Research on Social Computing and Big Data Analytics
(社群運算與大數據分析研究)

Time: 2016/11/17 (Thu) (15:30-17:30)
Place: 東吳大學資管研究所 <城中校區 教室：4303>
Host: 鄭麗珍 教授 (Prof. Li-chen Cheng)

Min-Yuh Day
戴敏育
Assistant Professor
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2016-11-17
戴敏育博士
(Min-Yuh Day, Ph.D.)
淡江大學資管系專任助理教授
中央研究院資訊科學研究所訪問學人
國立台灣大學資訊管理博士
Publications Co-Chairs, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013- )
Program Co-Chair, IEEE International Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2012- )
Workshop Chair, The IEEE International Conference on Information Reuse and Integration (IEEE IRI)
Outline

• Big Data Sentiment Analysis
• Social Computing
• International Research Collaboration and Mobility
Big Data
Sentiment Analysis
Big Data Analytics
Big Data 4 V

Volume

- **40 ZETTAYEBS** (43 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2005
- **6 BILLION PEOPLE** have cell phones
- **WORLD POPULATION**: 7 BILLION

Scale of Data

velocity

- **1 TB OF TRADE INFORMATION** during each trading session
- Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

Analysis of Streaming Data

Veracity

- **1 IN 3 BUSINESS LEADERS** don’t trust the information they use to make decisions
- **27% OF RESPONDENTS** in one survey were unsure of how much of their data was inaccurate

Uncertainty of Data

Variety

- 30 BILLION PIECES OF CONTENT are shared on Facebook every month
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users
- By 2014, it’s anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Different Forms of Data

The Four V’s of Big Data

- **It’s estimated that 2.5 QUINTILLION BYTES (2.3 TRILLION GIGABYTES)** of data are created each day
- Most companies in the U.S. have at least 100 TERABYTES (100,000 GIGABYTES) of data stored

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** (150 BILLION GIGABYTES)

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Source: [https://www-01.ibm.com/software/data/bigdata/](https://www-01.ibm.com/software/data/bigdata/)
Value
Big Data Technologies are Enabling a New Approach

![Diagram showing the relationship between data volume and response time, with categories such as Hadoop, data warehouses, in-memory databases, and event processing tools.](http://www.doclens.com/119898/think-1-13-big-datas-impact-on-analytics/)
Big Data Analytics and Data Mining
Stephan Kudyba (2014),
Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Big Data Sources

- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

- Middleware
- Extract
- Transform
- Load

Transformed Data

- Data Warehouse
- Traditional Format
- CSV, Tables

Big Data Platforms & Tools

- Hadoop
- MapReduce
- Pig
- Hive
- Jaql
- Zookeeper
- Hbase
- Cassandra
- Oozie
- Avro
- Mahout
- Others

Big Data Analytics Applications

- Queries
- Reports
- OLAP
- Data Mining

Raw Data

* Internal
* External
* Multiple formats
* Multiple locations
* Multiple applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools
- Queries
- Reports
- OLAP
- Data Mining

Data Mining
Big Data Analytics Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa(dp/149871093X
Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies

- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Conceptual Layer

- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Logical Layer

- Multivariate analysis
- Application specific task
- Data Mining

Integrated analysis

Physical Layer

- Social Data
- Hardware
- Software

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Business Intelligence (BI) Infrastructure

Operational Data

Historical Data

Machine Data

Web Data

Audio/Video Data

External Data

Data Mart

Data Warehouse

Hadoop Cluster

Extract, transform, load

Casual users
- Queries
- Reports
- Dashboards

Power users
- Queries
- Reports
- OLAP
- Data mining

"Deep learning."
Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Machine-learning technology powers many aspects of modern society: from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users’ interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning.

Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. For decades, conventional methods like decision trees and support vector machines were the state-of-the-art in many fields. They were developed specifically for the data in those fields, and could not be reapplied to new domains without substantial modifications.

Deep learning, in contrast, is a new approach that uses neural networks to represent and learn from data. This allows it to learn a rich representation, which often improves performance over traditional approaches. In addition, the neural network architecture itself can be learned from data, which means that the process of training the model can be automated.

Deep learning has been applied to a wide range of problems, including computer vision, speech recognition, natural language processing, and even autonomous driving. It has also been used for tasks such as image classification, object detection, and machine translation.

Overall, deep learning has had a profound impact on the field of artificial intelligence, and its potential applications are only beginning to be explored.
Sebastian Raschka (2015),
Python Machine Learning,
Packt Publishing
Sunila Gollapudi (2016),
Practical Machine Learning,
Packt Publishing

Machine Learning Models

- Deep Learning
- Association rules
- Decision tree
- Clustering
- Bayesian
- Kernel
- Ensemble
- Dimensionality reduction
- Regression Analysis
- Instance based

Source: Sunila Gollapudi (2016), Practical Machine Learning, Packt Publishing
Data Scientist
資料科學家

Deep Learning
Intelligence from Big Data

Source: https://www.vlab.org/events/deep-learning/
Big Data

- Mobile Sensors
- Social Media
- Video Surveillance
- Video Rendering
- Smart Grids
- Geophysical Exploration
- Medical Imaging
- Gene Sequencing

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Data Scientist: The Sexiest Job of the 21st Century

(Davenport & Patil, 2012)(HBR)

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren’t seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, “It was like arriving at a conference reception and realizing you don’t know anyone. So you just stand in the corner sipping your drink—and you probably leave early.”

Data Scientist Profile

- Technical
- Quantitative
- Curious and Creative
- Skeptical
- Communicative and Collaborative

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Key Roles for a Successful Analytics Project

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business User</td>
<td>Performs business analysis and identifies trends</td>
</tr>
<tr>
<td>Project Sponsor</td>
<td>Acts as liaison between business and analysts</td>
</tr>
<tr>
<td>Project Manager</td>
<td>Oversees the project and manages team members</td>
</tr>
<tr>
<td>Business Intelligence Analyst</td>
<td>Uses data to create predictive models</td>
</tr>
<tr>
<td>Database Administrator (DBA)</td>
<td>Manages the database and ensures data integrity</td>
</tr>
<tr>
<td>Data Engineer</td>
<td>Prepares the data for analysis</td>
</tr>
<tr>
<td>Data Scientist</td>
<td>Analyzes data and develops insights</td>
</tr>
</tbody>
</table>

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Key Outputs from a Successful Analytics Project

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Data Science vs.
Big Data vs. Data Analytics
Data Science vs. Big Data vs. Data Analytics

**WHAT ARE THEY?**

**Data Science** is a field that comprises of everything that related to data cleansing, preparation, and analysis.

**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.

**Data Analytics** involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

What are they used?

Data Science algorithms are used in industries like:
- Internet searches
- Search
- Recommenders
- Digital advertisements

Big Data is used in industries like:
- Financial Services
- Retail
- Communication

Data Analytics is used in industries like:
- Healthcare
- Travel
- Energy Management
- Gaming
Data Science
What are the Skills Required?

