807 - TEXT ANALYTICS

Massimo Poesio

Lecture 9: Text analytics for social media analysis

Text analytics-based social media analysis
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+, ...
• Content sharing platforms
  – YouTube, Instagram, Flickr, ...
• Blogs
  – Gizmodo, Boing Boing, ...
• Micro-blogs
  – Twitter, Weibo, ...
• Web user forums
  – Stackoverflow, CNET, Apple Support, ...
• Wikis
  – Wikipedia, Wiktionary ...
Social media analysis applications that rely on text analytics

- Health care applications
- Financial applications
- Predicting voting intentions
- Security and defence applications
- Disaster response applications
- NLP-based user modelling
- Applications for entertainment
- Social media monitoring
  - Event identification
  - Opinion mining and emotion analysis
  - Geo-location detection
  - Summarization
  - Machine translation

Health applications

- Many online platforms where people discuss their health:
  - Specialized forums, for various topics. The language is often informal and medical terms can be found, but most of the language is lay. Various kinds of information can be extracted automatically from such postings and discussions.
  - Opinions and arguments pro and cons topics such as: vaccinations, mammographies, new born genetic screening.
- Need privacy protection: detection of personal health information (PHI) such as names, dates of birth, addresses, health insurance numbers.
Financial applications

- Behavioral economics studies the correlation between public mood and economic indicators, and between financial news/rumors and stock exchange fluctuations.
- Recent studies show that using social media (Twitter, Sina weibo, Seeking Alpha) data to automatically measure public mood (rather than using expensive traditional polls) can be useful in financial applications.
  - Experiments were run on predicting stock market fluctuations for NASDAQ, New York Stock Exchange, DOW Jones, S&P 500, Shanghai Stock Exchange, Turkish Stock Exchange, etc.

Predicting voter intentions

- Need to detect messages about the desired topic or political entities of interest (using keyword search or text classification methods).
- Then use opinion detection/sentiment analysis techniques.
- Studies:
  - Automatic opinion polling given a comments written after voting, on the SodaHead social polling website.
  - Tjong Kim Sang and Bos (2012) used Twitter data to predict the 2011 Dutch Senate Election Results.
  - Bermingham and Smeaton (2011) used social media for prediction of the 2011 Irish General Election.
  - Several studies on US elections using congress debates,
  - political blogs and their comments, Twitter data, etc.
Security and defence applications

• Humans can read only a small part of the user-generated content in social media in order to detect possible threats to security and public safety (mentions of terrorist activities or extremist/radical texts). Automatic methods can detect messages that should be flagged as possible threats and forwarded to a human for further analysis.

• Forensic data mining for intrusion detection (Mohay et al., 2003)

• Military image classification based on text captions and tags.

• Information extraction from text. Key phrase search or classification of a text as being about a terrorism-related topic or not.

• Situation awareness. CRF classifiers on large amounts of textual maritime incident reports, to extract: vessel type, risk type, risk associates, a maritime general location, a maritime absolute location (latitude/longitude), date and time (Razavi et al., 2014)

Security and defence applications (cont’d)

• **Topic detection** in social media texts, Friendsfeed data, using multi-level LDA features (Razavi et al., 2013).

• **Location detection** from social media texts: Twitter user location (Li and Inkpen, 2015)

• **Emotion detection** from social media texts. Anger and sadness detection are of particular interest. Emotion classifiers (including anger and sadness) were tested on blog data (Ghazi et al., 2010), on LiveJournal data (Keshtkar and Inkpen, 2012) and other kinds of social media postings. Messages that express anger at high intensity levels could be flagged as possible terrorist threats.
Disaster response applications

- A sudden change in the topics discussed in social media in a region can indicate a possible emergency situation, for example a natural disaster such as an earthquake, fire, tsunami, or flooding.
- Social media messages can be used for spreading information about the evolution of the situation.
- New event detection or information about existing events.
- Experiments on tweets:
  - Extracting info about disaster response actions (Imran et al., 2013)
  - earthquake detector for Australia and New Zealand (Robinson et al., 2013)
  - detecting reported fires (Power et al., 2013)

