

Who will RT this?

Automatically Identifying and Engaging Strangers
on Twitter to Spread Information



Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle X. Zhou, **Jeffrey Nichols**
Utah State University & IBM Research – Almaden

jwnichols@us.ibm.com

 [@jwnichols](https://twitter.com/jwnichols)

Public Social Media Contains a Wealth of Information about Individuals...

Tweets

Retweeted by Kyumin Lee

Jalal Mahmud @jmahmud · Feb 24
Our paper on retweeter prediction #iui #iui2014 dl.acm.org/citation.cfm?i...

Expand

Retweeted by Kyumin Lee

Sue Moon (문수복) @sbmoon · Jan 30
정보과학회에서 전산분야성과지표로 왜 SCI를 서를 만들었습니다. 연구재단 미래부 기관평가 @nrf_sns @withmsip an.kaist.ac.kr/~sbr

Expand

Kyumin Lee @humanist0810 · Jan 14
@msbernst Congrats! You deserve it.

View conversation

Kyumin Lee @humanist0810 · Dec 13
@vandana_bvj very good. Enjoy your ne

View conversation

Kyumin Lee @humanist0810 · Dec 13
@vandana_bvj congrats! You deserve it

View conversation

Retweeted by Kyumin Lee

James Caverlee @TheRealCaverlee · Dec
Congrats to all our new graduates! cc:

Expand

Kyumin Lee @humanist0810 · Dec 4
@munmun10 congrats!

View conversation

Kyumin Lee @humanist0810 · Oct 18
@infolaber big time! G's stock price is c

Tweets

Jalal Mahmud @jmahmud · Feb 25
Our CSCW paper on inferring basic human value from dl.acm.org/citation.cfm?d... #CSCW2014

Expand

Jalal Mahmud @jmahmud · Feb 25
@marthalaldrige you are welcome to join next IUI, it

View conversation

Jalal Mahmud @jmahmud · Feb 25
@peterpaws how many papers are competing for top available in the website?

View conversation

Retweeted by Jalal Mahmud

Austin Kleon @austinkleon · Feb 25
Think you don't care about the California drought? CA all US fruits & veggies motherjones.com/environment

View summary

Retweeted by Jalal Mahmud

ACM CSCW @ACM_CSCW · Feb 19
CSCW 2015 will be March 14-18, 2015 in Vancouver, cscw.acm.org/2015/ #cscw2014 #cscw2015

Expand

Retweeted by Jalal Mahmud

Lora Aroyo @laroyo · Feb 25
John Riedl Session #IUI2014 Q&A; Business Rating F Services; Rating Bias & Crowdsourcing Annotations

Expand

Jalal Mahmud @jmahmud · Feb 24
Sorry to miss #iui2014 in person, looking forward to t experience

Tweets

Jeffrey Nichols @jwnichls · 10h
@wobbrockjo thoughts would appear and then be replaced with "wrong" and then disappear many times before something meaningful came out

View conversation

Jeffrey Nichols @jwnichls · 16h
.@wobbrockjo if I could control computers at the speed of thought, then I'd be spending even more time pressing "delete" than I do now

View conversation

Retweeted by Jeffrey Nichols

Leo Mutuku @C_Leo_patra · Feb 24
What Happens after the Olympics Leave Town? #resilientcities ...lientcities.rockefellerfoundation.org/blog/entry/wha...

Expand

Retweeted by Jeffrey Nichols

Chris Dixon @cdixon · Feb 24
In fact, one of the main advantages entrepreneurs have is a 5+ year time horizon vs big companies beholden to short term Wall Street types.

Expand

Retweeted by Jeffrey Nichols

Jeffrey P. Bigham @jeffbigham · Feb 23
i wonder if @netflix struck and leaked a deal to encourage backlash so net neutrality is reinstated ...

Expand

Retweeted by Jeffrey Nichols

Chris Dixon @cdixon · Feb 23
Innovation means running experiments. Some of the best will look silly at first. More parties who can block experiments = less innovation.

