



EMNLP

CONFERENCE ON EMPIRICAL METHODS
IN NATURAL LANGUAGE PROCESSING

2015 LISBON

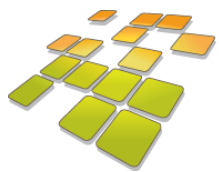
Tutorial

Applications of Social Media Text Analysis

September 18, 2015

Atefeh Farzindar, NLP Technologies Inc.

Diana Inkpen, University of Ottawa



NLP TECHNOLOGIES



uOttawa

Outline

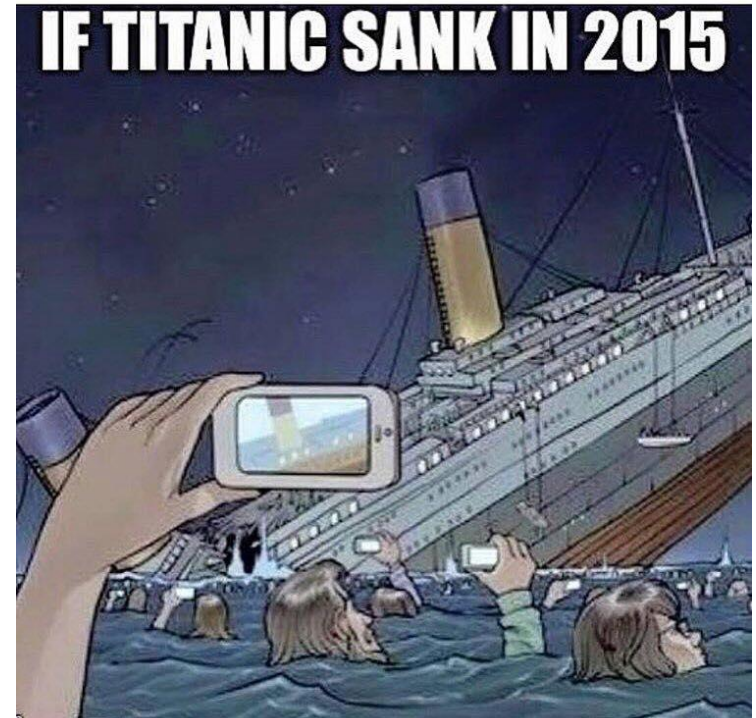
- Introduction
- Linguistic Preprocessing
- Applications & Tasks
 - Health care applications
 - Financial applications
 - Predicting voting intentions
 - Security and defence applications
 - Disaster response applications
 - NLP-based user modelling
 - NLP-based information visualisation
 - Applications for entertainment
 - Social media monitoring
 - Event identification
 - Opinion mining and emotion analysis
 - Geo-location detection
 - Summarization
 - Machine translation
 - Case study



uOttawa

Introduction

- Over the last few years, there has been a growing public and enterprise interest in social media
- Interests: the ability for users to **create and share** content via a variety of social media platforms



- The unprecedented volume and variety of **user-generated content** as well as the **user interaction** network
- New opportunities for understanding social behavior and building socially-aware systems.



What is a Social Network

- A **SOCIAL STRUCTURE** made up of
 - a set of actors (such as individuals or organizations) &
 - the relationships/interactions between these actors
- Social network perspective: to model the structure of a social group
 - how this **structure influences** other variables
 - how **structures change** over time



Social media platforms and their characteristics



Type	Characteristics	Examples
Social networks	A social networking website allows the user to build a web page and connect with a friend or other acquaintance in order to share user-generated content.	MySpace, Facebook, LinkedIn, Meetup, Google Plus+
Blog and Blog Comments	A blog is an online journal where the blogger can create the content and display it in reverse chronological order. Blogs are generally maintained by a person or a community. Blog comments are posts by users attached to blogs or online newspaper posts.	Huffington Post, Business Insider, Engadget and online journals
Microblogging	A microblog is similar to a blog with a limited post in the form of multimedia and other content.	Twitter, Tumblr, Plurk
Forums	An online forum is a place for members to discuss a topic by posting messages.	Online Discussion Community, phpBB, Developer Forums, Raising Children forum
Social Bookmarking	Services that allow users to save, organize and search links to various websites, and to share their bookmarks of web pages.	Delicious, Pinterest, Google Bookmarks
Wikis	These websites allow people to collaborate and add content or edit the information on community-based database.	Wikipedia, Wikitravel, Wikihow
Social News	Social news encourage their community to submit news stories, or to vote on the contents and share them.	Digg, Slashdot, Reddit
Media Sharing	A Web site that enables users to capture videos and picture or upload and share with others.	YouTube, Flickr, Snapchat, Instagram, vine

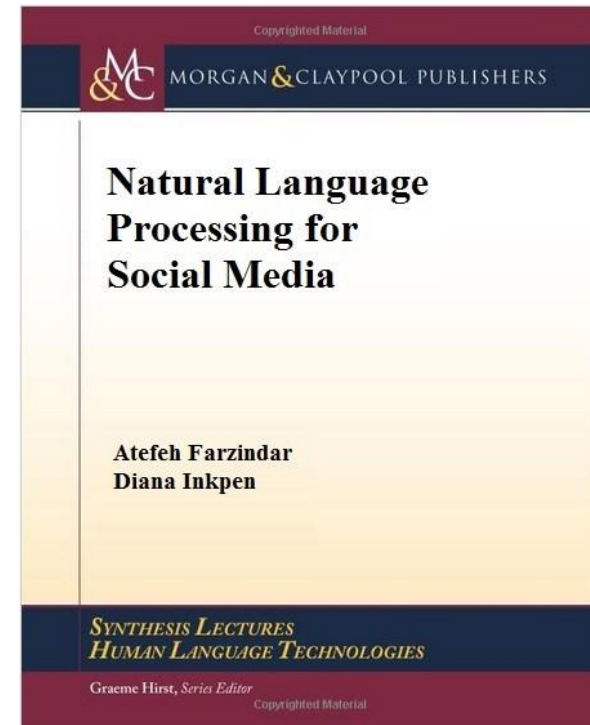
- Example of text analysis:
 - an article in the financial section of a newspaper
- Examples of NLP techniques:
 - Information extraction (e.g. person, location, organization)
 - categorization and clustering
 - automatic summarization
 - semantic search engine
 - statistical machine translation



Semantic analysis in social media



- NLP for social media content
 - Definition of Semantic Analysis in Social Media (SASM): Linguistic processing of social media messages enhanced with semantics and meta-data from the social networks.
- Book: Natural Language Processing for Social Media (upcoming book by Atefeh Farzindar and Diana Inkpen, Morgan & Claypool Publishers, 2015)
- Workshops EACL 2012, NAACL/HLT 2013, EACL 2014



Top news story



CNNMoney @CNNMoney - 59m

JUST IN: @Starbucks is offering free college tuition to its employees. Details: cnnmon.ie/1IE33IU By @ben_rooney



RETWEETS
218

FAVORITES
115



6:49 PM - 6 Apr 2015 · Details



Hide photo



uOttawa

Semantic Analysis in Social Media

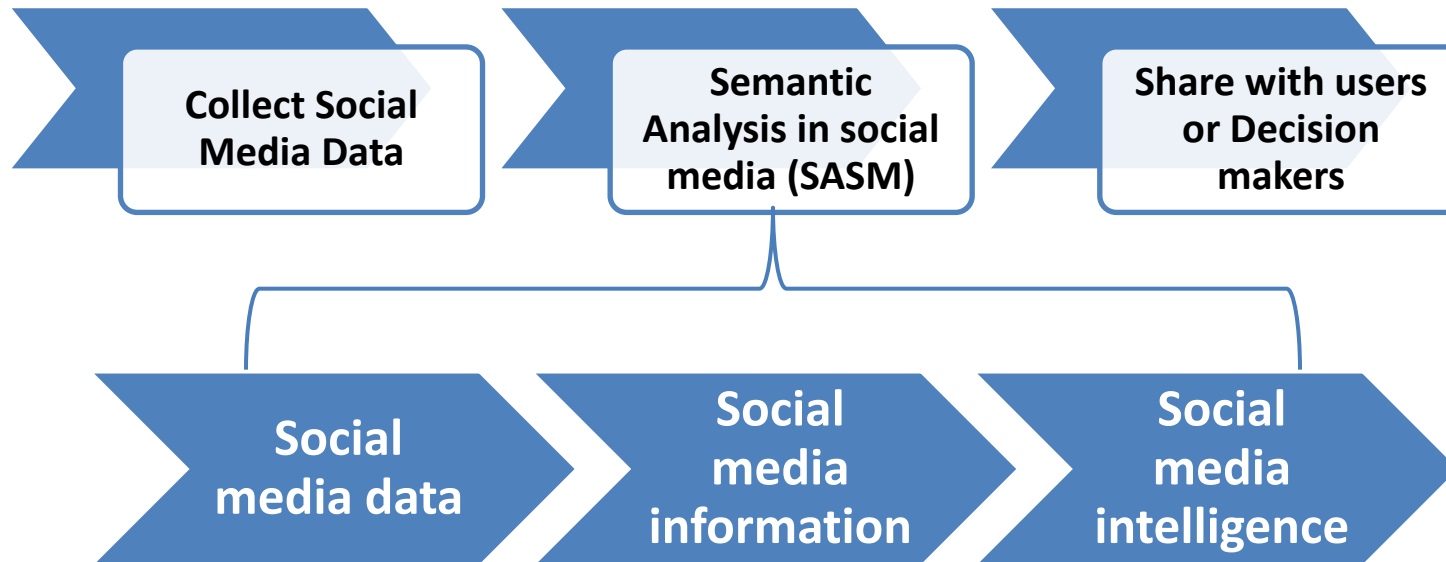


- NLP and Social Media (SM)
- Understand strategic, operational and tactical intelligence uses of social media.
- Development of automated tools and algorithms for monitoring, capturing, and analysing big data collected from social media for **behaviour prediction**.



uOttawa

From Big Data to Intelligence



A framework for semantic analysis in social media, where NLP tools transform the data into intelligence.



Properties of Social Media data

- **Social media data** is the collection of Open source information that can be obtained **publicly**
- Key properties:
 - Social, Real-time, Geo-spatially coded, Emotion, Neologisms, Credibility/rumors
- Non-structured text in many formats
- Written by different people in many languages and styles
- Written in everyday language
- Authors are not professional writers
- Pockets of sources in thousands of places on the www




Linguistic Pre-processing of Social Media Texts (Ch2)

- Natural Language Processing tools
 - Tokenizers
 - Part-of-speech taggers
 - Chunkers and parsers
 - Named entity recognizers
- Adaptation to social media text
 - Text normalization
 - Re-training NLP tools for social media texts
 - Existing NLP tools for English and their adaptation to social media text
- Multi-linguality and adaptation to social media texts
 - Language identification
 - Dialect identification



General NLP toolkits

- Stanford CoreNLP (Java) includes tokenization, POS tagging, named entity recognition, parsing, and co-reference. <http://nlp.stanford.edu/downloads/>
- Open NLP (Java) includes tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. <http://opennlp.apache.org/>
- FreeLing includes tools for English and several other languages: text tokenization, sentence splitting, morphological analysis, phonetic encoding, named entity recognition, POS tagging, chart-based shallow parsing, rule-based dependency parsing, nominal co-reference resolution, etc. <http://nlp.lsi.upc.edu/freeling/>
- NLTK is a suite of text processing libraries in Python for classification, tokenization, stemming, POS tagging, parsing, and semantic reasoning. <http://nltk.org/>
- GATE includes components such as parsers, morphology, POS tagging, and information retrieval and extraction components for various languages. The information extraction system (ANNIE) includes a named entity detector. 
<http://gate.ac.uk/>

NLP tools for social media

- Some components of these toolkits were re-trained for social media texts, such as the Stanford POS tagger by Derczynski et al. (2013), and the OpenNLP chunker by Ritter et al. (2011).
- GATE was fully adapted to social media text. A new module or plugin called TwitIE <https://gate.ac.uk/wiki/twitie.html> is available (Derczynski et al., 2013) for tokenization of Twitter texts, plus POS tagging, name entities recognition, etc.

