Exploiting Social Media for Natural Language Processing:
Bridging the Gap between Language-centric and Real-world Applications

Andrea Zielinski
&
Simone Paolo Ponzetto

Sofia, Bulgaria
August 4, 2013

Tutorial objectives

• Provide an introduction to the topic of Social Media and NLP, crucially including:
  – A review of Twitter and other micro-blogs’ main characteristics
  – Analyzing and extracting structured information from social media
  – High-end NLP applications exploiting them for real-world problems
  – Open issues

• Get you interested in working on this topic in the (near) future
  – Interesting research directions with many open issues
  – Many open areas to advance the state of the art
  – Real-world, high-impact applications leveraging large amounts of data
Tutorial outline I

- Social media and the wisdom of the crowd
  - Twitter and micro-blogging’s main characteristics
    - instant short-text messaging
    - a network of users
    - structured vs. non-structured features

Tutorial outline II

- Mining structured information from social media
  - Building a pre-processing pipeline for Twitter data
    - PoS tagging
    - Named Entity Recognition
  - Tackling difficult, high-level tasks
    - Modeling of Twitter conversation
    - Event detection
    - Topic tracking
Tutorial outline III

• Trend detection, social sensing and crisis management
  – Detecting breaking news/events from social media
    • consumer confidence
    • presidential job approval polls
    • stock market prices
    • flu epidemics
    • natural disasters like earthquakes, tsunami, etc.
  – real applications => end-user systems (DEMOS!)

PART 1
Twitter Basics
Introduction to Twitter

Twitter is a Microblogging Platform

– created in 2006
– global real-time communications platform
– offers users the ability to interact with various members in their community
– free messaging service, primarily internet-based, but also cell-phone compatible
– public messages, unless private direct message
– tweets appear as a stream of timely ordered messages

History of Social Media Networks

http://www.fredcavazza.net/2012/02/22/social-media-landscape-2012/
Twitter Statistics
June 2013

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of active registered Twitter users</td>
<td>554,750,000</td>
</tr>
<tr>
<td>Number of new Twitter users signing up everyday</td>
<td>135,000</td>
</tr>
<tr>
<td>Average number of tweets per day</td>
<td>58 million</td>
</tr>
<tr>
<td>Number of Twitter Users who use their phone to tweet</td>
<td>43%</td>
</tr>
<tr>
<td>Number of Twitter search queries</td>
<td>2.1 Billion</td>
</tr>
</tbody>
</table>

Twitter by Nation and Language

Top 20 countries in terms of Twitter accounts


http://www.webpronews.com/netherlands-twitter-activity-2012-02
Unique Tweet Identifier
Tweet_ID: 12296272736

Text of the Tweet
text: An early look at Annotation

Creation Date of Tweet
created_at: Wed May 23 06:01:13 +0000 2007

Time & Location Information from User Profile
Time_zone: Pacific Time (US & Canada)
Location: San Francisco/CA
Bounding_box:[[[-122.42284884, 37.76893497], [-122.3964, 37.76893497], [-122.3964, 37.78762897], [-122.42284884, 37.78762897]]

Point-coordinates provided by GPS
geo: nil
Basic Twitter Terms

**Direct Message:** Private message to a person a user is following.

**Replies:** A specific user TweetForUser can be directly addressed via the “@”-symbol as first symbol in the tweet. The message links to a previous message.

*Example:* @TweetForUser Yes, got it.

**Mentions:** In contrast to a reply, the message need not link to a previous message.

*Example:* Are you following @Username?

**Retweets:** Tweets originally posted by user AuthorOfTweet and forwarded to the Twitter community; preceded with “RT”

*Example:* RT @AuthorOfTweet Watch out! :{(}

Social Relations in Twitter

**Following-Relationship:**

- Twitter users subscribe to tweets of other users
- Relationship need not be reciprocal
- Twitter users are connected in a directed graph
Trending Topics

*Trending Topic* algorithm
- identifies popular topics on a daily basis
- based on frequent terms over time (≈ a week)
- tailored to individual users
  - based on the user’s location
  - based on the followings of a user

How does the information spread over the social network?

Research Questions:
- Who are the most popular/active/influential users in the network?
- What topics do users tweet about?
- Is twitter a timely provider of information?
- How many relationships are reciprocal and are users with a *following*-relationship more similar to each other (homophile) than those without based on the topics they are interested in?
Influence vs. Popularity

- **Popularity:**
  - A user’s popularity correlates with number of followers

- **Influence:**
  - A user’s influence correlates with audience engagements
    - Based on number of clicks, mentions, retweets w.r.t. user’s content

- **Most popular twitter users** (i.e., > 10,000 followers) are celebrities, politicians and the **media industry**.

The Social Ladder

![Social Ladder Diagram](image-url)

*Base: 17,874 US online adults (18+); 16,673 European online adults (18+)*

*Source: North American Technographics® Online Benchmark Survey, Q3 2011 US, Canadian European Information Technology Online Benchmark Survey Q3 2011*
Social Network Analysis

How Social Network is Evolving? - A Preliminary Study on Billion-scale Twitter Network \cite{Watanabe2012}

Social Network Analysis

\cite{Kwak2010}
Summary of Results:

Twitter diverges from social networks like Facebook:

- Twitter followers do not obey a power law
  - Twitter followers follow a fat-tailed distribution at the top end, but not as fat-tailed as a power law.

- Degree of separation is shorter
  - Milgram’s experiment = 6 vs. Twitter= 4.12

- Low degree of reciprocity
  - 67.6% are not followed by anyone, which might be due to users getting their "news" information from twitter.
  - Twitter is primarily a broadcasting medium and not a social network of users interacting with each other!

Topics in Social Networks

Topics related to the users’ interests derived from the twitter text using
- Terms (Term-Level Approach) and/or
- Semantic concepts (Category-Level Approach)

[from Kwak et al. 2010]

[from Java et al. 2010]
Example Tweets

@JayBosh1 haha I know I am :) oh no! I hope your ok and haven’t died :O I quite like earthquakes tho :P keep safe!!!xx

cuz u kno ppl. yu r wat u r.

Characteristics of Tweets

• Tweets are short messages (max. 140 chars)
  – Little context information available
• Informal language style
• Noisy text
  – Grammar and punctuation
  – Spelling and orthography
• Threading characteristics
  – Communicative acts (turn-taking, openings, closings, etc.)
NLP Analyses of Tweets
[from Liao et al. 2010]

Opportunities:
• Semi-structured, e.g. Social Network Structure, Timestamp, Geolocation
  ➢ Useful for topic, trend and event detection

Challenges:
• Little information content
  ➢ Filter out spam or ‘pointless babble’
• Informal, noisy text
  ➢ Text Normalization techniques required

Linguistic and Typographic Forms
[Ling & Baron 2007; Thurlow et al 2011]

• Avg. word (character) length of tweet ≈ 14 (65)
• Unorthodox linguistic forms, signal communicative immediacy:
  – Shortenings (i.e. missing end letters)
  – Contractions (i.e. missing middle letters)
  – G-clippings and other clippings (i.e. dropping final letter)
  – Acronyms and initialisms
  – Letter/number homophones
  – Intentional ‘misspellings’ and Typos
  – Accent stylizations
  – Logograms (e.g. xxxxxx and !!!!!)
  – Emoticons (e.g. 😊)
  – ‘Phonological approximations’
  – Onomatopoeic, exclamatory spellings (e.g. hahah!, arrrgh!, WOOHOO!, rrah, ahhh)
Out-Of-Vocabulary (OOV) words in Twitter, Facebook, SMS

