Semantic Analysis of Social Media

Timothy Baldwin

Talk Outline

What is Social Media?

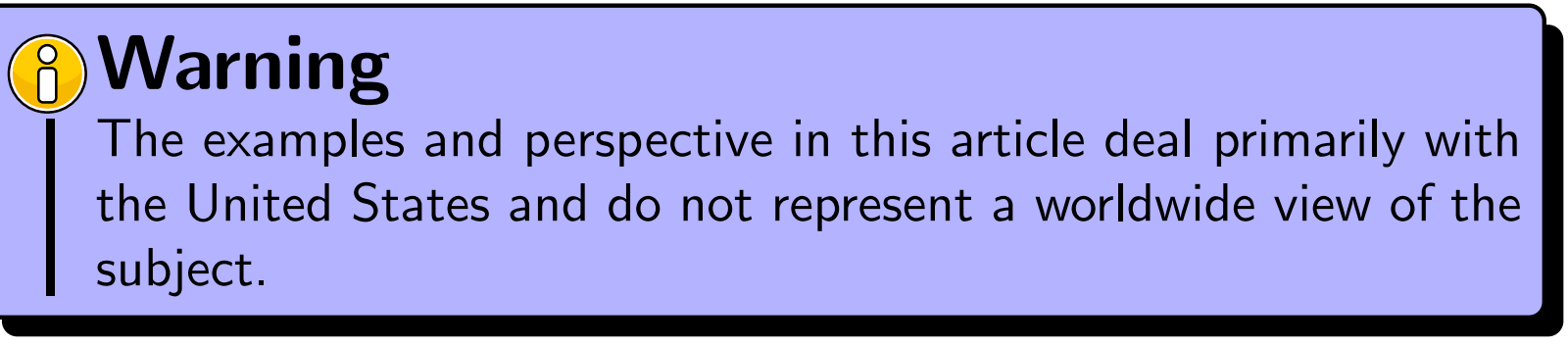
- According to Wikipedia (18/8/2014):

  Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."

What is Social Media?

- According to Wikipedia (18/8/2014):

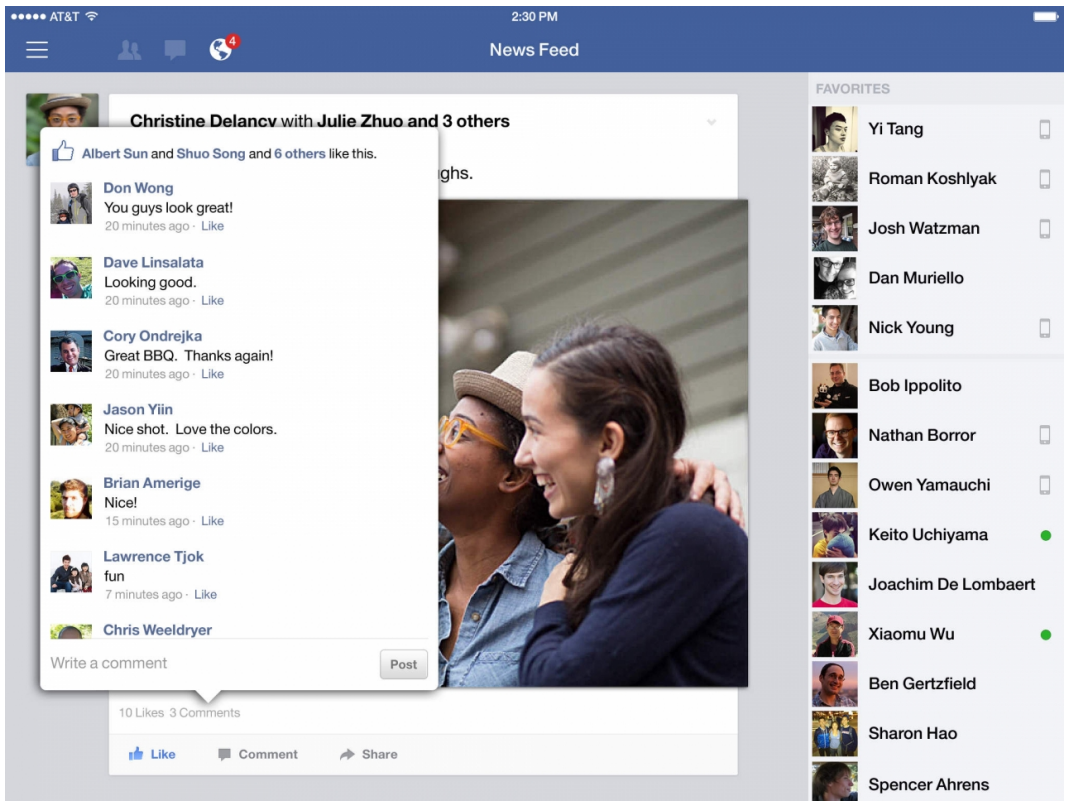
  Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."



Popular Forms of Social Media

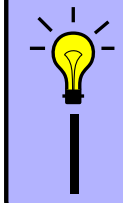
Social Networking sites

Facebook, Google+, ...



Source(s): `http://www.iclarified.com/images/news/33952/140980/140980-1280.jpg`

Popular Forms of Social Media

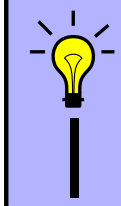


Content sharing sites

Instagram, Foursquare, Flickr, YouTube, ...

