Predicting Future Popularity of Events in Microblogging Platforms

Manish Gupta\(^1\), Jing Gao\(^2\), ChengXiang Zhai\(^1\), Jiawei Han\(^1\)

\(^1\)UIUC \hspace{1cm} \(^2\)SUNY

ASIST 2012
Abstract

• We introduce a novel problem of future popularity prediction of events for microblogging platforms and cast it as a multi-class categorization problem (classes correspond to different ranges of percentage change in popularity).

• We investigate into multiple machine learning and time series approaches using a variety of popularity, social and event features.

• Experimental results on two real datasets of 18382 events extracted from 133 million tweets show that our approach is effective, enabling the best classifier to achieve 74% accuracy in predicting the future popularity label.
Introduction

• Google Trends provides what is hot now in news and on web search.
• It talks neither about current social volume nor about predicted volume for the future.
Future Popularity Trends are Critical

• Twitter has 1620 tweets per second.
• Trendy events are discussed and can be summarized to represent public opinions.
• Future popularity trend of events can help
  – Predict box office revenues of a movie
  – News reporters and business analysts to focus on promising stories and viral marketing
  – Product managers to make decisions
Predicting Future Popularity of Events is Challenging

- Events on Twitter follow multiple types of popularity profiles.
- Twitter is a social network where content is user-generated. Hence, correlation to news volume may not be high.
- Propagation effect of the network plays a big role on Twitter.
- As new aspects of a news story unfold, discussions related to that story on Twitter gain new peaks of popularity.
Capturing in-Twitter and out-of-Twitter Dynamics as Features

• *Popularity* in terms of all tweets (*Pop*) as well as retweets (*RT*)
• *Ratios* (ratios of retweets and all tweets for same or previous time interval)
• Variety and depth of discussions as *Event* features which include number of sub-topics (*Aspects*) or frequent words (*Subordinate Words*)
• Out-of-Twitter popularity using *URLs*
• *Social* Features (*Friends* and *Followers*)
• Event *Category*
Definition of an Event

• An event $E$ relates to a real world event expressed using a set of words chosen from the vocabulary $V$. The set of words forming the event can be divided into two groups: core words $C_E$ and subordinate words $S_E$. Thus, $E = C_E \cup S_E$.

• A topic extracted from tweets belonging to a short time interval can be assumed to represent a coherent event.

• E.g., for the *Oil Spill Disaster* event, $C_E = \{bp, oil, spill\}$ while $S_E = \{deepwater, explosion, marine, life, scientists, static, kill, mud, cement, plug, 9m, gushed, barrels, gulf, mexico\}$.
Definition of Popularity of Events

• Contribution of the tweet m (containing n unique words) to the event e, c(m,e), is defined as

\[
\sum_{w \in m} I[w \in C_e \cup S_e] \frac{n}{n}
\]

• where I[.] is an indicator function which is 1 when the condition in parentheses is true, else 0.
• If \(c(m,e) \geq \psi\), we consider that the tweet m supports the event e. Popularity of an event e in a time interval t, \(P_t(e)\), is defined as the number of tweets in t, supporting the event e.
Future Event Popularity Trend Prediction Problem

• K-class classification problem
• Input: A set of candidate events $C$ which have high popularity in a time interval $t$.
• Output: A popularity trend class label ($k \in K$) corresponding to the transition from time interval $t$ to the time interval $t+1$ for each event $e \in C$.
• Problem: Predict the future popularity trend class label by effectively exploiting all the information available in terms of feature values.
Identifying Currently Popular (Candidate) Events

• Event Detection Model Choice
  – Latent Dirichlet Allocation (LDA) [Blei et al., 2003]
    • Needs number of topics as input.
    • Mixes multiple events together when used over a single document of all tweets in a time interval.
  – Phrase graph generation method [Sharifi et al., 2010]
    • Gives much importance to the position of the words.
    • Tweets may not contain exact phrases.
  – Twitter Trending Words
    • Focus on burst popularity.
    • Trend words may miss certain events if they have been popular for quite some time.
  – GroupBurst [Mathioudakis and Koudas, 2010]
    • Agglomerative clustering in the space of words.
    • May fail if reasonable number of tweets contain words from multiple events.
Our Event Detection Approach

Preprocess/Clean Tweets

Generate Stop-words & AvgWord Frequency

Find Core Words and Subordinate Words

Tweet Feeds

Currently Popular Events

Agglomerative Clustering of Events

Discover Aspects

8/29/2012
Our Event Detection Approach

• Finding core words (words which become suddenly popular in the tweet stream (bursty)).

  \[
  \text{Core Word Score: } CWS(w) = \frac{f(w)}{f_{avg}(w)}
  \]

• Subordinate words for the event are words that occur quite frequently with a core word in the current interval of the tweet stream.

