Challenges in developing opinion mining tools for social media

Diana Maynard, Kalina Bontcheva, Dominic Rout

University of Sheffield, UK

Introduction

- Information, thoughts and opinions are shared prolifically these days on the social web
- It can be difficult to get the relevant information out of such large volumes of data in a useful way
- Social web analysis is all about the users who are actively engaged and generate content
- Social networks are pools of a wide range of articulation methods, from simple "I like it" buttons to complete articles
- Along with entity, topic and event recognition, opinion mining forms the cornerstone for social web analysis

Key questions

- What are the opinions on crucial social events and the key people involved?
- How are these opinions distributed in relation to demographic user data?
- How have these opinions evolved?
- Who are the opinion leaders?
- What is their impact and influence?

What's the problem?

- Opinion mining is hard anyway, and it's harder in this case because:
  - we have lots of different text types and domains
  - we're processing social media, where language isn't used properly
  - we're processing multiple languages
  - we don't necessarily know what we're looking for

But there are lots of tools that do this already....

- Here are some examples:
  - Twitter sentiment http://twittersentiment.appspot.com/
  - Twends: http://twendz.waggeneredstrom.com/
  - Twittratr: http://twitrratr.com/
  - SocialMention: http://socialmention.com/

Venus Williams causes controversy...

Search for "Venus Williams"

