McCormick
Northwestern Engineering
Department of Electrical Engineering and Computer Science

Twitter Trending Topic Classification

Kathy Lee, Diana Palsetia, Ramanathan Narayanan,
Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary

ICDM 2011 Workshop on Optimization Based Methods for Emerging Data Mining Problems (OEDM'11)

Motivation

- Information explosion
  - 200 million tweets per day*
- Twitter provides trending topics
  - Most popular topics that people tweet about
- What is this trending topic about?
  - Hashtags, name of individual, words in other language, etc
  - Is this person a musician, artist, politician, or a sport man?

Trending Topics

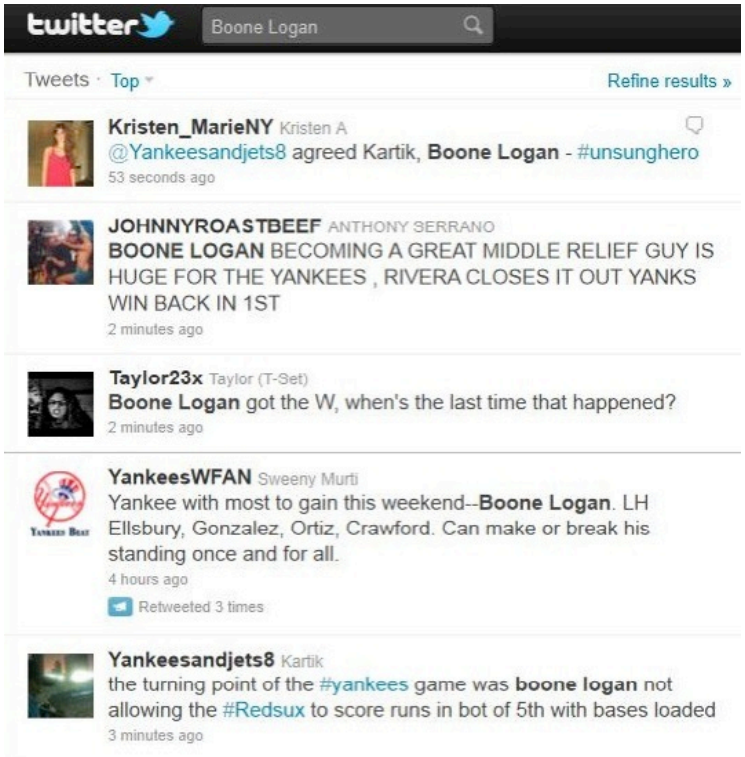
Trends: United States trends
Boone Logan
#MyYearofVIP
Barrett Jones
Outland
#itsalwayssunny
Ed Hochuli
Vaseline
Brett Keisel
#beyondsaredstraight
Gail Kim

* http://www.marketinggum.com/twitter-statistics-2011-updated-stats/

Extended Motivation

Trending Topics

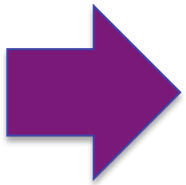
Trends: United States trends
Boone Logan
#MyYearofVIP
Barrett Jones
Outland
#itsalwayssunny
Ed Hochuli
Vaseline
Brett Keisel
#beyondsaredstraight
Gail Kim



Our Goal: Classify Trending Topics

Trending Topics

Trends: United States trends
Boone Logan
#MyYearofVIP
Barrett Jones
Outland
#itsalwayssunny
Ed Hochuli
Vaseline
Brett Keisel
#beyondsaredstraight
Gail Kim