• Aspects: Most frequent subsets of subordinate words are considered aspects (sub-topics).

• Events are detected by agglomerative clustering in the space of aspects.
Feature Set: Popularity Features (Pop)

• Breaking news events show clear distinguishable peaks.
  – E.g., “7.2 magnitude Turkey earthquake”
  – Short left duration and longer right duration.
• Predictable events often are not characterized by clear peaks.
  – E.g., “elections”, “christmas”, “weather changes”
  – Long left duration and an optional long right duration.
• About half of the events have more than one mode (local maxima) over their life cycle.
• Events differ a lot with respect to their highest popularity and also with respect to their duration.
• Features: Popularity for the past 24 hours (recent history at a fine level) and popularity across past 10 days (past history at a coarser level)
• We remove the daily trend of all events suffering from decrease in popularity levels in the night compared to that in the day time.

8/29/2012
Feature Set: Popularity Features (RT)

- Event with large number of initial original tweets but no retweets may not last long.
- Event with many original tweets and many followup retweets has surely caught momentum.
- Event life cycle
  - First phase: URLs are important.
  - Later phases: Friends, followers, retweets are important.
- Features
  - Retweet Popularity for the past 24 hours (recent history at a fine level) and popularity across past 10 days (past history at a coarser level)
  - $age_{\text{Pop}}$, $age_{\text{RT}}$, hour of day.
Feature Set: Ratios Features

- Relative change in popularity of event for the past few consecutive pairs of time intervals is crucial.

- $F_{\text{Pop:Pop}}$ – ratios of popularity across consecutive time intervals.

- $F_{\text{RT:RT}}$ – ratios of retweet popularity across consecutive time intervals.

- $F_{\text{RT:Pop}}$ – ratio of retweets to original tweets captures social impact of event.
Feature Set: Social Features (Followers, Friends)

• Twitter has a very vibrant social network
  – In our D2010 dataset, an average active user has 662 friends and 695 followers.
  – ~52% tweets contain user mentions

• Features
  – Followers for the past 10 hours.
  – Friends for the past 10 hours.
Feature Set: Out-of-Twitter Popularity (URLs)

• Events that are sensationalized by media (like Casey Anthony’s murder trial or Amanda Knox case) are expected to last for multiple days.
• Original tweets generally contain URLs which provide evidence for event.
• Number of unique URLs posted by Twitter users acts as a good proxy of the popularity of the event in out-of-Twitter world.
• As the real world event story develops, the number of unique URLs would continue to grow.
• Features: Number of posted URLs in past 10 hours.
Feature Set: Event Features (Aspects)

• A news story (event) in real world is generally composed of a smooth flow of different aspects.
• If there are a large number of aspects being discussed about the event, with a high probability, the event will last longer.
• If there are too many and relatively low frequency aspects being discussed (in a very short duration), the event may not be a coherent event and may just represent social gossip.
• A genuine news event would be characterized by a smooth increase in the number of aspects being discussed over time.
• Features: Number of aspects in the past 10 hours.
Feature Set: Event Features (Subordinate Words)

- Subordinate words of an event form the rich context of an event.
- Event with a sequence of relatively less number of subordinate words each with high frequency
  - Peaked event with short duration
  - Many highly negative or highly positive popularity changes
- Event with a large number of subordinate words
  - Many sub-topics with long duration
  - Many flat or small negative or positive popularity changes.
- Features: Number of Subordinate Words for the past 10 hours.

8/29/2012
Feature Set: Event Category

• Events belonging to categories like entertainment, sports and politics are generally short-lived because of constant supply of frequent fresh news.
• Events in some categories like technology last for longer time period.
• We detect category of event using correlation with news headlines. We categorize the events not found in news to Others category.
# Feature Set Summary