Two new toolkits were built especially for social media texts:

- 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 TweepoParser mentioned above.
- The UW Twitter NLP Tools (Ritter et al., 2011) contain the POS tagger and the annotated Twitter data. https://github.com/aritter/twitter_nlp



Semantic Analysis of Social Media Texts (Ch3)

- Geo-location detection
- Opinion mining and emotion analysis
- Event and topic detection
- Entity linking and disambiguation
- Summarization in social media
- Machine translation in social media

We will discuss some of these in the context of the applications that follow.



uOttawa

Applications of social media (Ch 4)



- Health care applications
- Financial applications
- Predicting voting intentions
- Security and defence applications
- Disaster response applications
- NLP-based user modelling
- NLP-based information visualisation for SM
- Applications for entertainment
- Media monitoring



uOttawa

Health care 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.



uOttawa

- 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 voting 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.
- Experiments:
 - 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.)

- **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.
- **Combine the three aspects!**



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:
 - Extracted info about disaster response actions (Imran et al., 2013)
 - developed an earthquake detector for Australia and New Zealand (Robinson et al., 2013)
 - detected reported fires (Power et al., 2013)

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



Estimating User's Location

based on the tweets of a user (Li and Inkpen, 2105)



- Two objectives:
 1. Classify a user into a state (out of 49) or one of the 4 regions in the country.
 2. Produce a pair of latitude/longitude.

Method:

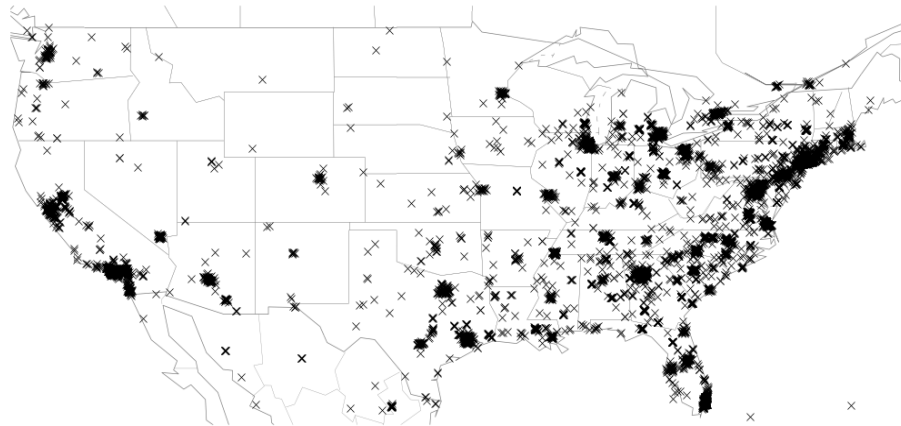
- Deep neural models: feedforward neural nets with 3 hidden layers; the output layer is different for the two objectives.
- Pre-training each hidden layer by treating them as denoising auto-encoders.



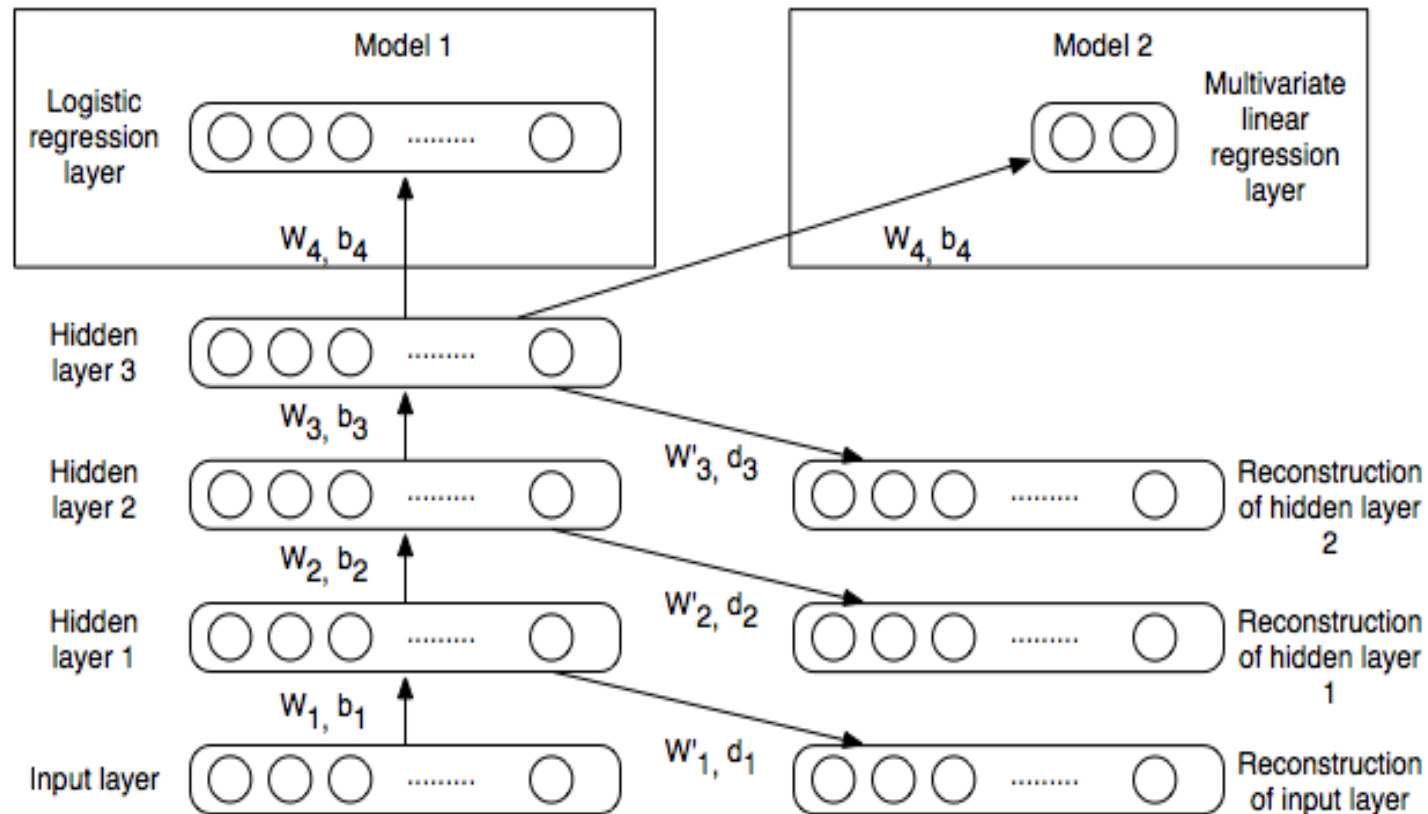
uOttawa

Estimating User Locations Dataset

- A publically-available dataset from **Eisenstein et al. (2010)**
- 380,000 tweets from 9,500 users, with geographic coordinates in the contiguous United States.



Estimating User Location Models



Estimating User Locations Experiments



- Baseline models:
- Obj. 1: SVM, Naive Bayes.
- Obj. 2: Multivariate linear regression (equivalent to our model without hidden layers).
- Pre-training each layer of auto-encoders.
- Training the whole neural net by **back-propagation**.
- Updating the parameters by **gradient descent**.
- **Dataset: 380,000 tweets from 9,500 users:
60% for training, 20% for validation, 20% for testing.**



uOttawa

Estimating User Locations

Results

- Objective 1:

	Model	Classification Accuracy(%)	
		Region (4-way)	State (49-way)
Eisenstein et al. (2010)	Geographical topic model	58	24
	Mixture of unigrams	53	19
	Supervised LDA	39	4
	Text regression	41	4
	K-nearest neighbors	37	2
Our models	SDA-1	61.1	34.8
	Baseline-Naive Bayes	54.8	30.1
	Baseline-SVM	56.4	27.5



Estimating User Locations Results

- Objective 2:

Model	Mean Error Distance(km)
Eisenstein et al. (2011)	845
SDA-2	855.9
Priedhorsky et al. (2014)	870
Roller et al. (2012)	897
Eisenstein et al. (2010)	900
Wing and Baldrige (2011)	967
Baseline-MLR	1268



Estimating User Locations:

More experiments: Roller 2014 dataset

- Contains 38 million tweets from 449,694 users, all from North America.
- 60% for training, 20% for validation and 20% for testing.

Model	Mean error (km)	Median error (km)	Acc. %
Roller et al. (2012)	860	463	34.6
Han et al. (2014)	NA	260	45
Han et al. (2014) using top 3% features (6420)	NA	NA	10
SDA-2	733	377	24.2

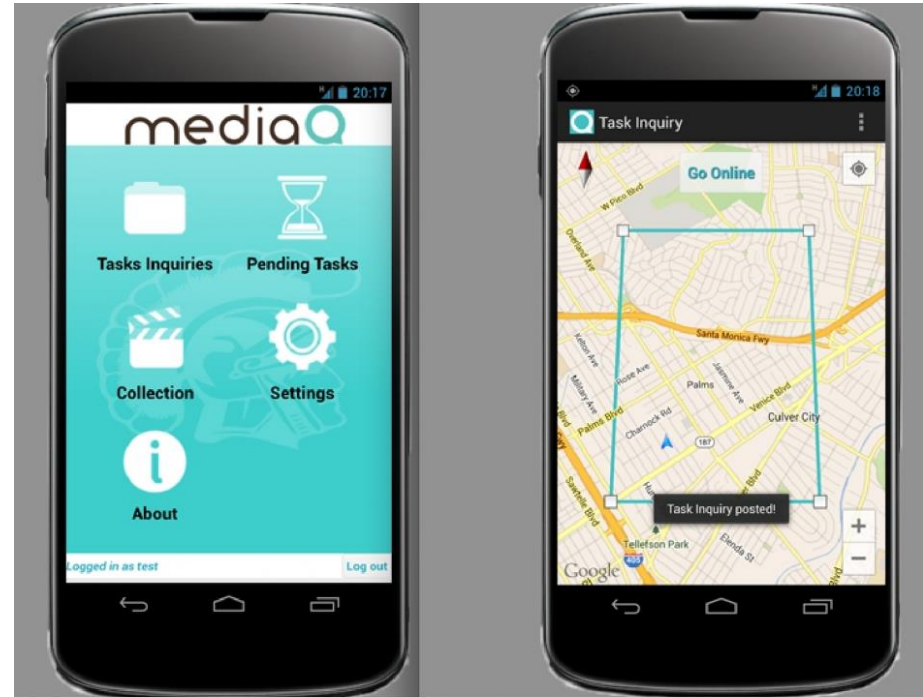


- Information visualisation for social media analysis has become important to explore the potential of social media data.
 - Challenges with the growing availability of big social data.
 - Data analytics and visualization tools are needed to represent the extracted topics and draw relationships between these elements.
- Applying NLP techniques on social media data give us the power of transforming the noisy data to structured information
 - but it is still difficult to discern meaning by extracting information piece by piece.
- The link inference and visualizing content would make the analysed information more apparent and meaningful to present to users and decision makers.



Visualisation – Ex. Geo-location detection

- Geo-location detection from social content, such as blog posts or tweets, is possible thanks to NLP methods.
- But the location itself is irrelevant. But the projection of location on the map, tracking the events on specific timeline and connecting with other name entity and sentiment analysis bring other dimension as “big picture” visualization.
- However, visualization provides an intuitive way to summarize the information in order to make it easier to understand and interaction.



