



NEW MEDIA AND SENTIMENT MINING

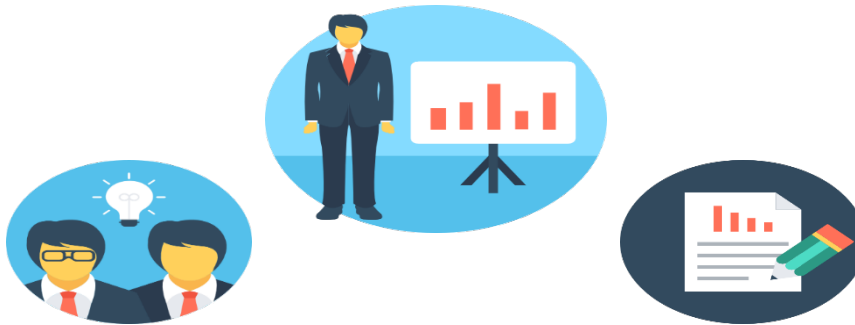
OVERVIEW OF SENTIMENT MINING

OVER
6,250 GRADUATE
ALUMNI

OFFERING OVER
150 ENTERPRISE IT, INNOVATION
& LEADERSHIP PROGRAMMES

TRAINING OVER
135,000 DIGITAL LEADERS
& PROFESSIONALS

Overview of sentiment mining



Eric Tham
Senior Lecturer & Consultant
isstyc@nus.edu.sg

Agenda

- Overview of sentiment mining
 - Definitions and levels of sentiment mining
- Industry applications
- Workshop on building a sentiment engine

Definition...

- *Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.*
- Generally speaking, sentiment mining seeks to obtain attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document.

Sentiment & Text Mining

- Text mining is a subset of data mining and is defined as the transformation of **unstructured data** into **answers** to **business questions**
- Sentiment mining largely understood as a classification technique presently.
 - Not just positive/ negative polarity though;
 - Can be expanded to include other text classification tasks.

Applications of sentiment mining

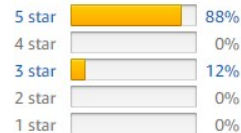
- Customer feedback and reviews
- Financial prediction
- Fraud detection of corporate misconduct
- Spam/ non spam
- Audit and legal analytics
- Political 'sentiment'
- Social credit system (new!)

1. Customers' reviews (eg Amazon)

Egs company websites, blogs, booking apps... (hungrygowhere, ieatigo ...)

Customer reviews

★★★★☆ 8
4.7 out of 5 stars ▾




Share your thoughts with other customers

Write a customer review

[See all 8 customer reviews ▸](#)

Top customer reviews

 Lucas N. Santos

★★★★☆ **Good overview on current topics**

January 15, 2009

Format: Hardcover | **Verified Purchase**

What I liked most about the book was the scratch I got when facing all the possibilities regarding data that is free available on the Internet. My interest area is crawling, and there is an exclusive chapter about it on the book. But as with all others chapters, it's only a bird's-eye view on the topic, so specifically the crawler part of the book wasn't of much use. In spite of it, my expectations were reached with the rest of the work, since I just wanted to be aware of what is happening today concerning Web data mining. I must note that, although chapters on relevant topics are small (more or less 30 and so pages) and surely don't cover all the nuances, the book comes with plenty of references for anyone who wants to dig further.

15 people found this helpful

Helpful

Not Helpful

Comment

Report abuse

 LOV

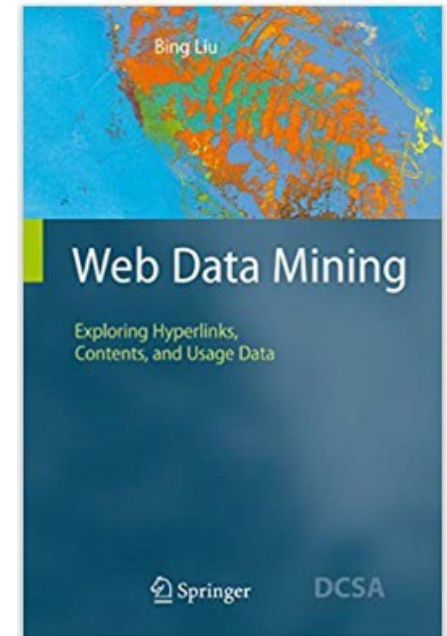
★★★★★ **A well-written book covering breadth and depth of web data mining**

February 9, 2015

Format: Hardcover | **Verified Purchase**

This is a very well-written book. It is also a highly ambitious project as the book covers breadth and depth of many web analytics and data mining/machine learning related topics. It is also written in a very accessible way but still delivers strong technical knowledge for technical audience. I don't think this book has much competition in this area (so far), and the author clearly is a real expert.