Observation:
- Context required for handling ambiguities
  - Example: “hw”: “how” versus “hw”: “homework”

Morpho-phonemic variations:
- Letter & Number (“b4” instead of “before”)
- Single Letter (“c” instead of “see”)
- Number substitution (“4” instead of “four”)
  - Example: “lув”: “love” versus “lув”: “laugh”

Out-Of-Vocabulary (OOV) words in Twitter, NYT, SMS

Observation:
- Context required for handling ambiguities
  - Example: “hw”: “how” versus “hw”: “homework”

Basic transformations:
- Character clipping (“runnin” instead of “running”)
- Single Character (“c” instead of “see”)
- Repeat Character (“thaaaank” instead of “thank”)
- Contractions (“y’all” instead of “you all”)
- You-to-u (“u” instead of “you”)
- Th-to-d (“dose” instead of “those”)
- Drop Vowels (“nd” instead of “and”)
- Prefix-style Compression (“bout” instead of “about”)

Results:
- Most OOVs in Twitter are abbreviations (due to length restriction) and typos
- Non-standard orthography used intentionally
- Twitter language depends on geo-location and user input device
Structured vs. unstructured resources

- **Machine-readable resources** exhibit different *degrees of formalization and amount of structure*, ranging from *raw text* to *full-fledged ontologies*.


---

Semi-structured resources

Micro-blogs: formalization/structure

degree of structure


Text vs. tweets

• Both unstructured in nature, i.e. streams of characters

• Tweet-specific features
  – Limited content (140-character limit)
  – Abbreviations and contracted language
  – Tags: Hashtags (#), user mentions (@), retweet markers (RT), smileys ;)


### Ontologies vs. tweets

- They can be both viewed as graphs
  - Knowledge graph vs. social graph

- Highly formalized semantic structured vs. unstructured, colloquial language

- Static vs. dynamic, continuously updated content

### Wikis vs. tweets

- Dynamic, up-to-date resources

- Based on a large Web user base

- Lack of an explicit semantics

- Open-domain content
Micro-blogging as yet another semi-structured resource

- **Unstructured content**
  - i.e. Short text

- **Structured features**
  - Social graph
  - Tags

---

The big picture

PART 2
NLP ALGORITHMS

NLP pipeline

• Involves a sequence of NLP processing components
  – Tokenizer
  – PoS tagger
  – Named Entity recognizer
  – Chunker
  – etc.
NLP pipeline and Twitter text

• Typically, pipeline components are trained on formal, “clean” documents like, e.g. newswire text from the Penn Treebank

• Application to micro-blogging text is typically poor (Finin et al., 2010) because of the highly unstructured nature of Twitter text

Applying NLP tools to Twitter text

• Finin et al. (2010) apply NLTK and the Stanford NER without retraining to Twitter data

• Example of errors
  – how come when george bush wanted to take out millions for the war congress had no problem ... but when obama wants money for healthcare the ...
  – RT @woodmuffin: "jay leno interviewing sarah palin: the seventh seal starts to show a few cracks"
Part-of-Speech tagging

• Task definition
  – “The process of assigning a part-of-speech or other lexical class marker to each word in a corpus” (Jurafsky and Martin)

• Standard information sources for PoS tagging
  – Tags of other words in the context
  – The word itself

• Standard approaches:
  – Rule-based Taggers
  – Statistical Taggers
    • E.g., HMM Tagger

Part-of-Speech Tagging for Twitter
[Gimpel et al., 2011]

• Main contributions
  – a PoS tagset for Twitter
  – a corpus of 1,827 PoS-annotated English tweets
  – statistical tagger using Twitter-specific features
Tweets with gold annotations

(a) @Gunservately @ obozo^ will go nuts when PA^ elects a Republican Governor next Tue^'. Can you say redistricting ?,

(b) Spending the day with mommymomma !

(c) lmao,..., s/o to the cool ass asian officer. 4 #1 not runnin my license and #2 not takin drunk to jail , Thank u God^', #amen#


PoS tagset for Twitter

Twitter PoS tagger

• Conditional Random Field (CRF)

• Base features
  – word type
  – word contains digits or hyphens
  – suffix features (up to length 3)
  – capitalization patterns

Twitter-specific features

• **TWORTH**: Twitter orthography
  – at-mentions, hashtags, and URLs
  – Regular expression-style rules

• **NAMES**: Frequently-capitalized tokes
  – Likelihood of capitalization of a token: \( \frac{N_{\text{cap}} + \alpha C}{N + C} \)
    • \( \alpha \): prior probability
    • \( C \): prior weight
Twitter-specific features

- **TAGDICT**: Traditional tag dictionary
  - tags that each word occurs with in the PTB conjoined with frequency rank

- **DISTSIM**: Distributional similarity
  - computed from truncated SVD of the successor and predecessor transition matrices

- **METAPH**: Phonetic normalization
  - coarse phonetic normalization of words based on rules to rewrite consonants and delete vowels

PoS tagging tweets: evaluation

- Dataset: 1,827 annotated tweets
- Comparison against the Stanford Tagger using a standard feature set

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>all features</td>
<td>89.37</td>
</tr>
<tr>
<td>base features</td>
<td>83.38</td>
</tr>
<tr>
<td>Stanford tagger</td>
<td>85.85</td>
</tr>
</tbody>
</table>
Improving Twitter PoS tagging with word clusters [Owoputi et al., 2013]

- Improve tagging accuracy with \textbf{large-scale unsupervised word clustering}

- Additional lexical features
  - \textbf{tag-word feature} from tags in PTB
  - \textbf{token-level gazetteer feature}:
    - Freebase lists of celebrities and video games
    - the Moby Words list of US Locations
    - lists of male, female, family and proper names

Unsupervised word clusters from tweets

- hierarchical 1,000 word clusters via \textit{Brown agglomerative clustering} [Brown et al., 1992] on a large set of unlabeled tweets

- Brown clustering
  - Begin with every word in its own cluster
  - Until we have one cluster
  - Merge two clusters so as to optimize the likelihood of a HMM with a one-class-per-lexical-type constraint

\[
p(w_1, w_2, \ldots, w_T) = \prod_{i=1}^{n} p(w_i | C(w_i)) p(C(w_i) | C(w_{i-1}))
\]
Unsupervised word clusters from tweets

Clusters are hierarchically arranged in a binary tree ⇒ **Use the root-to-leaf path as feature**


---

Improving Twitter PoS tagging: datasets

- **Oct27Test**: the original dataset from [Gimpel et al., 2011] of 1,827 tweets

- **Daily547**: one random English tweet from every day between January 1, 2011 and June 30, 2012
Improving Twitter PoS tagging: evaluation

<table>
<thead>
<tr>
<th>Feature set</th>
<th>OCT27TEST</th>
<th>DAILY547</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>91.60</td>
<td>92.80</td>
</tr>
<tr>
<td>with clusters; without tagdicts, namelists</td>
<td>91.15</td>
<td>92.38</td>
</tr>
<tr>
<td>without clusters; with tagdicts, namelists</td>
<td>89.81</td>
<td>90.81</td>
</tr>
<tr>
<td>only clusters (and transitions)</td>
<td>89.50</td>
<td>90.54</td>
</tr>
<tr>
<td>without clusters, tagdicts, namelists</td>
<td>86.86</td>
<td>88.30</td>
</tr>
<tr>
<td>Gimpel et al. (2011) version 0.2</td>
<td>88.89</td>
<td>89.17</td>
</tr>
<tr>
<td>Inter-annotator agreement (Gimpel et al., 2011)</td>
<td>92.2</td>
<td></td>
</tr>
<tr>
<td>Model trained on all OCT27</td>
<td></td>
<td>93.2</td>
</tr>
</tbody>
</table>


Named Entity Recognition

- **Task definition**
  - Classify entity mentions into coarse-grained semantic classes such as PERSON, LOCATION, ORGANIZATION, etc.