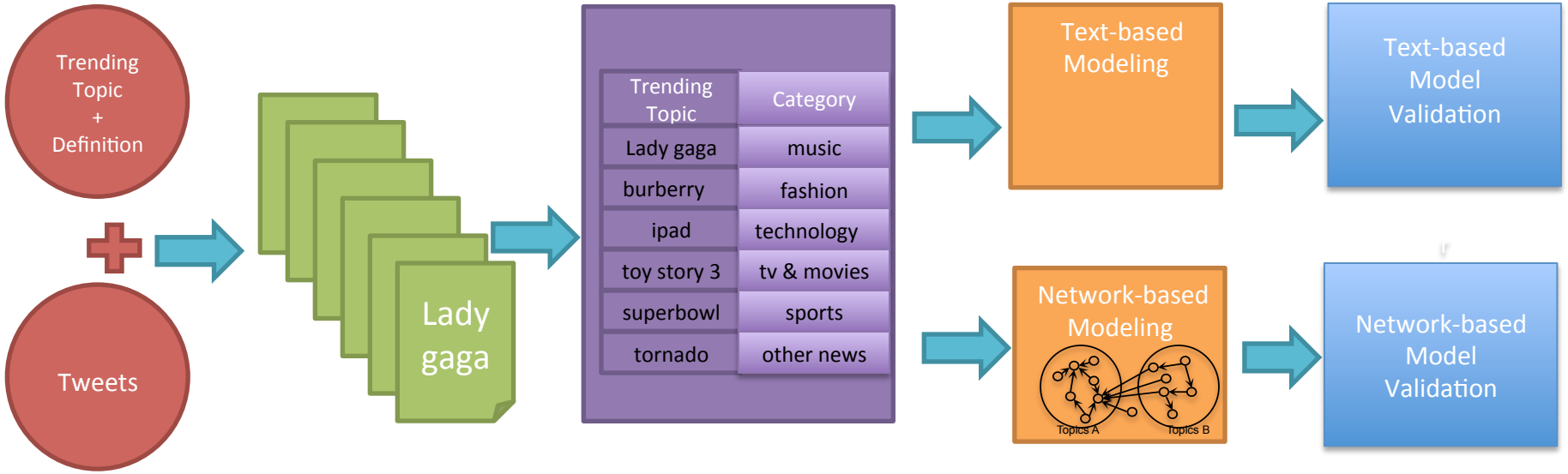
General Categories

- Business
- Health
- Music
- Politics
- Sports
- Science
- Technology

.
.
.

