Lecture 15:
Sentiment Analysis (aka Opinion Mining)

John Cartlidge

Department of Computer Science
University of Bristol

john.cartlidge@bristol.ac.uk

Sentiment Analysis

The process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, or service, etc., is positive, negative, or neutral.

Sentiment Analysis

The process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, or service, etc., is positive, negative, or neutral.

Sentiment analysis is important because, in the Internet age,…

- Opinions matter
- A lot!
- And everybody has one, and can easily tell everybody else

Opinions matter… a lot!

- 2015 - Competition and Markets Authority (CMA) estimates £23.31 billion a year of UK consumer spending is influenced by online reviews.
  - Travel & hotels (£14.38 bn)
  - Home improvements (£3.93 bn)
  - Electronic items (£3.13 bn)
  - Beauty and male grooming (£1.01 bn)
  - CDs, DVDs, & music (£0.5 bn)
  - Books (£0.36 bn)
- The UK accounts for roughly 2% of global GDP, so global consumer spending influenced by online reviews is likely to be many hundreds of billions of pounds / year


Online reviews and endorsements

Report on the CMA's call for information

https://www.gov.uk/government/consultations/online-reviews-and-endorsements

19 June 2015
CMA41

Reviews, Reputation, and Revenue: The Case of Yelp.com

Do online consumer reviews affect restaurant demand?

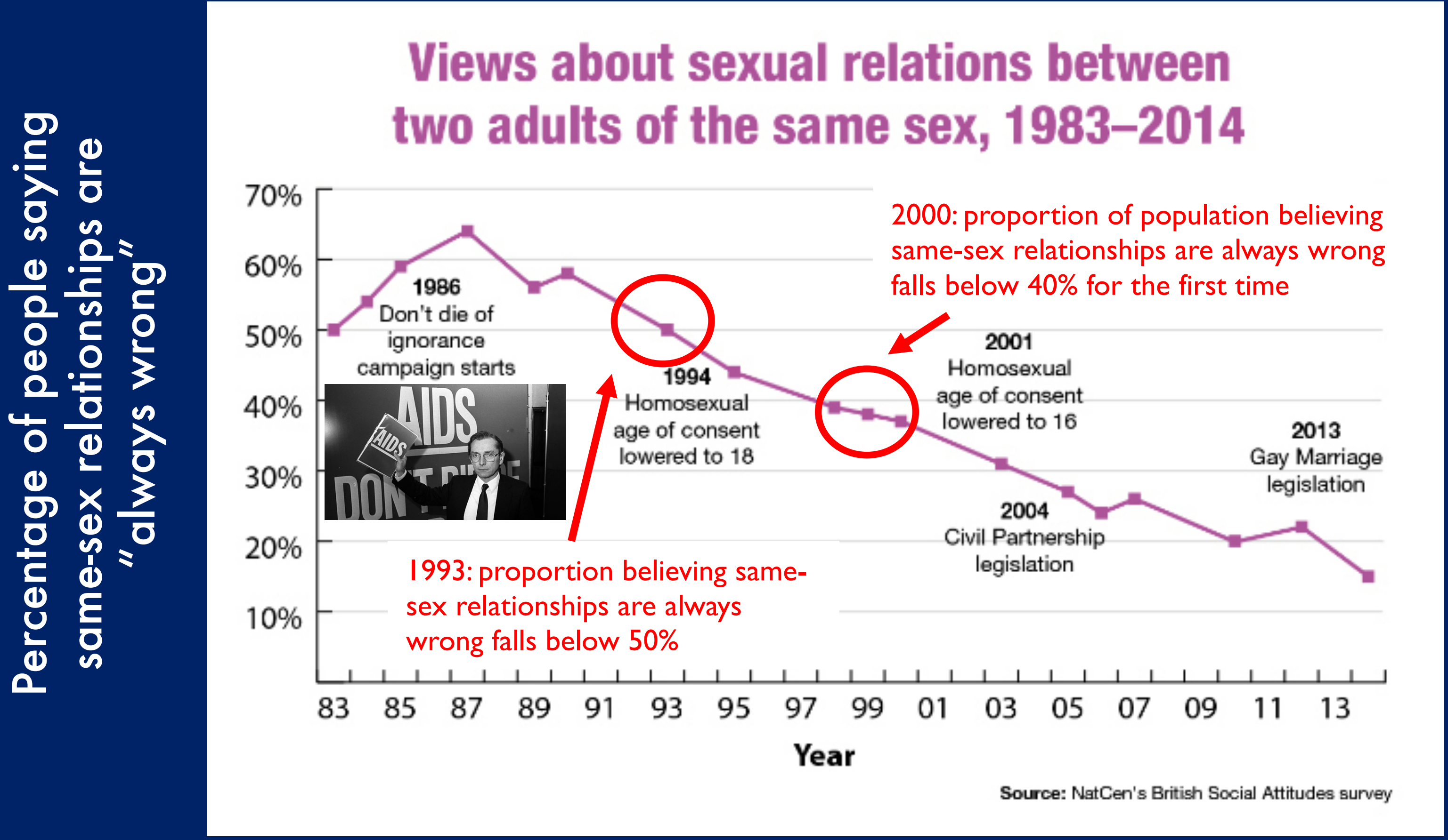
1. A one-star increase in Yelp rating leads to a 5-9% increase in revenue.
2. Effect is driven by independent restaurants; ratings do not affect restaurants with chain affiliation.
3. Chain restaurants have declined in market share as Yelp penetration has increased, suggesting that online consumer reviews substitute for more traditional forms of reputation.
4. Consumer response to a restaurant's average rating is affected by the number of reviews and whether the reviewers are certified as "elite" by Yelp.