- **Standard information sources in NER tagging**
  - Tags of other words in the context
  - The word itself
  - Its capitalization

- **Standard approaches**: 
  - Rule-based Taggers
  - Statistical Taggers
T-NER [Ritter et al., 2011]

- Off-the-shelf NER taggers perform poorly

- Plethora of distinctive, infrequent types
  - Bands, Movies, Products, etc...
  - Very little training data

- Very terse (often contain insufficient context)
- Inconsistent capitalization

Step 1: Learning NE boundaries

- Sequence Labeling Task
  - IOB encoding

- Conditional Random Fields

- Features:
  - Orthographic
  - Dictionaries
  - Contextual

<table>
<thead>
<tr>
<th>Word</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Mobile</td>
<td>B-ENTITY</td>
</tr>
<tr>
<td>to</td>
<td>O</td>
</tr>
<tr>
<td>release</td>
<td>O</td>
</tr>
<tr>
<td>Dell</td>
<td>B-ENTITY</td>
</tr>
<tr>
<td>Streak</td>
<td>I-ENTITY</td>
</tr>
<tr>
<td>7</td>
<td>I-ENTITY</td>
</tr>
<tr>
<td>on</td>
<td>O</td>
</tr>
<tr>
<td>Feb</td>
<td>O</td>
</tr>
<tr>
<td>2nd</td>
<td>O</td>
</tr>
</tbody>
</table>
## Features for learning NE boundaries

- **Orthographic**
  - capitalization - based on the output of a supervised classifier, T-CAP, which predicts whether or not a tweet is informatively capitalized

- **Dictionaries**
  - uses a set of type lists gathered from Freebase

- **Contextual**
  - words/PoS/chunks of neighboring words (based on in-domain-trained T-PoS and T-Chunk)

---

## Annotating Twitter Named Entity data

- Manually Annotated the 2,400 tweets with gold NE boundaries and 10 entity types
- Enables training on in-domain data
NER segmentation: evaluation

![Bar chart showing evaluation results for Stanford and T-NER.](image)

NER segmentation: evaluation

![Bar chart showing F1 scores for different methods.](image)
Step 2: Named Entity classification

• Given the previously identified boundaries, named entity still need to be classified into (Twitter-specific) coarse-grained semantic classes

  – PERSON, GEO-LOC, COMPANY, FACILITY, PRODUCT, BAND, SPORTTEAM, MOVIE, TV-SHOW, OTHER

Freebase: a large NE dictionary

• Freebase lists provide a source of supervision
• But entities often appear in many different lists, for example “China” could be:
  – A country
  – A band
  – A person (member of the band “metal boys”)
  – A film (released in 1943)

• 35% of the entities appear in more than one Freebase dictionary
• 30% of the entities do not appear in Freebase
Distant supervision with Topic Models

- Treat each entity string as a “document”
  - Associate a BOW with each entity, made up of words that co-occur with it

- Labeled LDA [Ramage et al. 2009]
  - Constrained Topic Model
  - Each entity is associated with a distribution over topics (constrained based on FB dictionaries)
  - Each topic is associated with a type (in Freebase)

Generative story

Seattle
- $P(\text{TEAM}|\text{Seattle}) = 0.6$
- $P(\text{LOCATION}|\text{Seattle}) = 0.4$
- Is a TEAM
- Is a LOCATION
- Victory
- Airport

Type 1: TEAM
- $P(\text{victory}|T1) = 0.02$
- $P(\text{played}|T1) = 0.01$
- ...

Type 2: LOCATION
- $P(\text{visiting}|T2) = 0.05$
- $P(\text{airport}|T2) = 0.02$
- ...

http://homes.cs.washington.edu/~aritter/twitter_ner.pptx
Topic Models: data/inference

- Gather entities and words which co-occur
  - Using about 60M status messages
- Used a set of 10 types from Freebase
  - Commonly occur in Tweets
  - Good coverage in Freebase
- Inference: Collapsed Gibbs sampling
  - Constrain types using Freebase
  - For entities not in Freebase, don’t constrain

Type lists

<table>
<thead>
<tr>
<th>Type</th>
<th>Top 20 Entities not found in Freebase dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRODUCT</td>
<td>nintendo ds lite, apple ipod, generation black, ipod nano, apple iphone, gb black, xperia, ipods, verizon media, mac app store, kde, ldi video, nokia n8, ipads, iphone/ipod, galaxy tab, samsung galaxy, playstation portable, nintendo ds, vps</td>
</tr>
<tr>
<td>TV-SHOW</td>
<td>pretty little, american skins, nof, order svu, greys, kktny, rhobh, parks &amp; recreation, parks &amp; rec, dawson’s creek, big fat gypsy weddings, big fat gypsy wedding, winter wipeout, jersey shores, idiot abroad, royle, jerseyshore, mr. sunshine, hawaii five-0, new jersey shore</td>
</tr>
<tr>
<td>FACILITY</td>
<td>voodoo lounge, grand ballroom, crash mansion, sullivan hall, memorial union, rogers arena, rockwood music hall, amway center, el mocambo, madison square, bridgestone arena, cat club, le poisson rouge, bryant park, mandalay bay, broadway bar, ritz carlton, mgm grand, olympia theatre, consol energy center</td>
</tr>
</tbody>
</table>

Classification results (gold segmentation)

- the most frequent class (PERSON)
- Only Freebase "monosemous" entities
- Vanilla MaxEnt classifier (4-fold cross-validation)
- co-training algorithm [Collins and Singer, 1999]

Segmentation + Classification

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>COTRAIN-NER (10 types)</td>
<td>0.55</td>
<td>0.33</td>
<td>0.41</td>
</tr>
<tr>
<td>T-NER(10 types)</td>
<td>0.65</td>
<td>0.42</td>
<td>0.51</td>
</tr>
<tr>
<td>COTRAIN-NER (PLO)</td>
<td>0.57</td>
<td>0.42</td>
<td>0.49</td>
</tr>
<tr>
<td>T-NER(PLO)</td>
<td>0.73</td>
<td>0.49</td>
<td>0.59</td>
</tr>
<tr>
<td>Stanford NER (PLO)</td>
<td>0.30</td>
<td>0.27</td>
<td>0.29</td>
</tr>
</tbody>
</table>

PLO = Person, Location, Organization (3 types)
Take-home messages

• Achieve state-of-the-art performance by:
  – annotating a few, in-domain annotated data
  – design domain-specific features
  – generalize lexical features – e.g., Brown clustering
  – exploit external knowledge – e.g., Freebase

The story so far...