★★★★☆ ▾ 8 customer reviews
Look inside ↓



ISBN-13: 978-3540378815

ISBN-10: 3540378812

[Why is ISBN important? ▾](#)

2. Financial prediction

- Sentiment drives the stock market. It changes people's consumption habits which in turn drives their investment and savings tendencies. More on it on 4th day.

The Unbearable Lightness of Expectations of the Chinese Investor

Handbook of Sentiment Analysis in Finance (2015)

11 Pages • Posted: 11 May 2018

[Eric Tham](#)

EDHEC Business School

Date Written: November 5, 2015

Abstract

The Chinese equities markets witnessed wild swings in 2014-2015. The stock market's impact on the Chinese economy, and in turn on the Federal Reserve's interest rate policy is indirect but significant. The high internet penetration of the Chinese population - about 670 million and growth in retail trading accounts may make predispose the market to investor herding. In this paper, investor sentiment is separately derived through the textual analysis of newswires and the social blogs, which are the general types of information fed to rational arbitrageurs and retail noise traders respectively. Through a state space model of index returns on the two types of sentiment, it is shown that social blog sentiment and its time varying sensitivities are most accountable for the index swings in 2014/15. This sensitivity to the blog sentiment decreased in June 2015, correlating with less volatility in stock prices. It remains to be seen if the market is less sentiment driven now.

Keywords: Investor sentiment, State space model, Chinese stock market, Social media investing, Prospect Theory

JEL Classification: G02, C32, G11

Suggested Citation:

2. Financial - Credit prediction

- Start-ups are using digital footprint to predict credit profile of customers.
 - <https://www.economist.com/finance-and-economics/2016/11/17/just-spend>
 - <https://www.fdic.gov/bank/analytical/cfr/2018/wp2018/cfr-wp2018-04.pdf>
- Examples abound with Alipay, Sesame Credit in China and in Germany too



WORKING PAPER SERIES

On the Rise of the FinTechs—Credit Scoring using Digital Footprints

Tobias Berg
Frankfurt School of Finance & Management

Valentin Burg
Humboldt University Berlin

Ana Gombović
Frankfurt School of Finance & Management

Manju Puri
*Duke University
Federal Deposit Insurance Corporation
National Bureau of Economic Research*

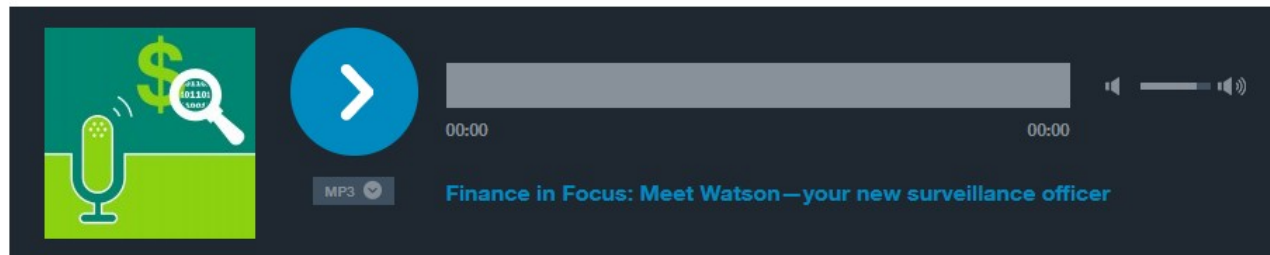
September 2018

3. Corporate misconduct

- Early warning of fraud and misconduct in chat messaging and emails.

Finance in Focus: Meet Watson—your new surveillance officer

NOVEMBER 11, 2016 | 12:29



Questionable behavior and employee misconduct have cost financial institutions more than [\\$200 billion in fines](#) and settlements since 2009. The extreme cost of this misconduct is proving to be a daunting challenge for financial services institutions. Although surveillance offers a means of protecting investors, consumers and firms from fraud and misconduct, current surveillance methods are falling short. Accordingly, financial institutions are searching for new ways of addressing this ever-present problem.