- Motivation
- **Method Overview**
- Data Set
- Methods
- Results
- Conclusion

System Architecture



- Motivation
- Method Overview
- **Data Set**
- Methods
- Results
- Conclusion

Building Training Set

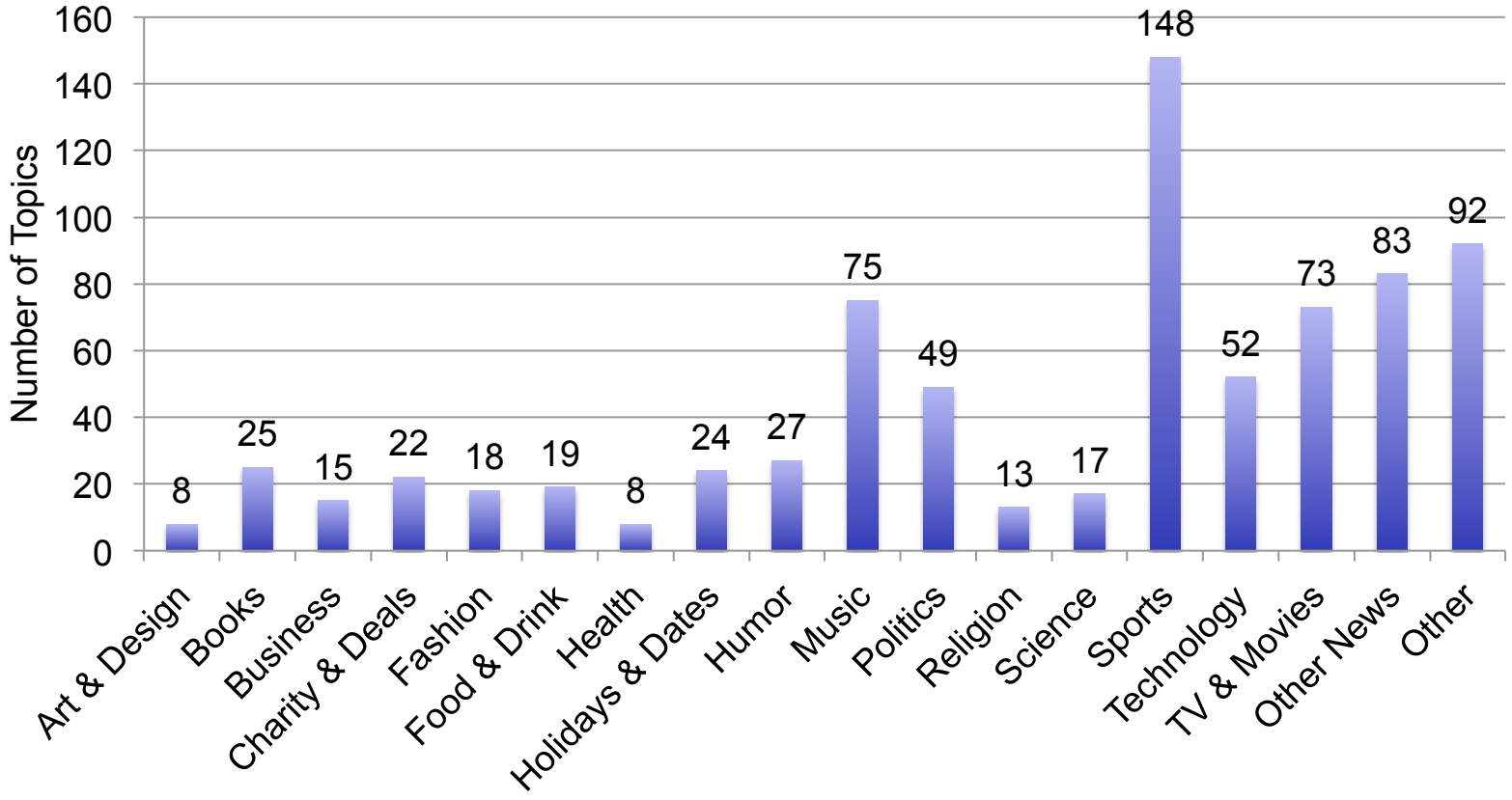
- 23000 trending topics

  (topics trended February 2010 – July 2011)
- Downloaded trend definition and tweets while each of 23000 topics was trending
- Random subset of 1000 topics
- Removed topics without trend definitions

Labeling

- 2 annotators labeled each topic
- 3rd annotator intervened in case of disagreement
- Removed topics that were labeled differently by all 3 annotators
- 768 trending topics in final training set
- Find 5 similar topics to 768 topics
- Labeled 3005 topics in total

Distribution of training data



- Motivation
- Method Overview
- Data Set
- **Methods**
  - **Text-based classification**
  - Network-based classification
- Results
- Conclusion

Document

Text-based data classification

- Bag-of-Words Text Classification

1. Preprocessing
   - Remove hyperlinks
2. Apply string-to-word vector filter
   - Remove symbols and stop words
   - Transform tokens into TF-IDF (term-frequency inverse-document-frequency) weight
3. Apply various classification models
   - Naïve Bayes, Naïve Bayes Multinomial, and SVM