Luca, Michael. "*Reviews, Reputation, and Revenue: The Case of Yelp.com*", Harvard Business School Working Paper, No. 12-016, September 2011. (Revised March 2016.)

https://www.hbs.edu/faculty/Pages/item.aspx?num=41233

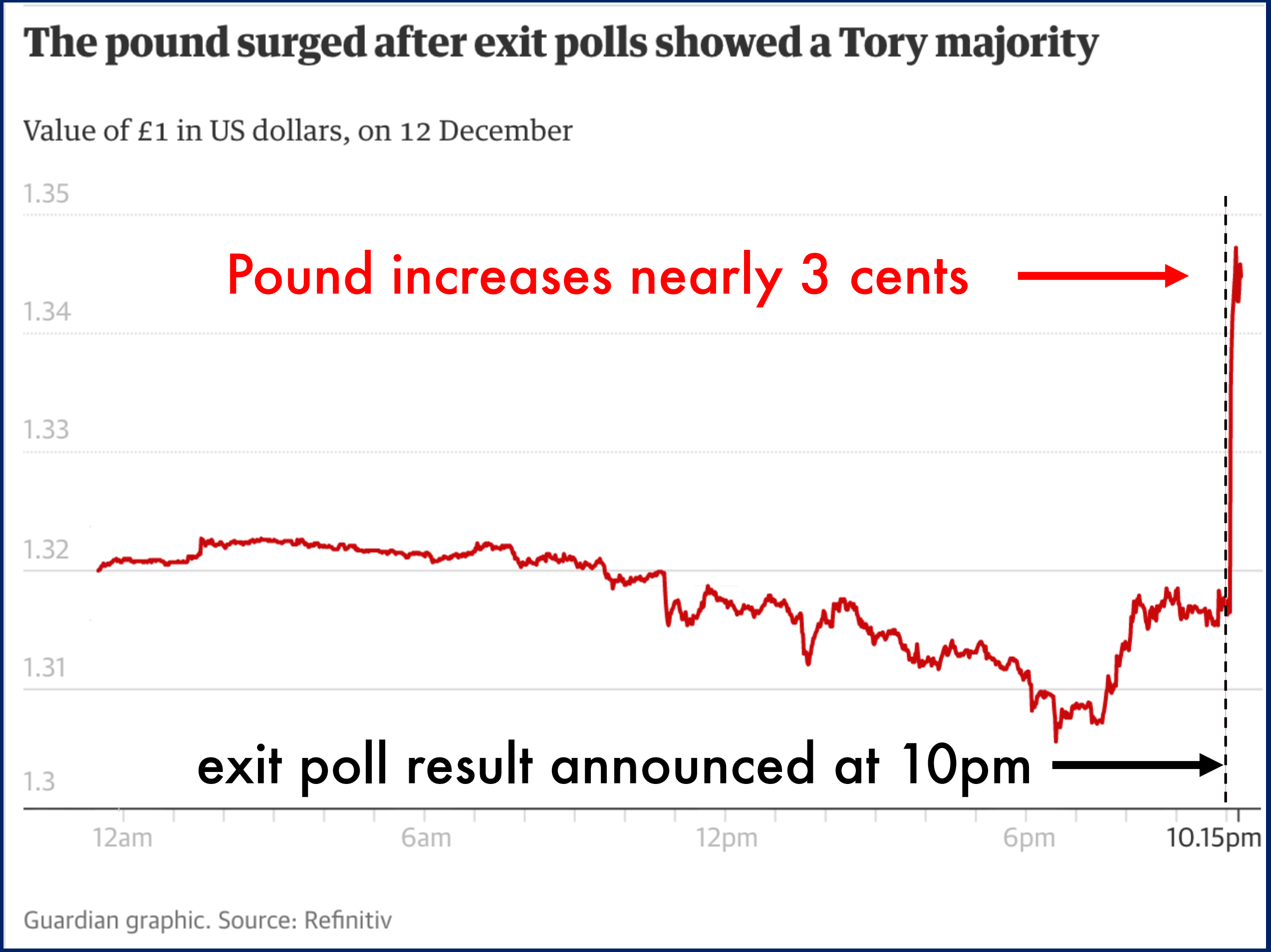
Public opinion: driving policy in the UK