arcomem

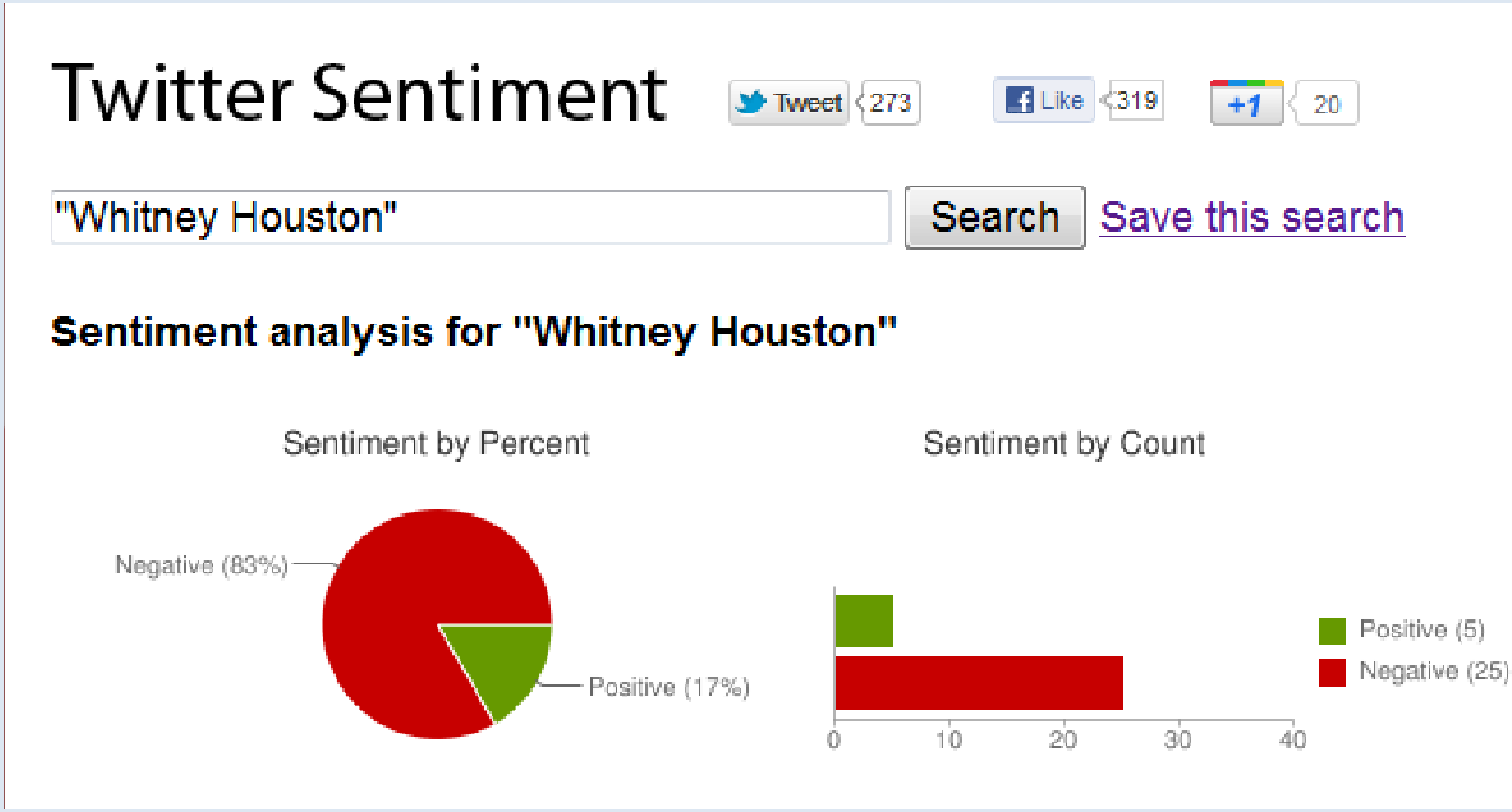
Using existing Twitter sentiment apps

- Easy to search for opinions about famous people, brands and so on
- Hard to search for more abstract concepts, perform a non-keyword based string search
- e.g. to find opinions about Venus Williams' dress, you can only search on “Venus Williams” to get hits

Why are these sites unsuccessful?

- They don't work well at more than a very basic level
- They mainly use dictionary lookup for positive and negative words
- They classify the tweets as positive or negative, but not with respect to the keyword you're searching for
- First, the keyword search just retrieves any tweet mentioning it, but not necessarily about it as a topic
- Second, there is no correlation between the keyword and the sentiment: the sentiment refers to the tweet as a whole
- Sometimes this is fine, but it can also go horribly wrong

Whitney Houston wasn't very popular...



Or was she?

Tweets about: "Whitney Houston"

bazzyboy25: **Whitney houston**...too soon? #CelebritiesThatLookLikeTheyStank
Posted 5 minutes ago

TeghanSimone: Radio playing **Whitney Houston**.. I swear I'm about to cry... So sad
Posted 5 minutes ago

JB3LL: hoes about to get **whitney houston**'d tonight! #TheWalkingDead
Posted 5 minutes ago

derickaadamss: “@indreamville_: Twitter I'm curious who do you think had more problems Michael Jackson or **Whitney Houston**???” <<<< **Whitney Houston**!
Posted 5 minutes ago

charlottesteer4: Listening to **Whitney Houston** loveeeee songsss <3 she's amazing <3
Posted 5 minutes ago

DionneHeraty40: @Sbarry25 The reason why **Whitney Houston** died at only 41 http://t.co/JJKRDjbj
Posted 5 minutes ago

ShortySoooFine: #musicwasbestwhen legends like James brown, Michael Jackson, **Whitney Houston** still lived.
Posted 5 minutes ago

CarlmannJohnson: Pray for Bobby Brown!!! He lost his ex-wife **Whitney Houston** and his dad Herbert Brown... Prayers up for you!!
Posted 5 minutes ago

LonelySpaceman: Is it bad that I thought **Whitney Houston** was already dead?
Posted 5 minutes ago

eatmy_CHOCLATE: My aunt in there playing **Whitney Houston** making me sad
Posted 5 minutes ago

The results for this query are: Accurate Inaccur

Twittrater's view of the Olympics

- A keyword search for Olympics shows exactly how existing systems fail to cut the mustard
- Lookup of sentiment words is not enough if
  - they're part of longer words
  - they're used in different contexts
  - the tweet itself isn't relevant
  - they're used in a negative or sarcastic sentence
  - they're ambiguous

Applications

- Developed a series of initial applications for opinion mining from social media using GATE
- Based on previous work identifying political opinions from tweets
- Extended to more generic analysis about any kind of entity or event, in 2 domains
  - Greek financial crisis
  - Rock am Ring (German rock festival)
- Uses a variety of social media including twitter, facebook and forum posts
- Based on entity and event extraction, and a rule-based approach

