Social Media: Friend or Foe of Natural Language Processing?

Tim Baldwin
Talk Outline

1. Social Media and Natural Language Processing
2. Bringing the Data to NLP
3. Bringing NLP to the Data
4. Concluding Remarks
What is Social Media?

- According to Wiktionary (21/8/2012), social media is:
  Interactive forms of media that allow users to interact with and publish to each other, generally by means of the Internet.

- While social media sites have strong support for multimedia content, text is still very much a core data type.
Social Media Include ...

Social Networking sites
posts, friends/circles, “likes”, shares, events, photos, comments, geotags, ...

Source(s): http://mashable.com/2011/02/04/facebook-7th-birthday/
Social Media Include ...

Micro-blogs

posts, followers/followeees, shares, hashtagging, geotags, ...

Source(s): http://itunes.apple.com/us/app/twitter/
Social Media Include ...

- Web user forums
  - posts, threading, followers/followees, ...

Source(s):

Social Media Include ...

Wikis

posts, versioning, linking, tagging, ...
Properties of Social Media Data

(NLP “ideal” → actuality)

- Edited text
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
Properties of Social Media Data

(NLP “ideal” → actuality)

<table>
<thead>
<tr>
<th>How different?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigram LM Perplexity:</td>
</tr>
<tr>
<td>BNC → Twitter₁ → Twitter₂ →</td>
</tr>
<tr>
<td>→ BNC</td>
</tr>
<tr>
<td>→ Twitter₁</td>
</tr>
<tr>
<td>→ Twitter₂</td>
</tr>
</tbody>
</table>
Properties of Social Media Data

(NLP “ideal” \(\rightarrow\) actuality)

- *Unedited text*
- *Static data*
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
- Streamed data
Properties of Social Media Data

(NLP “ideal” → actuality)

Challenges of Streaming Data
- require throughput guarantees
- batch vs. streamed processing of data (e.g. for topic modelling)
- potential need for “incremental” models
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Long(ish) documents; plenty of context
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
Properties of Social Media Data

(NLP “ideal” → actuality)

Document Context

Hard to adjust document-level priors when little context
Properties of Social Media Data

(NLP "ideal" $\rightarrow$ actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- All context is language context
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
Properties of Social Media Data

(NLP “ideal” → actuality)

? Priors, priors everywhere
user priors
user-declared metadata priors
location priors
social network-based priors
hashtag priors
timezone priors
implicit social networks (retweets, user mentions, ...)

...
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- Well-defined domain/genre
Properties of Social Media Data

\[(\text{NLP } \text{“ideal” } \rightarrow \text{ actuality})\]

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- All over the place
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- All over the place
- Sentence tokenisation
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Grammaticality
Properties of Social Media Data

(NLP “ideal” → actuality)

• Unedited text
• Streamed data
• Short documents; v. little context
• Little language, potentially lots of other context
• All over the place
• What’s a sentence?
• Yer what?
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Yer what?
- Most of what glitters is English (and if your method can handle one language, it can handle ’em all)
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Yer what?
- Anything goes — lots of languages, multilingual documents, ad hoc spelling, mix of language and markup ... language anarchy!
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
  - Is social media data too challenging for modern-day NLP?
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
  - Is social media data too challenging for modern-day NLP?
  - Are simple term search-based methods sufficient for social media analysis, i.e. is NLP overkill for social media?
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
  - Is social media data too challenging for modern-day NLP?
  - Are simple term search-based methods sufficient for social media analysis, i.e. is NLP *overkill* for social media?
- Is social media data the friend or foe of NLP?
Possible Ways Forward
Possible Ways Forward

- “Adapt” the data to the NLP tools through preprocessing of various forms
Possible Ways Forward

- “Adapt” the data to the NLP tools through preprocessing of various forms
- “Adapt” the NLP tools to the data through “domain” (de-)adaptation
Talk Outline

1. Social Media and Natural Language Processing
2. Bringing the Data to NLP
3. Bringing NLP to the Data
4. Concluding Remarks
Preprocessing

- Basic premise: the cleaner/richer the data, the easier it is to process/better quality the predictions that arise from it
- Overarching constraint: any preprocessing has to be able to keep pace with the torrent of streamed data ... although many of the models we use can be learned off-line
Language Identification: Task