Screenshots of MediaQ Android App

MediaQ is an online media management framework that includes functions to collect, organize, share, search, and trade user-generated mobile images and videos using automatically tagged geo-spatial metadata. (Kim et al. 2014)



uOttawa

- Kotval and Burns (2013) studied the visualization of entities within social media by analysing user needs.
- To understand user needs and preferences, authors developed fourteen social media data visualization concepts and conducted a user evaluation of these concepts by measuring the usefulness of the underlying data relationships and the usability of each data visualization concept.
- However, they reported a divergence and preferences for “big picture” visualization among users.



Applications for entertainment




- Media and entertainment industry has a big challenge facing social media
- Social media are changing users' expectations and their behaviour
- New approaches toward content creation, distribution, operations, technology, and user interaction
 - by online video, social media and mobile media to bring the information to the user and interact with them
- Serious issue: Since advertisers spend less on traditional paid media and require more resources for digital social media and e-marketing.



uOttawa

- Harry Potter's Facebook page recorded nearly 29 million "likes" during the run up to unveiling of the last film in the series.
- In the week leading to the July premiere, Harry Potter's Facebook page gained nearly 100,000 Facebook friends per day.



- **Social gaming:** Microsoft and Sony have made integrated video sharing a focus point of their next generation consoles.
- **Public relations agencies:** The traditional methods of 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.



- Sentiment analysis of Major events such as the Oscars
- Sentiment related to the movies premium
- Sinha et al. (2014) studied the Sentiment Analysis of Wimbledon Tweets by analysing a set of tweets of the Roger Federer and Novak “Nole” 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.



- The trusted measurement of movie and TV programming ranking is one of the important indicator regarding the popularity of a program or a movie in the entertainment industry.

NETFLIX

For example, Netflix, uses the popularity of a movie on Facebook as a proposed feature for consumers.

- Predicting TV audience ratings (Hsieh et al., 2013) , using the 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.



Privacy in social media (Ch5)

- Social media plays an important role in interactive relationships between individuals, organizations and societies.
 - Content shared in social media has an impact on privacy for end-users.
- Published information can also present some difficulties when circumstances change.
- Discussions and concerns about privacy regarding user misunderstandings, the bugs on development of social media platforms allowing unauthorized access, or lack of ethics in marketing.
 - Some privacy research focuses on concerns about data protection by establishing metrics, such as privacy scales, for evaluating those concerns (Wang et al., 2013).
 - However, there is little guidance or research study on how to protect information (ex. privacy in healthcare)



Democracy in social media (Ch5)



- Social media revolutionize liberal democracies and human rights.
 - political community and express democratic values for liberals, progressives, moderates and independents.
- In 2009, the Washington Times^[1] named Iran's Twitter revolution to protest against the rigged election in Iran and required permissions and news coverage. The Iranian election protests was a series of protests following the 2009 Iranian presidential election against the disputed victory of Iranian President Mahmoud Ahmadinejad and in support of opposition candidates Mir-Hossein Mousavi and Mehdi Karroubi. **#iranelection**.
 - After resident Ahmadinejad's victory, in many different cities around the world, Iranian protested against the "stolen election." Although many supporters including Iranian-Americans were not even eligible to vote, changed their Facebook profile picture to "Where is My Vote?"
- In view of recent mobilizations, social media has played a key role in the Arab Spring (2010-2012), which referred to the large-scale conflicts in Middle East and North Africa
- in Canada *Printemps érable* (Maple Spring) 2012, which was a series of protest and widespread student strikes against the provincial government in Quebec. Many researchers study the long-term evolution of US and European political systems via social media networks.

^[1] <http://www.washingtontimes.com/news/2009/jun/16/irans-twitter-revolution/>



Automatic Event Detection, Tracking, and Monitoring in Social Media



- **From industry perspective:** Social media data can dramatically improve business intelligence, for branding and awareness, customer/prospect engagement and improving customer service.
- **From investor perspective:** studying SM to understand situations, perform sentiment analysis and to be alerted against potential threats for Investment.



Monitoring Social Media

- Business Intelligence and Data Analytics
 - Monitoring marketing activities, consumer opinion, influencers, competitors, brands, investment, market prediction, etc.
- News
 - Monitoring, analysis, and aggregation of events from user-generated content.
- Security and Defense
 - Monitoring terrorist activities, crimes, threats, etc.

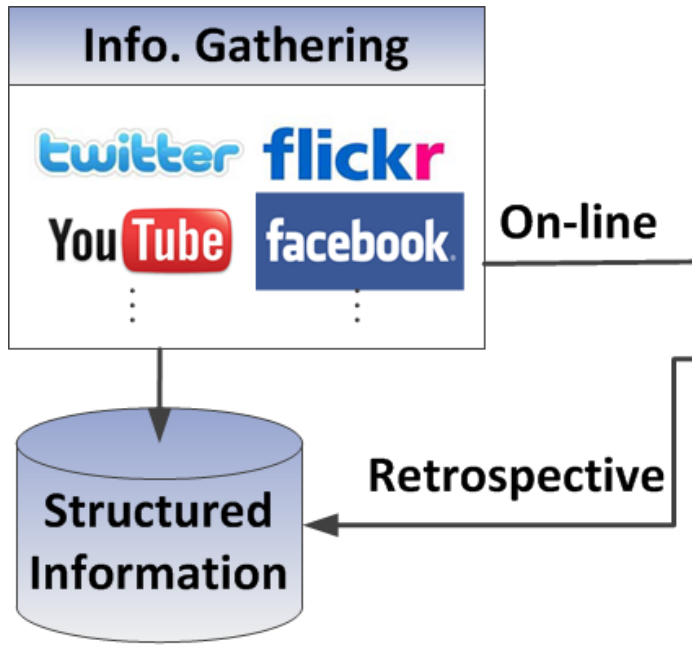


- Events can be defined as situations, actions or occurrences that happen in a certain location at a specific time
- An event is generally characterized by: **5W1H**
 - who? when? where? what? why? how?
- An “Event of interest” is domain dependent
 - Terrorist activities, sport tournaments, conference, trade show, natural disaster, etc.



- **Objective is “Event Detection”** - Discover new or previously unidentified events:
 - 1. Retrospective:** detection of 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





- Can be also categorized into **Unspecified** or **Specified** detection approaches:

1. Unspecified Event Detection

- No available information about the event
 - General military surveillance, first-story in news,...
- Requires **monitoring** the data stream (of multiple sources) for *trending or bursty* features, **grouping** features with identical trend into event, and ultimately **classifying** the event into type or topic



2. Specified Event Detection

- Specific information/features are known about the event:
 - Venue, time, type, description, etc.
 - Marketing activities of a specific competitor/brand
 - Specific military operation (with known target)
- These features can be exploited using Information Retrieval/Extraction techniques:
 - Filtering, query generation/expansion, clustering, information aggregation, etc.

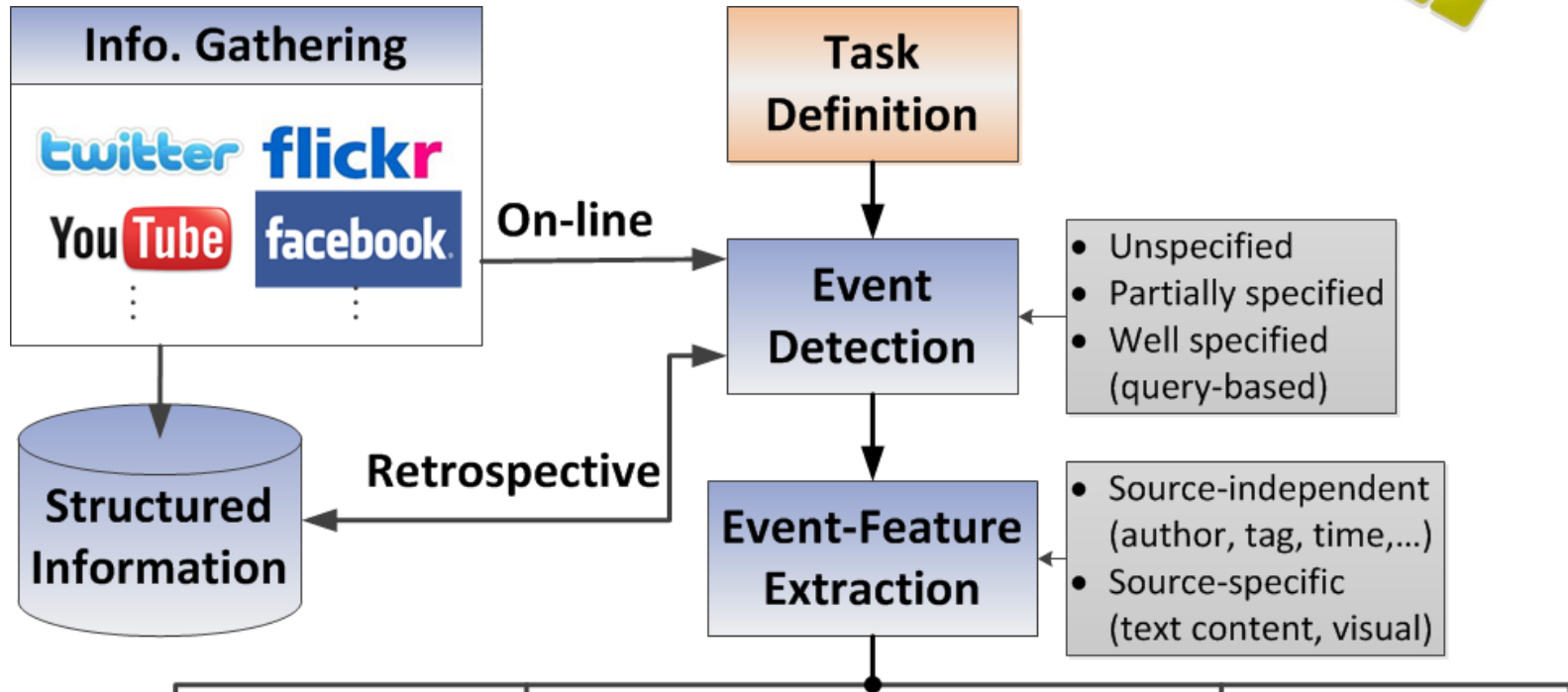


- **Source-independent** information
 - Provided as metadata, e.g.,
 - submitter, location, time, tag, geotag, description
- **Source-specific** information
 - Content information from the SM source, e.g.,
 - text: blogs, microblogs (Twitter), forums, etc.
 - images: flicker, picasa, etc.
 - videos: YouTube, metacafe, etc.



- Feature extraction for specific event/source
 - deducing specific knowledge related to the event of interest
 - From source-independent information
 - metadata
 - From source-specific information
 - text content, visual information from images or videos





- Analysis of consumer **positive/negative** tendency toward a product, brand, company, ..
- Analysis of social roles, popularity, values, etc. in social networks
- Methods: supervised learning (SVM) with various types of features.
- Lexical resources for opinion mining
 - SentiWordNet
 - Emotion lexicons (WordNet Affect, NRC lexicons)



Sentiment Analysis in Twitter

- 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 at SemEval 2014 <http://alt.qcri.org/semeval2014/task9/> and 2015 <http://alt.qcri.org/semeval2015/task10/>



uOttawa

Sentiment Analysis Task A: Classifying Target Word in Twitter Message (pos/neg/neutral)

(Poursepanj, Weissbock and Inkpen 2013)

System	SVM	MNB
Baseline	66.32%	66.32%
BOW features	66.32%	33.23%
BOW+text expansion	73.00%	80.36%

Accuracy results for task A by 10-fold cross-validation on the training data.

Method: ML after expanding target word in training data with definition and synonyms from SentiWordNet for the same semantic orientation.