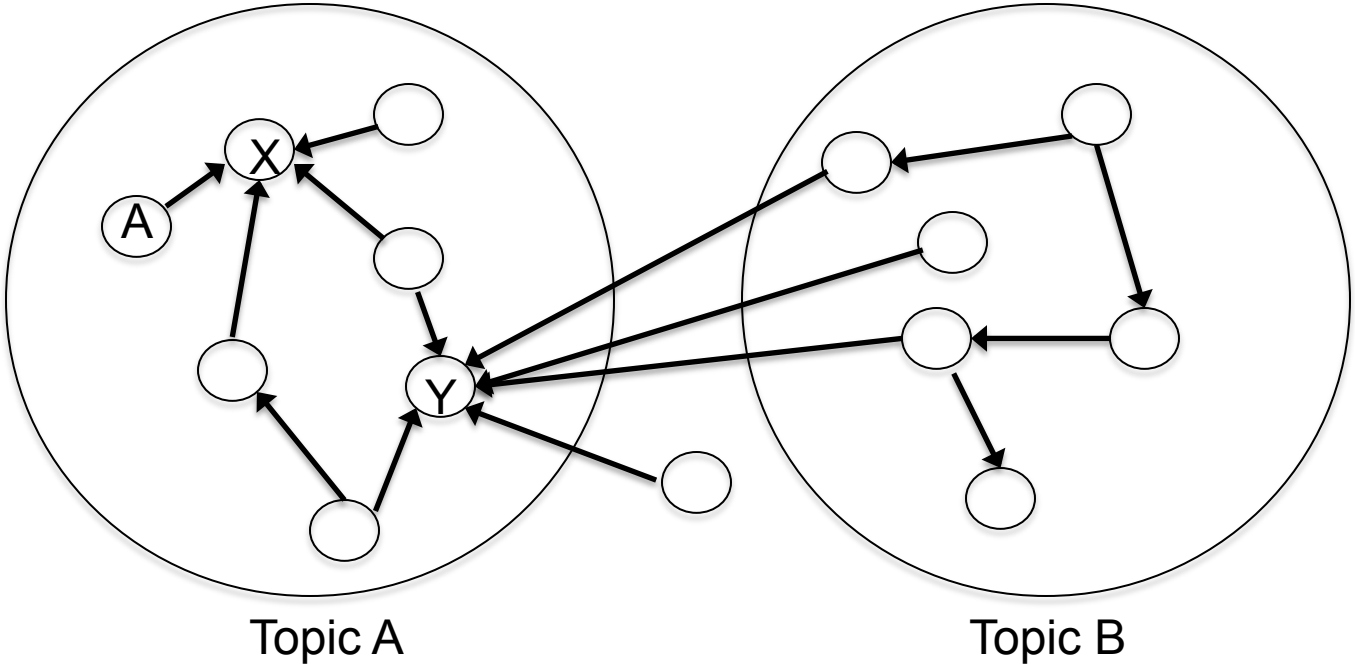
- Motivation
- Method Overview
- Data Set
- **Methods**
  - Text-based classification
  - **Network-based classification**
- Results
- Conclusion

Algorithm

- Finds topic-specific influential users using social network information
  - Friend-Follower relationship, tweet time, number of tweets, etc
- Take top 300 influential users for each topic
- Finds 5 most similar topics using the common influential users between two topics
- Classify a topic using categories of its similar topics