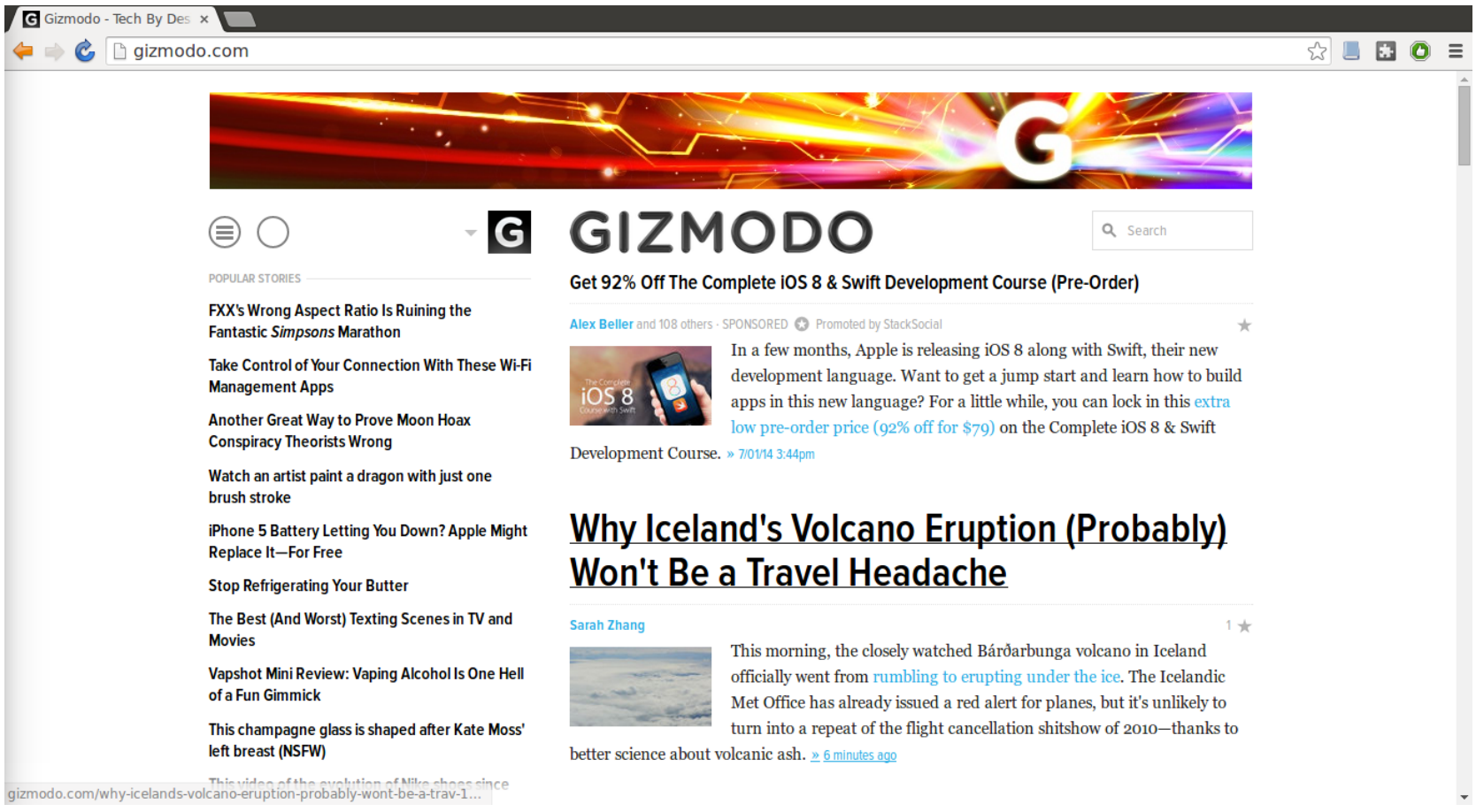
Source(s): `http://sanziro.com/2011/05/app-of-the-week-instagram.html`

Popular Forms of Social Media

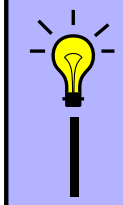


Blogs

Gizmodo, Mashable, Boing Boing, ...



Popular Forms of Social Media



Micro-blogs

Twitter, Weibo, Tumblr, ...

Source(s): `http://itunes.apple.com/us/app/twitter/`

Popular Forms of Social Media

Web user forums

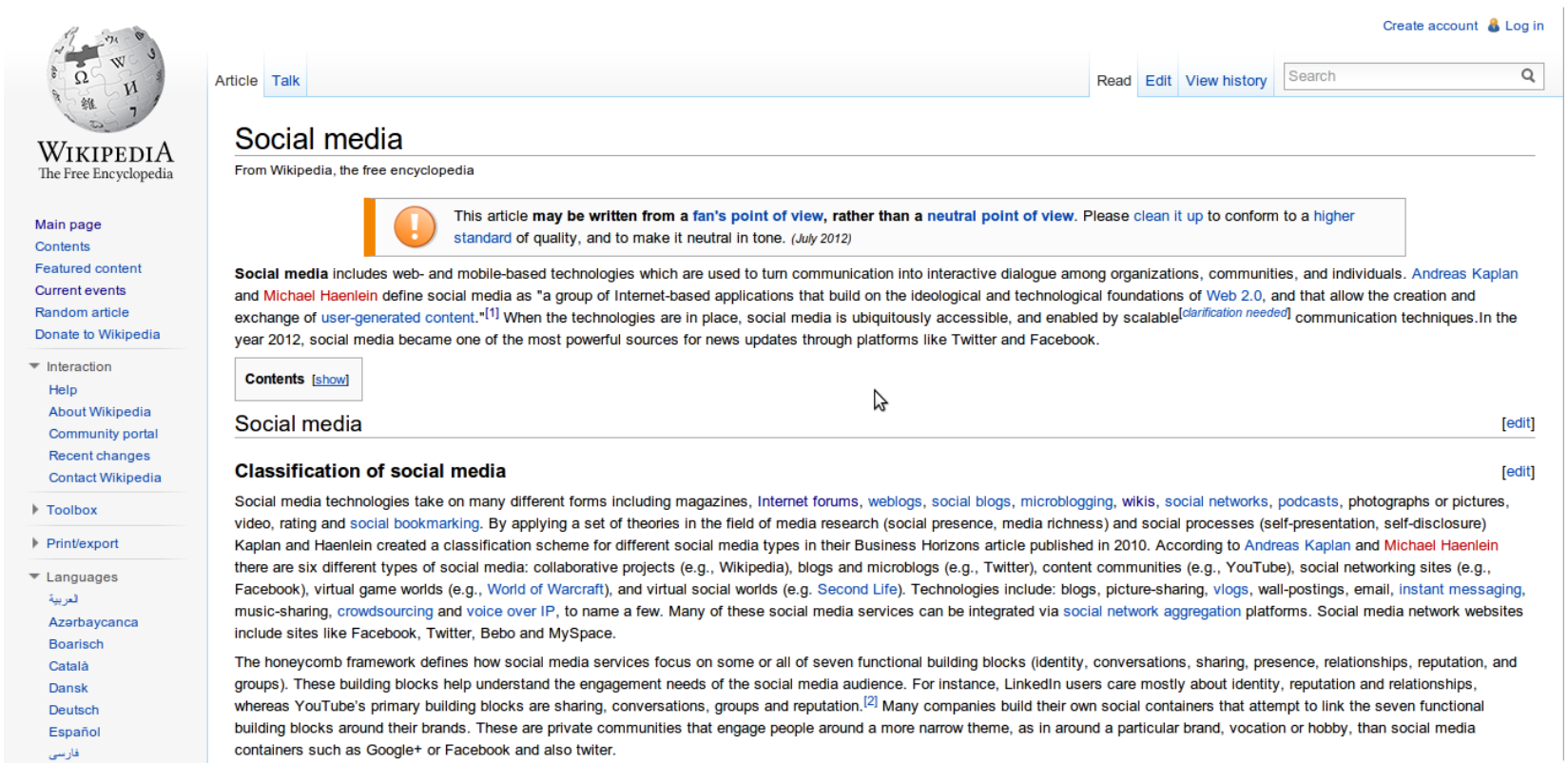
StackOverflow, CNET forums, Apple Support, ...

