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

Trending Topics

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

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

General Categories

- Business
- Health
- Music
- Politics
- Sports
- Science
- Technology
• Motivation
• **Method Overview**
• Data Set
• Methods
• Results
• Conclusion
System Architecture

Data Collection

Trending Topic + Definition

Tweets

Lady gaga

Labeling

Text-based Modeling

Network-based Modeling

Data Modeling

Text-based Model Validation

Network-based Model Validation

Machine Learning

Trending Topic | Category
---|---
Lady gaga | music
burberry | fashion
ipad | technology
toy story 3 | tv & movies
superbowl | sports
tornado | other news
• 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

\[ \text{Document} = \text{Trend Definition} + \text{Tweets} \]
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 Multinominal, 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

Network-based Classification
User similarity Model*

Network-based Classification

User similarity Model*

Topics A and B are more closely related than Topics A and C

If \(|A_{infl} \cap B_{infl}| > |A_{infl} \cap C_{infl}|\)

Network-based Classification

Topic “macbook” and 5 similar topics

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

<table>
<thead>
<tr>
<th>Similar Topic</th>
<th>Class of Similar Topic</th>
<th># Common Influential Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>iwork</td>
<td>technology</td>
<td>11</td>
</tr>
<tr>
<td>magic trackpad</td>
<td>technology</td>
<td>11</td>
</tr>
<tr>
<td>#landsend</td>
<td>charity &amp; deals</td>
<td>11</td>
</tr>
<tr>
<td>apple ipad</td>
<td>technology</td>
<td>11</td>
</tr>
<tr>
<td>mobileme</td>
<td>technology</td>
<td>10</td>
</tr>
</tbody>
</table>

**technology** = 11 + 11 + 11 + 10 = 43

**charity&deals** = 11
## Input to classifier

<table>
<thead>
<tr>
<th>Topic</th>
<th>technology</th>
<th>charity &amp; deals</th>
<th>books</th>
<th>music</th>
<th>fashion</th>
<th>tv &amp; movies</th>
<th>...</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>macbook</td>
<td>43</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>?</td>
</tr>
<tr>
<td>queen_rowling</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>...</td>
<td>?</td>
</tr>
<tr>
<td>lady_gaga</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>?</td>
</tr>
</tbody>
</table>

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

Accuracy (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB (TD)</td>
<td>53</td>
</tr>
<tr>
<td>NB (100, 500)</td>
<td>63.93</td>
</tr>
<tr>
<td>NB (100, 1000)</td>
<td>65.36</td>
</tr>
<tr>
<td>SVM (TD)</td>
<td>54</td>
</tr>
<tr>
<td>SVM (100,500)</td>
<td>61.76</td>
</tr>
<tr>
<td>SVM (100,1000)</td>
<td>59.81</td>
</tr>
<tr>
<td>NB (TD)</td>
<td>44.5</td>
</tr>
<tr>
<td>NB (100,1000)</td>
<td>45.31</td>
</tr>
<tr>
<td>NB (100,1000)</td>
<td>42.83</td>
</tr>
<tr>
<td>ZeroR</td>
<td>19.27</td>
</tr>
</tbody>
</table>
Network-based classification results

Accuracy (%)

- C 5.0: 70.96%
- K-Nearest Neighbor: 63.28%
- Support Vector Machine: 54.34%
- Logistic Regression: 53.45%
- ZeroR: 19.27%
• 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!