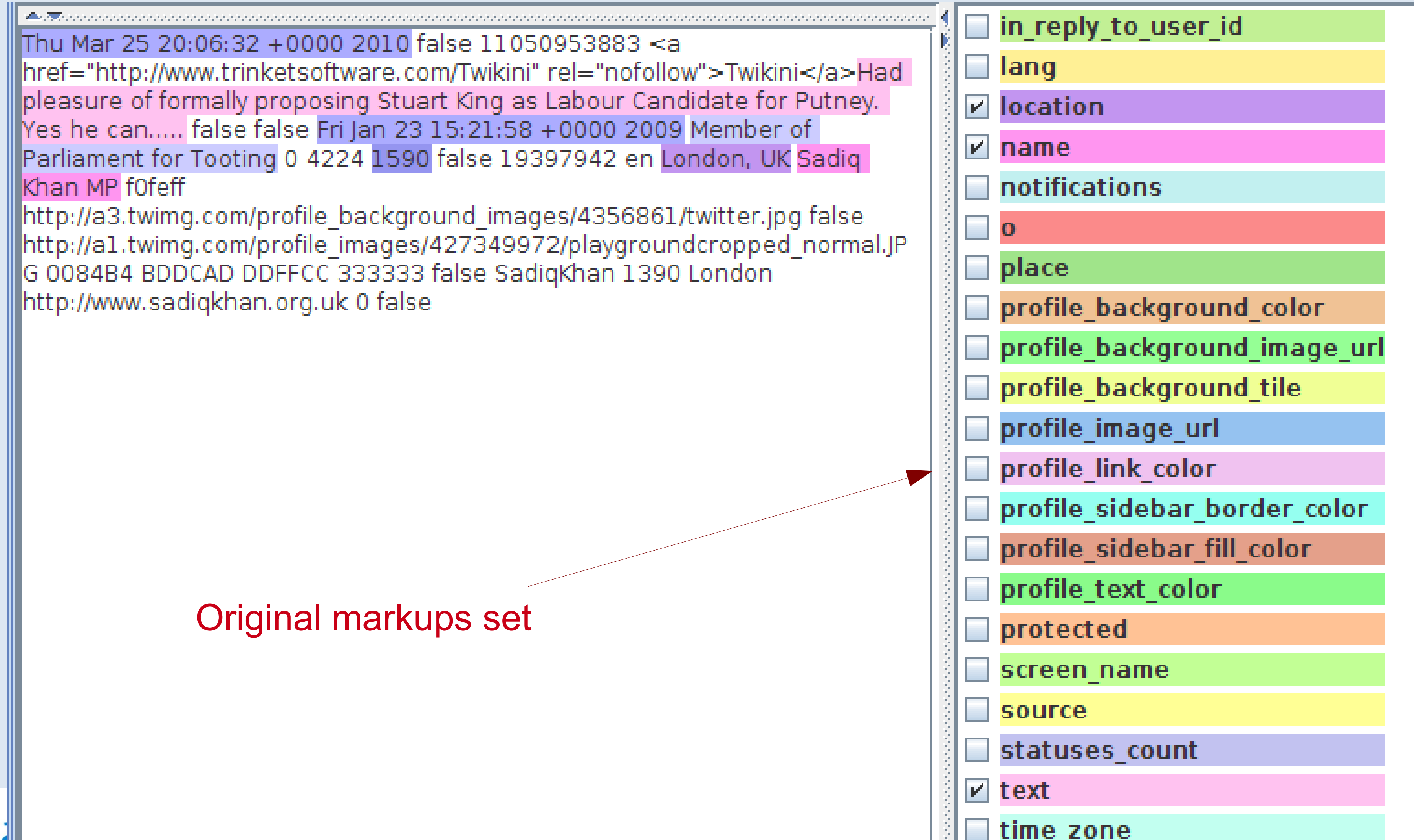
Why Rule-based?