Expand

Jeffrey Nichols @jwnichls · Feb 14
The Boar Board looks awesome. courtesy of @TheIonsShareSD

Public Social Media Contains a Wealth of Information about Individuals...

Can we harness this information for something useful?

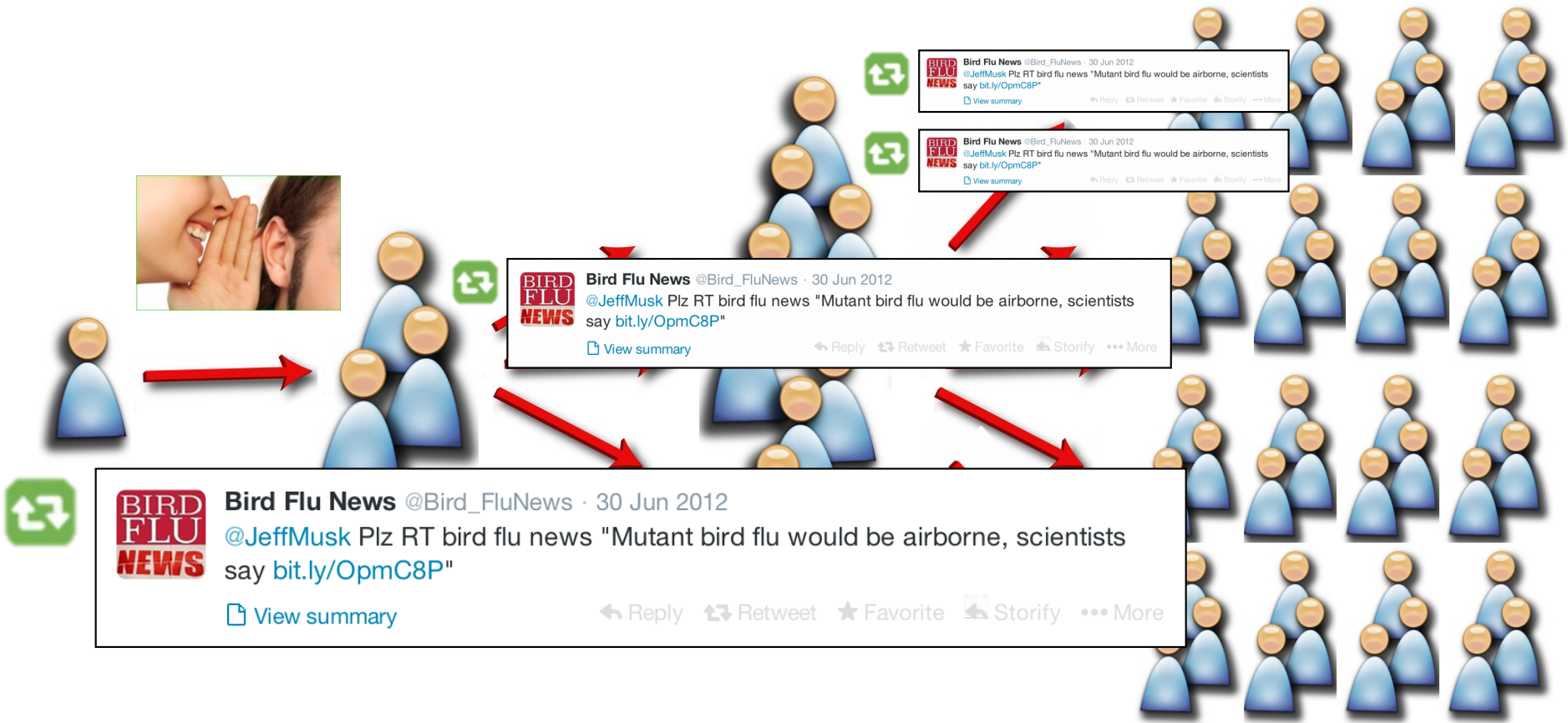
- Identify people to recruit to do various tasks
- Collect Information
- Spread Information

Today: Information Spreading



- Relevant marketing campaign messages
- Alerts and SOS messages in an emergency
- Etc.