Network-based Classification

Topic-specific Influential Users*

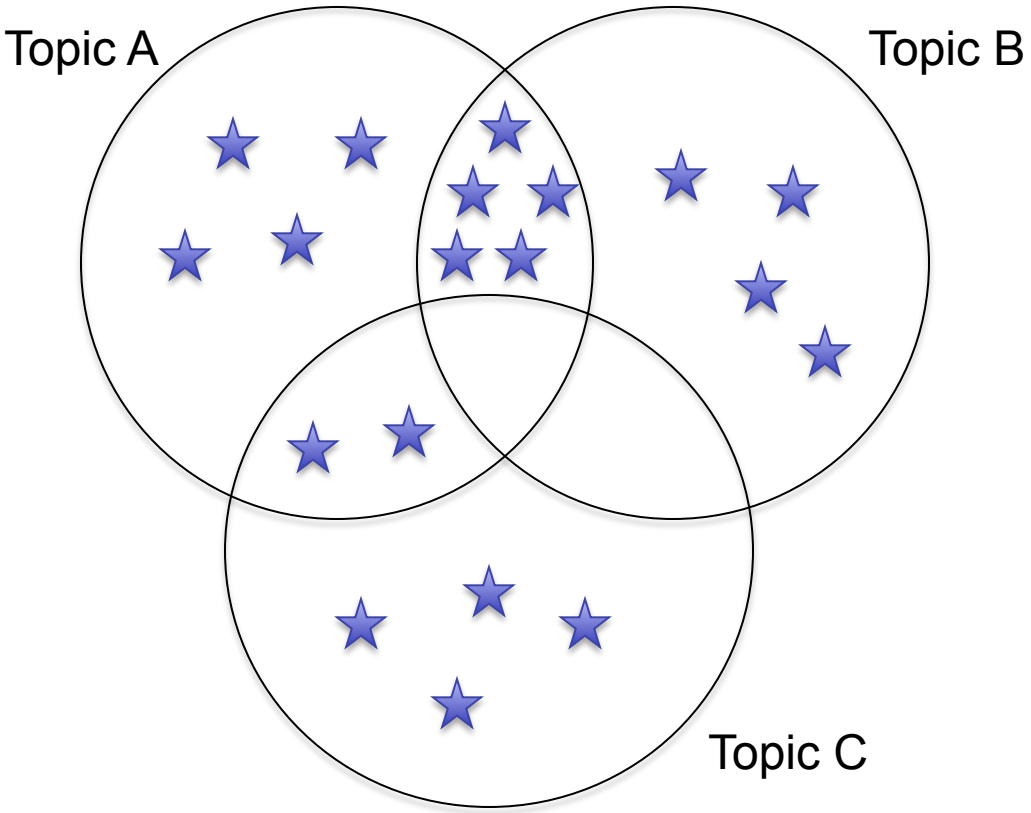


X is more influential than Y on Topic A

* R. Narayanan, “Mining Text for Relationship Extraction and Sentiment Analysis,” Ph.D. dissertation, 2010.

Network-based Classification

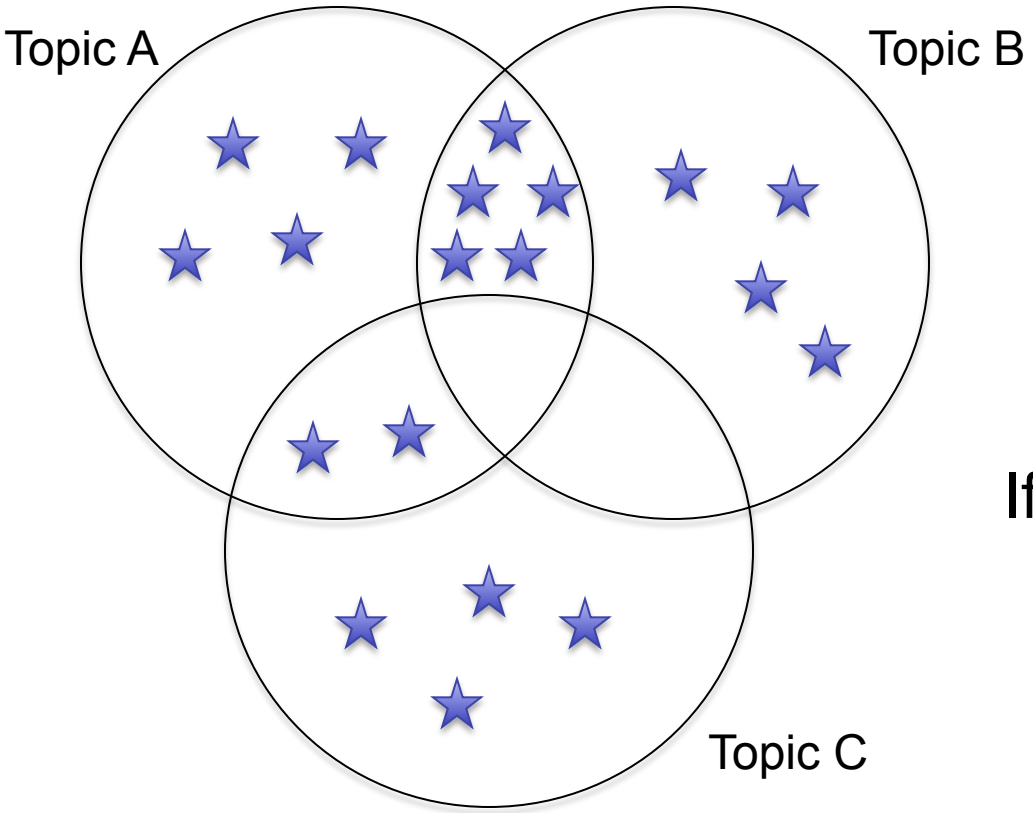
User similarity Model*



* R. Narayanan, “Mining Text for Relationship Extraction and Sentiment Analysis,” Ph.D. dissertation, 2010.