• PoS tagging and Named Entity Recognition are “only” pre-processing tasks

• We now turn to higher-level tasks and high-end applications
  – Discourse analysis
  – Event detection
  – Topic tracking
Unsupervised Modeling of Twitter Conversations [Ritter et al., 2010]

• Conversation make up 10-20% of Twitter

1. I'm going to the beach this weekend! Woo! And I'll be there until Tuesday. Life is good.
2. Enjoy the beach! Hope you have great weather!
3. thank you :) 

• Associate each utterance with a label describing its role in the conversation (dialogue/speech act)

---

Dialogue/speech acts in Twitter

1. I'm going to the beach this weekend! Woo! And I'll be there until Tuesday. Life is good.
2. Enjoy the beach! Hope you have great weather!
3. thank you :)
Discourse constraints

1. I'm going to the beach this weekend! Woo! And I'll be there until Tuesday. Life is good.
2. Enjoy the beach! Hope you have great weather!
3. thank you :)

Words indicate dialogue act

1. I'm going to the beach this weekend! Woo! And I'll be there until Tuesday. Life is good.
2. Enjoy the beach! Hope you have great weather!
3. thank you :)
Conversation specific topic words

1. I'm going to the beach this weekend! Woo! And I'll be there until Tuesday. Life is good.
2. Enjoy the beach! Hope you have great weather!
3. thank you :)

Content Modeling [Barzilay & Lee, 2004]

- Summarization
- Model order of events in news articles
  - Very specific topics: e.g. Earthquakes
- Model
  - Sentence-level HMM
    - states emit whole sentences
    - Learn parameters with EM

Where, when

Richter scale

Damage
Conversation Model


Conversation + Topic Model

Modeling Twitter conversations: evaluation

• Task: predict sentence order
• Generate all permutations of a conversation
  – Compute probability of each
  – How similar is the highest ranked to the original?
  – Measure permutation similarity with Kendall Tau
    • Counts number of swaps needed to get desired order

Conversation ordering: evaluation
Extracting Events and Event Descriptions from Twitter [Popescu et al., 2011]

• Problem formulation:
  – given a set of tweets (a so-called snapshot) about a specific entity
  – decide whether they describe a single central event focused around the entity

• Viewed as a supervised classification problem

Extracting events from Twitter: EventBasic

• Classify a snapshot based on a large set of Twitter-based and external features, e.g.
  – number of action verbs
  – entity buzziness in Twitter on the given day
  – entity buzziness in news on the given day

• Entity buzziness: \[
\frac{\sum_{i \in prev(s, N)} |s_i|}{N}
\]
Extracting Events from Twitter: *EventAboutness*

- Augment *EventBasic*’s feature set with *aboutness* information

  - Use a *document aboutness* system to rank the entities in a snapshot
  - Key idea: *events are about a few important entities*
  - Example features: *mean* and *std* of the 3 highest-scoring entities

Event detection from Twitter: evaluation

- Performance on event detection

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Avg P</th>
<th>AROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EventBasic</td>
<td>69.1</td>
<td>63.2</td>
<td>66.0</td>
<td>75.1</td>
<td>79.1</td>
</tr>
<tr>
<td>EventAboutness</td>
<td>70.2</td>
<td>64.1</td>
<td>67.0</td>
<td>75.2</td>
<td>78.8</td>
</tr>
</tbody>
</table>

- Performance on main entity extraction

<table>
<thead>
<tr>
<th>System</th>
<th>MRR</th>
<th>Prec@1</th>
<th>Prec@3</th>
<th>Prec@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>95.6</td>
<td>67.6</td>
<td>82.6</td>
<td>87.3</td>
</tr>
<tr>
<td>Aboutness</td>
<td>96.5</td>
<td>68.2</td>
<td>83.6</td>
<td>88.2</td>
</tr>
</tbody>
</table>

Source: Ana-Maria Popescu, Marco Pennacchiotti, and Deepa Paranjpe. Extracting events and event descriptions from Twitter. In Proc. of WWW-11
Extracting event description

• Given an event snapshot and its main entities
  – PoS tag the tweets (using Brill’s tagger)
  – Apply regular expressions

• Two tasks:
  – Action extraction
  – Opinion extraction

Extracting event description

• Action extraction:
  – ENTITY Verb NP
  – 77% grammatical, 68% appropriate summary

• Opinion extraction:
  – ENTITY (be | look | seem) (ADJP | NP)
  – I VP ENTITY
  – do sentiment-dictionary lookup
  – 85% grammatical
  – 78% spotted by the dictionary (84% accuracy)
Example event extraction and description

<table>
<thead>
<tr>
<th>Snapshot</th>
<th>Julia Roberts, 2010-01-28, Golden Globes attendance</th>
</tr>
</thead>
</table>
| Example Tweets | “Juli Roberts looks absolutely stunning! ..”  
“Iol juli Roberts is faddedddd”  
“I may have had one too many white Russians but doesn’t julia Roberts look like madge?”  
“#goldenglobes julia roberts presenting the best picture award 2 avatar. me sooo sad” |
| Main entities | julia roberts, golden globes |
| Audience opinions | + julia roberts : absolutely stunning  
-julia roberts : faddedddd  
+julia roberts : like madge  
+julia roberts : so kool |
| Main entities’ actions | julia roberts : presenting : best picture award  
-julia roberts : bustin : on nbc  
-julia roberts : sitting by : sir paul |

Source: Ana-Maria Popescu, Marco Pennacchiotti, and Deepti Paranjpe. Extracting events and event descriptions from Twitter. In Proc. of WWW-11

Open Domain Event Extraction from Twitter [Ritter et al., 2012]

- Extract large amounts of temporally anchored events from tweets

<table>
<thead>
<tr>
<th>Entity</th>
<th>Event Phrase</th>
<th>Date</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Jobs</td>
<td>died</td>
<td>10/6/2011</td>
<td>DEATH</td>
</tr>
<tr>
<td>iPhone</td>
<td>announcement</td>
<td>10/4/2011</td>
<td>PRODUCTLAUNCH</td>
</tr>
<tr>
<td>GOP</td>
<td>debate</td>
<td>9/7/2011</td>
<td>POLITICALEVENT</td>
</tr>
</tbody>
</table>
Event phrases

- Provide a description of the event
- Useful to categorize events into types

Dataset creation
- Annotated 1,000 tweets (19,484 tokens)
- Similar to EVENT tags in TimeBank
- Sequence-labeling problem
  - IOB Encoding
  - Conditional Random Fields

Learning to extract event phrases

- **Contextual features:**
  - POS tags
  - adjacent words
- **Dictionary Features:**
  - event terms from WordNet [Sauri et al., 2005]
  - Brown Clusters
- **Orthographic Features:**
  - prefixes
  - suffixes
Event phrase extraction: evaluation