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Symbolic Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop (35)</td>
<td>({P_{hour_0}(e), P_{hour_{-1}}(e), \ldots, P_{hour_{-23}}(e), P_{day_0}(e), ) (P_{day_{-1}}(e), \ldots, P_{day_{-9}}(e), age_{Pop}, hour_0})</td>
</tr>
<tr>
<td>RT (35)</td>
<td>({R_{hour_0}(e), R_{hour_{-1}}(e), \ldots, R_{hour_{-23}}(e), R_{day_0}(e), R_{day_{-1}}(e), \ldots, R_{day_{-9}}(e), age_{RT}})</td>
</tr>
<tr>
<td>Pop:Pop Ratios (34)</td>
<td>({PP_{h_0,\ldots,-1}(e), PP_{h_{-1},\ldots,-2}(e), \ldots, PP_{h_{-22},\ldots,-23}(e), PP_{d_0,\ldots,-1}(e), PP_{d_{-1},\ldots,-2}(e), \ldots, PP_{d_{-8},\ldots,-9}(e)})</td>
</tr>
<tr>
<td>RT:RT Ratios (34)</td>
<td>({RR_{h_0,\ldots,-1}(e), RR_{h_{-1},\ldots,-2}(e), \ldots, RR_{h_{-22},\ldots,-23}(e), RR_{d_0,\ldots,-1}(e), RR_{d_{-1},\ldots,-2}(e), \ldots, RR_{d_{-8},\ldots,-9}(e)})</td>
</tr>
<tr>
<td>RT:Pop Ratios (34)</td>
<td>({RP_{h_0}(e), RP_{h_{-1}}(e), \ldots, RP_{h_{-23}}(e), RP_{d_0}(e), RP_{d_{-1}}(e), \ldots, RP_{d_{-9}}(e)})</td>
</tr>
<tr>
<td>Followers (10)</td>
<td>({F_{hour_0}(e), F_{hour_{-1}}(e), \ldots, F_{hour_{-9}}(e)})</td>
</tr>
<tr>
<td>Friends (10)</td>
<td>({F_{hour_0}(e), F_{hour_{-1}}(e), \ldots, F_{hour_{-9}}(e)})</td>
</tr>
<tr>
<td>Aspects (10)</td>
<td>({A_{hour_0}(e), A_{hour_{-1}}(e), \ldots, A_{hour_{-9}}(e)})</td>
</tr>
<tr>
<td>Subordinate Words (10)</td>
<td>({S_{hour_0}(e), S_{hour_{-1}}(e), \ldots, S_{hour_{-9}}(e)})</td>
</tr>
<tr>
<td>URLs (10)</td>
<td>({U_{hour_0}(e), U_{hour_{-1}}(e), \ldots, U_{hour_{-9}}(e)})</td>
</tr>
<tr>
<td>Category (1)</td>
<td>Category</td>
</tr>
</tbody>
</table>
Learning Approaches (1)

• Time Series Models
  – Linear Regression
    • \[ x_t = \sum_{i=1}^{h} \beta_i x_i + \epsilon_t \]
  – Auto-Regression AR(p)
    • \[ x_t = c + \sum_{i=1}^{p} \psi_i x_{t-i} + \epsilon_t \]
  – Auto-Regressive Moving Average ARMA(p,q)
    • \[ x_t = c + \sum_{i=1}^{p} \psi_i x_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \]
  – Vector Auto-Regression (VAR)
    • \[ x_t = A_0 + \sum_{i=1}^{p} A_i x_{t-i} + \epsilon_t \]
Learning Approaches (2)

• Classification models
  – Support Vector Machines
  – K-Nearest Neighbors
  – Naïve Bayes
  – Decision Trees

• Hybrid Approaches
  – Regression models capture time dependencies.
  – Classification models learn across instances.
  – Hybrid models can use regression weights as features or to normalize feature values.
Dataset Details

• Twitter feeds (~133M tweets)
  – D2010 (Dec 2010)
  – D2011 (Mar 2011)

• Categorization done using news feeds from top ten news websites.

• A large percent of the INC instances belong to TopNews and Others categories, while many DEC3 instances belong to Sports and Others categories.

<table>
<thead>
<tr>
<th>Popularity Change</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than -75%</td>
<td>DEC3</td>
</tr>
<tr>
<td>-75% to -50%</td>
<td>DEC2</td>
</tr>
<tr>
<td>-50% to -25%</td>
<td>DEC1</td>
</tr>
<tr>
<td>-25% to 0%</td>
<td>FLAT</td>
</tr>
<tr>
<td>More than 0%</td>
<td>INC</td>
</tr>
</tbody>
</table>
## Basic Accuracy Results