Source(s): `http://tinyurl.com/pwk8p9j`

Popular Forms of Social Media

Wikis

Wikipedia, Wiktionary, ...



Create account Log in

Article Talk | Read Edit View history | Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
العربية
Azərbaycanca
Boarisch
Català
Dansk
Deutsch
Español
فارسی

Social media

From Wikipedia, the free encyclopedia

This article **may be written from a fan's point of view, rather than a neutral point of view**. Please clean it up to conform to a higher standard of quality, and to make it neutral in tone. *(July 2012)*

Social media includes web- and mobile-based technologies which are used to turn communication into interactive dialogue among organizations, communities, and individuals. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."[1] When the technologies are in place, social media is ubiquitously accessible, and enabled by scalable[*clarification needed*] communication techniques.In the year 2012, social media became one of the most powerful sources for news updates through platforms like Twitter and Facebook.

Contents [show]

Social media [edit]

Classification of social media [edit]

Social media technologies take on many different forms including magazines, Internet forums, weblogs, social blogs, microblogging, wikis, social networks, podcasts, photographs or pictures, video, rating and social bookmarking. By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure) Kaplan and Haenlein created a classification scheme for different social media types in their Business Horizons article published in 2010. According to Andreas Kaplan and Michael Haenlein there are six different types of social media: collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual social worlds (e.g. Second Life). Technologies include: blogs, picture-sharing, vlogs, wall-postings, email, instant messaging, music-sharing, crowdsourcing and voice over IP, to name a few. Many of these social media services can be integrated via social network aggregation platforms. Social media network websites include sites like Facebook, Twitter, Bebo and MySpace.

The honeycomb framework defines how social media services focus on some or all of seven functional building blocks (identity, conversations, sharing, presence, relationships, reputation, and groups). These building blocks help understand the engagement needs of the social media audience. For instance, LinkedIn users care mostly about identity, reputation and relationships, whereas YouTube's primary building blocks are sharing, conversations, groups and reputation.[2] Many companies build their own social containers that attempt to link the seven functional building blocks around their brands. These are private communities that engage people around a more narrow theme, as in around a particular brand, vocation or hobby, than social media containers such as Google+ or Facebook and also twiter.

Source(s): `http://en.wikipedia.org/wiki/Social_media`

Common Features of Social Media

- Posts

Common Features of Social Media

- Posts
- Social network (explicit or implicit)

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/voting/rating/...

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/voting/rating/...
- Author information, and linking to user profile features

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/voting/rating/...
- Author information, and linking to user profile features
- Streaming data

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/voting/rating/...
- Author information, and linking to user profile features
- Streaming data
- Aggregation/ease of access

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/voting/rating/...
- Author information, and linking to user profile features
- Streaming data
- Aggregation/ease of access
- Volume

@eltimster so what? #yawn

- OK, OK, but what's all this got to do with semantics?

@eltimster so what? #yawn

- OK, OK, but what's all this got to do with semantics?
- Basic question that I am asking in this talk:

  Lexical Semantic Analysis of Social Media

@eltimster so what? #yawn

- OK, OK, but what's all this got to do with semantics?
- Basic question that I am asking in this talk:

  (Lexical) Semantic Analysis of Social Media — Why Care?

@eltimster so what? #yawn

- OK, OK, but what's all this got to do with semantics?
- Basic question that I am asking in this talk:

  (Lexical) Semantic Analysis of Social Media — Why Care?

- Answer the question across three dimensions of social media analysis:
  1. content-based semantic analysis
  2. user-based semantic analysis
  3. network-based semantic analysis

Talk Outline

Content-based Semantic Analysis

- Content-based analysis = *base analysis on the content of social media posts*

Content-based Semantic Analysis

- Content-based analysis = *base analysis on the content of social media posts ... focusing primarily on the textual content, but don't forget the links*

Content-based Semantic Analysis

- Content-based analysis = *base analysis on the content of social media posts ... focusing primarily on the textual content, but don't forget the links*
- Superficial, hard-nosed answer as to why we should care about content-based semantic analysis:

Content-based Semantic Analysis

- Content-based analysis = *base analysis on the content of social media posts ... focusing primarily on the textual content, but don't forget the links*
- Superficial, hard-nosed answer as to why we should care about content-based semantic analysis:

  BECAUSE OTHERS CARE!