<table>
<thead>
<tr>
<th>Entity</th>
<th>Event Phrase</th>
<th>Date</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Jobs</td>
<td>died</td>
<td>10/6/2011</td>
<td>?</td>
</tr>
<tr>
<td>iPhone</td>
<td>announcement</td>
<td>10/4/2011</td>
<td>?</td>
</tr>
<tr>
<td>GOP</td>
<td>debate</td>
<td>9/7/2011</td>
<td>?</td>
</tr>
</tbody>
</table>

• We still need to type the extracted events
Categorizing event types

• Many different types
• Not sure what is the right set of types
• Set of types might change over time as different topics become more or less popular
• Different types might be appropriate for different groups of users

Unsupervised event type induction

• Latent Variable Models
  – Generative Probabilistic Models

• Advantages
  – Discovers types which match the data
  – No need to annotate individual events
  – Does not commit to a specific set of types
Each Event Phrase is modeled as a mixture of types

\[
P(\text{SPORTS} | \text{cheered}) = 0.6 \\
P(\text{POLITICS} | \text{cheered}) = 0.4
\]

Each Event Type is associated with a distribution over entities and dates

---

**Example event types**

<table>
<thead>
<tr>
<th>Label</th>
<th>Top 5 Event Phrases</th>
<th>Top 5 Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>tailgate - scrimmage - tailgating - homecoming - regular season</td>
<td>espn - ncaa - tigers - eagles - varsity</td>
</tr>
<tr>
<td>Concert</td>
<td>concert - presale - performs - concerts - tickets</td>
<td>taylor swift - toronto - britney spears - rihanna - rock</td>
</tr>
<tr>
<td>Perform</td>
<td>matinee - musical - priscilla - seeing - wicked</td>
<td>shrek - les mis - lee evans - wicked - broadway</td>
</tr>
<tr>
<td>TV</td>
<td>new season - season finale - finished season - episodes - new episode</td>
<td>jersey shore - true blood - glee - dvr - hbo</td>
</tr>
<tr>
<td>Movie</td>
<td>watch love - dialogue theme - inception - hall pass - movie</td>
<td>netflix - black swan - insidious - tron - scott pilgrim</td>
</tr>
<tr>
<td>Sports</td>
<td>inning - innings - pitched - homered - homer</td>
<td>mlb - red sox - yankees - twins - dl</td>
</tr>
<tr>
<td>Politics</td>
<td>presidential debate - osama - presidential candidate - republican debate - debate performance</td>
<td>obama - president obama - gop - cnn - america</td>
</tr>
</tbody>
</table>

Categorizing event types: evaluation

- **Dataset**
  - 500 (entity, date) pairs annotated with event types
  - Using types discovered by the topic model
- **Baseline:**
  - Supervised classification using 10-fold cross validation
  - Treat event phrases like bag of words

Topic Tracking in Tweet Streams
[Lin et al., 2012]

• The problem: tracking broad topics such as “baseball” and “fashion” in continuous streams of short texts, i.e. tweets

• The solution: use adaptive streaming language models
  – Use hashtags as proxies for topic labels
  – Train per-topic language models
  – Classify tweets based on perplexity

Language models

• Probability distribution of a word sequence

\[ P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_1 | w_2)P(w_3 | w_1, w_2)\ldots P(w_n | w_1 \ldots w_{n-1}) \]

• Unigram LM:

\[ P(w_n | w_1 \ldots w_{n-1}) \approx P(w_n) \]

• Bigram LM:

\[ P(w_n | w_1 \ldots w_{n-1}) \approx P(w_n | w_{n-1}) \]

• Classify based on perplexity threshold

\[ \text{Perplexity measures "surprise"} \]

\[ \text{pow} \left[ 2, -\frac{1}{N} \sum_{i=1}^{n} \log_2 P(w_i) \right] \]
Key issues in topic tracking

- **Recency**: need to keep track of recent events
- **Sparsity**: need to smooth

**General strategy** = integrate two components
- “Foreground model” to keep (recent) up-to-date statistics
- “Background model” to combat sparsity

**Key questions**
- How do we keep track of history?
- How do we smooth?

History and the topic tracking model

- **Heap’s Law**: vocabulary size grows unbounded with increasing amounts of text data
  - *must* “forget” some data to make the algorithm run in constant space

- History (i.e., recently encountered terms):
  - 1000 terms, 10000 terms...
  - Act like a “buffer”

- Different methods for maintaining context:
  - **Forget**: forget everything periodically
  - **Queue**: moving window (FIFO)
  - **Epoch**: throw away infrequent events periodically [Goyal et al., 2009]
Smoothing the per-topic LMs

Absolute Discounting

$$P(w) = \frac{\max(c(w;h) - \delta, 0)}{\sum_w c(w;h)} + \frac{\delta \cdot w}{\sum_w c(w;h)} P_b(w)$$

foreground  background

Jelinek-Mercer smoothing

$$P(w) = \lambda \frac{c(w;h)}{\sum_w c(w;h)} + (1 - \lambda) \cdot P_b(w)$$

foreground  background

Smoothing the per-topic LMs

Bayesian smoothing using Dirichlet priors

$$P(w) = \frac{c(w;h) + \mu \cdot P_b(w)}{\sum_w c(w;h) + \mu}$$

“Normalized” Stupid Backoff (Brants et al., EMNLP 2007)

$$P(w) = \begin{cases} 
\frac{1}{1 + \alpha} \cdot \sum_w c(w;h) & \text{if } c(w;h) > 0 \\
\frac{\alpha}{1 + \alpha} \cdot P_b(w) & \text{otherwise}
\end{cases} \quad = \text{foreground}
$$

$$P(w) = \frac{\alpha}{1 + \alpha} \cdot P_b(w) \quad = \text{background}$$
Topic tracking: evaluation

• Data
  – stream from October 1 to October 7, 2010
  – ~94m tweets per day, ~11m contain hashtags
  – 10 topics: #nfl, #apple, #teaparty
• Perform a separate run for each topic

Topic tracking: evaluation

• Initialize with the background model and an empty foreground model
• For each tweet
  – Discard if it does not contain an appropriate hashtag
  – Remove hashtag
  – Compute perplexity wrt model
  – Update the foreground model

• Intrinsic: compare perplexity between
  – Baseline “background” only
  – Different “background” + “foreground” combinations
• Extrinsic: R/P/F1 for different perplexity thresholds
Topic tracking: results

• **Smoothing**
  – Jelinek-Mercer achieves lowest perplexity

• **Context**
  – Longer is better, but shorter isn’t that bad
  – “Queue” works well, but “Forget” isn’t that bad

• Per topic perplexity varies a lot
• Adding “foreground” helps

Extrinsic evaluation: unigram vs. bigram

PART 3
APPLICATIONS

Overview of Twitter Applications

• Epidemic Surveillance
• Disaster Management
• Voting Prediction
• Journalism
Overview of Twitter Applications

• **Epidemic Surveillance:**
  – Outbreak of an epidemic disease
  – Food contamination
  – Patient’s mood surveillance
  – Self medication of patients
  – ...

Overview of Twitter Applications

• **Disaster Management**
  – Outbreak of a natural disasters (earthquake, tsunami, ..) or other critical events (conflicts at festivals, ..)
  – How is the situation on the ground?
  – Do people require help?
  – How is the state of the roads? ..
  – ...

Overview of Twitter Applications

• Voting Prediction
  – How will people vote at the next election?
  – Do they favor a certain candidate?
  – What topics do they discuss?
  – What attitude do voters have w.r.t. certain topics?
  – ...