<table>
<thead>
<tr>
<th>Features</th>
<th>Method</th>
<th>D2010</th>
<th>D2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLs</td>
<td>SVM</td>
<td>60.58</td>
<td>63.14</td>
</tr>
<tr>
<td>Social (Followers + Friends)</td>
<td>SVM</td>
<td>61.48</td>
<td>61.64</td>
</tr>
<tr>
<td>Event (Aspects + Subordinate Words)</td>
<td>SVM</td>
<td>67.17</td>
<td>66.66</td>
</tr>
<tr>
<td>Pop</td>
<td>SVM</td>
<td>69.96</td>
<td>67.48</td>
</tr>
<tr>
<td>Only RT</td>
<td>SVM</td>
<td>63.61</td>
<td>61.46</td>
</tr>
<tr>
<td>Pop+RT</td>
<td>SVM</td>
<td>70.47</td>
<td>68.28</td>
</tr>
<tr>
<td>Ratios</td>
<td>SVM</td>
<td>72.75</td>
<td>73.21</td>
</tr>
<tr>
<td>Pop+RT+Ratios</td>
<td>SVM</td>
<td>72.81</td>
<td>73.43</td>
</tr>
<tr>
<td>Pop+RT+Ratios</td>
<td>Decision Trees</td>
<td>70.95</td>
<td>69.79</td>
</tr>
<tr>
<td>Pop+RT+Ratios</td>
<td>Naïve Bayes</td>
<td>69.42</td>
<td>66.5</td>
</tr>
<tr>
<td>All</td>
<td>SVM</td>
<td>73.54</td>
<td>74.23</td>
</tr>
<tr>
<td>Pop</td>
<td>Linear Regression</td>
<td>70.62</td>
<td>71.5</td>
</tr>
<tr>
<td>All</td>
<td>Linear Regression</td>
<td>70.66</td>
<td>69.64</td>
</tr>
<tr>
<td>Pop</td>
<td>AR(1)</td>
<td>71.21</td>
<td>70.81</td>
</tr>
<tr>
<td>Pop</td>
<td>AR(2)</td>
<td>69.3</td>
<td>69.21</td>
</tr>
<tr>
<td>Pop</td>
<td>ARMA(1,1)</td>
<td>69.68</td>
<td>68.96</td>
</tr>
<tr>
<td>Pop+RT+Ratios</td>
<td>VAR(2,5)</td>
<td>65.95</td>
<td>65.71</td>
</tr>
</tbody>
</table>

### Classified as → Actual↓

<table>
<thead>
<tr>
<th></th>
<th>DEC3</th>
<th>DEC2</th>
<th>DEC1</th>
<th>FLAT</th>
<th>INC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEC3</td>
<td>1</td>
<td>0.75</td>
<td>0.5</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>DEC2</td>
<td>0.75</td>
<td>1</td>
<td>0.75</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>DEC1</td>
<td>0.5</td>
<td>0.75</td>
<td>1</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td>FLAT</td>
<td>0.25</td>
<td>0.5</td>
<td>0.75</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>INC</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.75</td>
<td>1</td>
</tr>
</tbody>
</table>

- SVMs is the best learning method.
- Pop+RT+Ratios is the best set of features.
## Varying the Model Parameters

<table>
<thead>
<tr>
<th>Limit</th>
<th>D2010</th>
<th>D2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 days</td>
<td>72.81</td>
<td>73.43</td>
</tr>
<tr>
<td>24 hours</td>
<td>72.47</td>
<td>70.94</td>
</tr>
<tr>
<td>15 hours</td>
<td>71.30</td>
<td>70.22</td>
</tr>
<tr>
<td>10 hours</td>
<td>70.81</td>
<td>69.94</td>
</tr>
<tr>
<td>5 hours</td>
<td>70.55</td>
<td>69.67</td>
</tr>
<tr>
<td>2 hours</td>
<td>68.94</td>
<td>67.98</td>
</tr>
</tbody>
</table>

### Accuracy (%) when the History Size (i.e. #Features) is Limited

<table>
<thead>
<tr>
<th>% of data</th>
<th>D2010</th>
<th>D2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>72.65</td>
<td>70.93</td>
</tr>
<tr>
<td>20</td>
<td>73.00</td>
<td>72.38</td>
</tr>
<tr>
<td>40</td>
<td>73.91</td>
<td>72.95</td>
</tr>
<tr>
<td>60</td>
<td>73.22</td>
<td>72.74</td>
</tr>
<tr>
<td>80</td>
<td>73.28</td>
<td>72.66</td>
</tr>
<tr>
<td>100</td>
<td>72.81</td>
<td>73.43</td>
</tr>
</tbody>
</table>