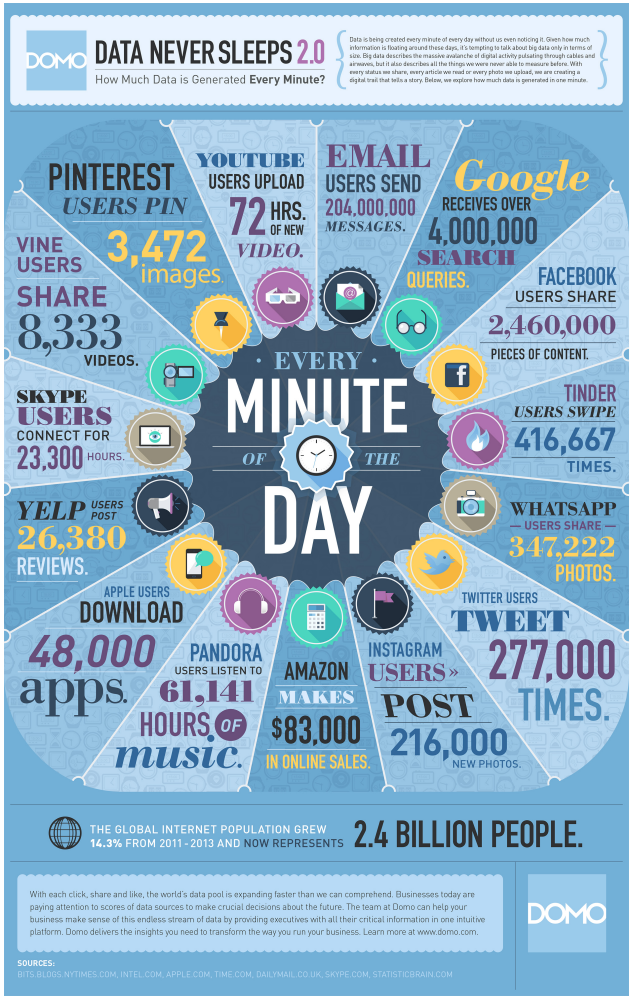
Content-based Semantic Analysis

- Content-based analysis = *base analysis on the content of social media posts ... focusing primarily on the textual content, but don't forget the links*
- Superficial, hard-nosed answer as to why we should care about content-based semantic analysis:

  BECAUSE OTHERS CARE!

  If we can put high-utility semantic data in the hands of social media analysts, people will use it (much to learn from the "outreach" successes of Sentiment Analysis et al.)

Semantic Analysis at Scale I



Source(s): `http://www.domo.com/learn/data-never-sleeps-2`

Semantic Analysis at Scale II

- **The good news**: social media content is generally plentiful, if you aren't picky about the data
  - $\Rightarrow$ great news for unsupervised models; potential challenges for scalability
- **The mixed news**:
  - the data requires quite a bit of "taming", in terms of:
    - the mix of language and topic, with heavy skewing toward particular languages and topics (*@justinbieber I'm tired write to you! But NEVER SAY NEVER! PT 18*)
    - orthography, although lexical normalisation helps out quite a bit [Baldwin et al., 2013]
  - documents are generally short (= limited textual context)

Pre-tagged Training Data Galore #kinda

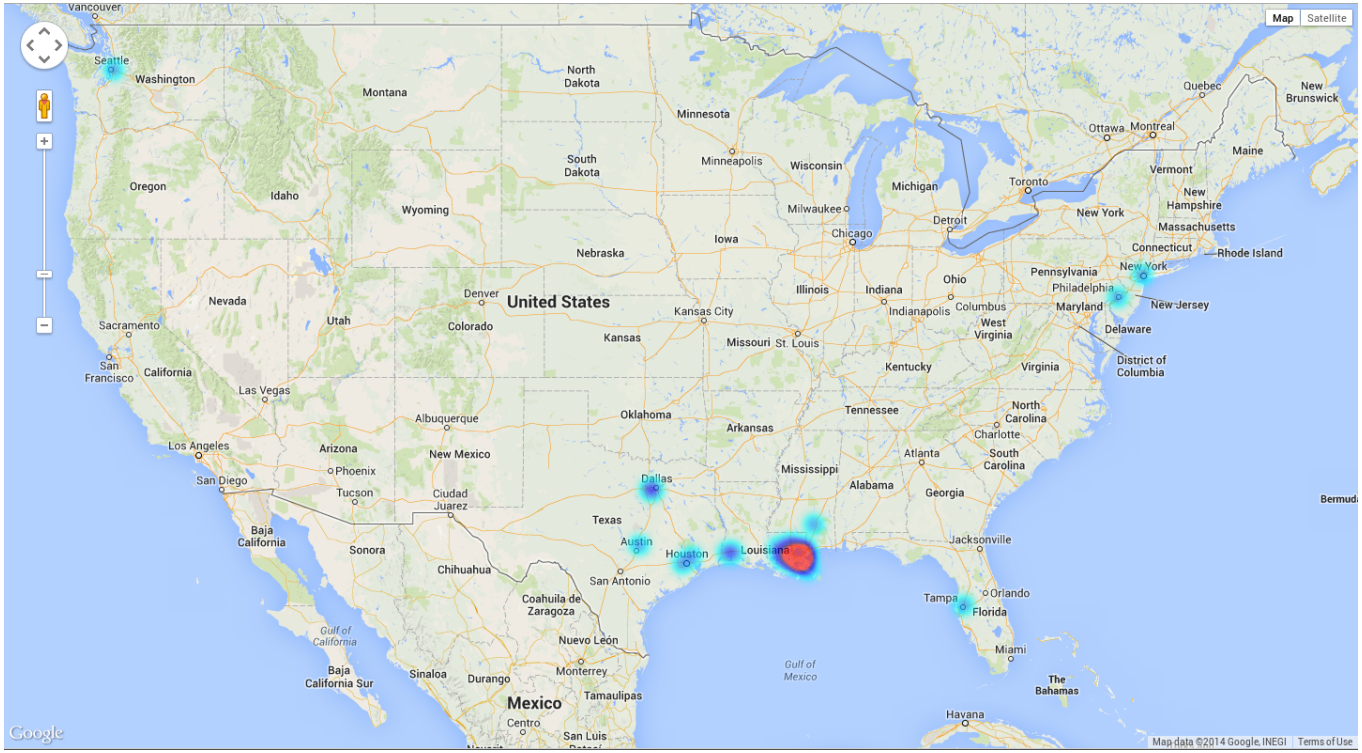
- Social media data is rife with user-provided (silver-standard) metalinguistic labels [Davidov et al., 2010a,b]:
  - hashtagging of sarcasm/irony and sentiment:
    (1) So glad to hear the police have everything under control in #Frerguson #sarcasm
    (2) Its been 3 days you guys sent us a broken laptop. No communication from your team . Feeling cheated. #FAIL
  - comments on images/videos (e.g. *Very atmospheric*)
  - free-text metadata associated with images/videos (e.g. *Dublin's cityscape seen over the river Liffey ...*)
  - social tagging of documents/images (e.g. *Ireland*)
  - geotagging of documents/images [Eisenstein et al., 2010, Wing and Baldridge, 2011, Han et al., 2014, Doyle, 2014]