The British Social Attitudes (BSA) survey: how public opinion drives policy in the UK



https://blogs.lse.ac.uk/impactofsocialsciences/2017/01/31/the-british-social-attitudes-survey-how-public-opinion-drives-policy-in-the-uk/

Financial markets: reaction to exit poll



UK General Election, Dec 12th 2019.

Kylie Jenner Tweets; *and the World Listens*

Kylie Jenner
@KylieJenner

29.7M followers

sooo does anyone else not open Snapchat anymore? Or is it just me... ugh this is so sad.

9:50 PM · Feb 21, 2018 · Twitter for iPhone

66K Retweets **351.3K** Likes

SNAP (SNAP) STOCK NYSE

+ ADD
SHARE

6.92 USD -0.56 (-7.49%) 03:23:04 PM EDT BTT

Prev. Close	7.48	Market Cap (USD)	7.77 B	Day Low	Day High	52 Week Low	52 Week High
Open	7.29	Volume (Qty.)	1,940,291	6.92	7.40	6.92	22.15

larkets Insider

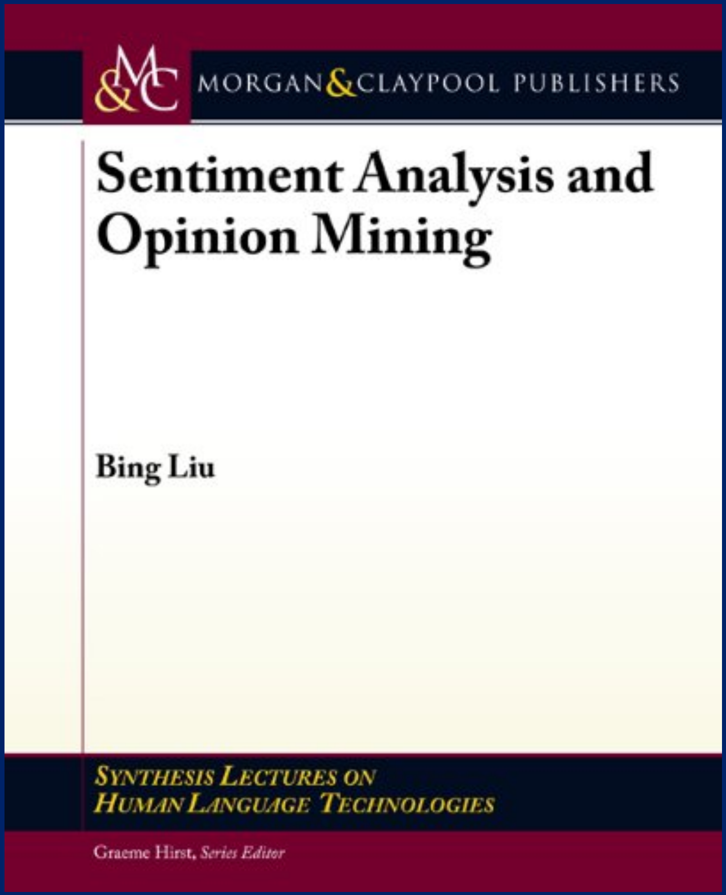
- Snap shares fell more than 7% Tuesday and have plunged about 63% since Kylie Jenner tweeted her displeasure with the Snapchat app's redesign.