### Accuracy (%) when the Amount of Training Data is Varied

<table>
<thead>
<tr>
<th>Sub-dataset</th>
<th>D2010</th>
<th>D2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>0to5</td>
<td>70.77</td>
<td>68.32</td>
</tr>
<tr>
<td>6to10</td>
<td>73.48</td>
<td>72.36</td>
</tr>
<tr>
<td>11to15</td>
<td>74.22</td>
<td>75.21</td>
</tr>
<tr>
<td>16to20</td>
<td>73.97</td>
<td>75.97</td>
</tr>
</tbody>
</table>

Accuracy for Pop+RT+Ratios Feature Set for Different Sub-Datasets based on Event Ranks
Varying the Train and Test Sets

<table>
<thead>
<tr>
<th>Train Set</th>
<th>Test Set</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2010</td>
<td>D2011</td>
<td>70.64</td>
</tr>
<tr>
<td>D2011</td>
<td>D2010</td>
<td>71.92</td>
</tr>
</tbody>
</table>

Cross Dataset Results

<table>
<thead>
<tr>
<th>Category</th>
<th>D2010</th>
<th>D2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>74.72</td>
<td>71.12</td>
</tr>
<tr>
<td>Top</td>
<td>73.47</td>
<td>72.55</td>
</tr>
<tr>
<td>Business</td>
<td>73.44</td>
<td>73.64</td>
</tr>
<tr>
<td>Others</td>
<td>73.42</td>
<td>72.00</td>
</tr>
<tr>
<td>Entertainment</td>
<td>72.38</td>
<td>70.37</td>
</tr>
<tr>
<td>Travel</td>
<td>71.76</td>
<td>67.53</td>
</tr>
<tr>
<td>Politics</td>
<td>71.34</td>
<td>71.71</td>
</tr>
<tr>
<td>Health</td>
<td>71.27</td>
<td>70.88</td>
</tr>
<tr>
<td>Sports</td>
<td>70.12</td>
<td>68.77</td>
</tr>
</tbody>
</table>

Accuracy (%) for Different Categories

<table>
<thead>
<tr>
<th>Train → Test ↓</th>
<th>Phase1</th>
<th>Phase2</th>
<th>Phase3</th>
<th>Phase4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase1</td>
<td>70.24,69.86</td>
<td>70.34,70.43</td>
<td>70.45,70.75</td>
<td>70,65.21</td>
</tr>
<tr>
<td>Phase2</td>
<td>66.12,67.97</td>
<td>72.35,71.44</td>
<td>76.92,74.73</td>
<td>75.22,73.5</td>
</tr>
<tr>
<td>Phase3</td>
<td>62.9,62.43</td>
<td>71.06,68.68</td>
<td>76.64,76.14</td>
<td>75.05,75</td>
</tr>
<tr>
<td>Phase4</td>
<td>60.01,59.8</td>
<td>67.82,66.12</td>
<td>74.36,73.85</td>
<td>77.18,75.49</td>
</tr>
</tbody>
</table>

Accuracy (%) for Events at Different Phases of Event Life (D2010, D2011)
Related Work (Event Detection)

• Epidemics [Lampos et al., 2010], wildfires, hurricanes, floods, earthquakes [Sakaki et al., 2010] and tornados.

• Twitter Trends words, Latent Dirichlet Allocation (LDA) [Blei et al., 2003] or Phrase Graph Generation Method [Sharifi et al., 2010]

• Aspect-based model of event detection: GroupBurst in Twitter Monitor [Mathioudakis and Koudas, 2010].
Related Work (Predictive Analysis on Twitter and Other Platforms)

- Box office forecasting of movies [Asur and Huberman, 2010]
- Predicting retweetability of tweets [Suh et al., 2010, Hong et al., 2011, Petrovic et al., 2011],
- Predicting for a pair of users, whether a tweet written by one will be retweeted by the other user [Zaman et al., 2010]
- Predicting information diffusion [Yang and Counts, 2010]
- Future popularity of social media content on Digg and Youtube [Szabo and Huberman, 2010, Lerman and Hogg, 2010].
Conclusions

• We explored the possibility of detecting news events from Twitter feeds.
• We proposed an event detection method tuned for the task.
• We studied the performance of a variety of URLs, Social, Event, Popularity, and Ratios features.
• Ratios turn out to be the best features, performing significantly better than “past tweets (Pop) or retweets (RT)”. Aspects and Subordinate Words related to the event are better than URLs or Social features for this task. On the other hand, in the initial phases of an event “life”, URLs act as a good predictor while Ratios features perform better in the later phases.
• SVMs work better than simple time series models that can account only for Pop features.
• Our models were observed to work reasonably well across events in different phases of their life and across categories.
References

• Yang, J. and Counts, S. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. ICWSM.
Thanks!