Example Task: Detection/Analysis of Localisms I

- **Task outline**: analyse the geographical spread of different terms based on geotagged data, and identify terms which have "low-entropy" localised geographic distributions
- **Approach (v1)**: for a pre-identified expression, analyse the geographical spread of use [Doyle, 2014]
- **Approach (v2)**: use feature selection methods to identify terms with a highly-localised geospread, based on 2D spatial analysis or discretisation of the data (e.g. into states or cities) [Cook et al., to appeara]

Example Task: Detection/Analysis of Localisms II

- **Example**: the term *buku*, identified using information gain ratio ratio over a set of North American tweets:



Source(s): Cook et al. [to appear]

Words of Caution on Pre-tagged Training Data

- Hashtags can be ambiguous/shift in meaning over time (e.g. *#acl2014*)
- Popular hashtags have a tendency to be spammed, and become less discriminating
- Not all possible metalinguistic labels are equally used, for good pragmatic reasons (cf. *#english*, *#bikemadmiddleagedaustralian*)
- Comments and metadata descriptions vary a lot in content, quality and relevance (not all comments are equal)
- Comments/social tags are notoriously patchy (not all posts are equally commented on/tagged)

Robustness and Semantic Parsing

- (Genuine) robustness has long been beyond the reach of NLP, but there is no data source better than social media text to test the robustness of an NLP tool:
  - the content is all over the place, documents are generally short, spelling and syntax are often "untamed", ...
- I would suggest that certain NLP tasks such as constituency parsing over social media text are a lost cause (Baldwin et al. [2013], although see Kong et al. [to appear] on dependency parsing Twitter), but that it's a natural target for semantic parsing:

(3) It's getting late the babe sleep guess I'll ko now kawaii ne! #fb

```
get_state'(late'())
sleep'(arg1=baby',trel=now,tense=pres)
sleep'(arg1=1p_sing,trel=now,tense=future)
```

Diachronic Analysis

- One of the benefits of streaming data is that it is timestamped, supporting diachronic analysis of the content, and opening up research on topics such as:
  - the detection of novel word senses [Cook et al., to appearb]
  - sense drift [Cook et al., 2013]
  - what senses "stick" (e.g. *swag* vs. *clamp*)
  - the rise (and fall) and use patterns of multiword expressions (MWEs) (e.g. *chick flick* vs. *myspace terms*)
  - (in combination with geotags) the geographical dispersal (over time) of words/senses/MWEs (e.g. *selfie*)

Trend Analysis

- Related to this, it is also possible to explore novel (lexical) semantic tasks with a dimension of time such as:
  - event/first story detection [Petrović et al., 2010]
  - trend/change-point analysis [Lau et al., 2012]
- Much of the work in this space has assumed a predefined event type, or done some variant of lexical "burstiness" analysis
- The semantics community can potentially offer much in terms of:
  - what is an event?
  - how should an event be represented/presented to a user?
  - how to represent/process uncertain/incomplete event information?
  - sense-sensitisation of burstiness/trend analysis

Content-based Semantic Analysis: Summary

- Content-based semantic analysis of social media – why care?
  - If we can generate high-utility semantic information, users will come
  - Possibilities/challenges for semantic analysis at scale ... but need to tame the data
  - Availability of silver-standard user/device-tagged data, e.g. hashtags, comments, free-text metadata, social tags, geotags
  - It's a great target for semantic parsing (and arguably terrible target for conventional syntactic parsing)
  - There are possibilities to carry out diachronic analysis of words/MWEs
  - There are opportunities to carry out trend analysis

Talk Outline

User Information I

- All I have said to now has ignored the fact that:
  (a) a myriad of people are posting the content

User Information I

- All I have said to now has ignored the fact that:
  (a) a myriad of people are posting the content
  (b) we generally know at least who the poster was, and in many cases also:
    - their name and "identity"
    - user-declared demographic/profiling information
    - what other posts they have made to the same site

User Information II

- Simply knowing the identity of the user opens up possibilities for user priors, e.g.:
  - analysis of per/cross-user sense usage patterns
  - user-biased semantic parsing, trend analysis, etc.
- In additionally knowing something about the messages associated with users (e.g. geotags) or the user themself (e.g. their technical proficiency), we can perform:
  - user profiling (e.g. user geolocation, language identification, user ethnicity, ...) [Bergsma et al., 2013]
  - message/question routing
  - user- and location-biased semantic parsing, trend analysis, etc.

Example Task 1: User-level Lexical Priors

Conventional text

- One sense per discourse [Gale et al., 1992]
- First-sense heuristic [McCarthy et al., 2004]

Example Task 1: User-level Lexical Priors

Conventional text

- One sense per discourse [Gale et al., 1992]
- First-sense heuristic [McCarthy et al., 2004]

Twitter

- One sense per tweeter?
  - documents are too small to consider applying one sense per discourse, but we can possibly address the lack of context with user-level sense priors

Example Task 1: User-level Lexical Priors

Conventional text

- One sense per discourse [Gale et al., 1992]
- First-sense heuristic [McCarthy et al., 2004]

Twitter

- One sense per tweeter?
  - documents are too small to consider applying one sense per discourse, but we can possibly address the lack of context with user-level sense priors
- First-sense heuristic?
  - shown to change substantially across domains, so not clear that it will work as well over Twitter

User-level Lexical Priors: Datasets

- Sense inventory: Macmillan Dictionary
- Target lemmas: 20 nouns ($\geq$ 3 senses)
- 4 datasets: $\{\text{TWITTER}, \text{UKWAC}\} \times \{\text{RAND}, \text{USER}\}$
  - UKWAC: more-conventional (web) text
  - RAND: random sample of usages from TWITTER/UKWAC
  - USER: 5 usages of a given word from each user (TWITTER) or document (UKWAC)
- 2000 items each: 100 usages of each noun

Source(s): Gella et al. [2014]

User-level Lexical Priors: Analysis

- Average proportion of users/documents using a noun in the same sense across all 5 usages
  - $\text{TWITTER}_{\text{USER}}$: 65%
  - $\text{UKWAC}_{\text{DOC}}$: 63%
- One sense per tweeter heuristic is as strong as one sense per discourse

Source(s): Gella et al. [2014]