- Language identification (langid) = prediction of the language(s) a given message is authored in

**Example**

*karena ada rencana ke javanet, maka siapkan link dolodan, di bookmark, ready to be a bandwidth killer.. siap siaplah javanet, im coming..
Language(s): ?*
Language Identification: Task

- Language identification (langid) = prediction of the language(s) a given message is authored in

Example

karena ada rencana ke javanet, maka siapkan link dolodan, di bookmark, ready to be a bandwidth killer.. siap siaplah javanet, im coming..

Language(s): MS,EN
Language Identification: Method

- Outline of the basic approach:

Source(s): Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]
Language Identification: Method

- Outline of the basic approach:
  1. represent each document as a set of byte $n$-grams of varying $n$
Language Identification: Method

- Outline of the basic approach:
  1. represent each document as a set of byte $n$-grams of varying $n$
  2. across a range of datasets, identify $n$-grams that are correlated with language and not dataset

$$\mathcal{LD}^{all}(t) = IG_{lang}^{all}(t) - IG_{domain}(t)$$
Language Identification: Method

- **Outline of the basic approach:**
  1. represent each document as a set of byte $n$-grams of varying $n$
  2. across a range of datasets, identify $n$-grams that are correlated with language and *not* dataset
  3. learn log likelihoods for each term and class from training data: $\log P(t_j|c_i)$

_Sources_: Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]
Language Identification: Method

- Outline of the basic approach:
  1. represent each document as a set of byte $n$-grams of varying $n$
  2. across a range of datasets, identify $n$-grams that are correlated with language and not dataset
  3. learn log likelihoods for each term and class from training data: $\log P(t_j|c_i)$
  4. classify a test document using multinomial naive Bayes over the $LD$ features

Source(s): Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]
Language Identification: Accuracy

- Comparative evaluation over pre-existing Twitter LangID datasets:

<table>
<thead>
<tr>
<th></th>
<th>langid.py</th>
<th>LangDetect</th>
<th>CLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy docs/s</td>
<td>ΔAcc</td>
<td>Slowdown</td>
</tr>
<tr>
<td>T-BE</td>
<td>0.941 367.9</td>
<td>-0.016</td>
<td>4.4×</td>
</tr>
<tr>
<td>T-SC</td>
<td>0.886 298.2</td>
<td>-0.038</td>
<td>2.9×</td>
</tr>
</tbody>
</table>

- Impact on bigram LM Perplexity:

<table>
<thead>
<tr>
<th></th>
<th>BNC→</th>
<th>Twitter-EN₁→</th>
<th>Twitter-EN₂→</th>
</tr>
</thead>
<tbody>
<tr>
<td>→BNC</td>
<td>185</td>
<td>1170 (-383)</td>
<td>1108 (-420)</td>
</tr>
<tr>
<td>→Twitter-EN₁</td>
<td>1528 (-2554)</td>
<td>215</td>
<td>416 (-471)</td>
</tr>
<tr>
<td>→Twitter-EN₂</td>
<td>1620 (-333)</td>
<td>469 (-469)</td>
<td>228</td>
</tr>
</tbody>
</table>

Source(s): Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]
Language Identification: Research Challenges

- We are very good at monolingual language identification, but what about multilingual documents?
Language Identification: Research Challenges

- We are very good at monolingual language identification, but what about multilingual documents?
  - multi-label language identification (*what language(s) is a document in*)
Language Identification: Research Challenges

- We are very good at monolingual language identification, but what about multilingual documents?
  - multi-label language identification (*what language(s) is a document in*)
  - language segmentation (*which parts of what messages correspond to what languages?*)
Language Identification: Research Challenges

- We are very good at monolingual language identification, but what about multilingual documents?
  - multi-label language identification (what language(s) is a document in)
  - language segmentation (which parts of what messages correspond to what languages?)
- How can we determine when we aren’t sure/don’t recognise the language(s)?
Lexical Normalisation: Task

• Lexical normalisation = “spell-correct” (English) messages to “canonical” lexical form:

Example

If you a Grl and you dont kno how to Cook yo bf should Leave you rite away

↓

If you a girl and you don’t know how to cook your boyfriend should leave you rite away

Source(s): Han and Baldwin [2011], Han et al. [2012], Gouws et al. [2011], Liu et al. [2011, 2012]
Lexical Normalisation: Method

- Outline of approach:

Source(s): Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Method

- Outline of approach:
  1. pre-learn (OOV,IV) word pairs from microblog data

Source(s): Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Method

- Outline of approach:
  1. pre-learn (OOV,IV) word pairs from microblog data
  2. lexical normalisation by simple dictionary lookup

Source(s): Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Method

- Outline of approach:
  1. pre-learn (OOV, IV) word pairs from microblog data
  2. lexical normalisation by simple dictionary lookup

- Learning the normalisation dictionary:

Source(s): Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Method

- **Outline of approach:**
  1. pre-learn (OOV, IV) word pairs from microblog data
  2. lexical normalisation by simple dictionary lookup

- **Learning the normalisation dictionary:**
  1. Extract (OOV, IV) pairs based on distributional similarity.

*Source(s):* Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Method

- **Outline of approach:**
  1. pre-learn (OOV, IV) word pairs from microblog data
  2. lexical normalisation by simple dictionary lookup

- **Learning the normalisation dictionary:**
  1. Extract (OOV, IV) pairs based on distributional similarity.
  2. Re-rank the extracted pairs by string similarity.

**Source(s):** Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Method

- **Outline of approach:**
  1. pre-learn (OOV, IV) word pairs from microblog data
  2. lexical normalisation by simple dictionary lookup

- **Learning the normalisation dictionary:**
  1. Extract (OOV, IV) pairs based on distributional similarity.
  2. Re-rank the extracted pairs by string similarity.
  3. Select the top-\(n\) pairs for inclusion in the normalisation lexicon.

*Source(s):* Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Results

- Lexical normalisation results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-dict</td>
<td>0.700</td>
<td>0.179</td>
<td>0.285</td>
</tr>
<tr>
<td>HB-dict</td>
<td>0.915</td>
<td>0.435</td>
<td>0.590</td>
</tr>
<tr>
<td>GHM-dict</td>
<td>0.982</td>
<td>0.319</td>
<td>0.482</td>
</tr>
<tr>
<td>HB-dict+GHM-dict+S-dict</td>
<td>0.847</td>
<td>0.630</td>
<td><strong>0.723</strong></td>
</tr>
</tbody>
</table>

Ultimately: dictionary combination works best

Source(s): Han and Baldwin [2011], Han et al. [2012]
Lexical Normalisation: Results

- Lexical normalisation results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-dict</td>
<td>0.700</td>
<td>0.179</td>
<td>0.285</td>
</tr>
<tr>
<td>HB-dict</td>
<td>0.915</td>
<td>0.435</td>
<td>0.590</td>
</tr>
<tr>
<td>GHM-dict</td>
<td>0.982</td>
<td>0.319</td>
<td>0.482</td>
</tr>
<tr>
<td>HB-dict+GHM-dict+S-dict</td>
<td>0.847</td>
<td>0.630</td>
<td><strong>0.723</strong></td>
</tr>
</tbody>
</table>

Ultimately: dictionary combination works best

- Impact on POS tagging:

<table>
<thead>
<tr>
<th>Tagger</th>
<th>Text</th>
<th>% accuracy</th>
<th># correct tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS_{Stanford}</td>
<td>original</td>
<td>68.4</td>
<td>4753</td>
</tr>
<tr>
<td>POS_{Stanford}</td>
<td>normalised</td>
<td>70.0</td>
<td>4861</td>
</tr>
<tr>
<td>POS_{twitter}</td>
<td>original</td>
<td>95.2</td>
<td>6819</td>
</tr>
<tr>
<td>POS_{twitter}</td>
<td>normalised</td>
<td>94.7</td>
<td>6780</td>
</tr>
</tbody>
</table>

Source(s): Han and Baldwin [2011], Han et al. [2012]
Other Instances of Preprocessing

- User/message geolocation
- Identification of “high-utility” messages
- Social media user profiling
- Credibility analysis
Talk Outline

