Credibility in Context
An Analysis of Feature Distributions in Twitter

John O’Donovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer
University of California
Santa Barbara, USA

Sibel Adalı
Rensselaer Polytechnic Institute
Troy, New York, USA

September 5th 2012, Amsterdam, The Netherlands. (SocialCom 2012)
Outline

• Background
  - Motivation
  - Research Questions
  - Contribution
  - Credibility
  - Related Work

• Features and Contexts

• Experimental Framework

• Results

• Conclusion
Motivation

• Growth of Social Media
  – User-Generated Content (UGC)
  – Information Overload

• “Credibility” models can help to identify useful information. They can leverage *historical* and *current* information available through social web APIs

• But… Indicators of credibility vary across contexts. There is a need for more adaptive models.
What is Credibility?

- Broad use, many different definitions:
  - Social (Golbeck, Ziegler), Cognitive (Gray, Todorov), Computational (Marsh, Josang), Psychological (Dellorcas, Erikson)

**Message-level Credibility**

A degree of believability that can be assigned to a tweet about a target topic, i.e.: an indication that the tweet contains believable information.

**Social Credibility**

The expected believability imparted on a user as a result of their standing in the social network, based on any and all available metadata.
Lots of useless information?

(Excerpt from mashable.com infographic)
Examples

• **Useless / Nonsensical Tweets**
  - “yo yo yo, looky here!!”

• **Spam Tweets**
  - ‘Have you heard millions of people are making $5k+/Mo from home? heres how...t.co/blah’

• **Credible / Newsworthy Tweets**
  - Great keynote by Todorov at #SocialCom2012 in #Amsterdam
  - #LADodgers commentator #vinscully back for another season!

• **Personal / Conversational**
  - @anTusail: thanks for the info!
## Related Work - *Credibility Evaluation*

<table>
<thead>
<tr>
<th>Classification-based</th>
<th>Supervised</th>
<th>Semi-supervised</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kang et al. 2012</td>
<td>Bian et al., 2009 Yin &amp; Tan, 2011</td>
<td>Gupta et al., 2011 Canini et al., 2011</td>
</tr>
<tr>
<td></td>
<td>Castillo et al., 2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graph Models</td>
<td>Agichtein et al., 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity-based Approaches</td>
<td>Juffinger et al. 2009, O’Donovan 2005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game Theory Models</td>
<td>Ghosh &amp; McAfee, 2011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Credibility Models

Social Model

Utility

RT-rate

Learning

Balance

Topic-Social Credibility

Content Model

Social Credibility

Credibility Model

Hybrid Model

Numeric Indicators

Positive Sentiment Factor

Negative Sentiment Factor

Sentiment Polarity

Number of intensifiers

Age of Profile

Number of popular topic-specific terms

Number of Uppercase Chars

Binary Indicators

Is Only Urls

Is a Retweet

Has a Question Mark

Has an Exclamation Mark

Has multiple Questions/

Has a positive emoticon

Has a negative emoticon

Social Features A, B C…

Content Features X, Y, Z…
Initial Experiments (Kang ‘12)

• Social model outperforms content-based and hybrid model
• Approximately 88.5% accuracy predicting manually labeled tweets using J48 Learner using our Social Model
• However, results varied greatly across different topics.
Research Questions

• How are the features that indicate credibility distributed in Twitter?
• How and why do they vary across different contexts?
• How do we use knowledge of feature distribution to create more adaptive, better performing credibility-based information filters?
Outline

- Background

- **Features and Contexts**
  - Terminology
  - The Twitter Graph
  - Details of Social, Content, and Hybrid Model

- Experimental Framework

- Results

- Conclusion
The Twitter Graph

- **Follower Group**
  - the people who receive my Twitter updates

- **Following Group**
  - the people I follow (their Twitter updates appear in my personal timeline)
The Twitter Graph

- **Follower Group**
  - the people who receive my Twitter updates

- **Following Group**
  - the people I follow (their Twitter updates appear in my personal timeline)
Slicing the Twitter Graph:

### Table I

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
<th># of Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diverse Topics</td>
<td>Diverse topics in Twitter; eg: #Romney #Facebook</td>
<td>8 different topics (see Table II)</td>
</tr>
<tr>
<td>Credibility</td>
<td>Manually provided assessments of tweets</td>
<td>Credible or non credible</td>
</tr>
<tr>
<td>Chain length</td>
<td>Mined retweet chains and classified based on length</td>
<td>Long or short</td>
</tr>
<tr>
<td>Dyadic pairs</td>
<td>Mined interpersonal interaction and classified</td>
<td>Dyadic or not dyadic</td>
</tr>
</tbody>
</table>
Feature Sets

Three classes of features were used: Social, Content-based and Behavioral/Dynamic.