Sentiment Analysis (aka Opinion Mining)

Make use of all the valuable online data: blogs, tweets, reviews, etc., to understand feelings, views, or opinions they express.

- For rapid business decision making (no need to conduct surveys and focus groups);
- For tracking customer concerns (indirect customer feedback);
- Spotting opinions that may impact a stock price;
- Predicting box-office performance of a new movie

A more difficult challenge than "topic modelling", which is identifying the topic of a document through text classification. Topics are often identifiable through keywords; while sentiment is often expressed in a more subtle manner.



Bing Liu, 2012, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers

[Pre-print available] https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Sentiment: Three levels of analysis

- Document-Level

  For a given document, e.g., a review, identify the overall attitude to the object under discussion.

  MacBook review: positive sentiment (polarity = 0.6)

- Sentence (or Phrase) Level

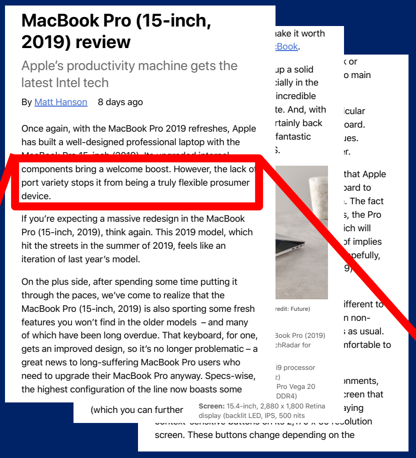
  For a given sentence, identify if it expresses a positive, negative or neutral (no) opinion

  Sentence: negative sentiment (polarity = -0.5)

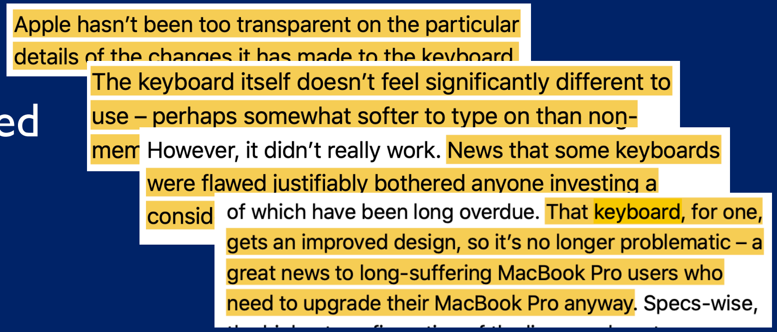
- Aspect (or Feature) Level

  For a given document, identify all opinions expressed regarding any aspect of any object.

  “Keyboard”: positive sentiment (polarity = 0.3)



However, the lack of port variety stops it from being a truly flexible prosumer device.