Moving surveillance out of the silo

Traditional surveillance systems are not working, in part because market manipulation techniques are constantly changing. Even



<https://www.ibmbigdatahub.com/podcast/finance-focus-meet-watson-your-new-surveillance-officer>

Enron emails corpus

- Good eg is the Enron corporate scandal.
- The open corpus (in Carnegie Mellon Univ.) has given rise to whole research study in NLP.
- Most of these research concludes that corporate scandals can be predicted through email correspondence ~
- Since then, it has resulted in a corporate culture shift in how companies communicate.

The Enron scandal drew attention to accounting and corporate fraud as its shareholders lost \$74 billion in the four years leading up to its bankruptcy, and its employees lost billions in pension benefits.

<https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/>

4. Audit and legal analytics

- Huge amounts of regulatory compliance
 - ~thousands new regulatory document. How to keep up? Opinion (clauses) mining needed.
 - Start-ups include Libryo



REGISTER NOW

DOWNLOAD
BROCHURE

DOWNLOAD
SPONSORSHIP
PROSPECTUS

CONTACT US

SUBSCRIBE TO US

About the Forum

Considered as the heartbeat of an organization, risk management and fraud management have the largest opportunity for incorporating and strengthening the use of AI. Clariden is proud to host the inaugural **Applying AI and Analytics For Better Risk, Fraud, Audit, Legal and Compliance Management Forum 2018** in Singapore from 27th – 29th August 2018. Join the stellar gathering of thought leaders and professionals from risk management.

Why You Must Join Us At This Forum

1. Stay ahead of the technology curve and boost your risk and fraud management capabilities in this disruptive innovation landscape
2. Learn and benchmark from organizations that have benefited from their successful AI deployment in their risk, fraud, compliance and legal framework

Featured Speakers



Chief Risk Officer
Singapore Pools



Eric Tham
Senior Lecturer & Consultant of
Analytics and Artificial Intelligence
National University of Singapore

4. Audit and legal analytics

- JP Morgan COIN (Contract intelligence)
 - <https://digital.hbs.edu/platform-rctom/submission/jp-morgan-coin-a-banks-side-project-spells-disruption-for-the-legal-industry/>
 - Review of monotonous repetitive legal documents. Save time and costs
- Helps that the credit contracts are large-scale and has low variability. May not work if it is too varied as per M&A or other documents.
- Works on the principle of entity and aspect that we will be studying in this course.

JP Morgan COIN: A Bank's Side Project Spells Disruption for the Legal Industry

By Legal ML

Student

POSTED NOV 13, 2018

Next:

[Empire State of Minds: Open Innovation in NYC](#)

5. Political sentiment

- US Presidential campaign – Trump vs Clinton

PACIS 2017 PROCEEDINGS

How Trump won: The Role of Social Media Sentiment in Political Elections

Chong Oh, *University of Utah*

Follow

Savan Kumar, *University of Utah*

Follow

Abstract

The outcome of the recent US Presidential Election of 2016 shocked and baffled many. Some claimed that social media may play a larger role in influencing the outcome that expected. This study examined Twitter messages containing political discussions with references to both Trump and Clinton to uncover insights about the role of social media sentiment in political elections. We adhere to the social media analytics (SMA) framework of Fan and Gordon (2014) and the sentiment analysis taxonomy of Abbasi, Chen, and Salem (2008) as a structure to extract positive and negative sentiment from the collected tweets during the pre-election period between Nov 3 and Nov 7. The first finding reveals that Trump has an overwhelmingly larger volumes of total, positive, and negative tweets over Clinton implying a higher volume of public discourse around Trump. Secondly, the propagation of negativism towards Clinton is much more than Trump although both candidates have increasingly more negative tweets days leading up to the Election Day of Nov 8. Finally, word clouds for both candidates reveal that the Twitter public are engrossed with more negative topics against Clinton than Trump. This study clarifies the role of social media sentiment, specifically in how Trump

Download

235 DOWNLOADS

Since September 11, 2017

PLUMX METRICS

SHARE



<https://aisel.aisnet.org/pacis2017/48/>

5. Political sentiment (cont'd)

- Russia's role in social media in Trump campaign using bots, fake Twitter accounts & FB advertisements
- Uses IRA (Internet Research Agency), friends of Russian Intelligence to increase clicks on target.

≡ FORTUNE

LEADERSHIP • DONALD TRUMP

How Russians Used Social Media to Boost the Trump Campaign, According to Robert Mueller's Indictment



6. Social Credit System

CHINA'S SOCIAL CREDIT SYSTEM

It's been dubbed the most ambitious experiment in digital social control ever undertaken. The Chinese government plans to launch its Social Credit System nationally by 2020.

WHAT'S THE AIM?