<table>
<thead>
<tr>
<th>Name</th>
<th>% Present</th>
<th>Average score</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>100.00</td>
<td>610.64</td>
<td>Social</td>
</tr>
<tr>
<td>listed_count</td>
<td>100.00</td>
<td>11.82</td>
<td>Social</td>
</tr>
<tr>
<td>status_count</td>
<td>100.00</td>
<td>554.49</td>
<td>Social</td>
</tr>
<tr>
<td>status_rt_count</td>
<td>100.00</td>
<td>10.17</td>
<td>Social</td>
</tr>
<tr>
<td>favourites_count</td>
<td>100.00</td>
<td>57.96</td>
<td>Social</td>
</tr>
<tr>
<td>followers</td>
<td>100.00</td>
<td>295.15</td>
<td>Social</td>
</tr>
<tr>
<td>followings</td>
<td>100.00</td>
<td>315.03</td>
<td>Social</td>
</tr>
<tr>
<td>fose_ratio</td>
<td>100.00</td>
<td>5.81</td>
<td>Social</td>
</tr>
</tbody>
</table>
# Feature Sets

**Content-based**

<table>
<thead>
<tr>
<th>Name</th>
<th>% Present</th>
<th>Average score</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>char</td>
<td>100.00</td>
<td>120.55</td>
<td>Content</td>
</tr>
<tr>
<td>word</td>
<td>100.00</td>
<td>18.69</td>
<td>Content</td>
</tr>
<tr>
<td>question</td>
<td>7.95</td>
<td>0.10</td>
<td>Content</td>
</tr>
<tr>
<td>excl</td>
<td>10.10</td>
<td>0.15</td>
<td>Content</td>
</tr>
<tr>
<td>uppercase</td>
<td>10.23</td>
<td>11.27</td>
<td>Content</td>
</tr>
<tr>
<td>pronoun</td>
<td>92.84</td>
<td>4.22</td>
<td>Content</td>
</tr>
<tr>
<td>smile</td>
<td>42.24</td>
<td>0.02</td>
<td>Content</td>
</tr>
<tr>
<td>frown</td>
<td>1.81</td>
<td>0.43</td>
<td>Content</td>
</tr>
<tr>
<td>url</td>
<td>14.17</td>
<td>0.42</td>
<td>Content</td>
</tr>
<tr>
<td>retweet</td>
<td>8.71</td>
<td>0.74</td>
<td>Content</td>
</tr>
<tr>
<td>sentiment_pos</td>
<td>71.51</td>
<td>1.53</td>
<td>Content</td>
</tr>
<tr>
<td>sentiment_neg</td>
<td>59.07</td>
<td>1.23</td>
<td>Content</td>
</tr>
<tr>
<td>sentiment</td>
<td>74.20</td>
<td>0.29</td>
<td>Content</td>
</tr>
<tr>
<td>num_hashtag</td>
<td>42.09</td>
<td>0.83</td>
<td>Content</td>
</tr>
<tr>
<td>num_mention</td>
<td>19.25</td>
<td>0.25</td>
<td>Content</td>
</tr>
<tr>
<td>tweet_type</td>
<td>100.00</td>
<td>1.10</td>
<td>Content</td>
</tr>
<tr>
<td>ellipsis</td>
<td>2.11</td>
<td>0.29</td>
<td>Content</td>
</tr>
<tr>
<td>news</td>
<td>5.13</td>
<td>2.03</td>
<td>Content</td>
</tr>
</tbody>
</table>
# Feature Sets