Document Level Sentiment Analysis: A Supervised Learning Example

- In 2002, Pang et al. examined the effectiveness of applying supervised machine learning techniques to the sentiment classification problem
  - Domain: movie reviews on Internet Movie Database (IMDb)
  - Aim: to classify a review as 'positive' or 'negative' towards its target
  - Supervised Learning: requires a tagged corpus, (a "ground truth") to learn from. IMDb allows reviewers to express their overall opinion using a 5-star rating. For the training corpus, reviews are tagged as positive if receiving 4-5 stars, and negative if receiving 1-2 stars. Therefore, no hand-labelling required. Neutral (3 star) reviews were ignored.
  - Features: "bag of words" approach. Simply use the frequency (i.e., the count), or presence (i.e., a binary flag 1 or 0) of words in the review as input (context-free).
  - ML Methods: Naïve Bayes (NB), Support Vector Machine (SVM)
  - Performance Accuracy: NB (81% best), SVM (82.9% best). Better than performance of a human-generated feature set (69% best). But not as high as reported on other non-opinion machine learning tasks, due to the difficulty of sentiment classification

---

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. EMNLP '02, Vol. 10. https://www.aclweb.org/anthology/W02-1011.pdf

Document Level Sentiment Analysis: An Unsupervised Learning Example (1/2)

- First introduced in 2002 by Turney, who examines the effectiveness of applying unsupervised machine learning techniques to the sentiment classification problem
  - Domain: reviews for banks, automobiles, movies, and travel destinations
  - Aim: to classify a review as 'thumbs-up' (+ve) or 'thumbs-down' (-ve) towards its target
  - Unsupervised Learning: No tagged/labelled data to learn from in a supervised fashion, so probabilities of word and phrase (co-)occurrences calculated using search engine results. Classification is predicted using *semantic orientation (SO),* calculated as *mutual information* between given phrases in the review and the word "excellent", minus the mutual information between the same phrases and the word "poor" (details on next slide…).
  - Performance Accuracy: Very high for banks (84%) and automobiles (80%). Not so good for movies (66%). *"It appears that movie reviews are difficult to classify, because the whole is not necessarily the sum of the parts… On the other hand, for banks and automobiles, it seems that the whole is the sum of the parts. Travel reviews are an intermediate case"*

Turney (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. https://www.aclweb.org/anthology/P02-1053.pdf

Document Level Sentiment Analysis: An Unsupervised Learning Example (2/2)

1. Extract *phrases* (consecutive words) containing adjectives or adverbs (which indicate subjective evaluation) followed by a noun (which provide context): e.g,: 'beautiful sound'; 'noisy engine', 'unpredictable plot')
2. Pointwise Mutual Information (PMI) for a pair of words calculates how strongly those words are associated semantically (PMI = 0 if words are independent)
3. Semantic Orientation (SO) of each *phrase* is calculated using PMI of the *phrase* with reference words "excellent" (often used in 5-star reviews) and "poor" (1-star reviews)
4. SO is estimated by issuing queries to a search engine, noting the number of hits (matching documents) where *phrase* is NEAR (within 10 words) to reference words "*excellent*" and "*poor*" (this is the log-odds ratio)
5. Calculate average SO of *phrases* in a review and classify as *thumbs-up* if average is +ve, *thumbs-down* if average is –ve

$$\mathrm{PMI}(word_1, word_2) = \log_2\left(\frac{\mathrm{p}(word_1 \;\&\; word_2)}{\mathrm{p}(word_1)\,\mathrm{p}(word_2)}\right)$$

$$\mathrm{SO}(phrase) = \mathrm{PMI}(phrase, \text{“excellent”}) - \mathrm{PMI}(phrase, \text{“poor”})$$

$$\mathrm{SO}(phrase) = \log_2\left(\frac{\mathrm{hits}(phrase\ \mathrm{NEAR}\ \text{“excellent”})\,\mathrm{hits}(\text{“poor”})}{\mathrm{hits}(phrase\ \mathrm{NEAR}\ \text{“poor”})\,\mathrm{hits}(\text{“excellent”})}\right)$$

Turney (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. https://www.aclweb.org/anthology/P02-1053.pdf

Aspect Level Sentiment Analysis

- A document may contain multiple expressions of sentiment regarding different 'aspects' (or features) of one or more objects.
- A "bag of words" (BoW) approach, which treats words as independent features (and therefore ignores context) will not perform well on the following:
  - *"The voice quality of this phone is amazing, but it is rather expensive."*
  - Object: phone, Aspects: (voice quality, +ve) (cost, -ve)
  - Two aspects with opposing sentiment, BoW will average to give neutral sentiment.