The system intends to monitor, rate and regulate the financial, social, moral and, possibly, political behavior of China's citizens - and also the country's companies - via a system of punishments and rewards. The stated aim is to "provide the trustworthy with benefits and discipline the untrustworthy."

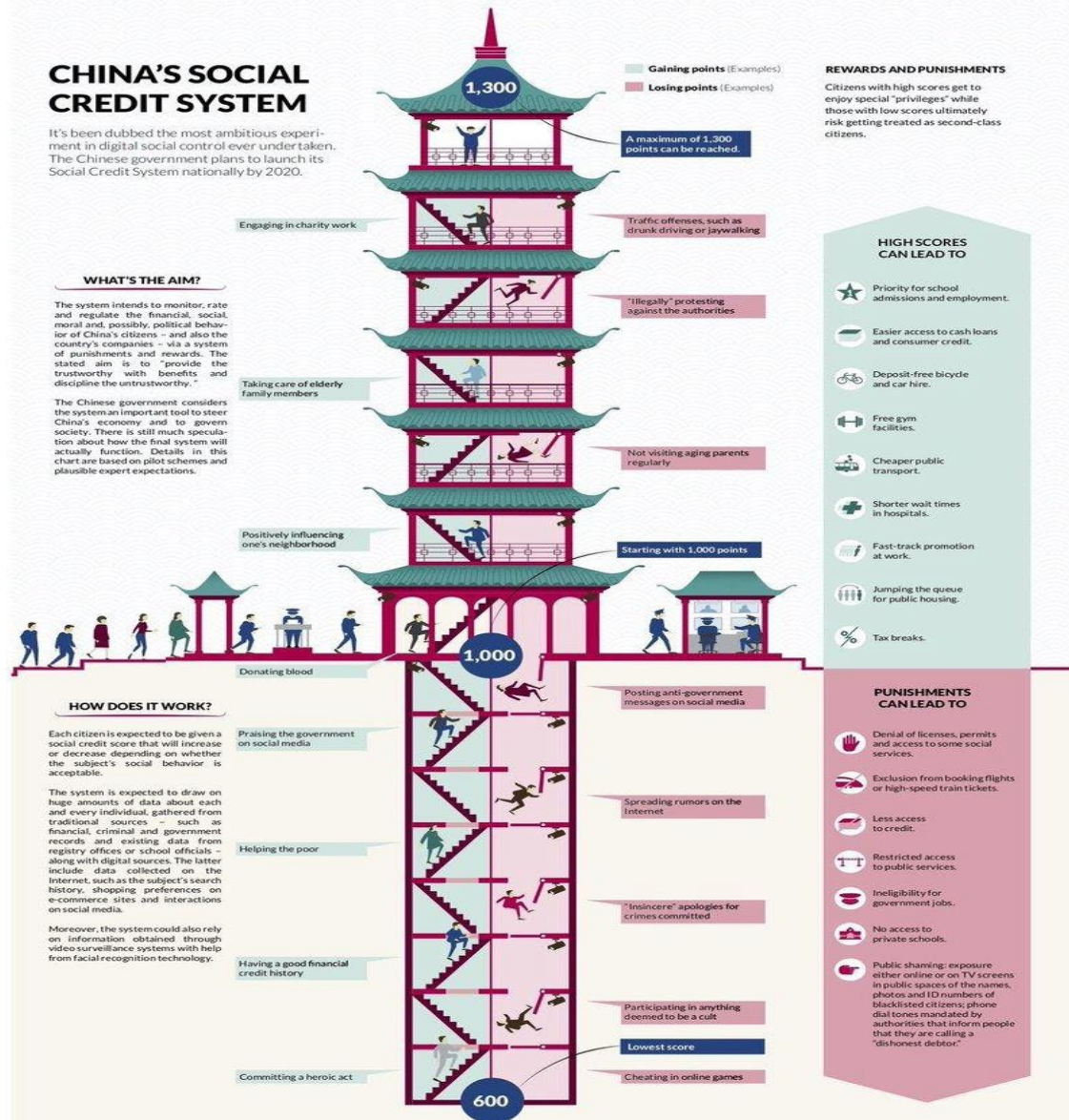
The Chinese government considers the system an important tool to steer China's economy and to govern society. There is still much speculation about how the final system will actually function. Details in this chart are based on pilot schemes and plausible expert expectations.

HOW DOES IT WORK?

Each citizen is expected to be given a social credit score that will increase or decrease depending on whether the subject's social behavior is acceptable.