System	Tweets	SMS
uOttawa system	0.6020	0.5589
Median system	0.7489	0.7283
Best system	0.8893	0.8837

Results for Task A for the submitted runs on test data (Average F-score for pos/neg).



uOttawa

Sentiment Analysis Task B: Classifying Twitter Messages (pos/neg/neutral)

(Pousepanj, Weissbock and Inkpen 2013)



System	SVM	MNB
Baseline	48.50%	48.50%
BOW features	58.75%	59.56%
BOW+ SentiWordNet	69.43%	63.30%
BOW+ extra features	82.42%	73.09%

Accuracy results for task B by 10-fold cross-validation on the training data.

System	Tweets	SMS
uOttawa submitted system	0.4251	0.4051
uOttawa new system	0.8684	0.9140
Median system	0.5150	0.4523
Best system	0.6902	0.6846

Results for Task B for the submitted runs on test data (Average F-score for pos/neg).

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 (!, !!, !!!, ...)



uOttawa

Emotion-annotated blog corpus

(Aman and Szpakowicz, 2008)



- 4090 annotated sentences (173 weblog posts) annotated by two judges with 6 emotions classes or no emotion.
- Inter-annotator agreement ranged from 0.6 to 0.79 for different emotions.
- Examples:

This was the best summer I have ever experienced. (*happiness*)

I don't feel like I ever have that kind of privacy where I can talk to God and cry and figure things out.

(*sadness*)



uOttawa

Hierarchical emotion classification for blog data

(Ghazi, Inkpen and Szpakowicz, 2010)



Two-levels:

non-emo

emo

happiness

sadness

fear

surprise

disgust

anger

Three-levels:

non-emo

emo

positive (happiness)

negative

sadness

fear

surprise

disgust

anger



uOttawa

Comparison of the three approaches for emotion classification (Overall accuracy)



- Flat classification **61.67%** (baseline 38%)

- Two-levels

non-emo happiness sadness fear surprise disgust anger

620 + 509 + 86 + 56 + 37 + 60 + 60

= 1428 / 2090 = **68.32%**

- Three-levels

non-emo happiness sadness fear surprise disgust anger

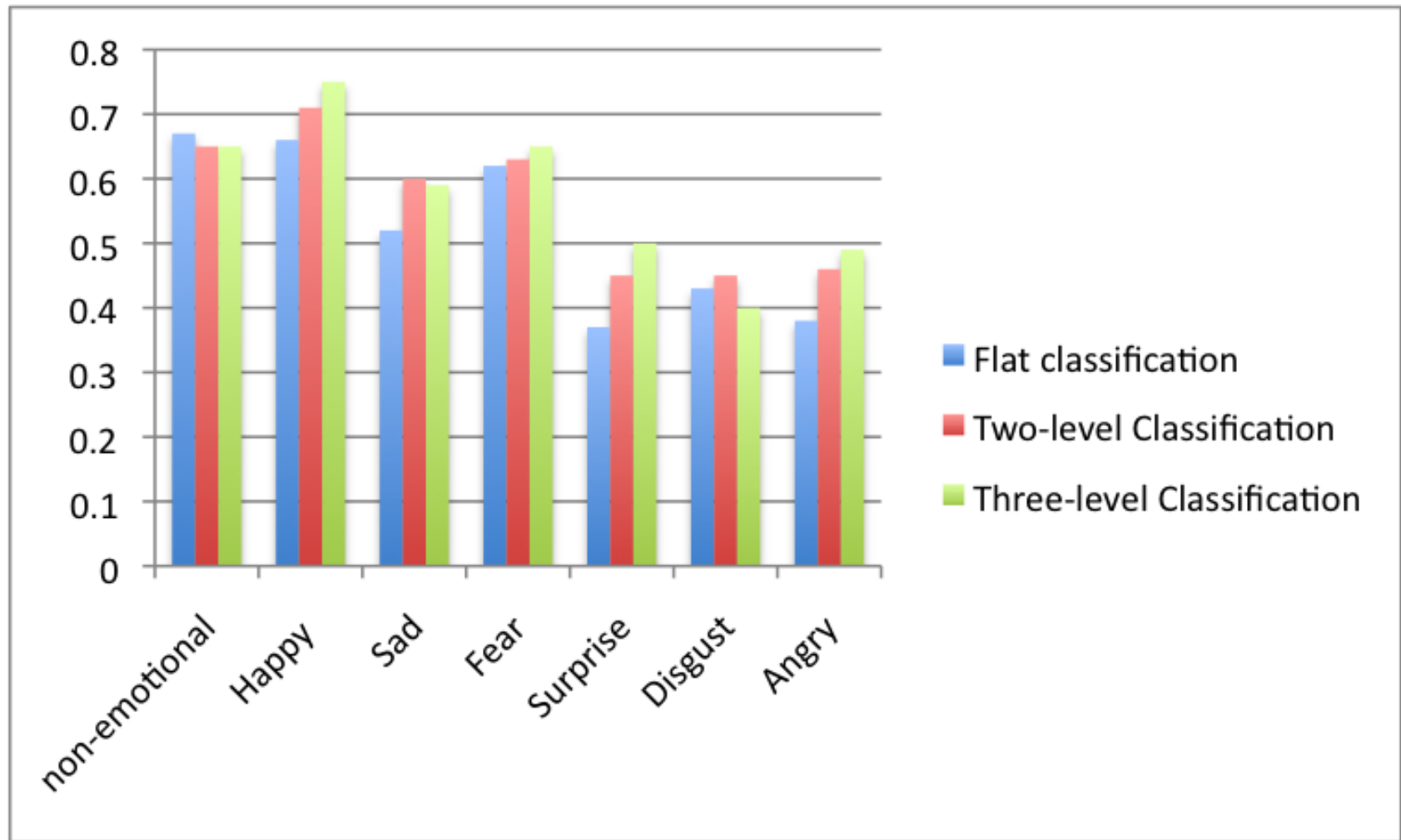
620 + 351 + 95 + 60 + 43 + 65 + 127

= 1361 / 2090 = **65.11%**



uOttawa

Comparison of the three approaches for emotion classification in blogs



- **Event identification or verification:** confirm specific properties of (or related to) the event
 - description, time, authorship, associated events, etc.
- Match and analyze information related to the event from different SM sources
 - e.g., match Twitter messages to flicker photos and Youtube videos using time, tags, location, etc. to verify some aspect of the event



- Geographic Localization of the event from multiple sources
 - Geotags could be readily available in metadata
 - Geolocation could be deduced from IP addresses of mobile devices
 - Location mentioned in the user-generated text content (geonames)
- Various precision levels: address, city, state/province, country (Inkpen et al., 2015).
- Combining the above improves accuracy



Location Entities Extraction

What are the difficulties?

- Two types of ambiguities (Amitay et al., 2004):
 1. Geo/non-geo ambiguity
 - E.g., Roberta, Georgia
 - Jackson, Mississippi
 - None, Italy
 2. Geo/geo ambiguity
 - E.g., London, UK v.s. London, ON v.s. London, TX
 - Ottawa, ON v.s. Ottawa, OH
 - American and Mexican cities named “China”
- Conclusion: Trivial keywords matching won't work.



Location Entities Extraction

Task/Data



- Geo/non-geo ambiguities
 - > Named Entity Recognition (NER)
(Types of locations)
 - > Sequential classification problem
- Conditional Random Fields (CRF): A sequential classifier suitable for NER tasks
- However, annotated data are required.
- A corpus of tweets with location entities annotated did not exist.
We created one! (Inkpen et al., 2015)
Available at <https://github.com/rex911/locdet>



uOttawa

Location Entities Extraction

Collecting tweets

- We used the Twitter API:
 - * query={iPhone, Android, Blackberry, Windows Phone, HTC, Samsung}
 - * language=English

- Example:

```
https://api.twitter.com/1.1/search/tweets.js  
on?q=iphone&lang=en&since_id=240126199840510  
00&count=100
```

- Results: over 20 million tweets from June 2013 to November 2013.



Location Entities Extraction

Manual annotation



- 1000 tweets for each search query
- Annotation schema: a simplified version of *spatialML* (Mani et al., 2008)
- We consider three types of locations:
 - » Countries (annotated as “country”)
 - » States and provinces in the U.S. and Canada (annotated as “SP”)
 - » Cities (annotated as “city”)



uOttawa

Location Entities Extraction

Manual annotation steps

1. Gazetteer matching.

- Gazetteer: a list of proper names such as people, organizations, and locations.
- We obtained a gazetteer of locations from GeoNames.
- Matching alternative names (**It's Twitter!**):
 - Acronym, airport code: ATL, LAX, PDX, NYC
 - Nicknames: Philly
 - Non-English names: Torontas, Losandzelosa
- **Matched by GATE's gazetteer matching module (Cunningham, 2002).**



Location Entities Extraction

Manual annotation steps

2. Manual filtering

- Ambiguities:

*Georgia —> American state, European country,
person name*

- Manual annotation: two annotators.
- Inter-annotator agreement: 88%
- Agreement between annotators and initial gazetteer matches: 56% and 47%, respectively.



Location Entities Extraction

Manual annotation examples

- Example:

Mon Jun 24 23:52:31 +0000 2013

<location locType='city'>Seguin </location> <location locType='SP'>Tx</location>

RT @himawari0127i: #RETWEET#TEAMFAIRYROSE #TMW #TFBJP #500aday
#ANDROID #JP #FF #Yes #No #RT #ipadgames #TAF #NEW #TRU #TLA #THF 51

Wed Sep 11 09:09:21 +0000 2013

Worldwide

BlackBerry lays off dozens of **<location locType='country'>US </location>** sales sta
: WSJ - Reuters Canada: MobileSyrup.comBlackBerry lays off dozens ...

<http://t.co/GZcO3H3wro>



Location Entities Extraction

Location entity detection



- Features:
 - Bag-of-Words (BOW)
 - Part-of-Speech (POS)
 - Left/right window (WIN)
 - Gazetteer (GAZ)
- **Tokenization and POS tags obtained by the Twitter NLP tool (Owoputi et al., 2013).**



uOttawa

Location Entities Extraction

An example of feature extraction

If this Big 4 plan works in Miami, they gotta sign Darko to a minimum deal for old time's sake

BOW	(1,0,0,...,0)	The first '1' indicates the activation of the token 'Miami'
POS	(1,0,0,...,0)	The first '1' indicates the activation of the POS tag 'proper noun'
WIN: left	(0,1,0,...,0)	The second '1' indicates the activation of the token 'in' on the left
WIN: right	(0,0,1,...,0)	The third '1' indicates the activation of the token ',' on the right
GAZ	(1)	The value '1' indicates that the token 'Miami' is in the gazetteer

Concatenating
BOW+POS

$$(1,0,0,...,0) \parallel (1,0,0,...,0)$$
$$=(1,0,0,...,0,1,0,0,...,0)$$


Location Entities Extraction Experiments



- Features: BOW + combinations of other feature sets
- Implemented with MinorThird package (Cohen, 2004), which provides a CRF module (Sarawagi and Cohen, 2004)
- 10-fold cross validation
- Metrics: **Precision, Recall** and **F-measure** at both *token* and *span* level:
 - » Example: Los Angles
Prediction: Angles



uOttawa

Location Entities Extraction Results

<i>Features</i>	<i>Country</i>		<i>SP</i>		<i>City</i>	
	<i>Token F</i>	<i>Span F</i>	<i>Token F</i>	<i>Span F</i>	<i>Token F</i>	<i>Span F</i>
BOW	0.88	0.87	0.84	0.84	0.71	0.68
BOW+POS	0.88	0.87	0.84	0.85	0.71	0.66
BOW+GAZ	0.88	0.87	0.84	0.85	0.80	0.78
BOW+WIN	0.88	0.88	0.84	0.85	0.78	0.76
BOW+POS+GAZ	0.88	0.87	0.85	0.86	0.81	0.78
BOW+WIN+GAZ	0.90	0.89	0.84	0.85	0.82	0.81
BOW+POS+WIN	0.88	0.88	0.85	0.85	0.79	0.80
BOW+POS+WIN+GAZ	0.90	0.90	0.85	0.85	0.83	0.81

Location Entities Extraction

Results and Discussion

1. Token level F and span level F are similar: most location names contain **one** word.
2. Using all features produces the best results (except for one case).
3. Using merely BOW features always produces the worst results.
4. For the “city” label, BOW features and GAZ features result in statistically significant improvements, POS features never do.
5. Differences are most obvious for the “city” label: much larger search space.
6. No pair of feature combinations yields statistically significant difference for the “SP” label.