Today: Information Spreading



Challenge: Low percentage of people respond to this task

- Can we predict who will retweet and direct requests only to them?
- Can we predict who will retweet more quickly?

Our Process

1. Data Collection
2. Feature Extraction
3. Feature Selection
4. Model Building
5. Evaluation

Ground-Truth Data Collection

Public Safety (location-based)

Bird Flu (topic-based)



Public Safety News @BayPublicSafety · 20 Jun 2012
@SARAHGAMBITCH Plz RT this public safety news "Medical emergency prompts 90-minute delays... bit.ly/Le9AuY"

Expand Reply Retweet Favorite Storify More



Bird Flu News @Bird_FluNews · 30 Jun 2012
@JeffMusk Plz RT bird flu news "Mutant bird flu would be airborne, scientists say bit.ly/OpmC8P"

[View summary](#) Reply Retweet Favorite Storify More

- Randomly selected users who tweeted from the San Francisco bay area (via geo-tags)
- Contacted **1,902** users
- 52 (**2.8%**) retweeted our message
- Message reached a total of **18,670** followers

- Randomly selected users who posted one of the following words in at least one tweet: “*bird flu*”, “*H5N1*” and “*avian influenza*”
- Contacted **1,859** users
- 155 (**8.4%**) retweeted our message
- Message reached a total of **184,325** followers



Public Safety News
@BayPublicSafety

Collect and send public safety news in the Bay Area. Please retweet the news to other residents in this area for their safety.
San Francisco, CA

150 TWEETS
8 FOLLOWING
28 FOLLOWERS



Bird Flu News
@Bird_FluNews

Collect and send bird flu news. Please retweet the news to your friends for their safety.

318 TWEETS
2 FOLLOWING
26 FOLLOWERS

Resulting Data

In total:

- Contacted 3,761 strangers
- 207 positive examples, 3554 negative examples

For each user we contacted, we collected:

- Twitter profile (screen name, tweet count, etc.)
- People they followed, followers,
- Up to 200 recent messages
- Ground truth (“retweeter” or “non-retweeter”)

Feature Extraction

Feature Categories

- Profile Features
- Social Network Features
- Personality Features
- Activity Features
- Past Retweeting Features
- Readiness Features

Feature Extraction

Feature Categories

- Profile Features
- Social Network Features
- Personality Features
- Activity Features
- Past Retweeting Features
- Readiness Features

Profile Features

- longevity (age) of an account
- length of screen name
- whether the user profile has a description
- length of the description
- whether the user profile has a URL

Social Network Features

- number of users following (friends)
- number of followers
- and the ratio of number of friends to number of followers

Feature Extraction

Feature Categories

- Profile Features
- Social Network Features
- **Personality Features**
- Activity Features
- Past Retweeting Features
- Readiness Features

Users' word usage has been found to predict their personality

- Linguistic Inquiry and Word Count (LIWC) dictionary
- Personality features derived from LIWC categories [Yarkoni 2010, Mahmud 2013]

Personality Features	Total Number	Examples
LIWC	68	Sadness, 1st Person Plural, Anxiety
Big Five	5	Agreeableness, Conscientiousness
Big Five Sub-Facets	30	Friendliness, Anxiety

Feature Extraction

Feature Categories

- Profile Features
 - Social Network Features
 - Personality Features
 - **Activity Features**
 - Past Retweeting Features
 - Readiness Features
- Number of status messages
 - Number of direct mentions (e.g., @johny) per status message
 - Number of URLs per status message
 - Number of hashtags per status message
 - Number of status messages per day during her entire
 - Account life (= total number of posted status messages / longevity)
 - Number of status messages per day during last one month
 - Number of direct mentions per day during last one month
 - Number of URLs per day during last one month
 - Number of hashtags per day during last one month

Feature Extraction

Feature Categories

- Profile Features
- Social Network Features
- Personality Features
- Activity Features
- Past Retweeting Features
- Readiness Features