The system is expected to draw on huge amounts of data about each and every individual, gathered from traditional sources - such as financial, criminal and government records and existing data from registry offices or school officials - along with digital sources. The latter include data collected on the Internet, such as the subject's search history, shopping preferences on e-commerce sites and interactions on social media.

Moreover, the system could also rely on information obtained through video surveillance systems with help from facial recognition technology.



TEXT: BERTELSMANN STIFTUNG; ILLUSTRATION: CHRISTOPHER CHEN; PHOTO: GETTY IMAGES/SHUTTERSTOCK

BertelsmannStiftung

6. Social Credit System (cont'd)

- Form of social mass surveillance starting in 2020
 - How will social media be used?
- Pros and cons
 - Data privacy and others? compared to the West?
 - Improve social behaviour?
- How will it be used?
 - _____
 - _____
 - _____
- Applies to businesses too

6. Writing marketing ads (and news)

- Persado start-up: AI copywriter (emotive words)
 - Part of [natural language generation \(NLU\)](#)
- [News article generation](#)
 - TYSONS CORNER, Va. (AP) — MicroStrategy Inc. (MSTR) on Tuesday reported fourth-quarter net income of \$3.3 million, after reporting a loss in the same period a year earlier.
 - MANCHESTER, N.H. (AP) — Jonathan Davis hit for the cycle, as the New Hampshire Fisher Cats topped the Portland Sea Dogs 10-3 on Tuesday.
- [Robotic Reporter](#): automatic news reporting

An innocent bystander was murdered in cold blood in Downtown Chicago. The words “innocent” and “murdered” and the phrase “in cold blood” are the uses of emotive language in this sentence.

What are the key aspects that you can identify to such automated writing?

A copywriter is someone who is paid to write “copy” – words designed to prompt action. Copywriting is always connected to the act of promoting or selling a business, organization, brand, product, or service, which makes it, by definition, a form of marketing.

6. Writing marketing ads (and news)

Emotional:

The Persado Message Machine identifies which words and phrases evoke specific emotions and understands which emotions deliver the strongest appeal.

Formatting:

Whether it's a matter of using bold versus an italicized font or optimizing images, Persado can identify differences in formatting then find the most powerful combination to maximize engagement.

Positioning:

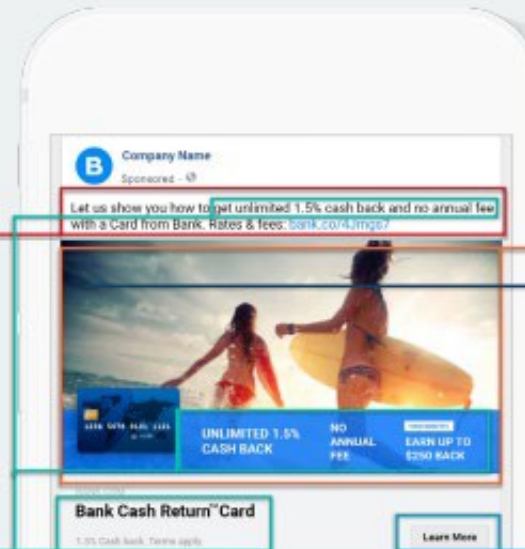
Changing the position of a particular word or phrase can make a difference in the performance of marketing creative.

Descriptive:

Descriptions are specific words and phrases related to product classifications, promotions, offers and services, and influence marketing campaign engagement.

Functional/CTA:

This refers to any part of the creative or message that directs the audience to take a specific action; a button that says '**Apply now**', text that includes '**call 1-800-###-####**', or a link to '**find a store**' are all calls to action.



Definition of sentiment mining

- Core tasks of sentiment mining. To be able to extricate quadruple (**o**, **t**, **t**, **s**)
- These are:
 - Opinion holder**: who expressed the opinion (o)
 - Opinion target**: Entity + (optionally) aspect about which the opinion was expressed (t)
 - Score**: +ve or -ve, or some *scale* & granularity (s)
 - Time**: when the opinion was expressed (t)