Location Entities Extraction

Location Disambiguation

- Rule-based heuristics
 1. Retrieving candidates
 2. Type filtering
 3. Checking adjacent locations
example: Ottawa, **ON**, **Canada**
 4. Checking global context
 5. Default location: most populated place



Location Entities Extraction Experiments and Results

- Further annotation of true locations in 300 tweets (from the previous 6000 tweets)
- Step 3 and step 4 can be deactivated

- Results:

<i>Deactivated steps</i>	<i>Accuracy</i>
<i>None</i>	<i>95.5%</i>
<i>Adjacent locations</i>	<i>93.7%</i>
<i>Global context</i>	<i>98.2%</i>
<i>Adjacent locations + Global context</i>	<i>96.4%</i>



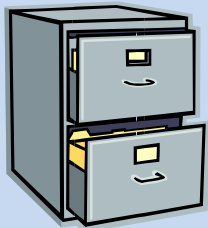
- There are many approaches for translation in which each approach needs some resources.
 - Statistical Machine Translation needs parallel corpus
 - Rule base approach needs extraction of rules



Translation Memories



Translated Archives



Dictionnaires/glossaries



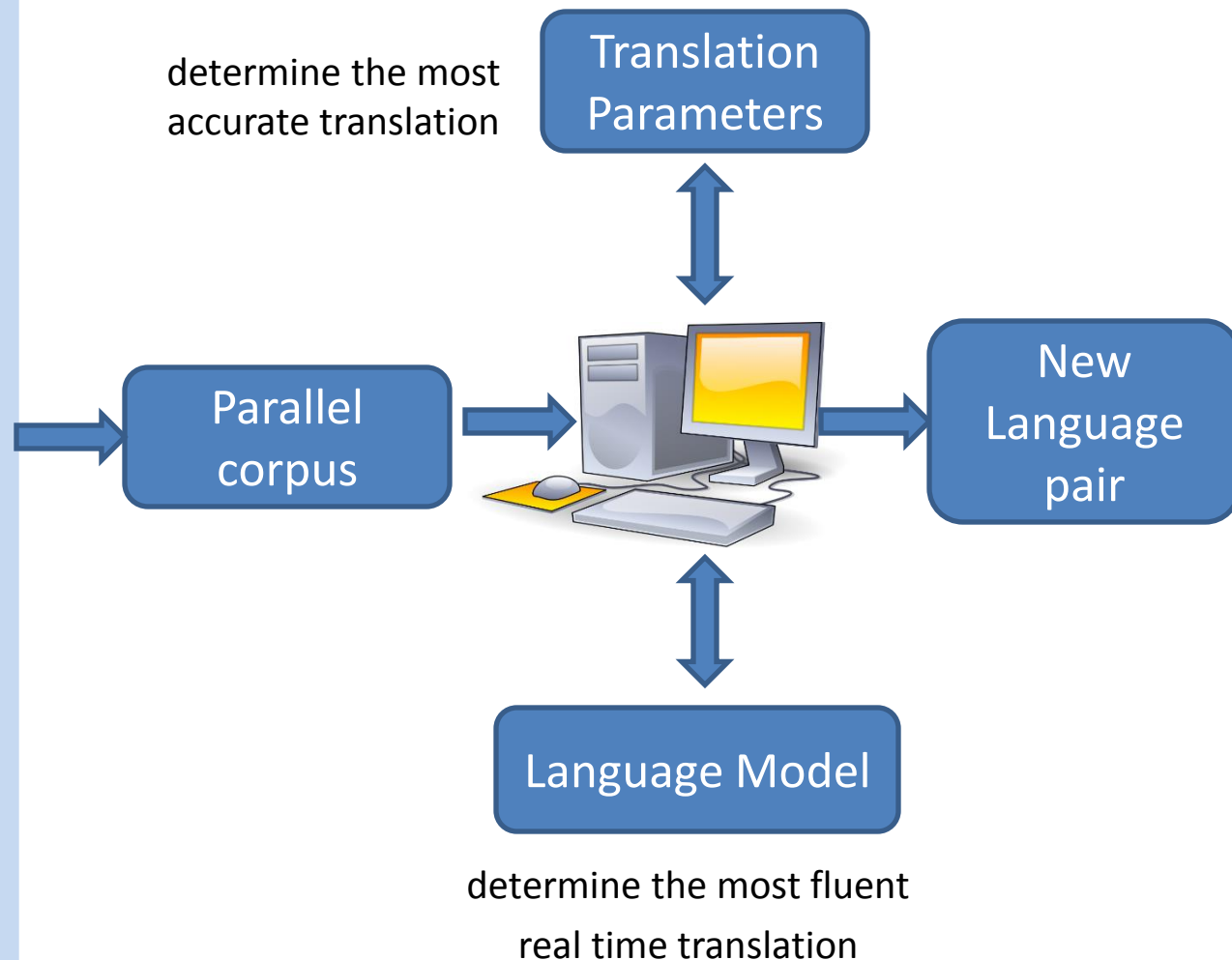
Internet



Human Translations



Statistical Machine Translation



- **Arabic** is the **official language** of **22 countries**.
- **Arabic** is spoken by **more than 500 M** people.
- Arabic Forms:
 - Modern Standard Arabic (MSA) : written form of Arabic, formal communication, language of education, found in news paper, books, etc
 - Arabic Dialects (ADs): spoken language, daily informal communication, more than 22 dialects, more than one dialects live side by side with MSA in one country.

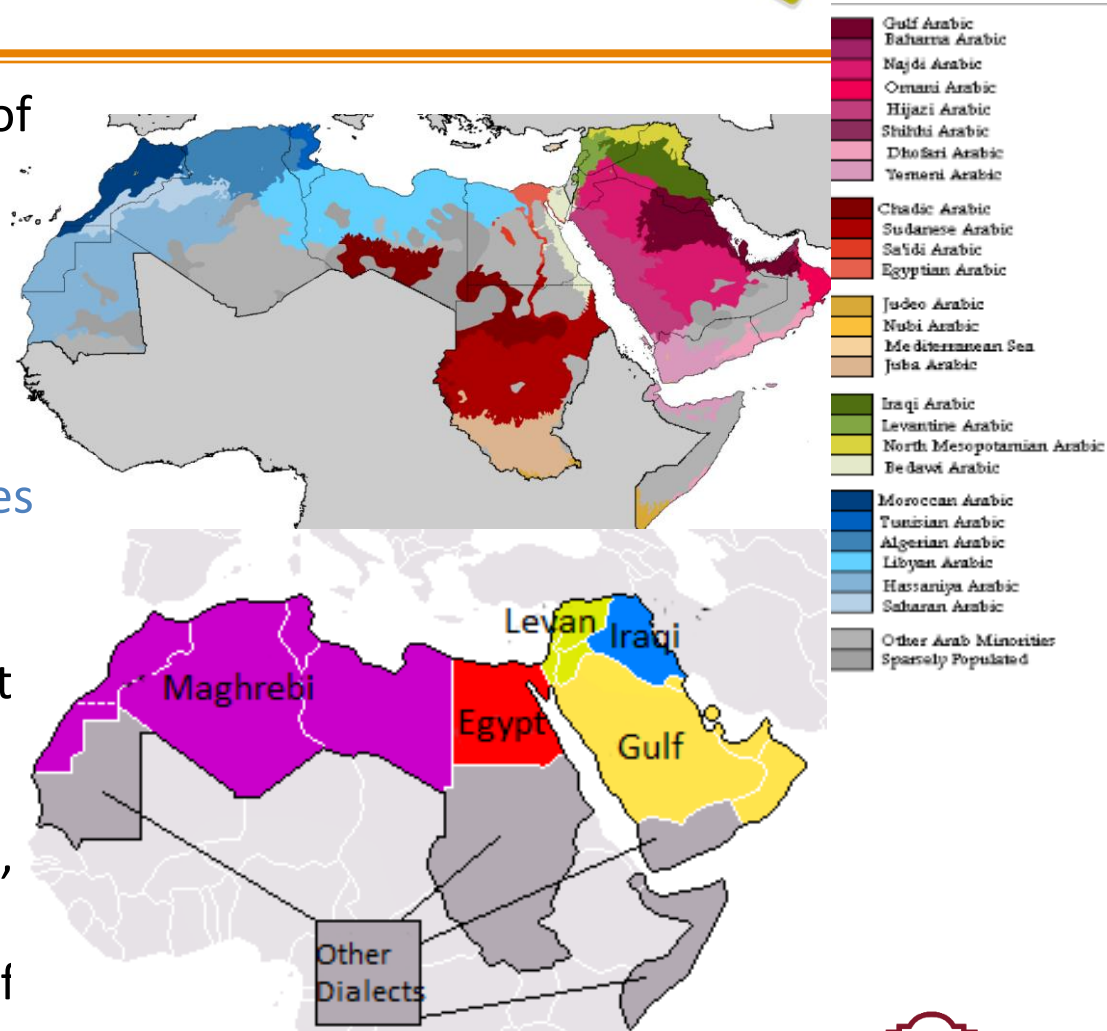


- Most NLP tools deal with Modern Standard Arabic (MSA).
- Social media(SM) bring new challenge to NLP technologies.
- The language used SM is a combined form of Formal (MSA) and Informal (Dialects).
- Tools for MSA are not suitable for Dialect.
 - Ref: Sadat, F., F. Kazemi, and A. Farzindar, "Automatic identification of Arabic dialects in social media."



ADs and MSA

- **Shared** considerable number of features:
 - semantic
 - syntactic
 - morphological
 - lexical
- But there are **many differences** among them in all the above features.
- Recently, attention was given to Arabic Dialect existed on the web in **social networks sites** such as chats, microblog, Blog, forums.
- Which is the target research of **sentiment analysis** and **opinion extraction**.



Creating tools for analyzing Arabic social media.

- **Identification of languages in the social media:** specifically the Arabic language among the languages with Arabic script (Persian, Urdu, etc.)
- **Identification of different dialects of Arabic:** five categories (Egyptian, Maghrebi, Mashreki, Levantine, etc.)
- **Mapping from any dialect to MSA:** establishing different rules for this mapping
- **Machine Translation** from Arabic dialects to English and French.



Translating Government Agencies' Tweet Feeds

- Timely warnings and emergency notifications to the public are important tasks of governments in public safety.
 - Ex. Environment Canada announced that weather warnings cannot be tweeted because official bilingualism in Canada has proved a barrier to weather warning tweets.
- Official Languages Act of Canada: official publications made by the Canadian government must be issued simultaneously in English and French.
- Includes the material published on Twitter by more than 100 government agencies and bodies and politicians, including the Prime Minister.



Gouvernement du Canada



This screenshot shows the Twitter profile of Santé Canada. The header includes the profile picture (a Canadian flag with "SANTÉ CANADA" text), the name "Santé Canada", and the handle "@SanteCanada". The bio states: "Santé Canada (Gouvernement du Canada): une source crédible de renseignements pertinents et à jour. Canada http://www.sante.gc.ca". It shows 2,117 tweets, 60 following, and 5,929 followers. The left sidebar contains navigation links: "Tweet to Santé Canada", "Channel:", "Tweets", "Following", "Followers", "Favorites", "Lists", and "Similar to Santé Canada". The main feed displays three tweets from Santé Canada, all dated 22 Oct, regarding face paint, electrical connectors, and bicycle recalls.