- Although ML applications are typically used for Opinion Mining, this task involves documents from many different text types, genres, languages and domains
- This is problematic for ML because it requires many applications trained on the different datasets, and methods to deal with acquisition of training material
- Aim of using a rule-based system is that the bulk of it can be used across different kinds of texts, with only the pre-processing and some sentiment dictionaries which are domain and language-specific

GATE Application

- Structural pre-processing, specific to social media types
- Linguistic pre-processing (including language detection), NE, term and event recognition
- Additional targeted gazetteer lookup
- JAPE grammars
- Aggregation of opinions
- Dynamics

Structural pre-processing on Twitter



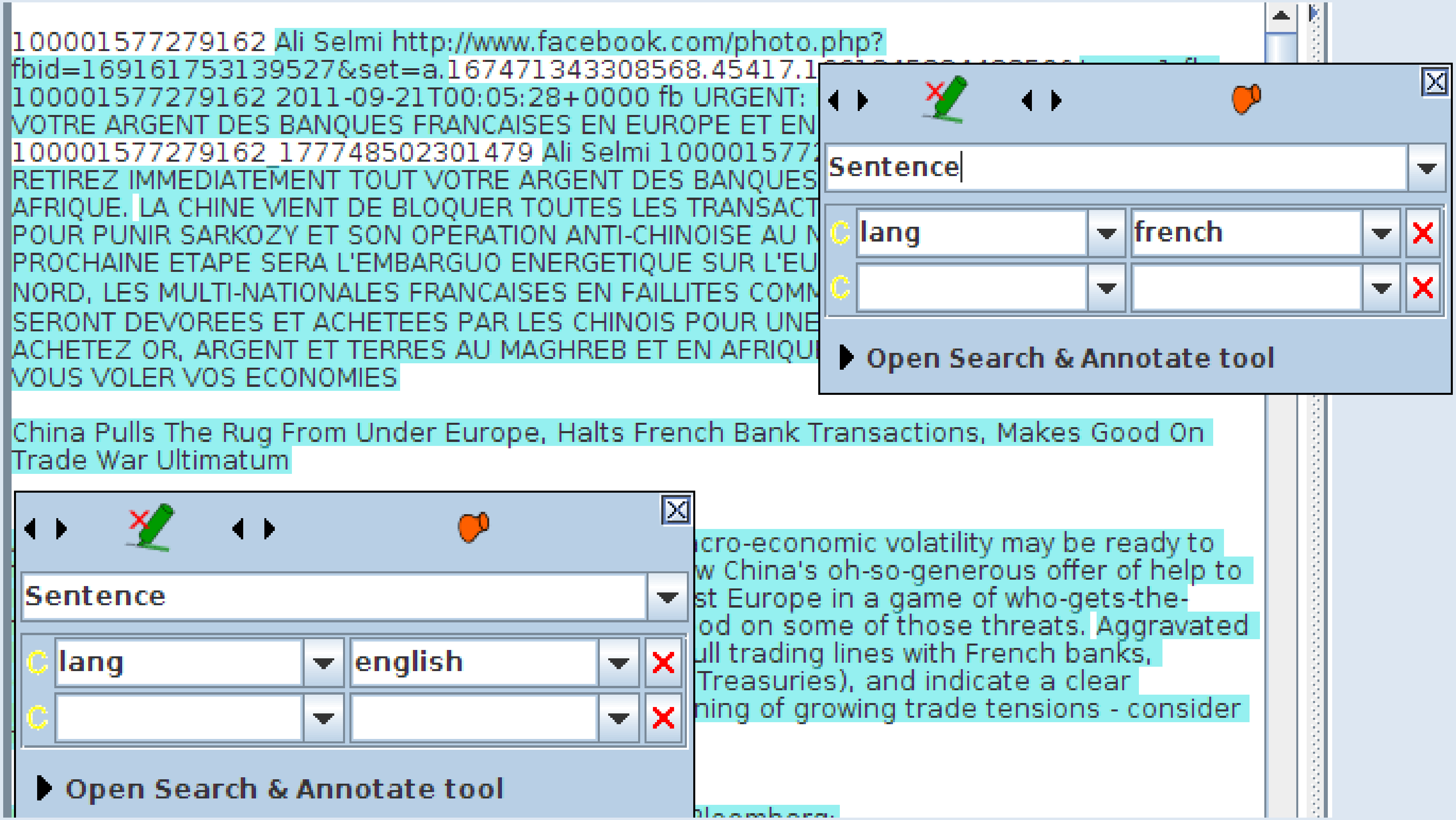
Linguistic pre-processing

- Language identification (per sentence) using TextCat
- Standard tokenisation, POS tagging etc using GATE
- Modified versions of ANNIE and TermRaider for NE and term recognition
- Event recognition using specially developed GATE application (e.g. band performance, economic crisis, industrial strike)

