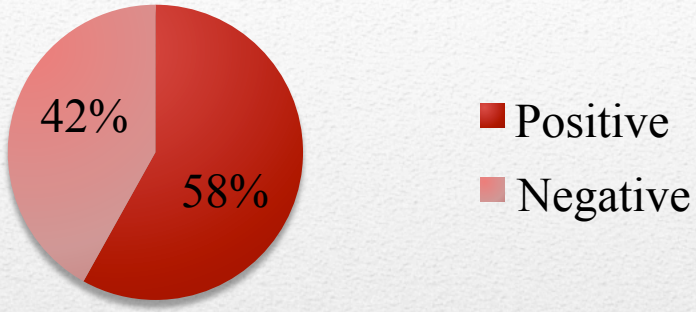
# Evaluation

# Results – Three Reports

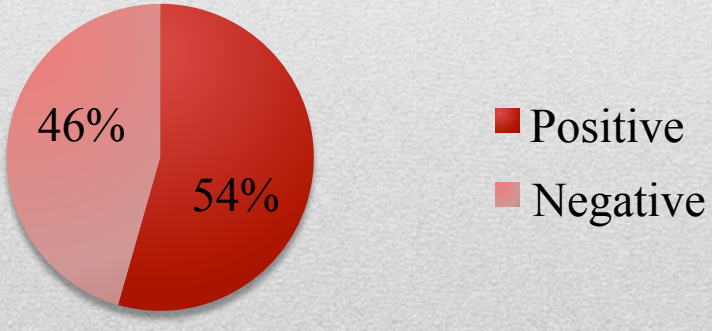
**Report of the Government Work**

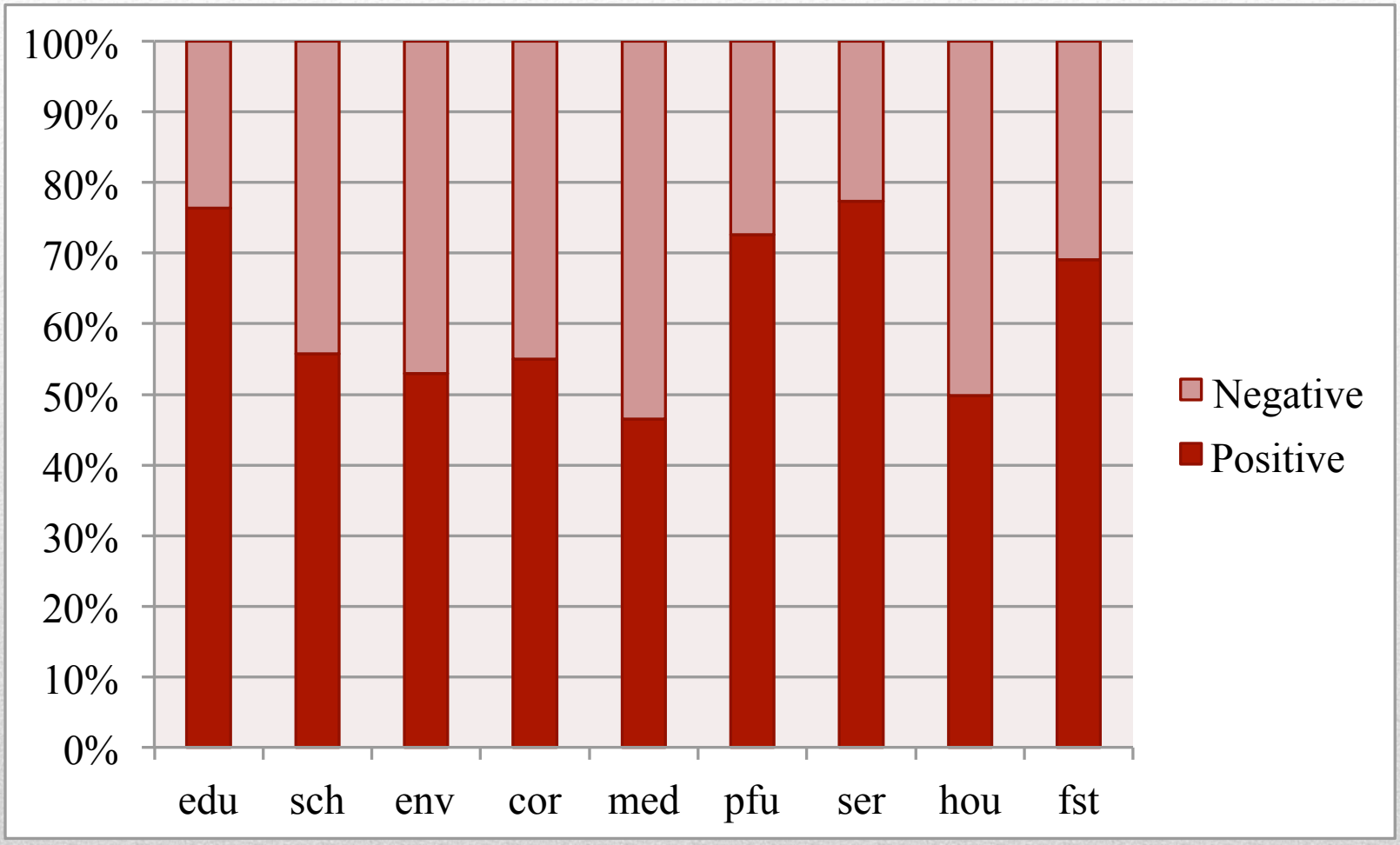



**Report of the Supreme People's Court**




**Report of the Superme People's Procuratorate**






# Results – Popular Issues

# Future Work

- Want to explore diachronic changes on the same issues. But currently did not have time to extract the previous years' Weibo data.
- Not only add more features, but add more sentiments as happy, sad, or angry to the system.
- If in the future only do the political analysis, should specify the training data to be related topics, instead of all the general data.
- Try SVM classifier.