Analysis: Pairwise Agreement

	Partition	Agreement (%)
Gale et al. (1992)	document	94.4
$\text{TWITTER}_{\text{USER}}$	user	95.4
$\text{TWITTER}_{\text{USER}}$	—	62.9
$\text{TWITTER}_{\text{RAND}}$	—	55.1
$\text{UKWAC}_{\text{DOC}}$	document	94.2
$\text{UKWAC}_{\text{DOC}}$	—	65.9
$\text{UKWAC}_{\text{RAND}}$	—	60.2

Source(s): Gella et al. [2014]

Analysis: Pairwise Agreement

	Partition	Agreement (%)
Gale et al. (1992)	document	94.4
$\text{TWITTER}_{\text{USER}}$	user	95.4
$\text{TWITTER}_{\text{USER}}$	—	62.9
$\text{TWITTER}_{\text{RAND}}$	—	55.1
$\text{UKWAC}_{\text{DOC}}$	document	94.2
$\text{UKWAC}_{\text{DOC}}$	—	65.9
$\text{UKWAC}_{\text{RAND}}$	—	60.2

Source(s): Gella et al. [2014]

Analysis: Pairwise Agreement

	Partition	Agreement (%)
Gale et al. (1992)	document	94.4
TWITTER$_{\text{USER}}$	user	95.4
TWITTER$_{\text{USER}}$	—	62.9
TWITTER$_{\text{RAND}}$	—	55.1
UKWAC$_{\text{DOC}}$	document	94.2
UKWAC$_{\text{DOC}}$	—	65.9
UKWAC$_{\text{RAND}}$	—	60.2

Source(s): Gella et al. [2014]

Analysis: Pairwise Agreement

	Partition	Agreement (%)
Gale et al. (1992)	document	94.4
$\text{TWITTER}_{\text{USER}}$	user	95.4
$\text{TWITTER}_{\text{USER}}$	—	62.9
$\text{TWITTER}_{\text{RAND}}$	—	55.1
$\text{UKWAC}_{\text{DOC}}$	document	94.2
$\text{UKWAC}_{\text{DOC}}$	—	65.9
$\text{UKWAC}_{\text{RAND}}$	—	60.2

Source(s): Gella et al. [2014]

User-level Lexical Priors: Other Analysis

- Comparing $\text{TWITTER}_{\text{RAND}}$ and $\text{UKWAC}_{\text{RAND}}$:
  - First-sense tagging is less accurate in Twitter data
    - $\text{TWITTER}_{\text{RAND}}$: 45.3%
    - $\text{UKWAC}_{\text{RAND}}$: 55.4%
  - Sense distributions are less skewed on Twitter
    - sense entropy lower for $\text{UKWAC}_{\text{RAND}}$ for 15 nouns
  - 8/20 nouns have different first senses
  - More "Other" senses in Twitter data
    - $\text{TWITTER}_{\text{RAND}}$: 12.3%
    - $\text{UKWAC}_{\text{RAND}}$: 6.6%

Source(s): Gella et al. [2014]

Example Task 2: User Geolocation

- What is the most likely geolocation for a message/user?

? Example

- Posts:
  - *Currently seated in the drunk people section. #sober*
  - *RT SFGiants: Sergio Romo's scoreless steak is snapped at 21.2 innings as he allows 1 run in the 8th. #SFGiants still hold 2-1 lead.*
  - *kettle corn guy featured on sportscenter!! #Sfgiants*
- User location: ?

Example Task 2: User Geolocation

- What is the most likely geolocation for a message/user?

? Example

- Posts:
  - *Currently seated in the drunk people section. #sober*
  - *RT SFGiants: Sergio Romo's scoreless steak is snapped at 21.2 innings as he allows 1 run in the 8th. #SFGiants still hold 2-1 lead.*
  - *kettle corn guy featured on sportscenter!! #Sfgiants*
- User location: **Fresno, CA**

User Geolocation: Approach

- Construct training/test data by identifying users with a certain volume of geotagged tweets, centred around a particular locale
- Approach the task via text classification over the meta-document that is the combination of (geotagged) tweets from that user: demo [Han et al., 2013]
- Challenges:
  - label set semantics: ideally continuous 2D representation
  - classifier output: ideally PDF over 2D space rather than discrete [Priedhorsky et al., 2014]
  - label set size (even assuming discrete representation, 3000+ cities in Han et al. [2014])
  - training set size (millions+ of training instances)
  - "currency" of the model (ideally want to update the model dynamically)

User Geolocation: Findings to Date

- The choice of class representation and approach to feature selection has a larger impact on results than the choice of model
- Including non-geotagged tweets boosts results (training and test)
- Pre-partitioning users by language improves results appreciably
- User metadata is a better predictor of location than the body of the posts from a user (esp. user-declared location, but self-description, timezone and real name also boost accuracy)
- Models stagnate over time
- Networks are much more effective than content ...

Source(s): Roller et al. [2012], Jurgens [2013], Han et al. [2014]

Example Task 3: Joint Discourse and Semantic Analysis I

- And just to prove that there's more to social media than Twitter: thread classification of web user forum threads (e.g. *has the information need of the initiator been resolved?*), based on the content of posts in the thread

Example Task 3: Joint Discourse and Semantic Analysis II

Debian VS. Red Hat

UserA Post1	Ive been using Red Hat for along time now ... But I hear a lot of fuss about Debian ... I like apt-get a lot ... which of those CDs do I need? ...
UserB Post2	if you like apt-get, you only need disk 1, everything else you need, you can just apt-get it.
UserA Post3	... Is that going to be an obvious option in the installer or do I have to just select the minimal stuff and then do a dist upgrade?
UserB Post4	there is a spot where you choose ftp or http sites for downloading files ... At the end of the installer, there is ... After this you are left with ...
UserC Post5	I mostly use a minimal boot CD (based on bf2.4) to install Debian ... Use it to install the base system, then apt-get or dselect to get whatever you need ...