## Behavioral / Dynamic

<table>
<thead>
<tr>
<th>Name</th>
<th>% Present</th>
<th>Average score</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>average balance of conversation</td>
<td>100.00</td>
<td>0.32</td>
<td>Behavioral</td>
</tr>
<tr>
<td>average number of friends in timeline</td>
<td>100.00</td>
<td>2086.28</td>
<td>Behavioral</td>
</tr>
<tr>
<td>average spacing between statuses in seconds in timeline</td>
<td>100.00</td>
<td>21959.07</td>
<td>Behavioral</td>
</tr>
<tr>
<td>average text length in timeline</td>
<td>100.00</td>
<td>104.52</td>
<td>Behavioral</td>
</tr>
<tr>
<td>average general response time</td>
<td>100.00</td>
<td>3.27</td>
<td>Behavioral</td>
</tr>
<tr>
<td>average number of messages per conversation</td>
<td>100.00</td>
<td>4.34</td>
<td>Behavioral</td>
</tr>
<tr>
<td>average trust value in conversation</td>
<td>100.00</td>
<td>0.10</td>
<td>Behavioral</td>
</tr>
<tr>
<td>fraction of statuses in timeline that are retweets</td>
<td>100.00</td>
<td>0.55</td>
<td>Behavioral</td>
</tr>
</tbody>
</table>
Outline

• Background
• Features and Contexts
• **Experimental Framework**
  - Crawler System
  - Data
  - Credibility Assessments
• Results
• Conclusion
Crawling Strategy

All Tweets in \( x : T_x \):

\[ t_1, t_2, \ldots, t_i \]

All Users: \( U_x \):

\[ U_1, U_2, \ldots, U_m \]

All Followers \( F_e(U_n) \):

\[ U_1, U_2, \ldots, U_k \]

All Followings \( F_o(U_n) \):

\[ U_1, U_2, \ldots, U_n \]

All Tweets from \( U_x \) outside of Topic \( x \): \( U(T) \)

\[ t_1, t_2, \ldots, t_j \]

0.8M User Profiles
1.7M Tweets
Segmenting based on Credibility

- Crawler System
  - Tweets and Metadata
  - User Survey
  - Annotated Tweets
  - Credibility Modeling
  - Credibility Scores
  - Accuracy Analysis
  - Social Model best performed

Survey ‘Trust Modeling in Microblogs’
University of California, Santa Barbara
Department of Computer Science
FourEyes Lab, http://lab.cs.ucsb.edu

At the Four Eyes Lab, we are conducting an investigation into credibility and trust metrics in Microblogs. In order for us to evaluate our models for predicting credible information in Twitter, we need to collect some ground truth data. Your assessment of the “credibility” of the tweets below will help us to perform this study. Thanks a lot for helping us out!

<table>
<thead>
<tr>
<th>Topic</th>
<th>#Libya</th>
<th>#Facebook</th>
<th>#Obama</th>
<th>#Japanquake</th>
<th>#LondonRiots</th>
<th>#Hurricane</th>
<th>#Egypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweets</td>
<td>28M</td>
<td>37M</td>
<td>5M</td>
<td>4M</td>
<td>4M</td>
<td>5M</td>
<td>36M</td>
</tr>
<tr>
<td>@Mentions</td>
<td>94M</td>
<td>62M</td>
<td>39M</td>
<td>25M</td>
<td>30M</td>
<td>35M</td>
<td>73M</td>
</tr>
<tr>
<td>@Replies</td>
<td>126K</td>
<td>708K</td>
<td>358K</td>
<td>131K</td>
<td>52K</td>
<td>114K</td>
<td>217K</td>
</tr>
<tr>
<td>Normal Tweets</td>
<td>37K</td>
<td>433K</td>
<td>162K</td>
<td>67K</td>
<td>26K</td>
<td>32K</td>
<td>49K</td>
</tr>
</tbody>
</table>

Overview of 7 topic-specific data collections mined from the Twitter streaming API.
Method

• Algorithm used
  – Use Weka3 toolkit
  – Train a J48(C4.5) Decision Tree Algorithm
  – 70:30 train-test ratio (both kept separate)
  – 10 Fold Cross Validation
Segmenting based on Topics