  - *"Apple is doing very well in this bad economy"*
  - Object: Apple Inc., Aspects: (general, +ve)
  - Object: Economy, Aspects: (general, -ve)
  - Two objects, each with different sentiment. BoW will not distinguish.

  - "*The photo quality is not good."*
  - Object: unknown, Apsects: (photo quality, -ve)
  - "not" is a sentiment shifter, BoW will likely be more influenced by "good".

Aspect Level Sentiment Analysis

Abstract problem:

For a given document (corpus), identify every quintuple of sentiment expressed:

1. The object

   (e.g., 'Apple iPhone')

2. The aspect

   (e.g., 'Price', 'Audio Quality')

3. The sentiment expressed

   (+, -, possibly a measure of intensity)

4. Who holds the sentiment
5. When it was expressed

Problem decomposition:

Aspect Extraction:

'The voice quality of this phone is amazing.'

-> (Phone, 'voice quality')

'I love this phone!'

-> (Phone, GENERAL)

Grouping Aspects into Categories

Find *synonyms* for aspects

(e.g., "battery life" and "battery power")

Aspect Sentiment Classification

(any approach – example next slide…)

Extract Entity, Opinion Holder, Time

Metadata can help (e.g., Tweet timestamp)

Aspect Level Sentiment Classification

A simple algorithm:

1. Mark sentiment words and phrases using a lexicon.
   - Positive words/phrases assigned +1, negatives -1.
2. Identify sentiment shifters: (e.g., ‘not’, ’never’, ‘cannot’)
   - Swap sentiment values for shifted words.
3. Identify ‘but’ phrases.
   - Heuristic states that if the sentiment on one side cannot be identified, it is considered opposite to the other side.
4. Sum sentiment scores, weighted by word distance from aspect word
   - Gives overall +/- sentiment.

Example Applications of Twitter Sentiment

1. Predicting box office revenues of movies using Tweet rates (number of tweets about a movie per hour, in week prior to movie release)
   - Liner regression model has very strong correlation with gross returns ($R^2 = 0.8$)
   - Predictions outperform the Hollywood Stock Exchange (HSX) information market
   - Sentiment content of tweets only improves revenue prediction *after* movie release

   – (Asur and Huberman, 2010)

2. (Not?) Predicting election results using Twitter sentiment
   - 2009 German federal election results successfully predicted by associating sentiment towards politicians as indicating voters' political preference - (Tumasjan et. al., 2011)
   - However, the same method was applied to 2010 US congressional elections, finding no correlation between voters' preference and electoral outcome, leading to suspicion that the original result was a lucky coincidence - (Gayo-Avello et al. (2011)

- Asur and Huberman, 2010, Predicting the future with social media. https://arxiv.org/pdf/1003.5699.pdf
- Tumasjan et al. (2010) Election forecasts with Twitter: How 140 characters reflect the political landscape. http://dx.doi.org/10.2139/ssrn.1833192
- Gayo-Avello et al. (2011), Limits of Electoral Predictions Using Twitter. https://pdfs.semanticscholar.org/3ef4/3622a68e717228ce0698246e427339a4bbdc.pdf

Example Applications of Twitter Sentiment

3. Predicting stock market movements using Twitter "*mood*"
   - Tagged tweets during 2008 with six mood states: Calm, Alert, Sure, Vital, Kind, and Happy. The topic of the tweet is ignored (i.e., tweets may not mention stocks or trading, etc.)
   - Average across all tweets to generate daily scores for each mood state, and compare with directional movements in the Dow Jones Industrial Average (DJIA) over the next 7 days
   - Identified correlation between "Calm" and a rise in DJIA 3 days later. Reported directional accuracy of 86.7%! (Data set was never released)

– Bollen et al. (2011)

- Bollen, Mao, and Zeng (2011), Twitter Mood Predicts the Stock Market. Journal of Computational Science 2(1): 1–8. https://arxiv.org/pdf/1010.3003.pdf
- Lachanski and Pav (2017), Shy of the Character Limit: "Twitter Mood Predicts the Stock Market" Revisited. https://risk.princeton.edu/img/Lachanski2017.pdf

