Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Alexander Pak, Patrick Paroubek

Université Paris-Sud 11, LIMSI-CNRS
Microblogging

Microblogging = posting small blog entries

Eg.: “@alex: I'm presenting now my paper at LREC'10”

Platforms:
  • Twitter
twitter
  • Tumblr
tumblr
  • Plurk
Twitter

Realtime results for LREC

@dj_satoru RT @sassano: LREC 2010 Proceedings http://www.lrec-conf.org/proceedings/lrec2010/index.html
6 minutes ago from TweetDeck

@nlpbot #NLBP: In Malta with #hblanca for the #LREC conference. Several hundreds researchers in Natural Language Processing...
http://is.gd/cfUgg
about 3 hours ago from tweet hopper

@anitawaard #LREC is on! I'm there like hundred other researchers.
about 4 hours ago from Seesmic

@cdeloupy #LREC Opening session. Roberto Cencioni: NLP can do better in terms of credibility, critical mass and visibility... true enough!
about 4 hours ago from Seesmic

@cdeloupy In Malta with @hblanca for the #LREC conference. Several hundreds researchers in Natural Language Processing. Let's meet old friends.
about 5 hours ago from Seesmic

Search tip
Use an @-sign (@) immediately before a username to find mentions of (and replies to) a user. Example: @mashable will find tweets referencing the user "mashable".
Twitter

Twitter - social network for publishing short messages (tweets)

1 tweet contains:
  maximum: 140 character
  in average: 1 sentence

More than 1 billion tweets per month
Twitter for opinion mining

People are expressing their opinions in tweets

Eg.: “CelineBG: @itsRyanButler u should come to Malta (europe) it's below Italy..we have sun nearly all year round =) we have amazing beaches =) follow me”

Twitter is multilingual

More than 14 billion tweets

Twitter API for data retrieval
Corpus collection

Use emoticons as noisy sentiment labels

Positive tweets with :) =) :D
@mia_jones oh lovely! I'm heading to Malta & Italy next week!! Can't wait :) 

Negative tweets with :( :'( :( 
Supposed to be flying tonight, now stuck in Malta until Thursday. Homesick :( 

Use newspapers' tweets for neutral texts
@nytimes: Iron Man Defeats Robin Hood at North American Box Office
Corpus analysis

Collected 300'000 positive, negative and neutral tweets

Distribution of word frequencies is Zipfian

Use TreeTagger for POS tagging
Objective vs. subjective tweets

Utterances indicate subjective texts

Objective texts contain more common and proper nouns

Subjective texts contain more personal pronouns

Twitter as a Corpus for Sentiment Analysis and Opinion Mining (A. Pak and P. Paroubek)
**Objective vs. subjective tweets**

<table>
<thead>
<tr>
<th>Verb Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBN</td>
<td>Verbs in objective texts are usually in the 3rd person</td>
</tr>
<tr>
<td>VBP</td>
<td>Authors write about themselves or address the audience</td>
</tr>
<tr>
<td>VBD</td>
<td>Past participle is used for stating facts</td>
</tr>
<tr>
<td>VBD</td>
<td>Modal verbs are used to express emotions</td>
</tr>
<tr>
<td>JJR</td>
<td>Comparative adjectives state facts</td>
</tr>
<tr>
<td>JJS</td>
<td>Superlative adjectives express emotions</td>
</tr>
</tbody>
</table>

*Twitter as a Corpus for Sentiment Analysis and Opinion Mining (A. Pak and P. Paroubek)*
Positive vs. negative tweets

- Negative tweets often contain past tense (VBN, VBD)
- Superlative adjectives may indicate positive tweets
- Positive tweets more often contain possessive endings
Building a classifier

Use the corpus to train a sentiment classifier

Use Naïve Bayes classifier

2 types of features: n-grams and POS

Bigrams showed the best performance

Handle negations by attaching negation particle
Eg.: I do not like fish: I do+not, do+not like, not+like fish
Building a classifier

Use hand annotated tweets for evaluation:

Positive: 108
Negative: 75
Neutral: 33
Total: 216
Increasing accuracy

Classify tweets with high confidence of precision

Other tweets are left as “undecided”

“decision” = ratio of classified tweets

Select n-grams with high salience (ignore n-grams with same frequency in all three sets)
### Increasing accuracy

<table>
<thead>
<tr>
<th>N-gram</th>
<th>Salience</th>
</tr>
</thead>
<tbody>
<tr>
<td>So sad</td>
<td>0.975</td>
</tr>
<tr>
<td>Miss my</td>
<td>0.972</td>
</tr>
<tr>
<td>So sorry</td>
<td>0.962</td>
</tr>
<tr>
<td>Love your</td>
<td>0.961</td>
</tr>
<tr>
<td>I'm sorry</td>
<td>0.96</td>
</tr>
<tr>
<td>Sad I</td>
<td>0.959</td>
</tr>
<tr>
<td>I hate</td>
<td>0.959</td>
</tr>
<tr>
<td>Lost my</td>
<td>0.959</td>
</tr>
<tr>
<td>Have great</td>
<td>0.958</td>
</tr>
</tbody>
</table>

Twitter as a Corpus for Sentiment Analysis and Opinion Mining (A. Pak and P. Paroubek)
Results

Comparison of n-gram order

Impact of negation attachment

Twitter as a Corpus for Sentiment Analysis and Opinion Mining (A. Pak and P. Paroubek)
Prototype

Results for query: «Malta»

1. Wear makeup to school, leave your hair down, plus ur mixed guys and girls, u have a cafeteria! U rarely find a mixed school in Malta we only
   @TeamJemixDemix created at Wed, 19 May 2010 13:01:26 +0000

2. We’re just putting the finishing touches on the Malta Live magazine - looking forward to sharing our insider’s guide to Malta with you guys!
   @MaltaLive created at Wed, 19 May 2010 11:44:38 +0000

3. Just received confirmation that my graduation ceremony will be held in Malta instead of London. I’m so disappointed!!!! :( 
   @jesmond created at Wed, 19 May 2010 11:37:29 +0000

4. @C8000Malta That’s superb!! All Malta is so proud of you guys! Can’t wait to hear all your stories up on Everest...
   @MyMaltaInfo created at Wed, 19 May 2010 11:19:30 +0000

   @Malta created at Wed, 19 May 2010 11:03:56 +0000

   @article24th created at Wed, 19 May 2010 10:55:57 +0000

7. If I was Cameron I’d be bit worried about Clegg giving too many speeches...but is UK really most centralised country in Europe bar Malta?
   @iankatz1000 created at Wed, 19 May 2010 10:30:18 +0000

Twitter as a Corpus for Sentiment Analysis and Opinion Mining (A. Pak and P. Paroubek)
Conclusion

Twitter can be used as a sentiment-labeled corpus

Naive-Bayes with bigram and POS features can perform a precise sentiment classification

Future work: collect more tweets, form a multilingual corpus
Thank you!

Twitter as a Corpus for Sentiment Analysis and Opinion Mining (A. Pak and P. Paroubek)