• Journalism
  – What topics are discussed on twitter?
  – Is there a trending topic?
  – Is the informant who sent the tweet trustworthy?
  – ..
Trending Now: Using Social Media to Predict and Track Disease Outbreaks


Trending Now: Using Social Media to Predict and Track Disease Outbreaks
Charles W. Interzit
Additional article information

site: www.nowtrendingchallenge.com

Mar 16, 2012 - Jun 01, 2012

Winner of NowTrendingChallenge: MappyHealth

How our App Works

Requirements:

- Show trending list of diseases
- Frequency count of tweets on health topics for a 24 hour period
- Selection of a user-defined geographic area of interest
- Use Illness Taxonomy
Based on pure Frequency counts

Center for Disease Control and Prevention (CDC)

- Ground truth data: Center for Disease Control and Prevention (CDC)

- Disadvantages:
  - Small Scale
    Influenza patient data covers not all clinics from all cities
  - Time Delay
    1–2 week reporting lag
Google Flu Trends

Google Flu Trends (GFT) is a novel Internet-based influenza surveillance system that uses search engine query data to estimate influenza activity and is available in near real time.

• GFT: By aggregating Google search queries on specific search terms that are good indicators of flu, accurate predictions can be made about its activity.

• Ginsberg et al. (2008) report a high Pearson correlation coefficient of 0.97 compared to CDC.

http://www.google.org/flutrends/

Google Queries Logs vs. Tweets

Pros & Cons:
• Tweets have higher signal-to-noise ratio
• Tweets provide higher volume of data
• Tweets longer than query logs
  • e.g., ‘fever’, ‘headache’ are polysemous
• Google search logs not publicly available

Basic Issues for queries logs and tweets:
• High media coverage
• Data sampling
Quantitative Approaches to Flu Detection

Task:
Given two plots of time series:
Are the trends the same?
Is there a relationship between number of tweets on flu and number of flu patients?

Pearson Correlation Coefficient - quantifies the strength of the linear relationship between a pair of variables.

INFLU kun: An Influenza Surveillance System for Japan

[Aramaki, 2011]

- *INFLU kun* is based on Machine Learning techniques to classify positive vs. negative tweets related to influenza epidemics.
- Correlation with statistics from Infection Disease Surveillance Center (IDSC)
  - 0.89 in non-excessive news period (early stage of influenza)

Basic Algorithm:
- Keyword-based filter with “influenza”
- SVM based on BoW with restricted context window of size 6 (f-measure: 76%)
- Positive Annotation of a tweet implies
  - >1 person with influenza in same location (according to user-profile)
  - tweet in present tense or recent past

[Demo: http://mednlp.jp/influ]
dizie - Disease Information Extraction from Tweets

• *Dizie* tracks 6 syndrome classes for major cities worldwide.

**Basic Algorithm:**
- Keyword Terms based on BioCaster Ontology and annotated tweets (2,000 tweets/syndrome)
- SVM and NB based on BoW tested (NB: f-measure 84%-89%)

[http://born.nii.ac.jp/_dev/static/trends](http://born.nii.ac.jp/_dev/static/trends)

---

Surveillance Results for different Approaches

[Drezde 2013]

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>2009</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keywords</td>
<td>0.970</td>
<td>0.646</td>
</tr>
<tr>
<td>Flu Classifier</td>
<td>0.970</td>
<td>0.519</td>
</tr>
<tr>
<td>Google Flu Trends</td>
<td>0.970</td>
<td>0.897</td>
</tr>
</tbody>
</table>
Flu Detection

**PIPELINE CLASSIFIERS**

- Four steps using supervised machine learning + NLP
  - Step 1: Identify health tweets
  - Step 2: Identify flu related
  - Step 3: Awareness vs. Infection
  - Step 4: Self vs. Other

Flu Detection

**STEP 1: HEALTH TWEETS**

- Method: supervised machine learning to label tweets
  - health related vs. not-health related
- Labeled data: 5,128 tweets using MTurk
- SVM classifier: unigrams, bigrams, etc.
  - Tuned for .90 precision
- Run on targeted stream of Tweets
**STEP 2: FLU TWEETS**

- Method: supervised machine learning to label tweets:
  about flu vs. not about flu/media
- Labeled data: 11,990 tweets using MTurk and hand corrected
- MaxEnt classifier, tuned for precision

**STEP 3: INFECTION VS. AWARENESS**

- Labeled tweets for:
  - Infection: reports an influenza infection
  - Awareness: shows awareness of the flu
  - MaxEnt classifier, tuned for precision
Flu Detection

STEP 4: SELF VS OTHER

Labeled tweets for:
- Self: the author has the flu
- Other: another person has the flu
- MaxEnt classifier, tuned for precision

Part of Speech Templates

defined on sequences of words, word classes, PoS

<table>
<thead>
<tr>
<th>Part of Speech Template</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronoun/Last Noun (Other)</td>
<td>I am worried that my son has the flu</td>
</tr>
<tr>
<td>Pro-Drop (Self)</td>
<td>Getting the flu.</td>
</tr>
<tr>
<td>Numeric Reference (Awareness)</td>
<td>So many people dying from the flu, I’m scared!</td>
</tr>
<tr>
<td>Pronoun/Noun (Awareness)</td>
<td>I had feared the flu</td>
</tr>
<tr>
<td>Flu-Noun before Verb (Awareness)</td>
<td>Flu can be dangerous</td>
</tr>
<tr>
<td>Noun in Question (Awareness)</td>
<td>Do you think that there will be an outbreak of flu?</td>
</tr>
</tbody>
</table>
Impact of NLP Features

Correlations for Flu Season 2011/2012

<table>
<thead>
<tr>
<th>Feature Removed</th>
<th>A / I</th>
<th>S / O</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-grams</td>
<td>0.6701</td>
<td>0.8440</td>
</tr>
<tr>
<td>Word Classes</td>
<td>0.7735</td>
<td>0.8549</td>
</tr>
<tr>
<td>Stylometry</td>
<td>0.8011</td>
<td><strong>0.8522</strong></td>
</tr>
<tr>
<td>Pronoun/Last Noun</td>
<td>0.7976</td>
<td>0.8534</td>
</tr>
<tr>
<td>Pro-Drop</td>
<td>0.7989</td>
<td>0.8523</td>
</tr>
<tr>
<td>Numeric Reference</td>
<td>0.7988</td>
<td>0.8530</td>
</tr>
<tr>
<td>Pronoun/Verb</td>
<td>0.7987</td>
<td>0.8530</td>
</tr>
<tr>
<td>Flu Noun Before Verb</td>
<td>0.7987</td>
<td>0.8526</td>
</tr>
<tr>
<td>Noun in Question</td>
<td>0.8004</td>
<td>0.8534</td>
</tr>
<tr>
<td>Subject, Object, Verb</td>
<td>0.8005</td>
<td>0.8541</td>
</tr>
</tbody>
</table>

Table 3: F1 scores after feature ablation.
Situational Awareness for Crisis Management

Earthquake in Haiti, 2010

Tohoku Tsunami, Japan 2011

Earthquake/Tsunami Detection

- Ground truth data: Earthquake Portals provide access to a broad range of earthquake data from all over the world, e.g., Centre Sismologique Euro-Mediterraneeen (CSEM-EMSC http://www.emsc-csem.org/) or USGS (http://earthquake.usgs.gov/).