DATA SCIENTIST
- In-depth knowledge in SAS and/or R
- Python coding
- Hadoop platform
- SQL database/coding
- Working with unstructured data

BIG DATA SPECIALIST
- Analytical skills
- Creativity
- Mathematics and
- Statistical skills
- Computer science
- Business skills

DATA ANALYST
- Programming skills
- Statistical skills
- Mathematics
- Machine learning skills
- Data wrangling skills
- Communication and Data Visualization skills
- Data Intuition

Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine

Social Media

Source: http://hungrywolfmarketing.com/2013/09/09/what-are-your-social-marketing-goals/
Emotions

Love

Joy

Surprise

Anger

Sadness

Fear

Example of Opinion: review segment on iPhone

“I bought an iPhone a few days ago.
It was such a nice phone.
The touch screen was really cool.
The voice quality was clear too.
However, my mother was mad with me as I did not tell her before I bought it.
She also thought the phone was too expensive, and wanted me to return it to the shop. …”
Example of Opinion: review segment on iPhone

“(1) I bought an ______ a few days ago.
(2) It was such a nice phone.
(3) The touch screen was really cool.
(4) The voice quality was clear too.
(5) However, my mother was mad with me as I did not tell her before I bought it.
(6) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

How consumers think, feel, and act

Maslow’s Hierarchy of Needs

1. Physiological Needs
   (food, water, shelter)

2. Safety Needs
   (security, protection)

3. Social Needs
   (sense of belonging, love)

4. Esteem Needs
   (self-esteem, recognition, status)

5. Self-actualization Needs
   (self-development and realization)

Maslow’s hierarchy of human needs
(Maslow, 1943)

Maslow’s Hierarchy of Needs

- **Physiological needs:**
  - food, water, warmth, rest

- **Safety needs:**
  - security, safety

- **Belongingness and love needs:**
  - intimate relationships, friends

- **Esteem needs:**
  - prestige and feeling of accomplishment

- **Self-actualization:**
  - achieving one’s full potential, including creative activities

Source: http://sixstoriesup.com/social-psyche-what-makes-us-go-social/
Social Media Hierarchy of Needs

Maslow’s Hierarchy of Needs:

1. Physiological Needs: Air, sleep, food, hunger, thirst, warmth
2. Safety & Security: Shelter, protection, safety & stability
3. Love & Belonging: Love, affection, family, & relationships
4. Esteem: Self-esteem, status, reputation, personal fulfillment, building a solid image, engaging in conversations, voicing your expertise
5. Self Actualization: Optimization & monetization, personal branding, community building, structure, existence (presence)

Social Media Hierarchy of Needs - by John Antonios

Source: http://2.bp.blogspot.com/_Rta1VZliMk/TPavcanFtfl/AAAAAAAAACo/OBGnRL5arSU/s1600/social-media-heirarchy-of-needs1.jpg
Social Media Hierarchy of Needs

- Physiological
  - Basic human needs
  - Employment
  - Friendship and family
  - Respect of and by others
  - Creativity & sense-making
  - Self actualization
  - Love/Belonging
  - Safety
  - Esteem
  - Self actualization
The Social Feedback Cycle
Consumer Behavior on Social Media

Marketer-Generated
Awareness
Consideration
Purchase

User-Generated
Use
Form Opinion
Talk

Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement
The New Customer Influence Path

Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement
Architectures of Sentiment Analytics
Bing Liu (2015), Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press

http://www.amazon.com/Sentiment-Analysis-Opinions-Sentiments-Emotions/dp/1107017890
Sentiment Analysis Architecture

Sentiment Classification Based on Emoticons

Tweeter

Tweeter Streaming API 1.1

Tweet preprocessing

Based on Positive Emotions

Based on Negative Emotions

Generate Training Dataset for Tweet

Positive tweets

Negative tweets

Test Dataset

Training Dataset

Classifier

Feature Extraction

Positive

Negative

Lexicon-Based Model

- Preassembled Word Lists
- Generic Word Lists
- Merged Lexicon

Tokenized Document Collection → Sentiment Scoring and Classification: Polarity → Sentiment Polarity

Sentiment Analysis Tasks

- Opinionated Document
- Subjectivity Classification
- Sentiment Classification
- Opinion holder extraction
- Object/Feature extraction
## Sentiment Analysis vs. Subjectivity Analysis

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Subjectivity Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Subjective</td>
</tr>
<tr>
<td>Negative</td>
<td>Objective</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
</tr>
</tbody>
</table>
Levels of Sentiment Analysis

- Word level Sentiment Analysis
- Sentence level Sentiment Analysis
- Document level Sentiment Analysis
- Feature level Sentiment Analysis

Sentiment Analysis

Tasks
- Subjectivity Classification
- Sentiment Classification
- Review Usefulness Measurement
- Opinion Spam Detection
- Lexicon Creation
- Aspect Extraction
- Polarity Determination
- Vagueness resolution in opinionated text
- Multi- & Cross-Lingual SC
- Cross-domain SC

Approaches
- Machine Learning based
- Lexicon based
- Hybrid approaches
- Ontology based
- Non-Ontology based

Sentiment Classification Techniques

- **Sentiment Analysis**
  - **Machine Learning Approach**
    - Supervised Learning
      - Decision Tree Classifiers
      - Linear Classifiers
      - Rule-based Classifiers
      - Probabilistic Classifiers
        - Support Vector Machine (SVM)
        - Neural Network (NN)
        - Deep Learning (DL)
        - Naïve Bayes (NB)
        - Bayesian Network (BN)
        - Maximum Entropy (ME)
  - Unsupervised Learning
    - Dictionary-based Approach
      - Statistical
    - Corpus-based Approach
      - Semantic

Sentiment Analysis and Opinion Mining

- Computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views, emotions, ets., expressed in text.
  - Reviews, blogs, discussions, news, comments, feedback, or any other documents

Research Area of Opinion Mining

- Many names and tasks with difference objective and models
  - Sentiment analysis
  - Opinion mining
  - Sentiment mining
  - Subjectivity analysis
  - Affect analysis
  - Emotion detection
  - Opinion spam detection

Sentiment Analysis

• Sentiment
  – A thought, view, or attitude, especially one based mainly on emotion instead of reason

• Sentiment Analysis
  – opinion mining
  – use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text
Applications of Sentiment Analysis

• Consumer information
  – Product reviews

• Marketing
  – Consumer attitudes
  – Trends

• Politics
  – Politicians want to know voters’ views
  – Voters want to know politicians’ stances and who else supports them

• Social
  – Find like-minded individuals or communities
**Sentiment detection**

- How to interpret features for sentiment detection?
  - Bag of words (IR)
  - Annotated lexicons (WordNet, SentiWordNet)
  - Syntactic patterns

- Which features to use?
  - Words (unigrams)
  - Phrases/n-grams
  - Sentences
Problem statement of Opinion Mining

- Two aspects of abstraction
  - Opinion definition
    - What is an opinion?
    - What is the structured definition of opinion?
  - Opinion summarization
    - Opinion are subjective
      - An opinion from a single person (unless a VIP) is often not sufficient for action
    - We need opinions from many people, and thus opinion summarization.