Past Retweeting Behavior

- Number of retweets per status message: R/N
- Average number of retweets per day
- Fraction of retweets for which original messages are posted by strangers who are not in her social network

Readiness Based on Previous Activity

- Tweeting Likelihood (Day)
- Tweeting Likelihood (Hour)
- Entropy of Tweeting Likelihood (Day)
- Entropy of Tweeting Likelihood (Hour)
- Tweeting Steadiness
- Tweeting Inactivity

Predicting Retweeters

Training and Test Sets:

- Each dataset (public safety and bird flu) was randomly split to training set (2/3 data) and testing set (1/3 data)

5 Predictive Models

- Random Forest, Naïve Bayes, Logistic Regression, SMO (SVM) and AdaboostM1

Handling Class Imbalance

- Used both over-sampling (SMOTE) and weighting approaches (cost-sensitive approach)

Feature Selection

Computed χ^2 value for each feature in training

Feature Group	Significant Features (bolded is common to both data sets)
Profile	the longevity of the account
Social-network	following ratio of number of friends to number of followers
Activity	 URLs per day direct mentions per day hashtags per day status messages status messages per day during entire account life status messages per day during last one month
Past Retweeting	 retweets per status message retweets per day
Readiness	Tweeting Likelihood of the Day Tweeting Likelihood of the Day (Entropy)
Personality	7 LIWC features: Inclusive , Achievement, Humans, Time, Sadness, Articles, Nonfluencies 1 Facet feature: Modesty

21 Features Selected by χ^2 in Publish Safety Dataset

Feature Group	Significant Features (bolded is common to both data sets)
Profile	the length of description has description in profile
Activity	 URLs per day direct mentions per day hashtags per day URLs per status message direct mentions per status message hashtags per status message
Past Retweeting	 retweets per status message retweets per day URLs per retweet message
Readiness	Tweeting Likelihood of the Hour (Entropy)
Personality	34 LIWC features: Inclusive , Total Pronouns, 1st Person Plural, 2nd Person, 3rd Person, Social Processes, Positive Emotions, Numbers, Other References, Occupation, Affect, School, Anxiety, Hearing, Certainty, SZensory Processes, Death, Body States, Positive Feelings, Leisure, Optimism, Negation, Physical States, Communication 8 Facet features: Liberalism, Assertiveness, Achievement Striving, Self-Discipline, Gregariousness, Cheerfulness, Activity Level, Intellect 2 Big5 features: Conscientiousness, Openness

46 Features Selected by χ^2 in Bird Flu Dataset

Activity, personality, readiness and past retweeting feature groups have more significant power.
Six significant features (bolded names) are common to both sets.

Evaluating Retweeter Prediction

Only the significant features are used for prediction

Classifier	AUC	F1	F1 of Retweeter
Basic			
Random Forest	0.638	0.958	0
Naïve Bayes	0.619	0.939	0.172
Logistic	0.640	0.958	0
SMO	0.500	0.96	0
AdaBoostM1	0.548	0.962	0.1
SMOTE			
Random Forest	0.606	0.916	0.119
Naïve Bayes	0.637	0.923	0.132
Logistic	0.664	0.833	0.091
SMO	0.626	0.813	0.091
AdaBoostM1	0.633	0.933	0.129
Cost-Sensitive (Weighting, showing the best results in each model)			
Random Forest	0.692	0.954	0.125
Naïve Bayes	0.619	0.93	0.147
Logistic	0.623	0.938	0.042
SMO	0.633	0.892	0.123
AdaBoostM1	0.678	0.956	0.133

Prediction accuracy (Public Safety)