<table>
<thead>
<tr>
<th>Set Name</th>
<th>Core Tweeters</th>
<th>Core Tweets</th>
<th>$F_o$ and $F_e$ (overlapped)</th>
<th>$F_o$ and $F_e$ (distinct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libya</td>
<td>37K</td>
<td>126K</td>
<td>94M</td>
<td>28M</td>
</tr>
<tr>
<td>Superbowl</td>
<td>191K</td>
<td>227K</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Romney</td>
<td>226K</td>
<td>705K</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Facebook</td>
<td>433K</td>
<td>217K</td>
<td>62M</td>
<td>37M</td>
</tr>
<tr>
<td>EnoughIsEnough</td>
<td>85K</td>
<td>129K</td>
<td>13M</td>
<td>4M</td>
</tr>
<tr>
<td>Egypt</td>
<td>49K</td>
<td>217K</td>
<td>73M</td>
<td>36M</td>
</tr>
<tr>
<td>Earthquake</td>
<td>67K</td>
<td>131K</td>
<td>15M</td>
<td>5M</td>
</tr>
</tbody>
</table>

**TABLE II**
Overview of 7 topic-specific data collections mined from the Twitter streaming API.
Segmenting based on Behavior:

• For our experiments, a “dyadic pair” is a conversation between two twitter users that contains at least three messages. Tweets from such conversations make up the “dyadic pair” data set.
Outline

- Background
- Experimental Framework
- Credibility Models
- Results
  - Results
  - Credibility Predictions
  - Location and Devices
- Conclusion
Results: Retweet Chains

- Longer Tweets and tweets with URLs tend to be retweeted more frequently.
Results: Features Across Topics
Results: Credibility Distribution

- Analyzed feature distribution across credible and non-credible sets of tweets.
- E.g. Long tweets are usually more credible.
- E.g. Negative sentiment occurred more in tweets that were tagged as “not credible”.

![Chart showing credibility distribution for various features such as word, tweet_type, char, pronoun, url, sentiment_pos, sentiment_neg, retweet, frown, news, uppercase, ellipsis, num_hashtag, sentiment, num_mention, question, excl, smile. The chart displays the average occurrence score on a scale from 0 to 1 for each feature, with separate bars for mean(neg) and mean(pos).]
Results: Dyadic Pairs

- Analyzed sets of tweets that were part of pairwise conversations with at least three messages
- Conversational tweets tended to be shorter
- More use of uppercase terms in non-conversational tweets
- More retweet tags in conversational tweets
Results: Feature Utility Scores

- Computed the utility of each feature based on occurrence across all contexts in our experiments.
- Most useful features include tweet length, sentiment, url, use of uppercase.
Results: Per-Topic Features

- Analyzed how often our credibility indicators occurred in each of our topic-based slices.
- Credibility indicating features tended to be used more in emergency and unrest situations.
- Interestingly, less credibility-indicating features in the political data set “#Romney”.
Location and Devices

- Analysis on the Crawled Data Set shows the Distribution of **Frequent Information Sources** and **Topics**.

Word cloud showing distribution of popular terms in the Libya data set.
Word cloud showing origin of tweets in the Libya data set.
Outline

- Background
- Features and Contexts
- Experimental Framework
- Results

**Conclusion**
- Research Question (revisited)
- Conclusion
- Future Work
Future Work

- Integration of distribution knowledge into credibility-based filtering algorithms.
- Analysis of behavioral patterns for groups of features (a correlation-based analysis).
- Cognitive modeling of users while interacting with data from different filters.
Conclusion

- How are the features that indicate credibility distributed in Twitter?
- Feature distribution changes substantially across different slices of the network. (Dyadic, Topic-based, Chain-based segmentations)
- How/Why do they vary across different contexts?
- Many influencing factors. For example, strong indicators tend to occur more frequently in conversational tweets, and in topics about emergency or social unrest situations
Thank you!
Overview of Experimental Framework

- Diverse Network Data Sources
- Experimental Workbench
  - Cognitive Modeling Components (CMU)
  - Human Factors Analyses (SA Technologies)
- Credibility Filtering Pipeline
  - Large Scale. (UIUC)
  - Medium Scale. (UCSB)
  - Human Scale (UCSB)

Alignment

Data

Control Pipeline

Credibility in Context

Experimentation

TasteWeights
Social Impact

- As of February 2010