What is an opinion?

• Id: Abc123 on 5-1-2008 “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, ...”

• One can look at this review/blog at the
  – Document level
    • Is this review + or -?
  – Sentence level
    • Is each sentence + or -?
  – Entity and feature/aspect level

**Entity and aspect/feature level**

- **Id: Abc123 on 5-1-2008** “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …”

- **What do we see?**
  - Opinion targets: entities and their features/aspects
  - Sentiments: positive and negative
  - Opinion holders: persons who hold the opinions
  - Time: when opinion are expressed

Two main types of opinions

• **Regular opinions**: Sentiment/Opinion expressions on some target entities
  – **Direct opinions**: sentiment expressions on one object:
    • “The touchscreen is really cool.”
    • “The picture quality of this camera is great”
  – **Indirect opinions**: comparisons, relations expressing similarities or differences (objective or subjective) of more than one object
    • “phone X is cheaper than phone Y.” (objective)
    • “phone X is better than phone Y.” (subjective)
• **Comparative opinions**: comparisons of more than one entity.
  – “iPhone is better than Blackberry.”

Subjective and Objective

• **Objective**
  – An objective sentence expresses some **factual information** about the world.
  – “I *returned* the phone yesterday.”
  – Objective sentences can implicitly indicate opinions
    • “The *earphone broke* in two days.”

• **Subjective**
  – A subjective sentence expresses some **personal feelings** or beliefs.
  – “The voice on my phone was *not so clear*”
  – Not every subjective sentence contains an opinion
    • “I wanted a phone with *good voice quality*”

• ➜ **Subjective analysis**

### Sentiment Analysis vs. Subjectivity Analysis

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Subjectivity Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Subjective</td>
</tr>
<tr>
<td>Negative</td>
<td>Objective</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
</tr>
</tbody>
</table>
A (regular) opinion

• Opinion (a restricted definition)
  – An opinion (regular opinion) is simply a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion holder.

• Sentiment orientation of an opinion
  – Positive, negative, or neutral (no opinion)
  – Also called:
    • Opinion orientation
    • Semantic orientation
    • Sentiment polarity

Entity and aspect

• Definition of Entity:
  – An *entity* *e* is a product, person, event, organization, or topic.
  – *e* is represented as
    • A hierarchy of components, sub-components.
    • Each node represents a components and is associated with a set of attributes of the components.

• An opinion can be expressed on any node or attribute of the node.

• Aspects(features)
  – represent both components and attribute.

Opinion Definition

• An opinion is a quintuple
  \((e_j, a_{jk}, s_{ijkl}, h_i, t_l)\)
  where
  – \(e_j\) is a target entity.
  – \(a_{jk}\) is an aspect/feature of the entity \(e_j\).
  – \(s_{ijkl}\) is the sentiment value of the opinion from the opinion holder on feature of entity at time.
    \(s_{ijkl}\) is +ve, -ve, or neu, or more granular ratings
  – \(h_i\) is an opinion holder.
  – \(t_l\) is the time when the opinion is expressed.

• \((e_j, a_{jk})\) is also called opinion target
Terminologies

• **Entity**: object
• **Aspect**: feature, attribute, facet
• **Opinion holder**: opinion source
• **Topic**: entity, aspect

• **Product features, political issues**

Subjectivity and Emotion

• Sentence subjectivity
  – An objective sentence presents some factual information, while a subjective sentence expresses some personal feelings, views, emotions, or beliefs.

• Emotion
  – Emotions are people’s subjective feelings and thoughts.

Classification Based on Supervised Learning

• Sentiment classification
  – Supervised learning Problem
  – Three classes
    • Positive
    • Negative
    • Neutral

Opinion words in Sentiment classification

• topic-based classification
  – topic-related words are important
    • e.g., politics, sciences, sports

• Sentiment classification
  – topic-related words are unimportant
  – opinion words (also called sentiment words)
    • that indicate positive or negative opinions are important,
      e.g., great, excellent, amazing, horrible, bad, worst

Features in Opinion Mining

• **Terms and their frequency**
  – TF-IDF

• **Part of speech (POS)**
  – Adjectives

• **Opinion words and phrases**
  – *beautiful, wonderful, good, and amazing* are positive opinion words
  – *bad, poor, and terrible* are negative opinion words.
  – opinion phrases and idioms, e.g., *cost someone an arm and a leg*

• **Rules of opinions**

• **Negations**

• **Syntactic dependency**

A Brief Summary of Sentiment Analysis Methods

<table>
<thead>
<tr>
<th>Study</th>
<th>Analysis Task</th>
<th>Sentiment Identification Method</th>
<th>Level</th>
<th>Sentiment Aggregation Method</th>
<th>Level</th>
<th>Nature of Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu and Li, 2011</td>
<td>Polarity</td>
<td>ML (Probabilistic model)</td>
<td>Snippet</td>
<td>Sum</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Li and Wu, 2010</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td>Sum</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Thelwall et al., 2010</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Sentence</td>
<td>Max &amp; Min</td>
<td>Snippet</td>
<td>Range</td>
</tr>
<tr>
<td>Boiy and Moens, 2009</td>
<td>Both</td>
<td>ML (Cascade ensemble)</td>
<td>Sentence</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Chung 2009</td>
<td>Polarity</td>
<td>Lexicon</td>
<td>Phrase</td>
<td>Average</td>
<td>Sentence</td>
<td>Valence</td>
</tr>
<tr>
<td>Wilson, Wiebe, and Hoffmann, 2009</td>
<td>Both</td>
<td>ML (SVM, AdaBoost, Rule, etc.)</td>
<td>Phrase</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Zhang et al., 2009</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Sentence</td>
<td>Weighted average</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Abbasi, Chen, and Salem, 2008</td>
<td>Polarity</td>
<td>ML (GA + feature selection)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Subrahmanian and Reforgiato, 2008</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td>Rule</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Tan and Zhang 2008</td>
<td>Polarity</td>
<td>ML (SVM, Winnow, NB, etc.)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Airoldi, Bai, and Padman, 2007</td>
<td>Polarity</td>
<td>ML (Markov Blanket)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Das and Chen, 2007</td>
<td>Polarity</td>
<td>ML (Bayesian, Discriminate, etc.)</td>
<td>Snippet</td>
<td>Average</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Liu et al., 2007</td>
<td>Polarity</td>
<td>ML (PLSA)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Kennedy and Inkpen, 2006</td>
<td>Polarity</td>
<td>Lexicon/Rule, ML (SVM)</td>
<td>Phrase</td>
<td>Count</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Mishne 2006</td>
<td>Polarity</td>
<td>Lexicon</td>
<td>Phrase</td>
<td>Average</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Liu et al., 2005</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td>Distribution</td>
<td>Object</td>
<td>Range</td>
</tr>
<tr>
<td>Mishne 2005</td>
<td>Polarity</td>
<td>ML (SVM)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Popescu and Etzioni 2005</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Efron 2004</td>
<td>Polarity</td>
<td>ML (SVM, NB)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Wilson, Wiebe, and Hwa, 2004</td>
<td>Both</td>
<td>ML (SVM, AdaBoost, Rule, etc.)</td>
<td>Sentence</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Nigam and Hurst 2004</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Chunk</td>
<td>Rule</td>
<td>Sentence</td>
<td>Valence</td>
</tr>
<tr>
<td>Dave, Lawrence, and Pennoke, 2003</td>
<td>Polarity</td>
<td>ML (SVM, Rainbow, etc.)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Nasukawa and Yi 2003</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td>Rule</td>
<td>Sentence</td>
<td>Valence</td>
</tr>
<tr>
<td>Yi et al., 2003</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td>Rule</td>
<td>Sentence</td>
<td>Valence</td>
</tr>
<tr>
<td>Yu and Hatzivassiloglou 2003</td>
<td>Both</td>
<td>ML (NB) + Lexicon/Rule</td>
<td>Phrase</td>
<td>Average</td>
<td>Sentence</td>
<td>Valence</td>
</tr>
<tr>
<td>Pang, Lee, and Vaithyanathan 2002</td>
<td>Polarity</td>
<td>ML (SVM, MaxEnt, NB)</td>
<td>Snippet</td>
<td></td>
<td></td>
<td>Valence</td>
</tr>
<tr>
<td>Subasic and Huettner 2001</td>
<td>Polarity</td>
<td>Lexicon/Fuzzy logic</td>
<td>Phrase</td>
<td>Average</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
<tr>
<td>Turney 2001</td>
<td>Polarity</td>
<td>Lexicon/Rule</td>
<td>Phrase</td>
<td>Average</td>
<td>Snippet</td>
<td>Valence</td>
</tr>
</tbody>
</table>