Network-based Classification

User similarity Model*

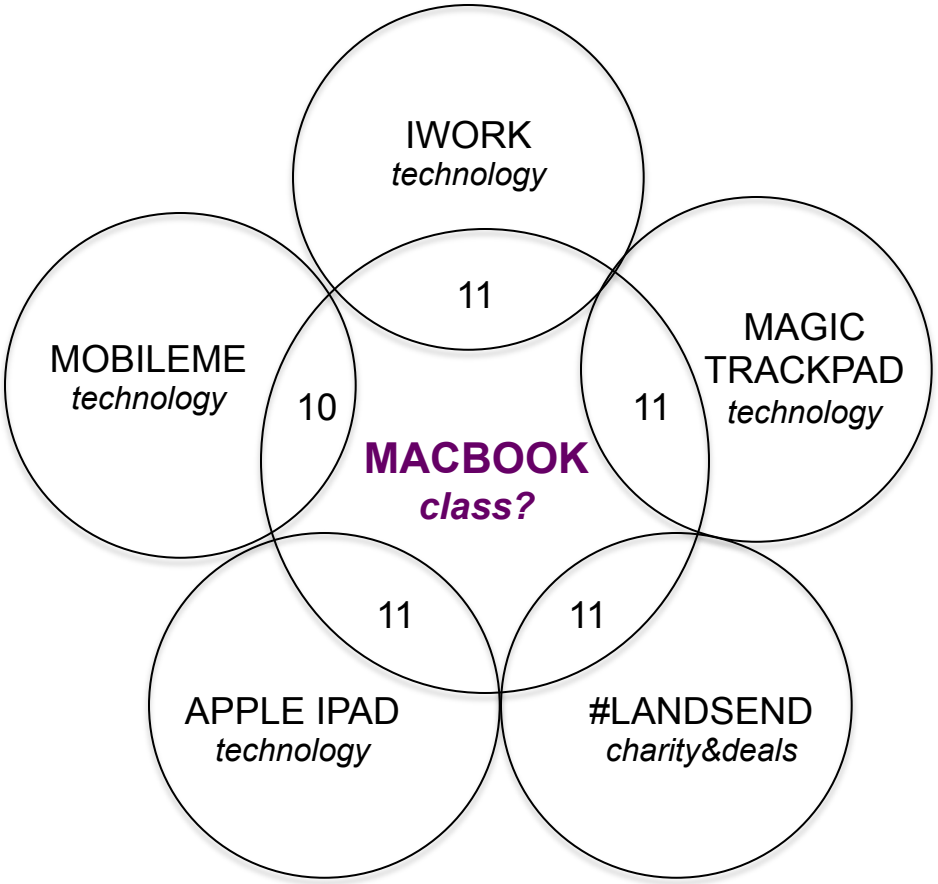


Topics A and B
are more closely related
than Topics A and C
If $|A_{infl} \cap B_{infl}| > |A_{infl} \cap C_{infl}|$

* R. Narayanan, “Mining Text for Relationship Extraction and Sentiment Analysis,” Ph.D. dissertation, 2010.

Network-based Classification

Topic "macbook" and 5 similar topics



Numbers in diagram **: number of common influential users** between topic "macbook" and the similar topic

Similar Topic	Class of Similar Topic	# Common Influential Users
iwork	technology	11
magic trackpad	technology	11
#landsend	charity & deals	11
apple ipad	technology	11
mobileme	technology	10

technology = 11 + 11 + 11 + 10 = 43
charity&deals = 11

Input to classifier

Topic	technology	charity &deals	books	music	fashion	tv &movies	…	Class
macbook	43	11	0	0	0	0	…	**?**
queen_rowling	0	0	30	0	0	10	…	**?**
lady_gaga	0	0	0	40	0	0	…	**?**