arcomem

SEVENTH FRAMEWORK PROGRAMME

Language ID with TextCat



Basic approach for opinion finding

- Find sentiment-containing words in a linguistic relation with entities/events (opinion-target matching)
- Use a number of linguistic sub-components to deal with issues such as negatives, irony, swear words etc.
- Starting from basic sentiment lookup, we then adjust the scores and polarity of the opinions via these components

SEVENTH FRAMEWORK PROGRAMME

Sentiment finding components

- **Flexible Gazetteer Lookup**: matches lists of affect/emotion words against the text, in any morphological variant
- **Gazetteer Lookup:** matches lists of affect/emotion words against the text only in non-variant forms, i.e. exact string match (mainly the case for specific phrases, swear words, emoticons etc.)
- **Sentiment Grammars**: set of hand-crafted JAPE rules which annotate sentiments and link them with the relevant targets and opinion holders
- **RDF Generation**: create the relevant RDF-XML for the annotations according to the data model (so they can be used by other components)

Opinion scoring

- Sentiment gazetteers (developed from sentiment words in WordNet) have a starting "strength" score
- These get modified by context words, e.g. adverbs, swear words, negatives and so on

Challenges imposed by social media

- **Language**: specific pre-processing for Twitter. use shallow analysis techniques with back-off strategies; incorporate specific subcomponents for swear words, sarcasm etc.
- **Relevance**: topics and comments can rapidly diverge. Solutions involve training a classifier or using clustering techniques
- **Target identification**: use an entity-centric approach
- **Contextual information**: use metadata for further information, also aggregation of data can be useful

Short sentences, e.g. tweets

- Social media, and especially tweets, can be problematic because sentences are very short and/or incomplete
- Typically, linguistic pre-processing tools such as POS taggers and parsers do badly on such texts
- Even things like language identification tools can have problems
- The best solution is to try not to rely too heavily on these tools
  - Does it matter if we get the wrong language for a sentence?
  - Do we actually need full parsing?
  - Can we use other clues when POS tags may be incorrect?

Dealing with incorrect English

- Frequent problem in any NLP task involving social media
- Incorrect capitalisation, spelling, grammar, made-up words (eg swear words, infixes)
- Some specific pre-processing
- Backoff strategies include
  - using more flexible gazetteer matching
  - using case-insensitive resources (but be careful)
  - avoiding full parsing and using shallow techniques
  - using very general grammar rules
  - adding specialised gazetteer entries for common misspellings, or using co-reference techniques

Tokenisation

- Plenty of “unusual”, but very important tokens in social media:
  - @Apple – mentions of company/brand/person names
  - #fail, #SteveJobs – hashtags expressing sentiment, person or company names
  - :-(, :-), :-P – emoticons (punctuation and optionally letters)
  - URLs
- Tokenisation key for entity recognition and opinion mining
- A study of 1.1 million tweets: 26% of English tweets have a URL, 16.6% - a hashtag, and 54.8% - a user name mention [Carter, 2013].

Example

#WiredBizCon #nike vp said when @Apple saw what http://nikeplus.com did, #SteveJobs was like wow I didn't expect this at all.