Example Applications of Twitter Sentiment

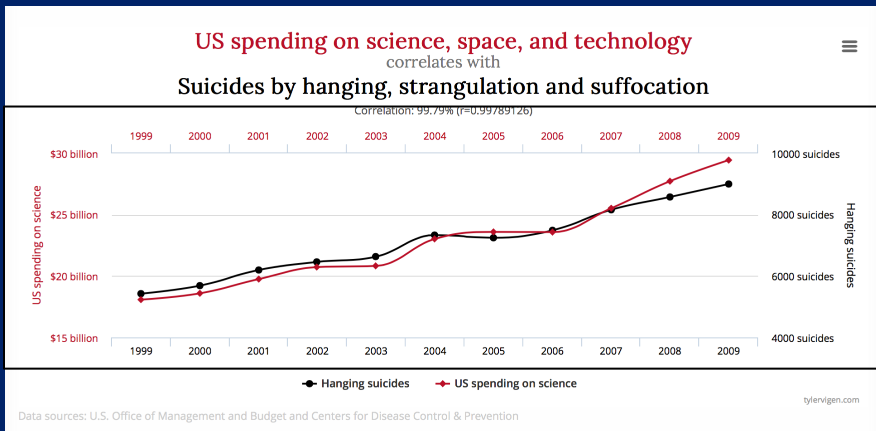
3. Predicting stock market movements using Twitter "*mood*"

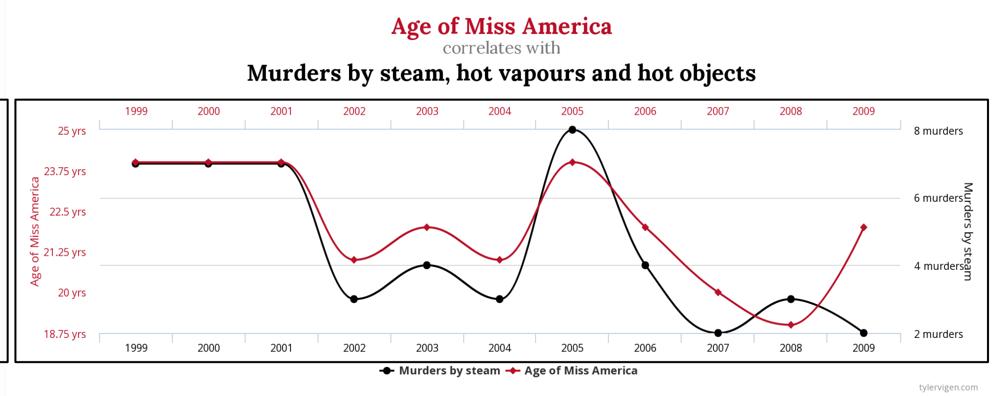
- Soon after, London based investment boutique Derwent Capital Markets launched a hedge fund to exploit this finding, with Bollen et al. as consultants. It folded within months!
- Lachanski and Pav provide a comprehensive critique of Bollen et. al, showing look-ahead bias in the normalisation process, unrepeatability of the results, and findings that contradict strong evidence in markets that effects should vanish quickly rather than appear after 3 days

– Lachanski and Pav (2011)

- Bollen, Mao, and Zeng (2011), Twitter Mood Predicts the Stock Market. Journal of Computational Science 2(1): 1–8. https://arxiv.org/pdf/1010.3003.pdf
- Lachanski and Pav (2017), Shy of the Character Limit: "Twitter Mood Predicts the Stock Market" Revisited. https://risk.princeton.edu/img/Lachanski2017.pdf