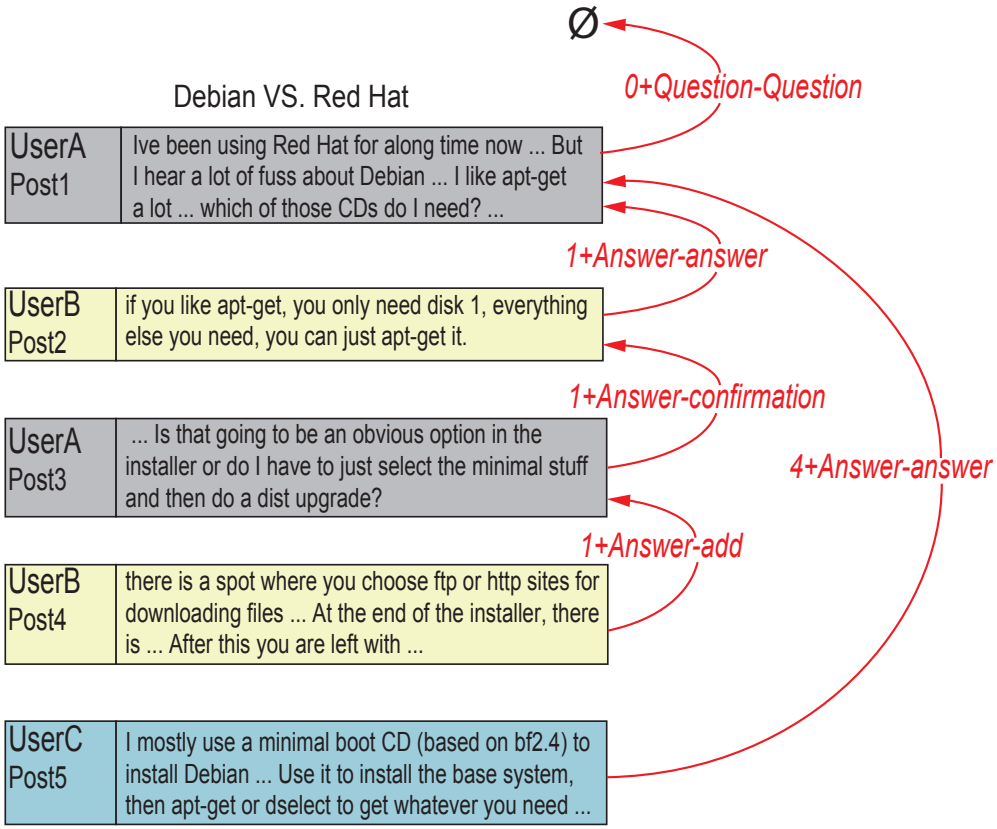
Example Task 3: Joint Discourse and Semantic Analysis III

Debian VS. Red Hat

UserA Post1	Ive been using Red Hat for along time now ... But I hear a lot of fuss about Debian ... I like apt-get a lot ... which of those CDs do I need? ...
UserB Post2	if you like apt-get, you only need disk 1, everything else you need, you can just apt-get it.
UserA Post3	... Is that going to be an obvious option in the installer or do I have to just select the minimal stuff and then do a dist upgrade?
UserB Post4	there is a spot where you choose ftp or http sites for downloading files ... At the end of the installer, there is ... After this you are left with ...
UserC Post5	I mostly use a minimal boot CD (based on bf2.4) to install Debian ... Use it to install the base system, then apt-get or dselect to get whatever you need ...

Solved?

Example Task 3: Joint Discourse and Semantic Analysis IV



User-based Semantic Analysis: Summary

- User-based semantic analysis of social media – why care?
  - much to be gained from inclusion of user priors in semantic analysis ("personalised semantic analysis", e.g. one sense per tweeter)
  - user-level aggregation as enabler for user-level analysis (e.g. user geolocation)
  - user identify powerful in understanding the information/discourse structure of threads on user forums, contributing to thread-level semantic analysis
  - vast untapped space of possibilities waiting to be explored by semanticists ...

Talk Outline

Network-based Semantic Analysis

- The final piece in today's puzzle is (user) network data, in the form of:
  - followers/followees
  - user interactions
  - reposting of content
  - shared hashtags
  - likes/favourites/starring/voting/rating/...

    ⋮
- Underlying assumption of "homophily" = *similars attract* (or less commonly "heterophily" = *similars repel*), as basis of propagating labels across network of users

Network-only Models

- It is possible to perform classification based on the network alone, e.g.:
  - **label propagation**: starting with a small number of labelled nodes in a graph, iteratively label other nodes based on the majority label of their neighbours [Zhu and Ghahramani, 2002, Jurgens, 2013]
- For tasks such as user geolocation, network-based models have been shown to be far more effective than content-based models

Combining Content and Network Analysis I

- Lots of possibilities for combining content- and network-analysis:
  - **nearest neighbour**: starting with a small number of geolocated users, iteratively geolocate other users based on the geolocations of their closest neighbour(s), based on content similarity (e.g. user-declared location or post similarity) [Jurgens, 2013]
  - generate the network based on content similarity, and perform network-based analysis [Burfoot et al., 2011]
  - generate network-based features (e.g. co-participation or reply-to features), and incorporate into content-based classification [Fortuna et al., 2007]
- Also possibility of performing user classification using joint network and content analysis, e.g. Thomas et al. [2006], Burfoot et al. [2011]

Combining Content and Network Analysis II

- Some well-known approaches to combining content and network analysis are:
  - **iterative classification** [Bilgic et al., 2007]:
    1. apply base classifiers to a text-based representation of each instance (e.g. the posts of a given user)
    2. expand the feature representation of each user through the incorporation of relational features
    3. retrain and reapply the classifier; repeat step 2 until class assignments stabilise
  - **dual classification**:
    1. generate base classifications for each instance based on: (a) content-based classifiers; and (b) network-based classifiers
    2. normalise the combined predictions, and decode the content- and network-based classifications using collective classification [Sen et al., 2008]

Where it really Starts Getting Interesting ...

- **Scaling it up**: much algorithmic work to be done in scaling up (higher-end) network analysis and joint content + network methods to social media-based social networks
- **Heterogeneous networks**: while there is a large body of literature based on "first-order" social networks, much less on combining multiple heterogeneous networks of different semantics (e.g. social network vs. content similarity ($\times n$) vs. repost vs. hashtag sharing vs. favouriting vs. ...)
- **Dynamic network modelling**: also much less work on dynamic network and content analysis, and interpreting the semantics of network and content change

Complications in Using Networks

- Difficulty in getting access to "first-order" social network data from sites such as Twitter and Facebook
- Extreme difficulty in getting access to diachronic network data
- Sparsity of networks based on co-participation, reply-to, etc.
- Noisiness of networks based on content similarity

Network-based Semantic Analysis: Summary

- Network-based semantic analysis of social media – why care?
  - simple network-based methods far superior to content-based methods in some instances
  - combined network- and content-analysis has been shown to be superior to just network or just content analysis in a number of contexts
  - increasing interest in combining network- and content-based analysis from the network analysis community; who better than this community to lead that effort?