1. Social Media and Natural Language Processing
2. Bringing the Data to NLP
3. Bringing NLP to the Data
4. Concluding Remarks
Instances of Social Media-adapted NLP tools

- CMU Twitter POS tagger: Twitter-tuned, coarse-grained POS tagset
- Self-training parser adaptation for social media data
- Named Entity Recognition for Twitter

Source(s): Gimpel et al. [2011], Foster et al. [2011], Ritter et al. [2011]
The Grand Challenge

- Social media data is highly temporal in nature, and models constantly need updating/de-adaptation.
- Often in social media analysis, people are interested in finding the unknown (e.g. novel event types, new products).
Talk Outline

1. Social Media and Natural Language Processing
2. Bringing the Data to NLP
3. Bringing NLP to the Data
4. Concluding Remarks
Friend or Foe?

If as NLPers we cherish a challenge, there is no question that social media is our friend. If we simplistically apply models trained on "traditional" datasets to social media, it is very much a foe... and evermore shall be so!

Social media also opens up immediate opportunities in terms of integrated multimodal analysis (links to image, video content); if we can harness this content, social media is again our friend (more context/better disambiguation).
Friend or Foe?

- If as NLPers we cherish a challenge, there is no question that social media is our friend.
Friend or Foe?

- If as NLPers we cherish a challenge, there is no question that social media is our friend.
- If we simplistically apply models trained on “traditional” datasets to social media, it is very much a foe ... and evermore shall be so!
Friend or Foe?

- If as NLPers we cherish a challenge, there is no question that social media is our friend.
- If we simplistically apply models trained on “traditional” datasets to social media, it is very much a foe ... and evermore shall be so!
- Social media also opens up immediate opportunities in terms of integrated multimodal analysis (links to image, video content); if we can harness this content, social media is again our friend (more context/better disambiguation).
NLP and Social Media

- Is NLP overkill for social media analysis?

Source(s): Ritterman et al. [2009], Sakaki et al. [2010]
NLP and Social Media

- Is NLP overkill for social media analysis?
- Much of the work on social media analysis is based on analysis of a pre-defined trend (e.g. election outcome prediction, flu outbreak tracking, earthquake detection)

Source(s): Ritterman et al. [2009], Sakaki et al. [2010]
NLP and Social Media

• Is NLP overkill for social media analysis?
• Much of the work on social media analysis is based on analysis of a pre-defined trend (e.g. election outcome prediction, flu outbreak tracking, earthquake detection)
  ... and perhaps NLP is overkill

Source(s): Ritterman et al. [2009], Sakaki et al. [2010]
NLP and Social Media

- Is NLP overkill for social media analysis?
- Much of the work on social media analysis is based on analysis of a pre-defined trend (e.g. election outcome prediction, flu outbreak tracking, earthquake detection)
  
  ... and perhaps NLP is overkill

- That is not to say there aren’t a myriad of applications which can’t be described with simple keywords for which NLP is vital (e.g. novel event detection, disaster management)

Source(s): Ritterman et al. [2009], Sakaki et al. [2010]
NLP and Social Media

- Is NLP overkill for social media analysis?
- Much of the work on social media analysis is based on analysis of a pre-defined trend (e.g. election outcome prediction, flu outbreak tracking, earthquake detection)
  
  ... and perhaps NLP is overkill
- That is not to say there aren’t a myriad of applications which can’t be described with simple keywords for which NLP is vital (e.g. novel event detection, disaster management)
  
  ... and perhaps the bottleneck is instead NLP accessibility

Source(s): Ritterman et al. [2009], Sakaki et al. [2010]
Final Words

- Social media is hip ... but also big and hairy, and poses both challenges and opportunities for NLP
- Ongoing work on a myriad of technologies/tasks relating to social media analysis, progressively making social media more “NLP accessible”
- There is plenty to be done ... come and join us!
Taking Credit for a Cast of Thousands

- This is joint work with Paul Cook, Bo Han, Aaron Harwood, Shanika Karunasekera, Su Nam Kim, Marco Lui, David Martinez, Joakim Nivre, Richard Penman, Li Wang, ...
References I


References II


References III