Classifier	AUC	F1	F1 of Retweeter
Basic			
Random Forest	0.707	0.877	0.066
Naïve Bayes	0.670	0.834	0.222
Logistic	0.751	0.878	0.067
SMO	0.500	0.876	0
AdaBoostM1	0.627	0.878	0.067
SMOTE			
Random Forest	0.707	0.819	0.236
Naïve Bayes	0.679	0.724	0.231
Logistic	0.76	0.733	0.258
SMO	0.729	0.712	0.278
AdaBoostM1	0.709	0.837	0.292
Cost-Sensitive (Weighting, showing the best results in each model)			
Random Forest	0.785	0.815	0.296
Naïve Bayes	0.670	0.767	0.24
Logistic	0.735	0.742	0.243
SMO	0.676	0.738	0.256
AdaBoostM1	0.669	0.87	0.031

Prediction accuracy (Bird Flu)

We use Random Forest for all following experiments.

Comparison with Two Baselines

Baselines

Random people contact

- Randomly select and ask a sub-set of qualified candidates

Popular people contact

- Sort candidates in our test set by their follower count in the descending order

Approach	Retweeting Rate in Testing Set	
	Public Safety	Bird flu
Random People Contact	2.6%	8.3%
Popular People Contact	3.1%	8.5%
Our Prediction Approach	13.3%	19.7%

Comparison of retweeting rates

Live Experiment

- To validate the effectiveness of our approach in a live setting, we used our recommender system to test our approach against the two baselines
- Randomly selected 426 candidates who had recently tweeted about “bird flu” in July 2013
- Each approach selected top 100 candidates based on its criteria

Approach	Retweeting Rate
Random People Contact	4%
Popular People Contact	9%
Our Prediction Approach	19%

Comparison of retweeting rates in live experiment

To wrap up...

- We have presented a **feature-based prediction model** that can automatically identify the right individuals at the right time on Twitter
- We have also described a **time estimation model**
- In the experiments, our approaches **doubled** the retweeting rates over the two baselines
- With our time estimation model, our approach **outperformed** other approaches significantly
- Overall, our approach **effectively identifies qualified candidates** for retweeting a message within a given time window

Thanks!



For more information, contact:

Jeffrey Nichols

jwnichols@us.ibm.com

 [@jwnichls](https://twitter.com/jwnichls)



Why Retweet a Stranger's Request?

We randomly selected 50 people who retweeted and asked them why they chose to retweet (33 replied)

Main reasons to retweet our requested message

- Trustworthiness of the content
"Because it contained a link to a significant report from a reputable media news source"
- Content relevance
"Because it happened in my neighborhood"
- Content value
"my followers should know this or they may think this info is valuable"

Real-Time Retweeter Recommendation

The screenshot shows a Twitter interface with two panels. The left panel, labeled 'Tweets', displays three recommended candidates for retweeting. The right panel shows a user's retweeting request, which is highlighted in yellow. The candidates are:

- leaveherforde**: Low-Path Bird Flu Reported in Friesland Poultry - <http://ThePoultrySite.com> : Low-Path Bird Flu Reported in Friesla... <http://bit.ly/153F75E> (Score: 0.987157)
- disneyrunne**: Scientists say it's likely that the new bird flu in China "passed between humans" for the first time in March. <http://gaw.kr/Y4qBNHR> (Score: 0.466862)
- David_Wile**: Study says rare bird flu strain h7n9 may be transmissible by air <http://on.wsj.com/18s3mTk> (Score: 0.12816)

The right panel shows a user's retweeting request, which is highlighted in yellow:

@%s Plz RT bird flu news: Quick Update on Recent Bird Flu Vaccine <http://bit.ly/16bC81u>

@%s Plz RT bird flu news: Bird Flu viruses could evolve in nature <http://bit.ly/MQBASY>

@%s US journal prints controversial bird flu research <http://bit.ly/MkHAF5> Plz RT bird flu news

Below the request, there is a text input field containing the following text:

```
@leaveherforde Plz RT bird flu news: Quick Update on Recent Bird Flu Vaccine http://bit.ly/16bC81u
```

A "Send Tweet" button is located at the bottom of the right panel.

The interface of our retweeter recommendation system: (a) left panel: system-recommended candidates, and (b) right panel: a user can edit and compose a retweeting request.