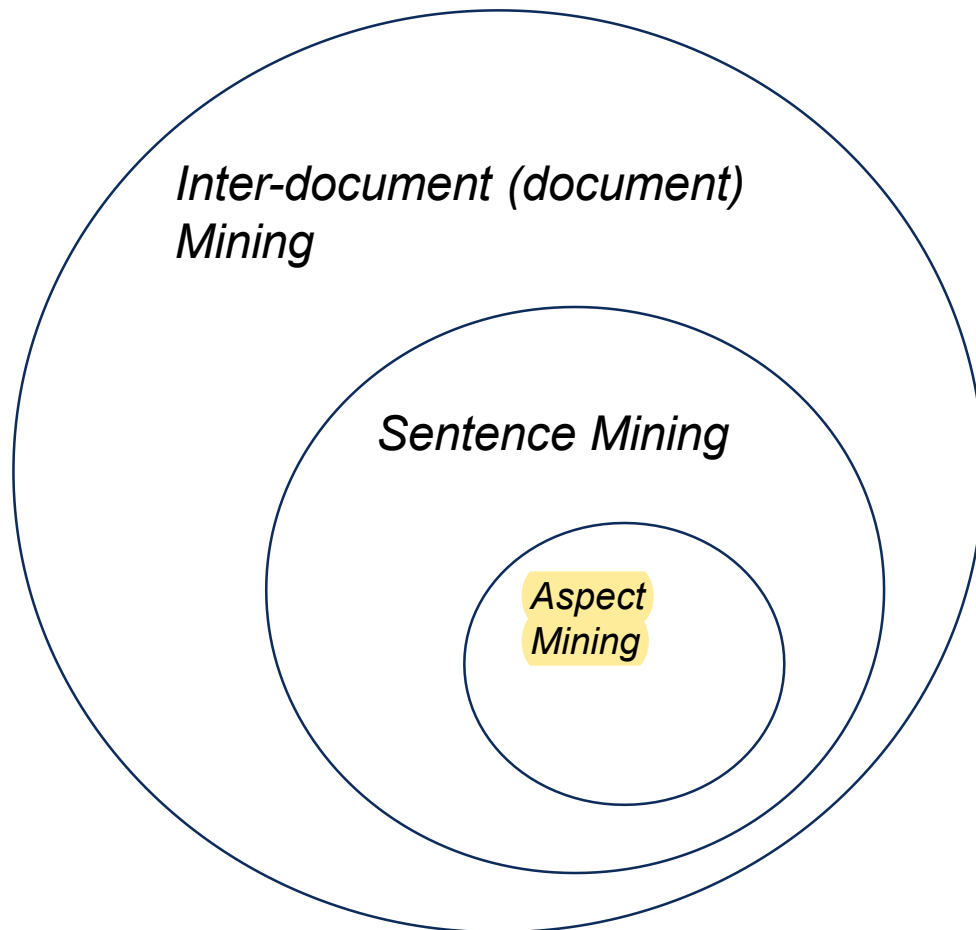
Data granularity is the level of detail considered in a model or decision making process or represented in an analysis report.

The time aspect (t) and opinion holder (o) are important when it comes to sentiment aggregation.