This screenshot shows the Twitter profile of Health Canada. The header includes the profile picture (a Canadian flag with "HEALTH CANADA" text), the name "Health Canada", and the handle "@HealthCanada". The bio states: "Health Canada (Government of Canada): Providing current, trusted and accurate information. Canada http://www.health.gc.ca". It shows 2,126 tweets, 137 following, and 27,681 followers. The left sidebar contains navigation links: "Tweet to Health Canada", "Channel:", "Tweets", "Following", "Followers", "Favorites", "Lists", and "Similar to Health Canada". The main feed displays three tweets from Health Canada, all dated 22 Oct, regarding face paint, electrical connectors, and bicycle recalls.

En : <https://twitter.com/HealthCanada>

Fr : <https://twitter.com/SanteCanada>

Gotti, F., P. Langlais, and A. Farzindar, "Translating Government Agencies' Tweet Feeds: Specificities, Problems and (a few) Solutions, 2013

© 2015 Atefeh Farzindar & Diana Inkpen



uOttawa

Bilingual Twitter feeds

Gotti et al. [2013]



Health Canada @HealthCanada

21 Nov

Did you know it's best to test for #radon in the fall/winter? bit.ly /QtHuyt #health #safety



Santé Canada @SanteCanada

21 Nov

L'automne/l'hiver est le meilleur moment pour tester le taux de radon. bit.ly/PJnfvh #santé #sécurité

Tokenize/normalize/serialize URLs

- Improved version of Twokenize (O'Connor et al. 2010)

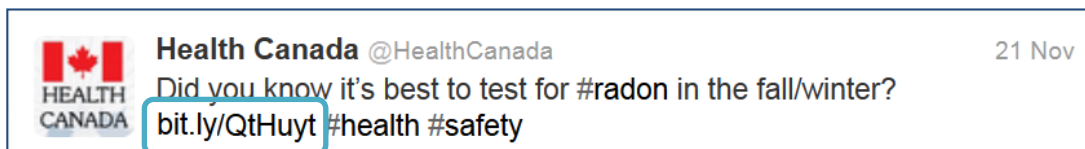
```
did you know it 's best to test for #radon in the fall /  
winter ? urlurlurlurlurlurl #health #safety
```



uOttawa

In-domain corpus: URL corpus

- Idea: mining the parallel text available in URLs mentioned in parallel tweets



*aligned
to*



uOttawa


Translation of #Hashtags

- Gotti, F., P. Langlais, and A. Farzindar, "Hashtag Occurrences, Layout and Translation: A Corpus-driven Analysis of Tweets Published by the Canadian Government" 2014
- The hashtags appear in either a tweet's prologue, announcing its topic, or in the tweet's text instead of traditional words, or in an epilogue.
- Out of Vocabulary (OOV) hashtag
- Which hashtag should be translate?
 - #siliconvalley-> #siliconvalley
 - gold valley (En)-> Vallee D'Or (fr)



Hashtags layout

epilogues
and
prologues

 **Health Canada** @HealthCanada · Aug 5
**#Recall: Disney Store recalls various
themed gadget pencil cases** ow.ly/QxxYU
#Health #Safety

RETWEETS
13

FAVORITE
1



4:11 PM - 5 Aug 2015 · Details



Inline hashtags



Health Canada @HealthCanada · Jul 31
Health Canada provides stable funding to **#FirstNations** and **#Inuit**
#health programs: ow.ly/Qle5A

RETWEETS
8

FAVORITES
5



1:16 PM - 31 Jul 2015 · Details



uOttawa

- Prove compact summaries of information for saving reading time.
- Reduce amount of data and number of features (hence time and memory complexity) for other tasks.
 - classification, clustering, ...



- linking news summarization to online and dynamic settings.
 - uses Web documents such as blogs, reviews, and news articles, to identify new information on a topic.
- Update summarization task TREC 2008: consists in building a short (100-word) summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles.
- TREC 2013 defined Temporal Summarization. Unexpected news events such as natural disasters represent a unique information access problem where the performance of traditional approaches deteriorates.



- Social media text is by definition social in nature
- Information about the relations between textual entries and interactions between users can be useful for summarization,
 - information about the structure and activity of the network itself
 - additional sources of information to aid in summarization of the content of the network
 - and as a target of summarization in and of itself.
 - Liu et al. [2012a] leveraged social network features to adapt a graph-based summarization approach to the task of summarizing tweets.

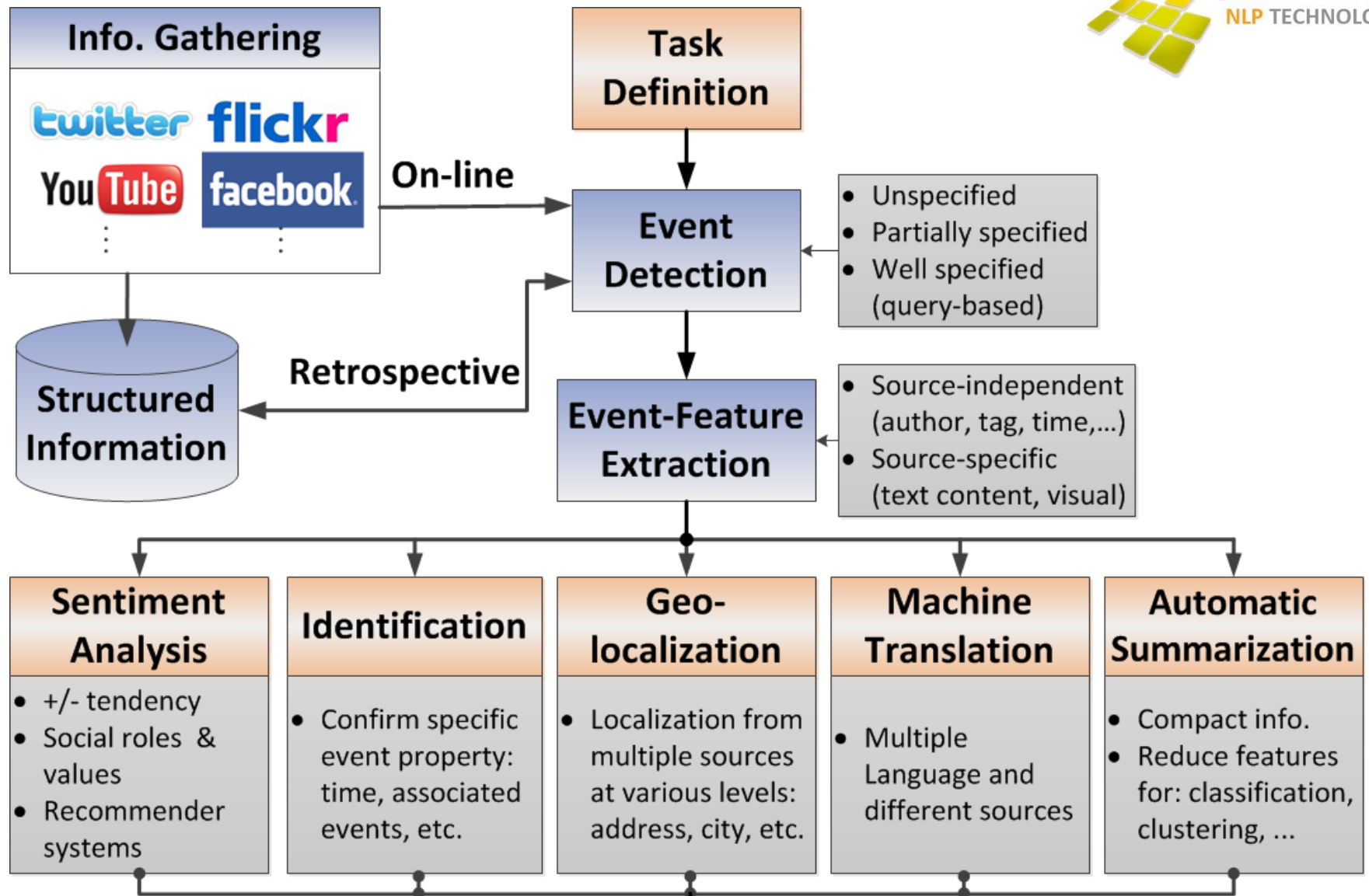


- Can help better understand opinion mining and sentiment analysis
- May target a general assessment of sentiment polarity regarding a particular product or service,
 - which can be invaluable for marketing or reputation management
 - e.g., "Do customers feel positive or negative regarding a particular brand or product?").
- May also target specific query-based information,
 - such as "Which particular features do customers like best about a given product?"



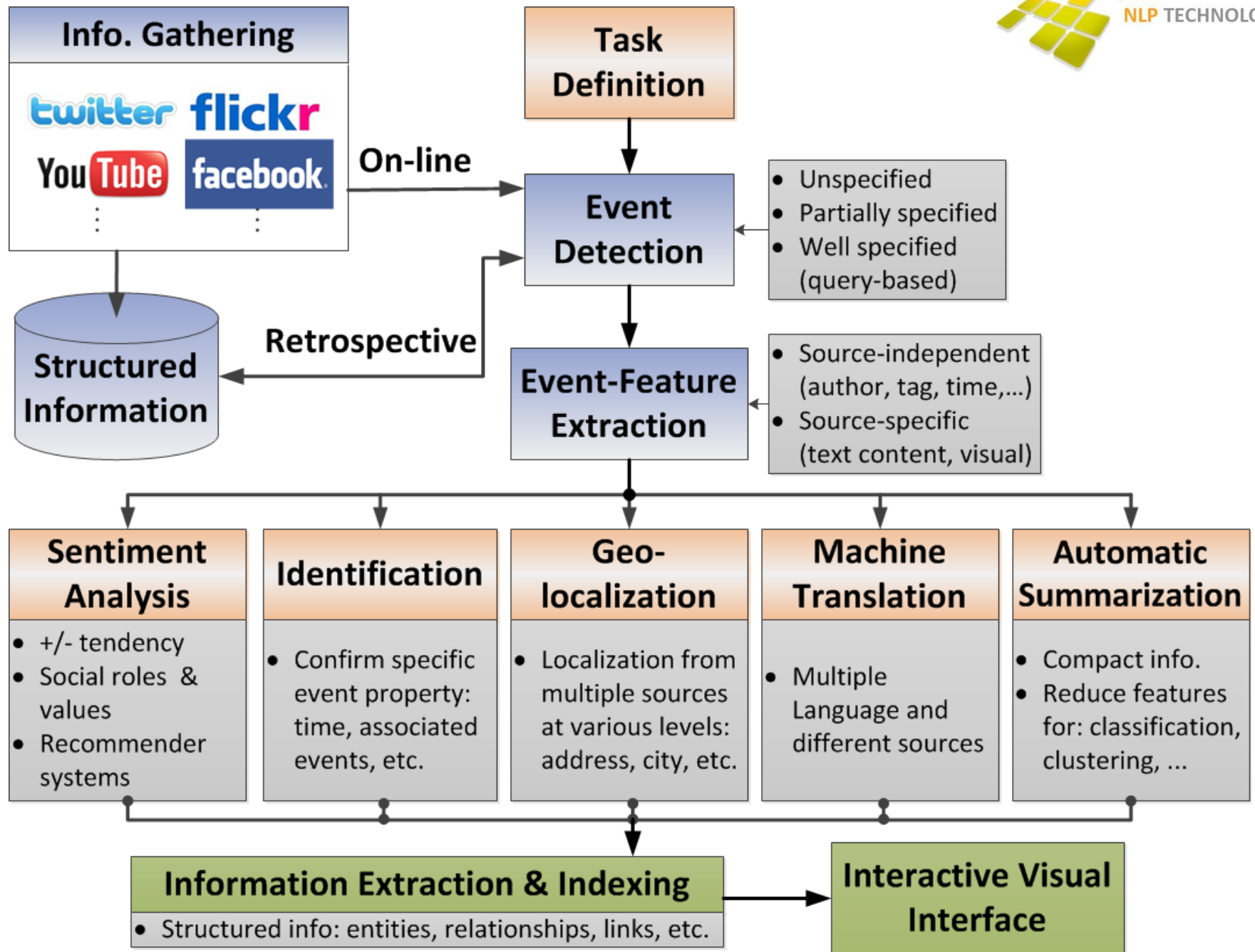
- Event summarization seeks to extract social media text representative of some real-world event.
- Here, an event can be broadly defined as any occurrence unfolding over some spatial and temporal scope (Farzindar and Khreich, 2013).