(Both = Subjectivity and Polarity; ML = Machine Learning; Lexicon/Rule = Lexicon enhanced by linguistic rules)

Word-of-Mouth (WOM)

• “This book is the best written documentary thus far, yet sadly, there is no soft cover edition.”

• “This book is the best written documentary thus far, yet sadly, there is no soft cover edition.”

This book is the best written documentary thus far, yet sadly, there is no soft cover edition.

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>This</td>
<td>DT</td>
</tr>
<tr>
<td>book</td>
<td>NN</td>
</tr>
<tr>
<td>is</td>
<td>VBZ</td>
</tr>
<tr>
<td>the</td>
<td>DT</td>
</tr>
<tr>
<td>best</td>
<td>JJS</td>
</tr>
<tr>
<td>written</td>
<td>VBN</td>
</tr>
<tr>
<td>documentary</td>
<td>NN</td>
</tr>
<tr>
<td>thus</td>
<td>RB</td>
</tr>
<tr>
<td>far</td>
<td>RB</td>
</tr>
<tr>
<td>,</td>
<td></td>
</tr>
<tr>
<td>yet</td>
<td>RB</td>
</tr>
<tr>
<td>sadly</td>
<td>RB</td>
</tr>
<tr>
<td>,</td>
<td></td>
</tr>
<tr>
<td>there</td>
<td>EX</td>
</tr>
<tr>
<td>is</td>
<td>VBZ</td>
</tr>
<tr>
<td>no</td>
<td>DT</td>
</tr>
<tr>
<td>soft</td>
<td>JJ</td>
</tr>
<tr>
<td>cover</td>
<td>NN</td>
</tr>
<tr>
<td>edition</td>
<td>NN</td>
</tr>
</tbody>
</table>
Conversion of text representation

<table>
<thead>
<tr>
<th>Word Vector (WV)</th>
<th>Polarity Score Vector (PSV)</th>
<th>Microstate Sequence (MS)</th>
<th>Probability Distribution (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This</td>
<td>0, 0</td>
<td>0</td>
<td>P(“1”)=3/17</td>
</tr>
<tr>
<td>book</td>
<td>0, 0</td>
<td>0</td>
<td>P(“-1”)=3/17</td>
</tr>
<tr>
<td>is</td>
<td>0, 0</td>
<td>0</td>
<td>P(“0”)=11/17</td>
</tr>
<tr>
<td>the</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>best</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>written</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>documentary</td>
<td>0.75, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>thus</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>far</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>yet</td>
<td>0.375, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>sadly</td>
<td>0.375, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>there</td>
<td>0.0125, 0</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>is</td>
<td>0.25, 0.5</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>soft</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>cover</td>
<td>0, 0.75</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>edition</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0, 0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Example of SentiWordNet**

<table>
<thead>
<tr>
<th>POS</th>
<th>ID</th>
<th>PosScore</th>
<th>NegScore</th>
<th>SynsetTerms</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>00217728</td>
<td>0.75</td>
<td>0</td>
<td>beautiful#1</td>
<td>delighting the senses or exciting intellectual or emotional admiration; &quot;a beautiful child&quot;; &quot;beautiful country&quot;; &quot;a beautiful painting&quot;; &quot;a beautiful theory&quot;; &quot;a beautiful party&quot;</td>
</tr>
<tr>
<td>a</td>
<td>00227507</td>
<td>0.75</td>
<td>0</td>
<td>best#1</td>
<td>(superlative of `good') having the most positive qualities; &quot;the best film of the year&quot;; &quot;the best solution&quot;; &quot;the best time for planting&quot;; &quot;wore his best suit&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00042614</td>
<td>0</td>
<td>0.625</td>
<td>unhappily#2 sadly#1</td>
<td>in an unfortunate way; &quot;sadly he died before he could see his grandchild&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00093270</td>
<td>0</td>
<td>0.875</td>
<td>woefully#1 sadly#3</td>
<td>lamentably#1 deplorably#1 in an unfortunate or deplorable manner; &quot;he was sadly neglected&quot;; &quot;it was woefully inadequate&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00404501</td>
<td>0</td>
<td>0.25</td>
<td>sadly#2</td>
<td>with sadness; in a sad manner; &quot;`She died last night,' he said sadly&quot;</td>
</tr>
</tbody>
</table>
The car is very old but it is rather not expensive.

The car is very **old** but it is rather **not** **expensive**.

The car is **very** **old** **but** it is **rather** **not** **expensive**.

Polarity Detection with SenticNet

The car is very old but it is rather not expensive.
The car is very old but it is rather not expensive.

Polarity Detection with SenticNet

Polarity Detection with SenticNet

Polarity Detection with SenticNet

Polarity Detection with SenticNet

Evaluation of Text Mining and Sentiment Analysis

• Evaluation of Information Retrieval
• Evaluation of Classification Model (Prediction)
  – Accuracy
  – Precision
  – Recall
  – F-score
Deep Learning for Sentiment Analytics
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
Stanford University, Stanford, CA 94305, USA
richard@socher.org, {aperelyg, jcchuang, ang}@cs.stanford.edu
{jeaneis, manning, cgpotts}@stanford.edu

Abstract

Semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of composition. To remedy this, we introduce a Sentiment Treebank. It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. To address them, we introduce the Recursive Neural Tensor Network. When trained on the new treebank, this model out-

Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (−−, −, 0, +, ++), at every node of a parse tree and capturing the negation and its scope in this sentence.
Recursive Neural Tensor Network (RNTN)

Recursive Neural Network (RNN) models for sentiment

Recursive Neural Tensor Network (RNTN)

Roger Dodger is one of the **most** compelling variations on this theme.