Table with 768 rows and 19 columns

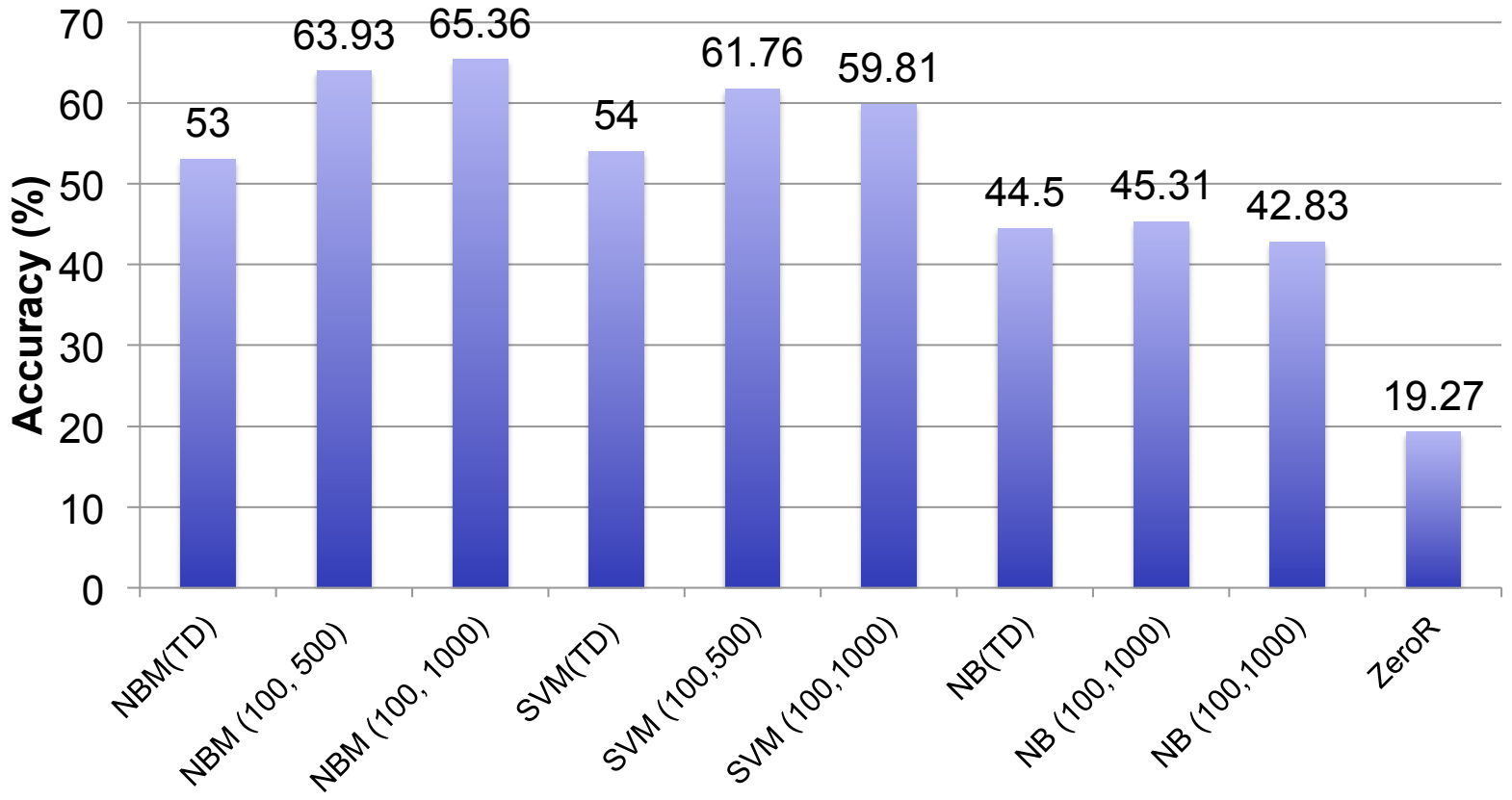
- Run various classifier
  - C5.0, K-Nearest Neighbor, SVM, Logistic Regression