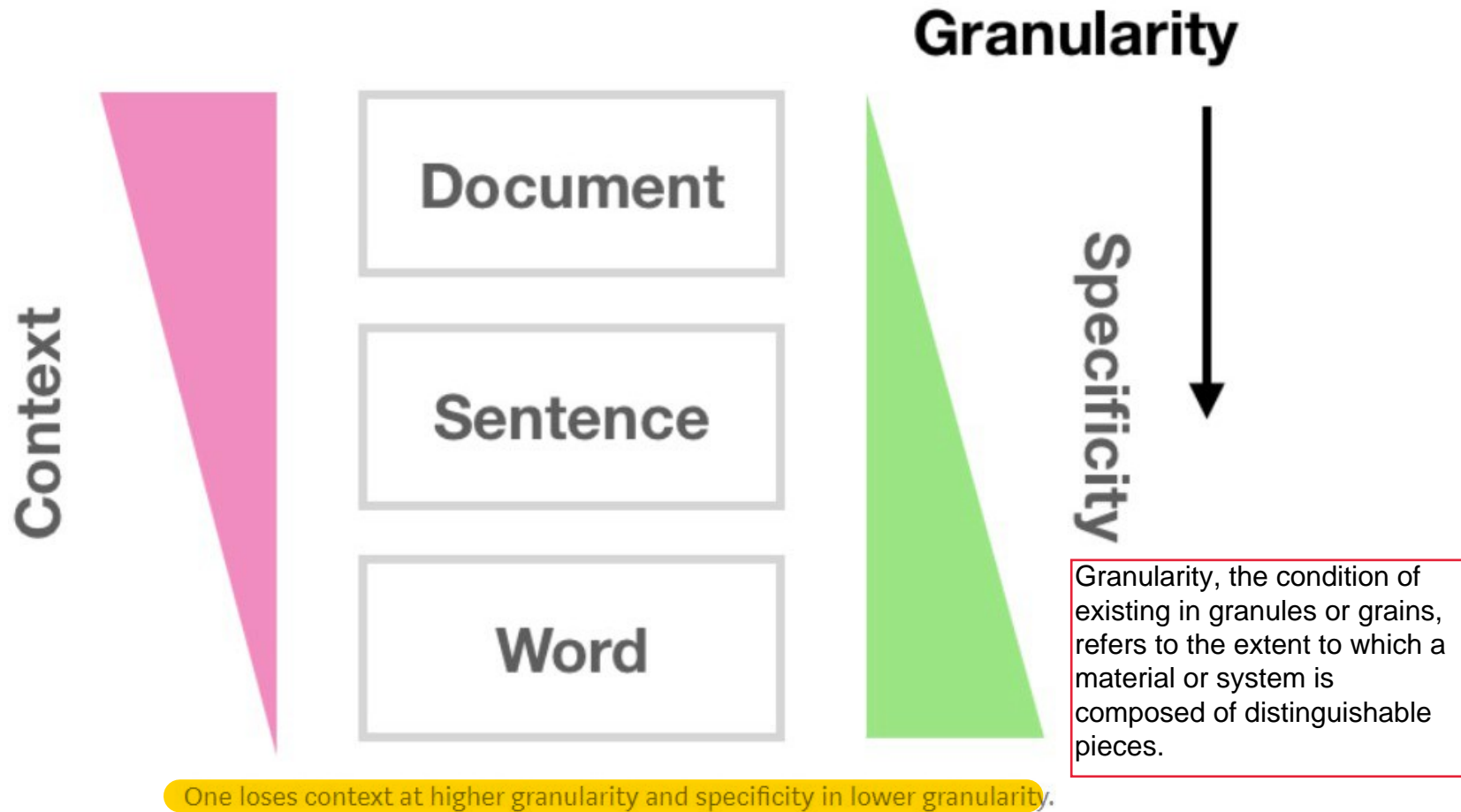
Why is sentiment analysis hard?

- Language nuances
 - **Rational Evaluation** express no emotion
 - Evaluation of aspects/features: how to *quantify* then?
 - E.g., “The camera takes 2 secs to start up”
 - Contrast with “The camera starts up very fast.”
 - **Emotional words** may not express a sentiment
 - E.g., “*Amazingly*, the stewardesses treat is as normal”
 - E.g., “The phone is an *amazing* birthday present”
 - Slang! Eg. can...
- Sarcasm/ irony
 - What do you mean?
 - What are you trying to say?
 - Donald Trump’s tweets
- Often a mix of intuition and science to do a good sentiment analysis engine. Every system is bespoke. made for a particular customer or user

Main levels of sentiment mining



Main levels of sentiment mining



Information Retrieval vs Sentiment Mining

- Information extraction (retrieval) vs entity and aspect mining?
 - What is the difference?
- Information retrieval – algorithms like PageRank amidst a sea of documents
- Entity and aspect mining – from a document/ paragraph/ phrasal level to determine relationships amongst the entities, their features and the adjoining entities.
- But the distinction is blurred at times. IE can be crafted at a (lower) paragraph level as well.

(Inter-) Document level sentiment mining

- Documents include blogs, articles, news etc. usually longer.
- Key issues include ‘too many items talked about’ and how to ‘link them’ together. Not wise to allocate a single polarity to the whole context of document.
- An approach is to model the general topic, or unearth the individual entities and relationships in document and do granular polarity assignment.

- Some domain problems are easier than the rest
 - Eg restaurants, movies and hotels because they are much studied and applied
 - Can you transfer know-how from one domain to another? Not with extreme difficulty!

In computational linguistics literature, ~88% accuracy in SemEval competitions is very good.

- Availability of meta-data
- Opinion spam
- Fake news
- Influencers
- Virality of posts

1. Availability of meta-data

- New media is about user generated content (UGC) data
- Meta-data includes 'likes', ratings (overall sentiment), author, author, date/ time etc. Use them.

Reviews at Wild Honey

[Add Review](#) 



CHUBBY BOTAK KOALA • 07 Oct 2011 • 406 Reviews

Around the world with Breakfast



This place has been buzzing with review since it opened in late 2009. During the weekend, it is notoriously famous for long waiting time. 76 reviews in HGW and counting, but this does not discourage me in penning my 2 cents. ...