Roger Dodger is one of the **least** compelling variations on this theme.

Roger Dodger is one of the most compelling variations on this theme.

Roger Dodger is one of the least compelling variations on this theme.

### Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) and for all nodes

<table>
<thead>
<tr>
<th>Model</th>
<th>Fine-grained</th>
<th>Positive/Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Root</td>
</tr>
<tr>
<td>NB</td>
<td>67.2</td>
<td>41.0</td>
</tr>
<tr>
<td>SVM</td>
<td>64.3</td>
<td>40.7</td>
</tr>
<tr>
<td>BiNB</td>
<td>71.0</td>
<td>41.9</td>
</tr>
<tr>
<td>VecAvg</td>
<td>73.3</td>
<td>32.7</td>
</tr>
<tr>
<td>RNN</td>
<td>79.0</td>
<td>43.2</td>
</tr>
<tr>
<td>MV-RNN</td>
<td>78.7</td>
<td>44.4</td>
</tr>
<tr>
<td>RNTN</td>
<td>80.7</td>
<td>45.7</td>
</tr>
</tbody>
</table>

## Accuracy of negation detection

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy Negated Positive</th>
<th>Accuracy Negated Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>biNB</td>
<td>19.0</td>
<td>27.3</td>
</tr>
<tr>
<td>RNN</td>
<td>33.3</td>
<td>45.5</td>
</tr>
<tr>
<td>MV-RNN</td>
<td>52.4</td>
<td>54.6</td>
</tr>
<tr>
<td>RNTN</td>
<td>71.4</td>
<td>81.8</td>
</tr>
</tbody>
</table>

Long Short-Term Memory (LSTM)

Source: https://cs224d.stanford.edu/reports/HongJames.pdf
# Deep Learning for Sentiment Analysis

**CNN** RNTN LSTM

<table>
<thead>
<tr>
<th>Model</th>
<th>Fine (5-class)</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCNN (Blunsom, et al. 2014)</td>
<td>0.485</td>
<td>0.868</td>
</tr>
<tr>
<td>RNTN (Socher, et al. 2013)</td>
<td>0.457</td>
<td>0.854</td>
</tr>
<tr>
<td>CNN-non-static (Kim, 2014)</td>
<td>0.480</td>
<td>0.872</td>
</tr>
<tr>
<td>CNN-multi-channel (Kim, 2014)</td>
<td>0.474</td>
<td>0.881</td>
</tr>
<tr>
<td>DRNN w. pretrained word-embeddings (Irsoy and Cardie, 2014)</td>
<td>0.498</td>
<td>0.866</td>
</tr>
<tr>
<td>Paragraph Vector (Le and Mikolov. 2014)</td>
<td>0.487</td>
<td>0.878</td>
</tr>
<tr>
<td>Dependency Tree-LSTM (Tai, et al, 2015)</td>
<td>0.484</td>
<td>0.857</td>
</tr>
<tr>
<td>Constituency Tree-LSTM (Tai, et al, 2015)</td>
<td>0.439</td>
<td>0.820</td>
</tr>
<tr>
<td>Constituency Tree-LSTM (Glove vectors) (Tai, et al, 2015)</td>
<td>0.510</td>
<td>0.880</td>
</tr>
<tr>
<td>Paragraph Vector</td>
<td>0.391</td>
<td>0.798</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.456</td>
<td>0.843</td>
</tr>
<tr>
<td>Deep Recursive-NN</td>
<td>0.469</td>
<td>0.847</td>
</tr>
</tbody>
</table>

Source: [https://cs224d.stanford.edu/reports/HongJames.pdf](https://cs224d.stanford.edu/reports/HongJames.pdf)
<table>
<thead>
<tr>
<th>Method</th>
<th>Data Set</th>
<th>Acc.</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine Learning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Movie reviews</td>
<td>86.40%</td>
<td>Pang, Lee[23]</td>
</tr>
<tr>
<td>CoTraining SVM</td>
<td>Twitter</td>
<td>82.52%</td>
<td>Liu[14]</td>
</tr>
<tr>
<td>Deep learning</td>
<td>Stanford Sentiment Treebank</td>
<td>80.70%</td>
<td>Richard[18]</td>
</tr>
<tr>
<td><strong>Lexical based</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corpus</td>
<td>Product reviews</td>
<td>74.00%</td>
<td>Turkey</td>
</tr>
<tr>
<td>Dictionary</td>
<td>Amazon’s Mechanical Turk</td>
<td>---</td>
<td>Taboada[20]</td>
</tr>
<tr>
<td><strong>Cross-lingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>Amazon</td>
<td>81.00%</td>
<td>Wan,X[16]</td>
</tr>
<tr>
<td>Co-Train</td>
<td>Amazon, ITI68</td>
<td>81.30%</td>
<td>Wan,X.[16]</td>
</tr>
<tr>
<td>EWGA</td>
<td>IMDb movie review</td>
<td>&gt;90%</td>
<td>Abbasi,A.</td>
</tr>
<tr>
<td>CLMM</td>
<td>MPQA, NTCIR, ISI</td>
<td>83.02%</td>
<td>Mengi</td>
</tr>
<tr>
<td><strong>Cross-domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active Learning</td>
<td>Book, DVD, Electronics, Kitchen</td>
<td>80% (avg)</td>
<td>Li, S</td>
</tr>
<tr>
<td>Thesaurus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Resources of Opinion Mining
Datasets of Opinion Mining

- **Blog06**
  - 25GB TREC test collection
  - [http://ir.dcs.gla.ac.uk/test collections/access to data.html](http://ir.dcs.gla.ac.uk/test collections/access to data.html)

- **Cornell movie-review datasets**

- **Customer review datasets**

- **Multiple-aspect restaurant reviews**
  - [http://people.csail.mit.edu/bsnyder/naacl07](http://people.csail.mit.edu/bsnyder/naacl07)

- **NTCIR multilingual corpus**
  - NTCIR Multilingual Opinion-Analysis Task (MOAT)

Lexical Resources of Opinion Mining

- SentiWordnet
  - http://sentiwordnet.isti.cnr.it/
- General Inquirer
  - http://www.wjh.harvard.edu/~inquirer/
- OpinionFinder’s Subjectivity Lexicon
  - http://www.cs.pitt.edu/mpqa/
- NTU Sentiment Dictionary (NTUSD)
  - http://nlg18.csie.ntu.edu.tw:8080/opinion/
- Hownet Sentiment
### Example of SentiWordNet