- Tokenising on white space doesn't work that well:
- Nike and Apple are company names, but if we have tokens such as #nike and @Apple, this will make the entity recognition harder, as it will need to look at sub-token level
- Tokenising on white space and punctuation characters doesn't work well either: URLs get separated (http, nikeplus), as are emoticons and email addresses

The GATE Twitter Tokeniser

- Treat RTs, emoticons, and URLs as 1 token each
- #nike is two tokens (# and nike) plus a separate annotation HashTag covering both. Same for @mentions
- Capitalisation is preserved, but an orthography feature is added: all caps, lowercase, mixCase
- Date and phone number normalisation, lowercasing, and other such cases are optionally done later in separate modules
- Consequently, tokenisation is faster and more generic

De-duplication and Spam Removal

- Approach from [Choudhury & Breslin, #MSM2011]:
- Remove as duplicates/spam:
  - Messages with only hashtags (and optional URL)
  - Similar content, different user names and with the same timestamp are considered to be a case of multiple accounts
  - Same account, identical content are considered to be duplicate tweets
  - Same account, same content at multiple times are considered as spam tweets

Normalisation

- "RT @Bthompson WRITEZ: @libbyabrego honored?! Everybody knows the libster is nice with it...lol...(thankkkks a bunch;))"
- OMG! I'm so guilty!!! Sprained biibii's leg! ARGHHHHHH!!!!!!
- Similar to SMS normalisation
- For some later components to work well (POS tagger, parser), it is necessary to produce a normalised version of each token
- BUT uppercasing, and letter and exclamation mark repetition often convey strong sentiment, so we keep both versions of tokens
- Syntactic normalisation: determine when @mentions and #tags have syntactic value and should be kept in the sentence, vs replies, retweets and topic tagging

arcomem

SEVENTH FRAMEWORK PROGRAMME

Irony and sarcasm

- *Life's too short, so be sure to read as many articles about celebrity breakups as possible.*
- *I had never seen snow in Holland before but thanks to twitter and facebook I now know what it looks like. Thanks guys, awesome!*
- *On a bright note if downing gets injured we have Henderson to come in.*

How do you know when someone is being sarcastic?

- Use of hashtags in tweets such as #sarcasm
- Large collections of tweets based on hashtags can be used to make a training set for machine learning
- But you still have to know which bit of the tweet is the sarcastic bit

To the hospital #fun #sarcasm

Man , I hate when I get those chain letters & I don't resend them , then I die the next day .. #Sarcasm

lol letting a baby goat walk on me probably wasn't the best idea. Those hooves felt great. #sarcasm

How else can you deal with it?

- Look for word combinations with opposite polarity, e.g. “rain” or “delay” plus “brilliant”

Going to the dentist on my weekend home. Great. I'm totally pumped. #sarcasm

- Inclusion of world knowledge / ontologies can help (e.g. knowing that people typically don't like going to the dentist, or that people typically like weekends better than weekdays.
- It's an incredibly hard problem and an area where we expect not to get it right that often
- Still very much work in progress for us

Evaluation

- Very hard to measure opinion polarity beyond positive/negative/neutral
- On a small corpus of 20 facebook posts, we identified sentiment-containing sentences with 86% Precision and 71% Recall
- Of these, the polarity accuracy was 66%
- While this is not that high, not all the subcomponents are complete in the system, so we would expect better results with improved methods for negation and sarcasm detection
- NE recognition was high on these texts: 92% Precision and 69% Recall (compared with other NE evaluations on social media)

Comparison of Opinion Finding in Different Tasks

Corpus	Sentiment detection	Polarity detection	Target assignment
Political Tweets	78%	79%	97.9%
Financial Crisis Facebook	55%	81.8%	32.7%
Financial Crisis Tweets	90%	93.8%	66.7%

Summary

- Ongoing work on adapting opinion-mining tools to social media
- Deal with multilinguality, ungrammatical English, and very short posts (tweets)
- Components for negation, swear words, sarcasm etc
- Promising initial evaluations
- Much further work still to come

Further information

- Work done in the context of the EU-funded ARCOMEM and TrendMiner projects
- ARCOMEM also includes analysis of multimedia information
- See http://www.arcomem.eu and http://www.trend-miner.eu for more details
- More information about GATE at http://gate.ac.uk
- More information about opinion mining see the LREC 2012 Tutorial "Opinion Mining: Exploiting the Sentiment of the Crowd"
- Module 12 of the GATE Training Course (new material after June 2012) https://gate.ac.uk/family/training.html