**Limitation of conventional sensors:**

- Sensor networks in general can identify **earthquake** events reliably but in some cases the magnitude can be ill-determined.
- For **tsunamis**, only a limited scale of instrumental observations (e.g., DART-like offshore buoys) is available.
- No Information on what has happened on the ground.
Toretter: A Framework for Earthquake and Tsunami Detection in Japan

[Sakaki, 2010]

- Toretter is based on a probabilistic state sequence model, trying to locate the centers of earthquakes based on GPS data or location information in the user’s profile.

Basic Algorithm:
- Keyword-based filter to find tweets that are indicative of a tsunami/earthquake event
- SVM filter based on a BoW (f-measure: 73% earthquake/82% shake)
- 96% of earthquakes (> JMA 3) detected
- 80% of earthquake (> JMA 3) detected timelier than Japan’s Meteorological Agency

Toretter: Probabilistic Time Model

[Sakaki, 2010]

- Probabilistic models for event detection from time-series data

Basic Idea:
- the data fits very well to an exponential function
- design the alarm of the target event probabilistically!
Toretter: Probabilistic Location Model

Epicenter detection from a series of ‘noisy’ spatial information, i.e. tweets based on user-profile location.

Tweet Earthquake Dispatch (TED)

- The Tweet Earthquake Dispatch (TED) system is managed by U.S. Geological Survey (USGS).

Basic Algorithm:
- Times series analysis
- Threshold (tweets/min) based on short-term average (1 minute) over long-term average (60 minutes)
Natural Language Processing to the Rescue?

For relief management during an earthquake event, tweets can contribute valuable information on situational awareness.

Linguistic Features

Features predicted from linguistic classifiers:

- Subjectivity: Is it subjective or objective?
- Register: Does it show formal or informal register?
- Style: Is it written in personal or impersonal style?

NLP To the Rescue

- Four steps using supervised machine learning + NLP
  - Step 1: Identify earthquake/tsunami tweets
  - Step 2: Identify Subjective vs. Objective
    - so proud to be from Oklahoma. The outpouring of support for those devastated by the fires is amazing
    - Here is a list of the earthquake survivors. You may find update from Hotel Montana: http://bit.ly/7o7mUk
  - Step 3: Formal vs. Informal
    - Staging for fires in Elk City Area is currently at Elk City FD
    - landed in Fargo today...locals say red river will crest at 43 feet...worse than 97 flood
  - Step 4: Personal vs. Impersonal
    - Our best hopes and wishes go out to the folks in Manitoba. As the Red River is about to crest.
    - The Red River at Fargo is 40.76 feet. 22.76 feet above flood stage. 0.66 feet above 1897 record. #flood09
TweetComP1

- The TweetComP1 system builds upon a system for tsunami detection and early warning, developed in the project Collaborative, Complex, and Critical Decision-Support in Evolving Crises (TRIDEC).
- The major challenge is to support practitioners working in the field and provide them with functionalities that fit into existing workflows.

**Basic Components:**
- Focused Crawling
- Trustworthiness analysis
- Geoparsing
- Multilingual filter for off-topic tweets
- Aberration detection
- Interactive visualization of geospatial data
- Alerting service

[Zielinski et al, 2013]

Command and control unit’s graphical user interface (CCUI)

---

The Tridec Project

TRIDEC focuses on new technologies for real-time intelligent information management in collaborative, complex critical decision processes in earth management.

**Key challenge:**
Construction of a communication infrastructure of interoperable services through which intelligent management of dynamically increasing volumes and dimensionality of information and data is efficiently supported; where groups of decision makers collaborate and respond quickly in a decision-support environment.

**Start and End Date:**
01.09.2012 – 31.08.2013
Further Scenario-Specific Challenges

- Tsunamis are *rare events*

Some figures:
- Tohoku, Japan, 2011
- Indian Ocean (Sumatra, Indonesia), 2009
- Salomones, 2007
- Java, 2006
- Indian Ocean (Sumatra, Indonesia), 2004
- Messina, Italy, 1908
- Portugal, Morocco, Ireland, 1755
- Ancient Greece (Crete, Santorini), 1410 B.C.E

- Scarcity of training data

TRIDEC Architecture
Focussed Crawling

Motivation
• To identify relevant and meaningful search terms that generate content related to the event in question.

Adaptive Crawling
Learn and add relevant tags or keywords in real-time.
• Monitor existing data being collected
• Find related hashtags in a sliding window
• Weight and run parallel collectors
• Discard unneeded streams
• Cluster analysis of new data collection streams in relation to old streams

Focussed Crawling
• Preselected tags or keyword require real-time knowledge
  – Free text nature of twitter = Miss important information
Trustworthiness Analysis

Social Network Graph and Statistics
- Generation of Network Graph
- Social Network Statistics (centrality measures, influence, etc.)

Real-time Extraction of User Classes
- Neo4J Graph Database
- Mahout/Hadoop Clustering for User Class generation

Geo-Parsing

Geo-location Controlled Vocabulary for Coastal Regions
- GooglePlaces API: Place names and geometry
- OpenStreetMap: Street names and geometry
- Geonames: Region names and geometry

Real-time Location Extraction on Twitter Feeds
- N-gram Based
- Named Entity Extraction: Matched places, streets and regions

Geospatial clustering and Visualization
- Hierarchical clustering: Matched location clusters and geometry
- OpenGIS database
- Geoserver real-time spatial presentation layer
- GoogleEarth and/or MapServer base mapping layer
Geo-Parsing

Multilingual Tweet Analysis

Motivation
- To filter out off-topic tweets for a variety of languages

Bootstrap language-specific monolingual classifiers
- Training:
  - Translate labeled training data using MT (i.e., Google Translate)
  - Feature Selection & Weighting
  - Train language-specific classifiers using ML
  - Train polylingual classifier as back-off

Run-Time
- Language Identification (LangDetect Library)
- Tokenization (CMU ARK Twitter NLP tools)
- Apply language-specific classifier
  - Consider geo-location info:
    - Prioritize local-language classifier
Early Event Detection

Motivation
- detect **aberration** when number of on-topic tweets *aggregated for a certain spatial region and time-span* exceeds an expected threshold

Aggregation of tweets
- Time-span Parameter set to, e.g., an interval of one minute
- Geo-location info in tweets mapped to HASC regions (Hierarchical Administrative Subdivisions)

Timeline for Philippines Tsunami Event, 2012

TweetComp1 Visualization

![TweetComp1 Visualization](http://tridec.server.de/geoserver/)
Twitter as a Voting Barometer for Elections

Voting Intention Polls

Ground truth data: Predicting voting intention polls for the specific country provide the baselines.

Methodology:

- Interviews with 1,000 residents per country
- Try to cover a broad spectrum of potential voters (sampling on the basis of demographic variables such as age, gender, race, education, etc.)
- Often telephone interviewing
Two Simple Approaches for Predicting Elections

Election polls are reflected in..