<table>
<thead>
<tr>
<th>POS</th>
<th>ID</th>
<th>PosScore</th>
<th>NegScore</th>
<th>SynsetTerms</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>00217728</td>
<td>0.75</td>
<td>0</td>
<td>beautiful#1</td>
<td>delighting the senses or exciting intellectual or emotional admiration; &quot;a beautiful child&quot;;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&quot;beautiful country&quot;; &quot;a beautiful painting&quot;; &quot;a beautiful theory&quot;; &quot;a beautiful party&quot;</td>
</tr>
<tr>
<td>a</td>
<td>00227507</td>
<td>0.75</td>
<td>0</td>
<td>best#1</td>
<td>(superlative of `good') having the most positive qualities; &quot;the best film of the year&quot;;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&quot;the best solution&quot;; &quot;the best time for planting&quot;; &quot;wore his best suit&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00042614</td>
<td>0</td>
<td>0.625</td>
<td>unhappily#2 sadly#1</td>
<td>in an unfortunate way; &quot;sadly he died before he could see his grandchild&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00093270</td>
<td>0</td>
<td>0.875</td>
<td>woefully#1 sadly#3 lamentably#1</td>
<td>deplorably#1 in an unfortunate or deplorable manner; &quot;he was sadly neglected&quot;; &quot;it was</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>woefully inadequate&quot;</td>
</tr>
<tr>
<td>r</td>
<td>00404501</td>
<td>0</td>
<td>0.25</td>
<td>sadly#2</td>
<td>with sadness; in a sad manner; &quot;`She died last night,' he said sadly&quot;</td>
</tr>
</tbody>
</table>
《知網》情感分析用詞語集（beta版）

• “中英文情感分析用詞語集”
  – 包含詞語約 17887

• “中文情感分析用詞語集”
  – 包含詞語約 9193

• “英文情感分析用詞語集”
  – 包含詞語 8945

中文情感分析用詞語集

<table>
<thead>
<tr>
<th>詞語類型</th>
<th>數量</th>
</tr>
</thead>
<tbody>
<tr>
<td>中文正面情感詞語</td>
<td>836</td>
</tr>
<tr>
<td>中文負面情感詞語</td>
<td>1254</td>
</tr>
<tr>
<td>中文正面評價詞語</td>
<td>3730</td>
</tr>
<tr>
<td>中文負面評價詞語</td>
<td>3116</td>
</tr>
<tr>
<td>中文程度級別詞語</td>
<td>219</td>
</tr>
<tr>
<td>中文主張詞語</td>
<td>38</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9193</strong></td>
</tr>
</tbody>
</table>
中文情感分析用詞語集

• “正面情感” 詞語
  — 如：
    愛，讚賞，快樂，感同身受，好奇，
    喝彩，魂牽夢縈，嘉許 ...

• “負面情感” 詞語
  — 如：
    哀傷，半信半疑，鄙視，不滿意，不是滋味兒
    ，後悔，大失所望 ...

中文情感分析用詞語集

• “正面評價”詞語
  - 如：
    不可或缺，部優，才高八斗，沉魚落雁，
    催人奮進，動聽，對勁兒 ...

• “負面評價”詞語
  - 如：
    醜，苦，超標，華而不實，荒涼，混濁，
    畸輕畸重，價值，空洞無物 ...

中文情感分析用詞語集

• “程度級別”詞語
  - 1. “極其 | extreme / 最 | most”
    • 非常，極，極度，無以倫比，最為
  - 2. “很 | very”
    • 多麼，分外，格外，著實
  - ...

• “主張”詞語
  - 1. {perception | 感知}
    • 感覺，覺得，預感
  - 2. {regard | 認為}
    • 認為，以為，主張

Social Computing
Social Network Analysis
Social Computing

• Social Network Analysis
• Link mining
• Community Detection
• Social Recommendation
Business Insights with Social Analytics
Analyzing the Social Web: Social Network Analysis
Devangana Khokhar (2015), *Gephi Cookbook*, Packt Publishing

Social Network Analysis (SNA)
Facebook TouchGraph
Graph Theory
Graph
Graph $g = (V, E)$
Vertex (Node)
Vertices (Nodes)
Edges
Arc
Arcs
Undirected Graph

Source: https://www.youtube.com/watch?v=89mxOdwpfxA
Directed Graph

Source: https://www.youtube.com/watch?v=89mXOdwpfxA
Measurements of Social Network Analysis
Exploratory Network Analysis

1. see the network

1st graph viz tool: Pajek (1996)
Vladimir Batagelj, Andrej Mrvar

2. interact in real time

Gephi prototype (2008)
group, filter, compute metrics…

3. build a visual language

size by rank, color by partition,
label, curved edges, thickness…

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf
Looking for a “Simple Small Truth”? What Data Visualization Should Do?

1. Make complex things **simple**
2. Extract **small** information from large data
3. Present **truth**, do not deceive

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf
Measurements
Looking for Orderness in Data

Make varying 3 cursors simultaneously to extract meaningful patterns

- at different levels
- on multiple dimensions
- at time scale

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf
“Zoom” cursor on Quantitative Data

Global
- connectivity
- density
- centralization

Local
- communities
- bridges between communities
- local centers vs periphery

Individual
- centrality
- distances
- neighborhood
- location
- local authority vs hub

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf
“Crossing” cursor on Quantitative Data

Social
- who with whom
- communities
- brokerage
- influence and power
- homophily

Semantic
- topics
- thematic clusters

Geographic
- spatial phenomena

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf
“Timeline” cursor on Temporal Data

Evolution of social ties
Evolution of communities
Evolution of topics

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf
# SNA Guideline

## # nodes

<table>
<thead>
<tr>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 100</td>
<td>lists + edges in bonus, focus on qualitative data</td>
</tr>
</tbody>
</table>
| 100 - 1,000 | easy to read, “obvious” patterns  
focus on entities (in context)  
metrics are tools to describe the graph (centrality, bridging…)  
links help to build and interpret categories of entities  
challenge: mix attribute crossing and connectivity |
| 1,000 - 50,000 | hard to read, problem of “hidden signals”:  
track patterns with various layouts and filtering  
focus on structures  
metrics are tools to build the graph (cosine similarity…)  
categories help to understand the structure  
challenge: pattern recognition |
| > 50,000 | require high computational power |

Degree

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Degree

A: 2
B: 4
C: 2
D: 1
E: 1

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Density

Edges (Links): 5
Total Possible Edges: 10
Density: 5/10 = 0.5

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Nodes (n): 10
Edges (Links): 13
Total Possible Edges: \( \frac{n \times (n-1)}{2} = \frac{10 \times 9}{2} = 45 \)
Density: \( \frac{13}{45} = 0.29 \)
Diameter

radius ($r$)

diameter ($d$)
Diameter
Diameter
Geodesic Path (Shortest Path)

A → I : Diameter = 4
Which Node is Most Important?
Centrality

• Important or prominent actors are those that are linked or involved with other actors extensively.

• A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts.

• The links can also be called ties. A central actor is one involved in many ties.

Social Network Analysis (SNA)

• Degree Centrality
• Betweenness Centrality
• Closeness Centrality
Degree Centrality
Social Network Analysis: Degree Centrality
Social Network Analysis: Degree Centrality

<table>
<thead>
<tr>
<th>Node</th>
<th>Score</th>
<th>Standardized Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>5/10 = 0.5</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>4/10 = 0.4</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1/10 = 0.1</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1/10 = 0.1</td>
</tr>
</tbody>
</table>
Betweenness Centrality
Betweenness centrality:

Connectivity

Number of shortest paths going through the actor
Betweenness Centrality

\[ C_B(i) = \sum_{j<k} g_{ik}(i) / g_{jk} \]

Where \( g_{jk} \) = the number of shortest paths connecting \( jk \)
\( g_{jk}(i) \) = the number that actor \( i \) is on.