User modelling

- Learn user profiles based on their social media behaviour (all the posts of a user).
- Modelling a user’s personality.
  - ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media and the hared tasks on Computational Personality Recognition 2014 and 2013.
  - Big Five model: extraversion, emotional stability, agreeableness, conscientiousness and openness to experience.
- Modelling the user’s health profile.
- Modelling the user’s political orientation.
- Modelling the user’s life events.
- Modelling the user’s location.
Applications in the media and the entertainment industry

- Social media are a big challenge for the media and entertainment industry
- Social media are changing users’ expectations and their behaviour
- New approaches toward content creation, distribution, operations, technology, and user interaction
- Serious issue: Since advertisers spend less on traditional paid media and require more resources for digital social media and e-marketing.

PR

- The traditional method used by public relations agencies is sending out press releases and waiting for the media to write about their event
- Social sharing press releases and creating social campaigns around customer case studies, publishing short videos on YouTube and choosing the best quotes to share on Twitter or Facebook
- Journalists rely heavily on Twitter, Facebook and other social media platforms to source and research stories.
Entertainment: movie / TV ranking

- Measurements of movie and TV programming ranking are one of the important indicator regarding the popularity of a program or a movie in the entertainment industry.
- Netflix now uses the popularity of a movie on Facebook as a proposed feature for consumers.
- Hsieh et al. (2013) predicted TV audience ratings using a back-propagation network and the number of posts, likes, comments and shares on the fan pages of various TV dramas to try to find their relationships to ratings.
  - Their result showed that using Facebook fan page data to perform ratings forecasts for non-broadcast programs should be feasible.

Sentiment analysis and the entertainment industry

- Sentiment analysis of Major events such as the Oscars
- Sentiment related to the movies premium
- Sinha et al. (2014) carried out a sentiment analysis of Wimbledon Tweets by analysing a set of tweets of the Roger Federer and Novak “Nolé” Djokovic semifinals match at Wimbledon 2012.
  - In the absence of textual metadata for annotating videos, they assumed that the live video coverage of an event and the time correlated textual microblog streams about the same event can act as an important source for such annotation.
  - The intensity of sentiment is used to detect peaks of sentiments towards players as well, and can tag best moments in the game.
Analyzing social media

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
Three types of semantic analysis of social media data

• content-based semantic analysis
• user-based semantic analysis
• network-based semantic analysis

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
- 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.)

Sentiment and emotion analysis in Twitter: SemEval

- A benchmark dataset was created for a shared task at SemEval 2013
  - http://www.cs.york.ac.uk/semeval-2013/task2/
  - 8000 tweets annotated with the labels: positive, negative, neutral and objective (no opinion).
  - Task A: Given a message that contains a marked instance of a word or phrase, the goal of Task A was to determine whether that instance is positive, negative or neutral in that context.
  - Task B was to classify the whole message as pos/neg/neutral
  - Example: 00032373000896513 15486118 lady gaga "positive" Wow!! Lady Gaga is actually at the Britney Spears Femme Fatale Concert tonight!!! She still listens to her music!!!! WOW!!!
- More editions have been held
Example of a system for SemEval 2013 (Pousepanj et al, 2013)

• Method: ML with extra features: the number of positive words and negative words from SentiWordNet, General Inquirer, and Polarity Lexicon; the number of emoticons; the number of elongated words; and the number of punctuation tokens (!, !!, !!!, ...)

• Results on Task B (accuracy):
  – Best system: 0.69 on tweets, 0.68 on SMS

Automatic event detection / tracking / monitoring in social media

• Events can be defined as situations, actions or occurrences that happen in a certain location at a specific time

• An event is generally characterized by: 5W1H – who? when? where? what? why? how?