[Read more](#)



Must Tries: English breakfast, Scandinavian breakfast

2. Opinion Spam (Fake) Detection

- Opinion spam refers to **non-genuine** opinion found on websites, reviews, etc.
 - Undeserving positive reviews – promote products
 - Malicious negative reviews – damage reputations
 - In general, “reviews” (positive or negative) where the “opinion holder” did not honestly and actually review the product.
 - E.g., a **positive** review of a **very good** book by someone who had not read the book would be opinion spam
- Opinion spam is a business
 - Pay for positive reviews for own business, negative reviews for competitors

Is this review fake or not?

I want to make this review in order to comment on the excellent service that my mother and I received on the Serenade of the Seas, a cruise line for Royal Caribbean. There was a lot of things to do in the morning and afternoon portion for the 7 days that we were on the ship. We went to 6 different islands and saw some amazing sites! It was definitely worth the effort of planning beforehand. The dinner service was 5 star for sure. One of our main waiters, Muhammad was one of the nicest people I have ever met. However, I am not one for clubbing, drinking, or gambling, so the nights were pretty slow for me because there was not much else to do. Either than that, I recommend the Serenade to anyone who is looking for excellent service, excellent food, and a week full of amazing day-activities!

How about this review?

This restaurant is good. The price is very reasonable, and so is the service. The location is also very good with good transportation. Oh my, I really loved the place and will come again. The food is also very good, very yummy and delicious. What shall I say more? I like the scenery and the ambience. Come again.

<https://www.inc.com/jessica-stillman/heres-how-to-spot-fake-online-reviews-with-90-perc.html>

<https://www.cs.uic.edu/~liub/FBS/fake-reviews.html>

2. Opinion Spam (Fake) Detection

- Most spam reviews detector works for only reviews generated by AI.
- Not quite work if it is written by humans. Hard to tell if he or she actually use the product (unless verified buyer)

3. Fake News in Social Media

Fake news often sensationalise which can impact people's opinions greatly.

Fake News: Lies spread faster on social media than truth does

People are quicker to repeat something that's wrong than something that's true

by Maggie Fox / Mar.09.2018 / 3:05 AM ET / Updated Mar.09.2018 / 8:51 PM ET

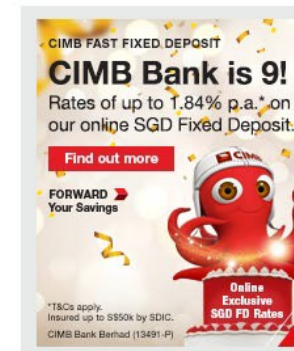
Singapore's proposed fake news law worries social media players

🕒 March 23, 2018 👤 News Centre 📁 News 💬 0

4. Influencers in social media

- Influencers have many followers setting trends in what people think or feel.
- Impacts sentiment aggregation – a core task of sentiment mining

Environment ministry pays social media influencers to spread word on climate change



5. Virality in social media

- Posts in social media go viral, **Bitcoin's soaring value was down to 'infected' buyers, economists say**

Barclays analysts compare speculation in digital currency to spread of infectious disease

- Eg in cryptocurrency trading – of 2017 was the huge amount posts.



▲ Economists said Bitcoin's peak before Christmas was probably the ultimate price that could ever be achieved.
Photograph: Dado Ruvic/Reuters

The rise of bitcoin has comparisons with the spread of an infectious disease, according to economists who argue the digital currency may have peaked in value as more consumers become immune to its appeal.

Summary

- Sentiment mining is a subset of text mining. It can be done at a document, sentence or even more granular aspect level.

- The issues regarding sentiment mining are non-trivial including most NLP-related tasks and other domain area issues.

Since the zero solution is the "obvious" solution, hence it is called a trivial solution. Any solution which has at least one component non-zero (thereby making it a non-obvious solution) is termed as a "non-trivial" solution.

- Its uses are extensive as long as there is text – in customer analytics, financial areas etc.
- New media has its own nuances in impacting sentiment – influencers, spam, virality and fake news.

References (Introduction)

- Liu Bing, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, May 2012, available for free download at <http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.html>

Recommended Readings

- Bo Pang & Lilian Lee, Opinion Mining and Sentiment Analysis, in *Foundations and Trends in Information Retrieval* Vol 2 No 1-2 (2008). Prepublication version available for free download at <http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>
- Liu Bing, Sentiment Analysis and Subjectivity, in *NLP Handbook*, 2nd Ed., eds: N.Indurkha & F.J.Damerou, 2010. Draft available for download at <http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf>

These references mainly describe the nuances that are particular to sentiment mining and are relatively 'outdated' (before ~2013). Still there are important principles to learn from it.