Normalized Betweenness Centrality

\[ C'_B(i) = C_B(i) / \left[ (n-1)(n-2)/2 \right] \]

Number of pairs of vertices excluding the vertex itself.

Source: https://www.youtube.com/watch?v=RXohUeNCJiU
Betweenness Centrality

A: Betweenness Centrality = 0

A: B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0
Betweenness Centrality

B: Betweenness Centrality = 5
Betweenness Centrality

C: Betweenness Centrality = 0

C:  
A→B: 0/1 = 0  
A→D: 0/1 = 0  
A→E: 0/1 = 0  
B→D: 0/1 = 0  
B→E: 0/1 = 0  
D→E: 0/1 = 0  

Total: 0
Betweenness Centrality

A: 0
B: 5
C: 0
D: 0
E: 0
Which Node is Most Important?
Which Node is Most Important?
Betweenness Centrality

\[ C_B(i) = \sum_{j<k} \frac{g_{ik}(i)}{g_{jk}} \]
Betweenness Centrality

A: Betweenness Centrality = 0

- B → C: 0/1 = 0
- B → D: 0/1 = 0
- B → E: 0/1 = 0
- C → D: 0/1 = 0
- C → E: 0/1 = 0
- D → E: 0/1 = 0

Total: 0
Closeness
Centrality
Social Network Analysis: Closeness Centrality

C: Closeness Centrality = $\frac{15}{9} = 1.67$
Social Network Analysis: Closeness Centrality

G: Closeness Centrality = 14/9 = 1.56
Social Network Analysis: Closeness Centrality

H: Closeness Centrality = 17/9 = 1.89
Social Network Analysis: Closeness Centrality

G: Closeness Centrality = 14/9 = 1.56
C: Closeness Centrality = 15/9 = 1.67
H: Closeness Centrality = 17/9 = 1.89
International Research Collaboration and Mobility
Application of SNA

Social Network Analysis of Research Collaboration in Information Reuse and Integration

Example of SNA Data Source

IRI 2010: Las Vegas, NV, USA


Reda Alhajj, James B. D. Joshi, Mei-Ling Shyu: Message from Program Co-Chairs. 1

Stuart Harvey Rubin, Shu-Ching Chen: Forward. 1

Lotfi A. Zadeh: Precisiation of meaning - toward computation with natural language. 1-4

Reda Alhajj, Shu-Ching Chen, Gongzhu Hu, James B. D. Joshi, Gordon K. Lee, Stuart Harvey Rubin, Mei-Ling Shyu, Lotfi A. Zadeh: Panel title: Critical need for funding of basic and applied research in large-scale computing. 1

Automation, Integration and Reuse across Various Apps

László István Etesi, André Csillaghy, Lin-Ching Chang: A message-based interoperability framework with application to astrophysics. 1-6

Awny Alnusair, Tian Zhao, Eric Bodden: Effective API navigation and reuse. 7-12

Manabu Ohta, Ryohei Inoue, Atsuhiro Takasu: Empirical evaluation of active sampling for CRF-based analysis of pages. 13-18

Qunzhi Zhou, Viktor K. Prasanna: Workflow management of simulation based computation processes in transportation domain. 19-24

Source: http://www.informatik.uni-trier.de/~ley/db/conf/iri/iri2010.html
Research Question

• RQ1: What are the scientific collaboration patterns in the IRI research community?

• RQ2: Who are the prominent researchers in the IRI community?

Methodology

• Developed a simple web focused crawler program to download literature information about all IRI papers published between 2003 and 2010 from IEEE Xplore and DBLP.
  – 767 paper
  – 1599 distinct author

• Developed a program to convert the list of coauthors into the format of a network file which can be readable by social network analysis software.

• UCINet and Pajek were used in this study for the social network analysis.

Top 10 prolific authors (IRI 2003-2010)

1. Stuart Harvey Rubin
2. Taghi M. Khoshgoftaar
3. Shu-Ching Chen
4. Mei-Ling Shyu
5. Mohamed E. Fayad
6. Reda Alhajj
7. Du Zhang
8. Wen-Lian Hsu
9. Jason Van Hulse
10. Min-Yuh Day

Data Analysis and Discussion

• Closeness Centrality
  – Collaborated widely

• Betweenness Centrality
  – Collaborated diversely

• Degree Centrality
  – Collaborated frequently

• Visualization of Social Network Analysis
  – Insight into the structural characteristics of research collaboration networks

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Closeness</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.024675</td>
<td>Shu-Ching Chen</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.022830</td>
<td>Stuart Harvey Rubin</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.022207</td>
<td>Mei-Ling Shyu</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.020013</td>
<td>Reda Alhajj</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>0.019700</td>
<td>Na Zhao</td>
</tr>
<tr>
<td>6</td>
<td>260</td>
<td>0.018936</td>
<td>Min Chen</td>
</tr>
<tr>
<td>7</td>
<td>151</td>
<td>0.018230</td>
<td>Gordon K. Lee</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>0.017962</td>
<td>Chengcui Zhang</td>
</tr>
<tr>
<td>9</td>
<td>1043</td>
<td>0.017962</td>
<td>Isai Michel Lombera</td>
</tr>
<tr>
<td>10</td>
<td>1027</td>
<td>0.017962</td>
<td>Michael Armella</td>
</tr>
<tr>
<td>11</td>
<td>443</td>
<td>0.017448</td>
<td>James B. Law</td>
</tr>
<tr>
<td>12</td>
<td>157</td>
<td>0.017082</td>
<td>Keqi Zhang</td>
</tr>
<tr>
<td>13</td>
<td>253</td>
<td>0.016731</td>
<td>Shahid Hamid</td>
</tr>
<tr>
<td>14</td>
<td>1038</td>
<td>0.016618</td>
<td>Walter Z. Tang</td>
</tr>
<tr>
<td>15</td>
<td>959</td>
<td>0.016285</td>
<td>Chengjun Zhan</td>
</tr>
<tr>
<td>16</td>
<td>957</td>
<td>0.016285</td>
<td>Lin Luo</td>
</tr>
<tr>
<td>17</td>
<td>956</td>
<td>0.016285</td>
<td>Guo Chen</td>
</tr>
<tr>
<td>18</td>
<td>955</td>
<td>0.016285</td>
<td>Xin Huang</td>
</tr>
<tr>
<td>19</td>
<td>943</td>
<td>0.016285</td>
<td>Sneh Gulati</td>
</tr>
<tr>
<td>20</td>
<td>960</td>
<td>0.016071</td>
<td>Sheng-Tun Li</td>
</tr>
</tbody>
</table>