- Provide structured information about the event and its associated events
 - entities, relationships, links, etc.
- Allow for efficient and accurate event retrieval
 - retrospective structural search from microblog data (Metzler, Cai and Hovy 2012)
- Aggregate information from multiple sources into a unified and interactive visual platform





TRANSLI™-SM

NLP-based social media analytics and





TRANSLI™ for Social Media

- Social Media Analytics and Monitoring
- A visual analytics system designed to help users use the power of social intelligence for news and other events.

TRANSLI™ - Login Page

[Home](#)[Contact us](#)[Log in](#)[English \(US\)](#) ▼

Username :

Bill

Password :

.....

OR

 Sign in with Twitter

 Log in with Facebook

Log in

[Create an account](#)

[Forgot your password?](#)

© Copyright NLP Technologies.

[Corporate site](#) [Follow us](#) [White paper](#)

TRANSLI™ - Create Event

[Home](#)[Browse events](#)[Create an event](#)[Contact us](#)

Welcome, [Bill Gates!](#)
[Log out](#)

[English \(US\)](#) ▼

Create an event

Event name

Ex: Sochi Winter Olympics

Event description

Search keywords

Enter one keyword per line.
Use quotation marks to define an exact expression.

Languages

Select each language you want collected tweets to be in.

- ☐ English
☐ French
☐ Spanish

Start date

2014-09-29

When the data collection has to start

End date

2014-10-06

When the data collection has to end.

Create event

TRANSLI™ - Browse Events

[Home](#)[Browse events](#)[Create an event](#)[Contact us](#)

Welcome, [Bill Gates!](#)

[Log out](#)

English (US) ▼

MS elects a new CEO

Coverage of the election of Satya Nadella as new CEO of Microsoft.

Tweets containing "Satya Nadella" Microsoft CEO

Messages from 2014-02-01 to 2014-03-03 16:14:17
Event closed.

[Delete](#)

Crimea invasion

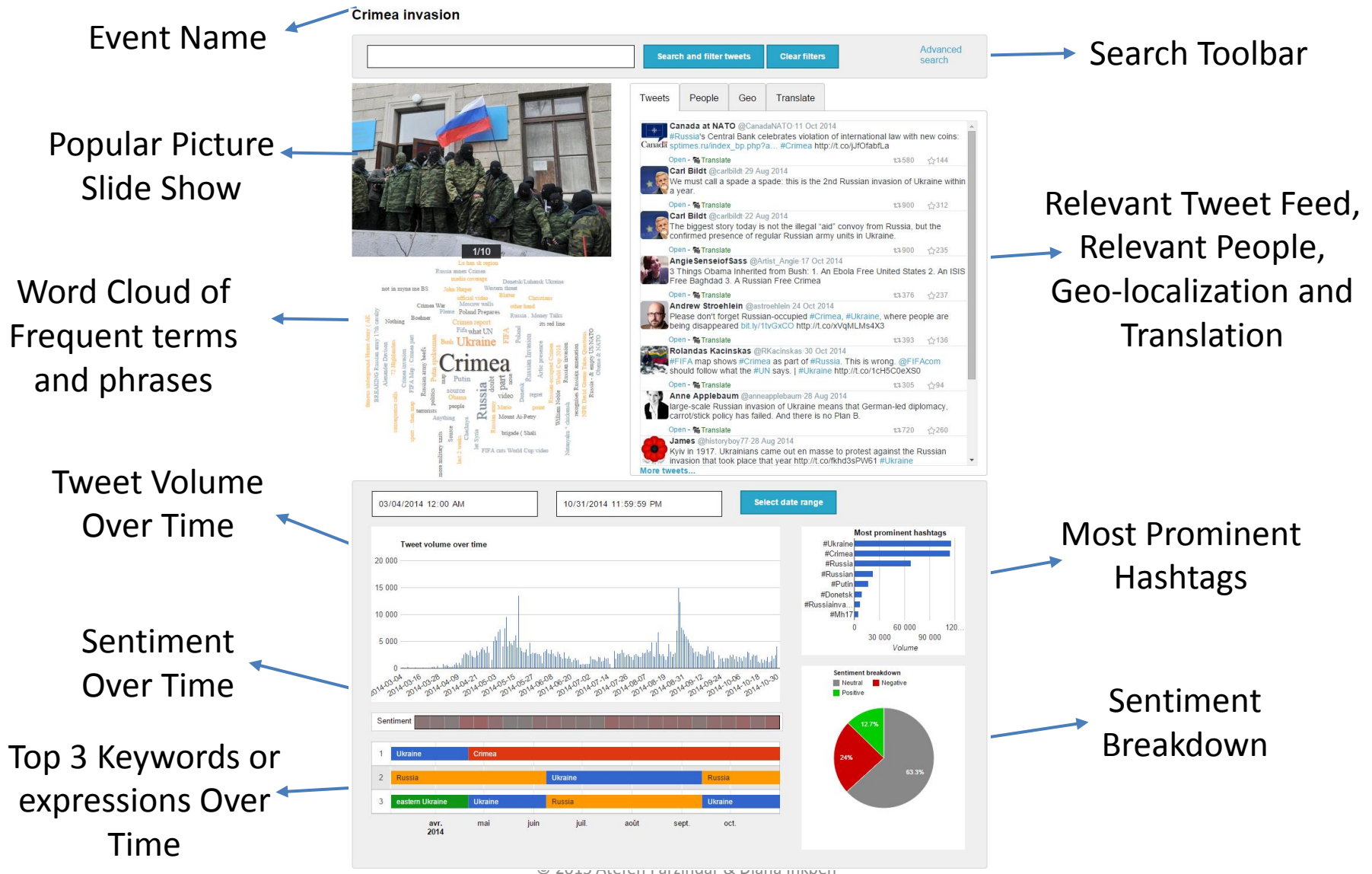
Invasion of Crimea by Russia

Tweets containing Crimea "Russian army" "Russian invasion" "invasion of Ukraine"

Messages from 2014-03-04 to 2014-08-12 09:04:59
Aggregated in 64 minutes intervals.

[Delete](#)

TRANSLI™ for Social Media



TRANSLI™ - Event Dashboard - Picture Pop-up



masakhwan @masakhwan · 09 Jun 2014
hihi lucu... RT @DarthPutinKGB: She said 'Who's a naughty little despot who stole #Crimea and re-wrote history?' <http://t.co/MYGn365OhD>
Open - Translate 10 0

Lou/ Scotto @lxuriousbastard · 14 Jun 2014
I can't, wait, what?! " @DarthPutinKGB: 'Who's a naughty little despot who stole #Crimea and re-wrote history?' <http://t.co/yPrgljZ3Ma>"
Open - Translate 10 0


The Power Of One. @france7776 · 09 Jun 2014
She said 'Who's a naughty little despot who stole #Crimea and re-wrote history?' <http://t.co/rwilWGg520> v @DarthPutinKGB
Open - Translate 14 3


Richard St Ruth @RSR108 · 09 Jun 2014
:-) RT @france7776 She said 'Who's a naughty little despot who stole #Crimea and re-wrote history?' <http://t.co/PzRIEenWZn> v @DarthPutinKGB


TRANSLI™ - Event Dashboard - People





Tweets People Geo Translate


**euromaidan** @euromaidantwit
Earn money from twitter <http://t.co/bQSEupJeuN>


**Видео Крым** @VideoCrimea
Свежее видео из Крыма


**Brian Brown, Ph.D.** @BrianBrownNet

**Blue Marble Times** @BlueMarbleTimes
The Momentum of Movements.

**Adin of Crimea** @RealCrimea
Cultural Communication and Travel Consultant, English Teacher, Freelance Writer, Wanderer, World Citizen.

**RJ** @JamesRon1980
Please do not consider responding to a tweet as spam - Retweet does not mean I agree - It's twitter - let's communicate - *** Peace for Ukraine ***

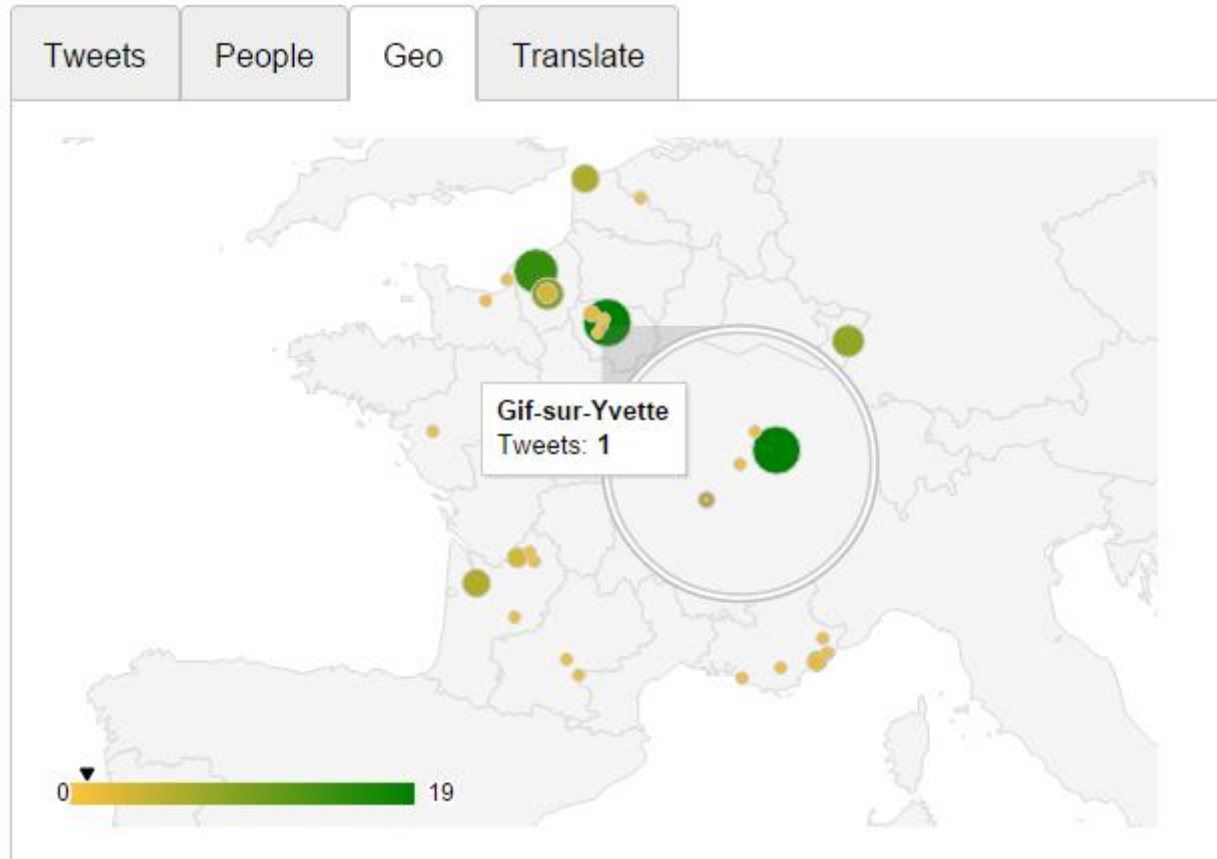
**Euromaidan** @ukrainetwit24
Earn money from twitter <http://t.co/bQSEupJeuN>

**GO! Russia** @GORussiaNews
Russia on Watchinga brings the latest national & state news & video to you real-time. Or keep with what's hot in #russia directly via <http://t.co/FFwMofYHDt>

TRANSLI™ - Event Dashboard - Geo



TRANSLI™ - Event Dashboard - Geo - Local



TRANSLI™ - Event Dashboard - Translation



Tweets

People

Geo

Translate

1914: WW1- the Russian army began to advance into E. Prussia to attack Germany and divert their resources relieving pressure on ally France

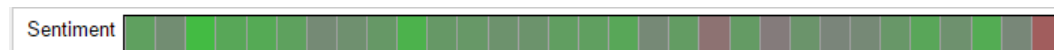
Translate

English to French ▼

1914: ww1-l'armée russe a commencé à progresser vers e. prussia d'attaquer l'allemagne et détourner leurs ressources levé la pression sur notre allié france

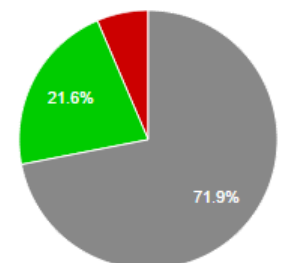
Dashboard – Sentiment Analysis

- Each publication is assigned a sentiment value of “positive”, “neutral” or “negative”:
 - Objective publications (reviews many points of view eg. News report) are always neutral
 - Subjective publications (one person's opinion) are rated as being positive or negative based on the words and expressions used in the text
- Sentiment averages are visualized on a timeline that represents the evolution of the sentiment on the topic, and in a global pie chart.