- Motivation
- Method Overview
- Data Set
- Methods
- **Results**
- Conclusion

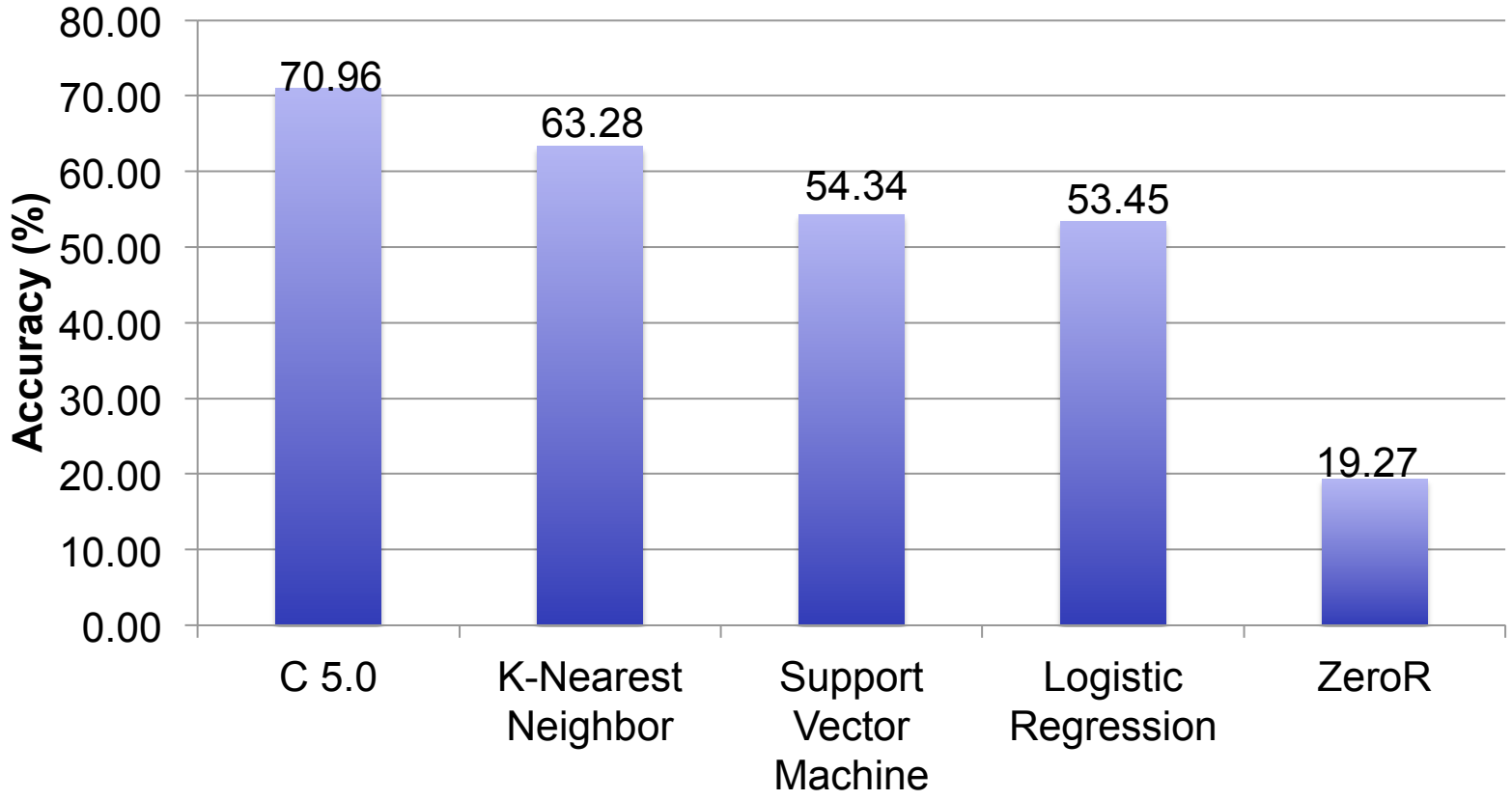
Experimental Setup

- TD: Trend Definition
- Model(x, y): classifier model used to classify a document consisting of x number of tweets per topic using y top frequent terms
  - e.g., NBM(100,1000)
    - Naïve Bayes Multinomial classifier
    - Document containing 100 tweets using
    - 1000 top frequent terms
- WEKA and SPSS modeler for classification
- 10-fold cross validation

Text-based Classification Results



Network-based classification results



- Motivation
- Method Overview
- Data Set
- Methods
- Results
- **Conclusion**

Key Contributions

- Use of social network structure for topic classification
- Good accuracy (65%) on Text-based classification
  - tweets are not grammatically structured (noisy)
- Network-based classifier (71%) outperforms text-based classifier

Future Work

- Integrate text-based classification and network-based classification
- Multi-labeling
  - topics could fall under more than one category
  - e.g., news about a famous actor's biography

Questions?

Thank you !