### Top 20 authors with the highest betweenness scores

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Betweenness</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.0000752</td>
<td>Stuart Harvey Rubin</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.0000741</td>
<td>Shu-Ching Chen</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.0000406</td>
<td>Taghi M. Khoshgoftaar</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
<td>0.0000385</td>
<td>Xingquan Zhu</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.0000376</td>
<td>Mei-Ling Shyu</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.0000296</td>
<td>Reda Alhajj</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>0.0000256</td>
<td>Xindong Wu</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>0.0000194</td>
<td>Chengcui Zhang</td>
</tr>
<tr>
<td>9</td>
<td>39</td>
<td>0.0000185</td>
<td>Wei Dai</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>0.0000107</td>
<td>Narayan C. Debnath</td>
</tr>
<tr>
<td>11</td>
<td>31</td>
<td>0.0000094</td>
<td>Qianhui Althea Liang</td>
</tr>
<tr>
<td>12</td>
<td>151</td>
<td>0.0000094</td>
<td>Gordon K. Lee</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>0.0000085</td>
<td>Du Zhang</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
<td>0.0000072</td>
<td>Baowen Xu</td>
</tr>
<tr>
<td>15</td>
<td>41</td>
<td>0.0000067</td>
<td>Hongji Yang</td>
</tr>
<tr>
<td>16</td>
<td>270</td>
<td>0.0000060</td>
<td>Zhiwei Xu</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>0.0000043</td>
<td>Mohamed E. Fayad</td>
</tr>
<tr>
<td>18</td>
<td>110</td>
<td>0.0000042</td>
<td>Abhijit S. Pandya</td>
</tr>
<tr>
<td>19</td>
<td>106</td>
<td>0.0000042</td>
<td>Sam Hsu</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>0.0000042</td>
<td>Wen-Lian Hsu</td>
</tr>
</tbody>
</table>

## Top 20 authors with the highest degree scores

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Degree</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.035044</td>
<td>Shu-Ching Chen</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.034418</td>
<td>Stuart Harvey Rubin</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.030663</td>
<td>Taghi M. Khoshgoftaar</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.028786</td>
<td>Reda Alhajj</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.028786</td>
<td>Wen-Lian Hsu</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>0.024406</td>
<td>Min-Yuh Day</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.022528</td>
<td>Mei-Ling Shyu</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>0.021277</td>
<td>Richard Tzong-Han Tsai</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>0.017522</td>
<td>Eduardo Santana de Almeida</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>0.017522</td>
<td>Roumen Kountchev</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>0.016896</td>
<td>Hong-Jie Dai</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>0.015645</td>
<td>Narayan C. Debnath</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>0.015019</td>
<td>Jason Van Hulse</td>
</tr>
<tr>
<td>14</td>
<td>25</td>
<td>0.013767</td>
<td>Roumiana Kountcheva</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>0.013141</td>
<td>Silvio Romero de Lemos Meira</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>0.013141</td>
<td>Vladimir Todorov</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>0.013141</td>
<td>Mariofanna G. Milanova</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>0.013141</td>
<td>Mohamed E. Fayad</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>0.012516</td>
<td>Chengcui Zhang</td>
</tr>
<tr>
<td>20</td>
<td>18</td>
<td>0.011890</td>
<td>Waleed W. Smari</td>
</tr>
</tbody>
</table>

Visualization of IRI (IEEE IRI 2003-2010)
co-authorship network (global view)

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"
Visualization of Social Network Analysis

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"
Visualization of Social Network Analysis

Visualization of Social Network Analysis

Tamkang University

NTCIR-12, 2016
NTCIR-11, 2014
NTCIR-10, 2013
NTCIR-9, 2011
IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-9 RITE

Department of Information Management
Tamkang University, Taiwan

Min-Yuh Day
Chun Tu
myday@mail.tku.edu.tw

NTCIR-9 Workshop, December 6-9, 2011, Tokyo, Japan
IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-10 RITE-2

Department of Information Management
Tamkang University, Taiwan

Min-Yuh Day
Chun Tu
Hou-Cheng Vong
Shih-Wei Wu
Shih-Jhen Huang

myday@mail.tku.edu.tw

NTCIR-10 Conference, June 18-21, 2013, Tokyo, Japan
IMTKU Question Answering System for World History Exams at NTCIR-12 QA Lab2

Department of Information Management
Tamkang University, Taiwan

myday@mail.tku.edu.tw
教育部資通人才培育計畫
社群運算與巨量資料
課程四大模組

(1)「社群媒體」(Social Media) (政治大學)
(2)「資料科學」(Data Science) (政治大學)
(3)「分析技術」 (Analytics Technology) (高雄大學) (淡江大學)
(4)「領域應用」(Domain Application) (淡江大學) (政治大學)
1. 「社群媒體」(Social Media) (政治大學)

- 探討社群媒體和資料分析的概念，以個案方式教學
2. 「資料科學」 (Data Science) (政治大學)

- 探討 Data Thinking 和 EDA 等，與 DSP 或痞客邦合作
3. 「分析技術」 (Analytics Technology)
   (高雄大學) (淡江大學)

• 列舉重要的分析方法，包括社會網絡分析，文字探勘分析技術簡介。
  - * 社會網絡分析 (高雄大學)
  - * 社會網絡量測 (高雄大學)
  - * 社會網絡分析工具 (高雄大學)
  - * 文字探勘分析技術簡介 (淡江大學)
4. 「領域應用」 (Domain Application) (淡江大學) (政治大學)

• 區分 Domain Knowledge ，聚焦探討各種商業行銷和輿情分析等
  –* 社群媒體行銷分析 (淡江大學)
  –* 社群媒體情感分析 (淡江大學)
Summary

• Big Data Sentiment Analysis
• Social Computing
• International Research Collaboration and Mobility
References

  http://www.cs.uic.edu/~liub/WebMiningBook.html

• Jennifer Golbeck (2013), Analyzing the Social Web, Morgan Kaufmann. 
  http://analyzingthesocialweb.com/course-materials.shtml

• Sentinel Visualizer, http://www.fmsasg.com/SocialNetworkAnalysis/

References

  http://www.cs.uic.edu/~liub/WebMiningBook.html
- Bing Liu (2013), Opinion Spam Detection: Detecting Fake Reviews and Reviewers, 
  http://www.cs.uic.edu/~liub/FBS/fake-reviews.html
- Wiltrud Kessler (2012), Introduction to Sentiment Analysis, 
  http://www.ims.uni-stuttgart.de/~kesslewd/lehre/sentimentanalysis12s/introduction_sentimentanalysis.pdf
References

• Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In the 26th International Conference on Computational Linguistics (COLING), Osaka. 2016.


• Steven Struhl (2015), Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence (Marketing Science), Kogan Page

• Bing Liu (2015), Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press
Research on Social Computing and Big Data Analytics
(社群運算與大數據分析研究)

Time: 2016/11/17 (Thu) (15:30-17:30)
Place: 東吳大學資管研究所 <城中校區 教室 : 4303>
Host: 鄭麗珍 教授 (Prof. Li-chen Cheng)

Min-Yuh Day
戴敏育
Assistant Professor
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2016-11-17