• An “Event of interest” is domain dependent – Terrorist activities, sport tournaments, conference, trade show, natural disaster, etc
Event detection

- Objective: discover new or previously unidentified events
  1. Retrospectively: detect previously unidentified events from accumulated data
     – clustering techniques based on similarity measures
     – new events are distant from previous clusters
  2. Online: discovery of new events from continuous stream(s) in (near) real-time
     - statistical: frequency, n-grams, HMM, wavelets, ...
     - new events: bursty deviation from normal behavior

Thread classification in online forums

- 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
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 these 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 alias for downloading files... At the end of the installer, there is... After this you are left with...

UserA
Post5
I mostly use a minimal boot CD (based on 2.4) to install Debian... Use it to install the base system, then apt-get or dselect to get whatever you need...

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 these 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 alias for downloading files... At the end of the installer, there is... After this you are left with...

UserA
Post5
I mostly use a minimal boot CD (based on 2.4) to install Debian... Use it to install the base system, then apt-get or dselect to get whatever you need...

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 these 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 alias for downloading files... At the end of the installer, there is... After this you are left with...

UserA
Post5
I mostly use a minimal boot CD (based on 2.4) to install Debian... Use it to install the base system, then apt-get or dselect to get whatever you need...
Unanswered questions in online forums

Linguistic applications of content-based analysis: 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 appear]
  – 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)
Linguistic applications of content-based analysis: detection / analysis of localisms

- Task outline: analyse the geographical spread of different terms based on geotagged data, and identify terms which have 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 appear]

Detection / analysis of localisms

Example: distribution of the term *buku* based on a geotagged analysis of US tweets (Cook et al)
Semantic analysis at scale

• The good news: social media content is generally plentiful, if you aren’t picky about the data
  – 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 #Frergusson #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)
Word 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 social media analysis

- (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”, ...
  - (3) It’s getting late the babe sleep guess I’ll ko now kawaii ne! #fb
- Certain NLP tasks such as constituency parsing over social media text are probably a lost cause (Baldwin et al. [2013], although see Kong et al. [to appear] on dependency parsing Twitter),
- but like speech, this type of language use a natural application of semantic parsing
Content-based analysis of social media: 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

User-based analysis
User-based analysis

• All we 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

What we can learn about users

• 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 addition knowing something about the messages associated with users (e.g. geotags) or the user her/himself (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 user-oriented task: 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 user-oriented task: 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: [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 ...
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 computational linguists

Network-based approaches
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

More applications
Other NLP applications in social media

• Machine translation
• Summarization
  – E.g., SENSEI

Practical aspects
Social media analysis, practical - I

- 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]

NLP for social media: existing tools

- Language identification:
  – langid.py, CLD2, langdetect, TwitIE, polyglot, ...
- (English) tokenisation:
  – GATE, Twokenizer, Chris Potts’ tokeniser, ...
- (English) lexical normalisation:
  – UniMelb lexical normalisation dictionary, TextCleanser, TwitIE
- POS tagging:
  – GATE, ARK Twitter POS tagger, Twitter NLP, TwitIE
NLP for social media: existing tools

• NER:
  – Twitter NLP, TwitIE
• Message geolocation/geoparsing:
  – CMU GeoLocator
• User geolocation:
  – UniMelb Twitter user geolocator
• User profiling:
  – Bot or not, Twitter Clusters

Specialised tools

• TweetNLP is a Java-based tokenizer and POS tagger for Twitter text (Owoputi et al., 2013). It includes training data of manually labeled POS annotated tweets (that we noted above), a Web-based annotation tool, and hierarchical word clusters from unlabeled tweets. http://www.ark.cs.cmu.edu/TweetNLP/. It also includes the TweeboParser.
• The UW Twitter NLP Tools (Ritter et al., 2011) contain the POS tagger and the annotated Twitter data. https://github.com/aritter/twitter_nlp
Some datasets

• 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]
  – SENSEI Online forums summarization dataset (ask me)

Readings

• A. Farzindar & D. Inkpen, *Natural Language Processing for Social Media*, Synthesis Lectures on Human Language Technologies
Acknowledgments

• Some of the slides from Tim Baldwin’s tutorial on semantic analysis in social media at *SEM 2014
• Other slides from Farzinder & Inkpen’s tutorial at EMNLP 2015