- Relative frequency of mentions of political parties in Twitter
  - "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." [Tumasjan, Andranik, et al., ICWSM 10]

- Ratio of positive and negative tweets that mention a political party
  - "Predicting the 2011 Dutch Senate election results with twitter." [Sang, Erik Tjong Kim, and Johan Bos, Proceedings of the Workshop on Semantic Analysis in Social Media. ACL, 2012]

TrendMiner: A Large-Scale Cross-Lingual Trend Mining and Summarization of Real-time Media Streams

[TrendMiner is aiming at delivering portable real-time services for cross-lingual mining and summarization of real-time information from dynamic, large-volume media data streams.]

Basic System for extracting useful information from Twitter stream
- Map-reduce pipeline architecture
- Components for text processing
- Data sets and analysis of
  - Voting intention polls
  - Financial market prices
- Bag-of-words features and sentiment word lists

[www.trendminer-project.eu]
• **Language Detection**
  (Lui and Baldwin, Cross-Domain Feature Selection for Language Identification, IJCNLP 2011)

• **Part-of-Speech tagging**
  RT/D @MediaScotland/@ greeeat/!
  lvly/N by/P cameron/P on/P scott's/L indy/N #:)/indyref/#
  (Gimpel et al., Part-of-Speech Tagging for Twitter: Annotation, Features and Experiments, ACL, 2011)

• **Text Normalisation**
  RT @MediaScotland greeeat (great)!lvly (lovely) speech by cameron on scott's indy (independence) #:indyref
  (Han and Baldwin, Lexical Normalisation of Short Text Messages: makes a #twitter, NAACL 2011)

• **User influence**
  Using the Klout API, gives a score from 0-100 to each OSN user.

• **Correlation and Regression**
  Calculate feature correlation statistics against a time-series, and online training and predicting with regression models
• predicting voting intention polls
• strong baselines, realistic evaluation
• 2 different use cases (U.K. and Austria)

![Graph](www.trendminer-project.eu)

UK polls (240), 04/2010 – 02/2012

Ö. Polls (98), 01/2012 – 12/2012

• collection focused on all the data from users of Twitter
  40,000 UK users, random selection
  60M tweets
  1,200 Austrian users, selected by experts
  800K tweets
• preprocessing with the Trendminer Pipeline
  (Preotiuc et. al 2012)
• Vast number of users & content
  – explore word frequencies in space of users and tweets

• Social media & biased opinions
  – Focus on most ‘informative’ users

• Multi-Task Learning
  – Automatic selection of the most important terms and users

Results – UK case

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>BEN</th>
<th>BGL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conserv.</strong></td>
<td><strong>Labour</strong></td>
<td><strong>Lib Dem</strong></td>
</tr>
<tr>
<td>Avg. Value</td>
<td>2.272</td>
<td>1.663</td>
</tr>
<tr>
<td>Prev. Day</td>
<td>2</td>
<td>2.074</td>
</tr>
<tr>
<td>Linear Reg.</td>
<td>3.845</td>
<td>2.912</td>
</tr>
<tr>
<td>BEN</td>
<td>1.939</td>
<td>1.644</td>
</tr>
<tr>
<td>BGL</td>
<td>1.785</td>
<td>1.595</td>
</tr>
</tbody>
</table>

Results in RMSE (lower is better)
Citizen Journalism: Using Social Media for Breaking News Story Detection

<table>
<thead>
<tr>
<th>Party</th>
<th>Tweet</th>
<th>Score</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</td>
<td>PM in friendly chat with top EU mate, Sweden’s Fredrik Reinfeldt, before family photo</td>
<td>1.334</td>
<td>Journalist</td>
</tr>
<tr>
<td></td>
<td>Have Liberal Democrats broken electoral rules? Blog on Labour complaint to cabinet secretary</td>
<td>-0.991</td>
<td>Journalist</td>
</tr>
<tr>
<td>LAB</td>
<td>Blog Post Liverpool: City of Radicals Website now Live &lt;link&gt; #liverpool #art</td>
<td>1.954</td>
<td>Art Fanzine</td>
</tr>
<tr>
<td></td>
<td>I am so pleased to head Paul Savage who worked for the Labour group has been appointed the Marketing manager for the baths hall GREAT NEWS</td>
<td>-0.552</td>
<td>Political (Labour)</td>
</tr>
<tr>
<td>LBD</td>
<td>RT @user: Must be awful for TV bosses to keep getting knocked back by all the women they ask to host election night (via @user)</td>
<td>0.874</td>
<td>LibDem MP</td>
</tr>
<tr>
<td></td>
<td>Blog Post Liverpool: City of Radicals 2011 – More Details Announced #liverpool #art</td>
<td>-0.521</td>
<td>Art Fanzine</td>
</tr>
</tbody>
</table>
Opportunities as opposed to Traditional Media

- Wider Audience
  - 20% come across a news story through social media (Facebook, Twitter)
- Timely Information
  - The detection of the breaking news story is a major sales argument.

Note also: Citizens are not professional journalists!

Twitter Ahead of News

Twitter more timely than traditional media:
- Crash of Turkish Airlines flight 1951 in Amsterdam, February, 2009 (CNN, 2009)
- Result of the German presidential election, 2009

Coverage of topics in Twitter higher than in traditional media:
- Post-election riots in Tehran, Iran
- Terrorist attacks in Mumbai, 2009
Key journalistic needs when dealing with Social Media

Basic tools needed by journalists (cf., Diplaris, 2012)
- Real-time alerts
- Trustworthiness (of informants)
- Trend and Sentiment Detection
- Responsiveness
- Access to Contributors
- Verification

Vox Civitas: A Social Media Analytic Tool for Journalists

- Vox Civitas is an aggregation and analysis platform that collects data from Twitter around an event, tracks it over time, and filters the results into a robust, easily digestible, visual format.
- It is designed to help journalists and media professionals extract news value from large-scale aggregations of social media content around broadcast events.

Modules:
- Salient Keyword Extraction
- Sentiment Analysis
- Detection of unique tweets (Novelty)
- Detection of relevant tweets (similar to audio transcript)
- Visualization

http://sm.rutgers.edu/vox/
Conclusions

• We presented an introduction survey of recent work on exploiting micro-blogging data such as Twitter’s for complex AI and NLP tasks

• Take-home messages
  – Micro-blogging: a goldmine of data
  – Enables novel applications (e.g., social sensing)
  – Real-world, high-impact applications

• Open issues
  – Better exploiting of the social graph
  – More more applications!

References


Gupta, Manish, Peixiang Zhao, and Jiawei Han. "Evaluating event credibility on twitter." Proceedings of SIAM (2012).
References

Han, Bo, and Timothy Baldwin. "Lexical normalisation of short text messages: Makn sens a# twitter." Proceedings of ACL-11.

Han, Bo, Paul Cook and Timothy Baldwin. "Geolocation Prediction in Social Media Data by Finding Location Indicative Words", Proceedings of COLING-2012, 1045–1062


References


References


D. Maynard and K. Bontcheva and D. Rout. Challenges in developing opinion mining tools for social media. In Proceedings of *@NLP can u tag #user generated content?! Workshop at LREC 2012*.


Ralph Passarella, Atul Nakhasi, Sarah Bell, Michael J. Paul, Peter Pronovost, Mark Dredze. Twitter as a Source for Learning about Patient Safety Events. Annual Symposium of the American Medical Informatics Association (AMIA), 2012.


References