Sentiment breakdown

- Neutral
- Positive
- Negative



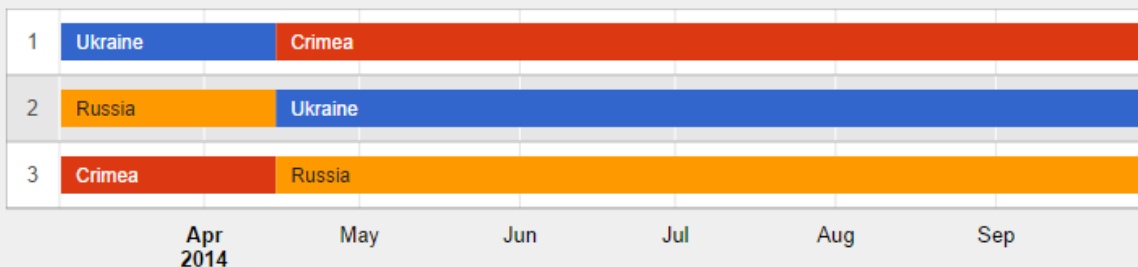
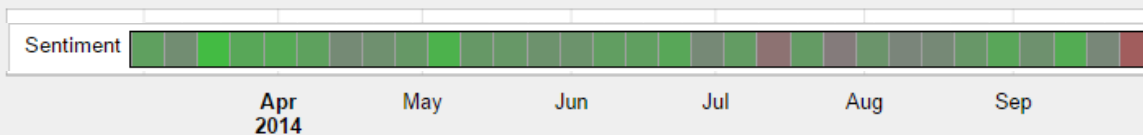
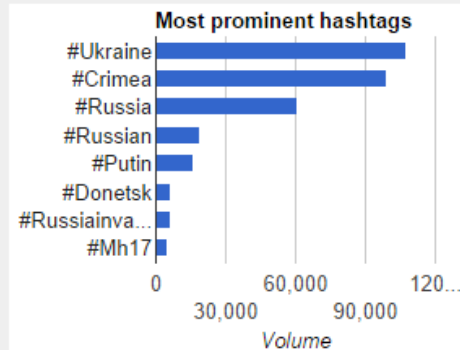
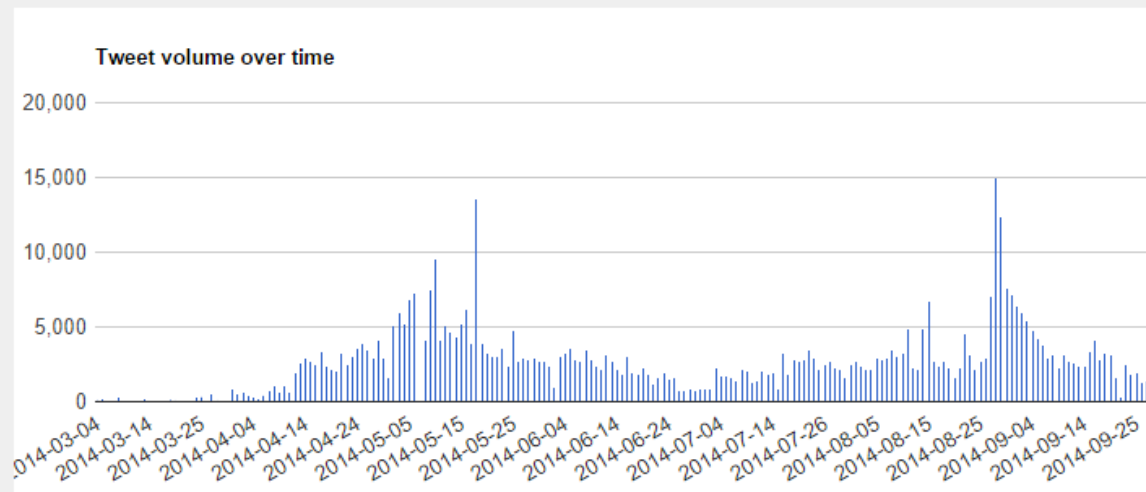
TRANSLI™ - Analytical Graphs



2014-03-04 12:00 AM

2014-09-29 03:14:23 PM

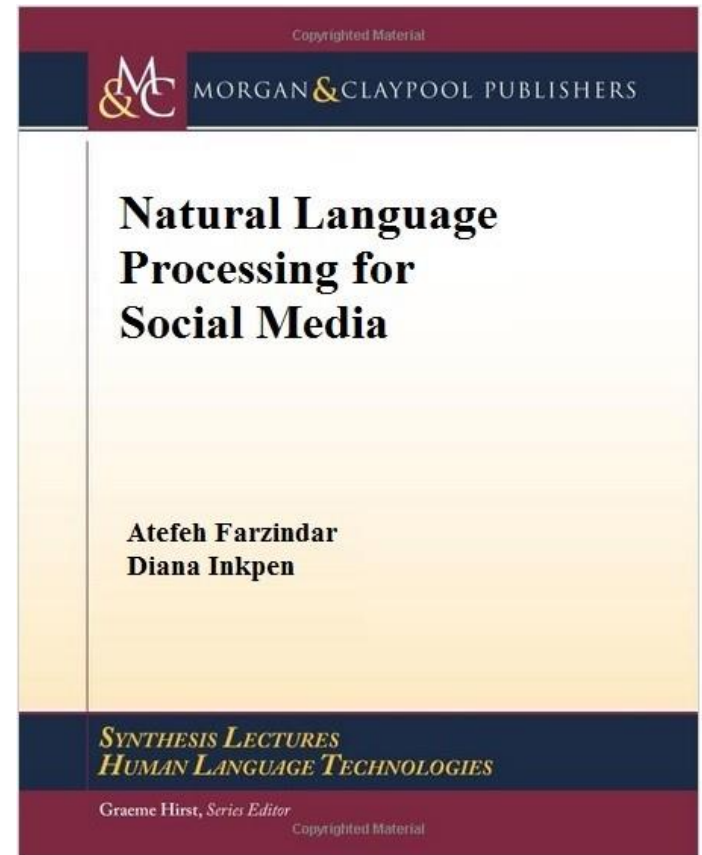
Select date range



Main Reference:

NLP for social media (2015)
Synthesis Lectures on
Human Language Technologies

Authors : Atefeh Farzindar and Diana Inkpen
Editor : Graeme Hirst, *University of Toronto*



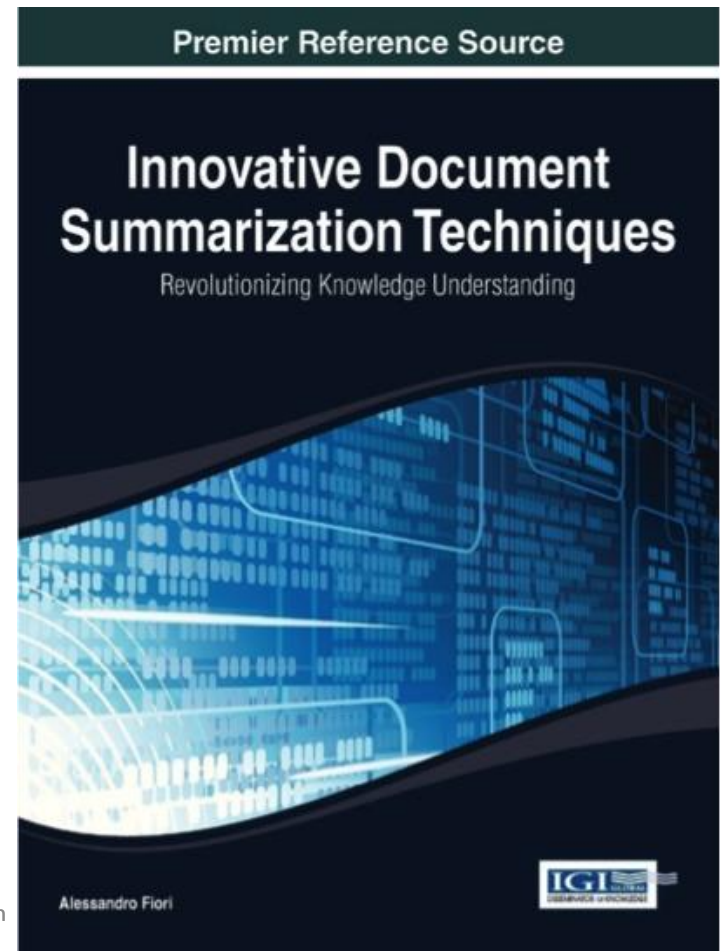
uOttawa

Reference: Book published by IGI, January 2014:



- Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding

Social Network Integration
in Document Summarization
(pages 139-162), by Atefeh
Farzindar



More references

- Atefeh Farzindar and Wael Khreich. A survey of techniques for event detection in Twitter. Computational Intelligence, 2013.
- Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical versus flat classification of emotions in text. In Proceedings of the NAACL HLT 2010 Workshop on Computational approaches to analysis and generation of emotion in text, pages 140–146, Los Angeles, CA, June 2010.
- Fabrizio Gotti, Philippe Langlais, and Atefeh Farzindar. Translating government agencies’ tweet feeds: Specificities, problems and (a few) solutions. In Proceedings of the Workshop on Language Analysis in Social Media, pages 80–89, Atlanta, Georgia, June 2013.
- Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. Location detection and disambiguation from Twitter messages. In Proceedings of CICLing 2015, LNCS 9042, pages 321–332, Cairo, Egypt, 2015.
- Ji Liu and Diana Inkpen. Estimating user locations on social media: A deep learning approach. In Proceedings of the NAACL 2015 Workshop on Vector Space Modeling for NLP, Denver, Colorado, 2015.
- Fatiha Sadat, Farzaneh Kazemi, and Atefeh Farzindar. Automatic identification of Arabic dialects in social media. In SoMeRA 2014: International Workshop on Social Media Retrieval and Analysis, 2014a.





Contact Information



 @nlptechnologies

Atefeh Farzindar
farzindar@nlptechnologies.ca

www.nlptechnologies.ca

Diana Inkpen
Diana.Inkpen@uottawa.ca

<http://www.site.uottawa.ca/~diana/>



uOttawa