Talk Outline

Tame your own Social Media Data

- Assuming you are interested in only certain languages, you will first need to carry out **language identification** [Lui and Baldwin, 2014]
- If you are after high recall and not interested in the "unknown", you can either ignore content with high OOV rates or look to **lexical normalisation** [Han and Baldwin, 2011, Eisenstein, 2013]
- If you are interested in regional analysis, you either need to make do with the subset of geotagged messages, or carry out your own **geolocation**
- For many semantic applications, you need to consider what is a "representative" sample of social media data, and possibly consider **user profiling** as a means of selecting/excluding certain users [Bergsma et al., 2013]

Key Resources

- **Language identification**: langid.py, CLD2, langdetect, TwitIE, polyglot
- **(English) tokenisation**: Twokenizer, Chris Potts' tokeniser
- **(English) lexical normalisation**: UniMelb lexical normalisation dictionary, TextCleanser, TwitIE
- **POS tagging**: ARK Twitter POS tagger, Twitter NLP, TwitIE
- **NER**: Twitter NLP, TwitIE
- **Message geolocation/geoparsing**: CMU GeoLocator
- **User geolocation**: UniMelb Twitter user geolocator
- **User profiling**: Bot or not, Twitter Clusters

Some Datasets to Get Going with

- **Sense-tagged social media datasets**:
  - lexical sample: Twitter [Gella et al., 2014]
  - supersense data: Twitter [Johannsen et al., to appear]
- **User geolocation**: CMU Geo-tagged Microblog Corpus [Eisenstein et al., 2010]
- **Web user forum thread and post analysis**: CNET thread dataset [Kim et al., 2010, Wang et al., 2012]

Talk Outline

Summary

- Social media opens up a myriad of new opportunities for semantic research, in terms of content analysis, potentially incorporating user and network information
- Research in the space is booming, much of it outside NLP ... the *SEM community has much to offer in leading/guiding the research agenda:

- These slides are available from: `http://people.eng.unimelb.edu.au/tbaldwin/pubs/starsem2014.pdf`

Summary

- Social media opens up a myriad of new opportunities for semantic research, in terms of content analysis, potentially incorporating user and network information
- Research in the space is booming, much of it outside NLP ... the *SEM community has much to offer in leading/guiding the research agenda:

  @semanticists get on board with social media analytics #nlproc

- These slides are available from: `http://people.eng.unimelb.edu.au/tbaldwin/pubs/starsem2014.pdf`

Acknowledgements

- My contributions in this space have been in collaboration with Paul Cook, Jey Han Lau, Bo Han, Marco Lui, Li Wang, Spandana Gella, Clint Burford, Su Nam Kim, Diana McCarthy, Nigel Collier, David Martinez and Dave Newman
- Thanks to the University of Melbourne NLP Group for valuable feedback on an earlier version of this talk

References I

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan, 2013.

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 1010–1019, Atlanta, USA, 2013. URL `http://www.aclweb.org/anthology/N13-1121`.

Mustafa Bilgic, Galileo Namata, and Lise Getoor. Combining collective classification and link prediction. In *ICDM Workshops*, pages 381–386, 2007.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 1506–1515, Portland, USA, 2011.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex 2013*, pages 49–65, Tallinn, Estonia, 2013.

Paul Cook, Bo Han, and Timothy Baldwin. Statistical methods for identifying local dialectal terms from corpora of GPS-tagged documents. *Dictionaries*, to appeara.

References II

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. Novel word-sense identification. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland, to appearb.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the 14th Conference on Natural Language Learning (CoNLL-2010)*, pages 107–116, Uppsala, Sweden, 2010a. URL http://www.aclweb.org/anthology/W10-2914.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Posters Volume*, pages 241–249, Beijing, China, 2010b. URL http://www.aclweb.org/anthology/C10-2028.

Gabriel Doyle. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the EACL (EACL 2014)*, pages 98–106, Gothenburg, Sweden, 2014. URL http://www.aclweb.org/anthology/E14-1011.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 359–369, Atlanta, USA, 2013. URL http://www.aclweb.org/anthology/N13-1037.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1277–1287, Cambridge, USA, 2010. URL http://www.aclweb.org/anthology/D10-1124.

References III

Blaz Fortuna, Eduarda Mendes Rodrigues, and Natasa Milic-Frayling. Improving the classification of newsgroup messages through social network analysis. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007)*, pages 877–880, Lisbon, Portugal, 2007.

William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237, 1992.

Spandana Gella, Paul Cook, and Timothy Baldwin. One sense per tweeter ... and other lexical semantic tales of Twitter. In *Proceedings of the 14th Conference of the EACL (EACL 2014)*, pages 215–220, Gothenburg, Sweden, 2014.

Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 368–378, Portland, USA, 2011.

Bo Han, Paul Cook, and Timothy Baldwin. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013): System Demonstrations*, pages 7–12, Sofia, Bulgaria, 2013.

Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

References IV

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Sgaard. More or less supervised super-sense tagging of Twitter. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, Dublin, Ireland, to appear.

David Jurgens. Thats what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, pages 273–282, Dublin, Ireland, 2013.

Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and linking web forum posts. In *Proceedings of the 14th Conference on Natural Language Learning (CoNLL-2010)*, pages 192–202, Uppsala, Sweden, 2010.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, to appear.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1519–1534, Mumbai, India, 2012.

References V

Marco Lui and Timothy Baldwin. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden, 2014. URL http://www.aclweb.org/anthology/W14-1303.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 280–287, Barcelona, Spain, 2004.

Sasa Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 181–189, Los Angeles, USA, 2010.

Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2014)*, pages 1523–1536, Baltimore, USA, 2014.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1500–1510, Jeju Island, Korea, 2012. URL http://www.aclweb.org/anthology/D12-1137.

References VI

Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29:93–106, 2008.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 327–335, Sydney, Australia, 2006.

Li Wang, Su Nam Kim, and Timothy Baldwin. The utility of discourse structure in identifying resolved threads in technical user forums. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2739–2756, Mumbai, India, 2012.

Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 955–964, Portland, USA, 2011. URL `http://www.aclweb.org/anthology/P11-1096`.

Xiaojin. Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.