Bad Science: correlation *does not imply* causation



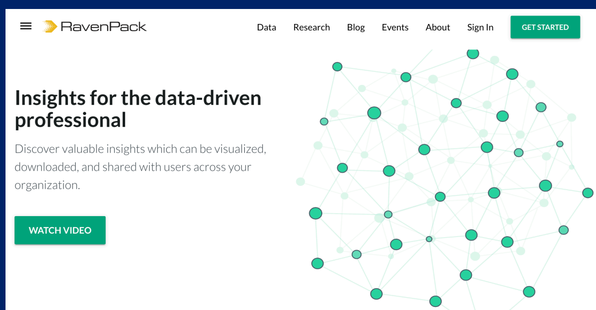


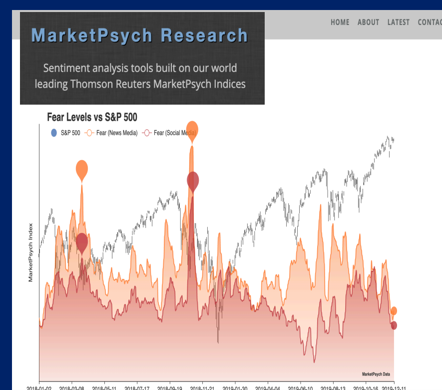
- Spurious correlations:
  - *if* you look for enough 95% confidence correlations in a dataset, you are likely to find one…
  - Bollen et al. (2010) compared 6 moods against 7 days of DJIA movements (i.e., 42 pairs), with no theoretical rationale behind the finding that "calm" correlates with DJIA movements in 3-days time.
- Related to the "Crud Factor":
  - *"In the social sciences and arguably in the biological sciences, everything correlates to some extent with everything else."* Meehl (1990)

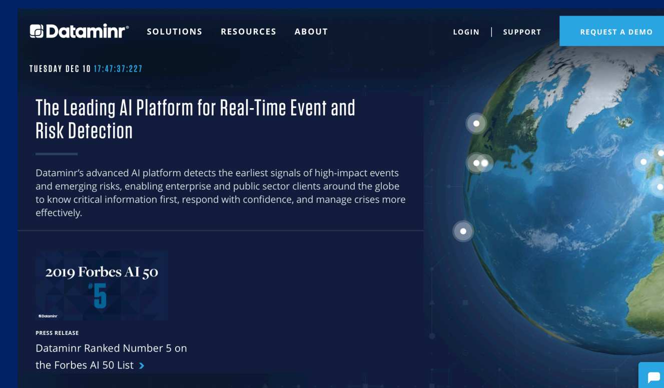
Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. Psychological Reports, 66, 195-244. https://doi.org/10.2466/pr0.1990.66.1.195

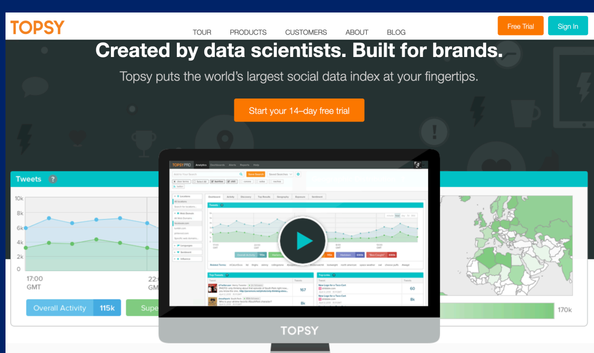
A false start, *not a false dawn*

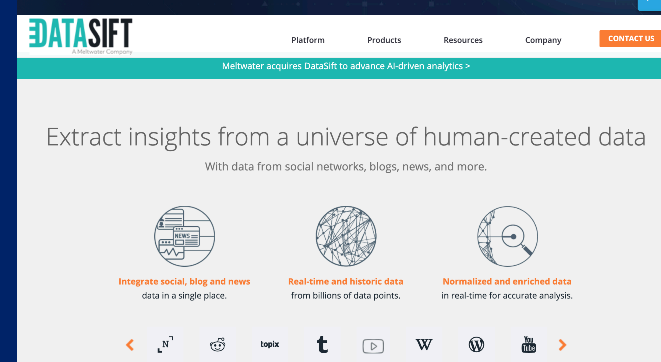
- Bollen et al.'s result may now be suspect, but this did *not* spell the end
- Social media and financial markets are now intricately intertwined, with an industry of third-party applications offering real-time Twitter firehose analytics to trading companies











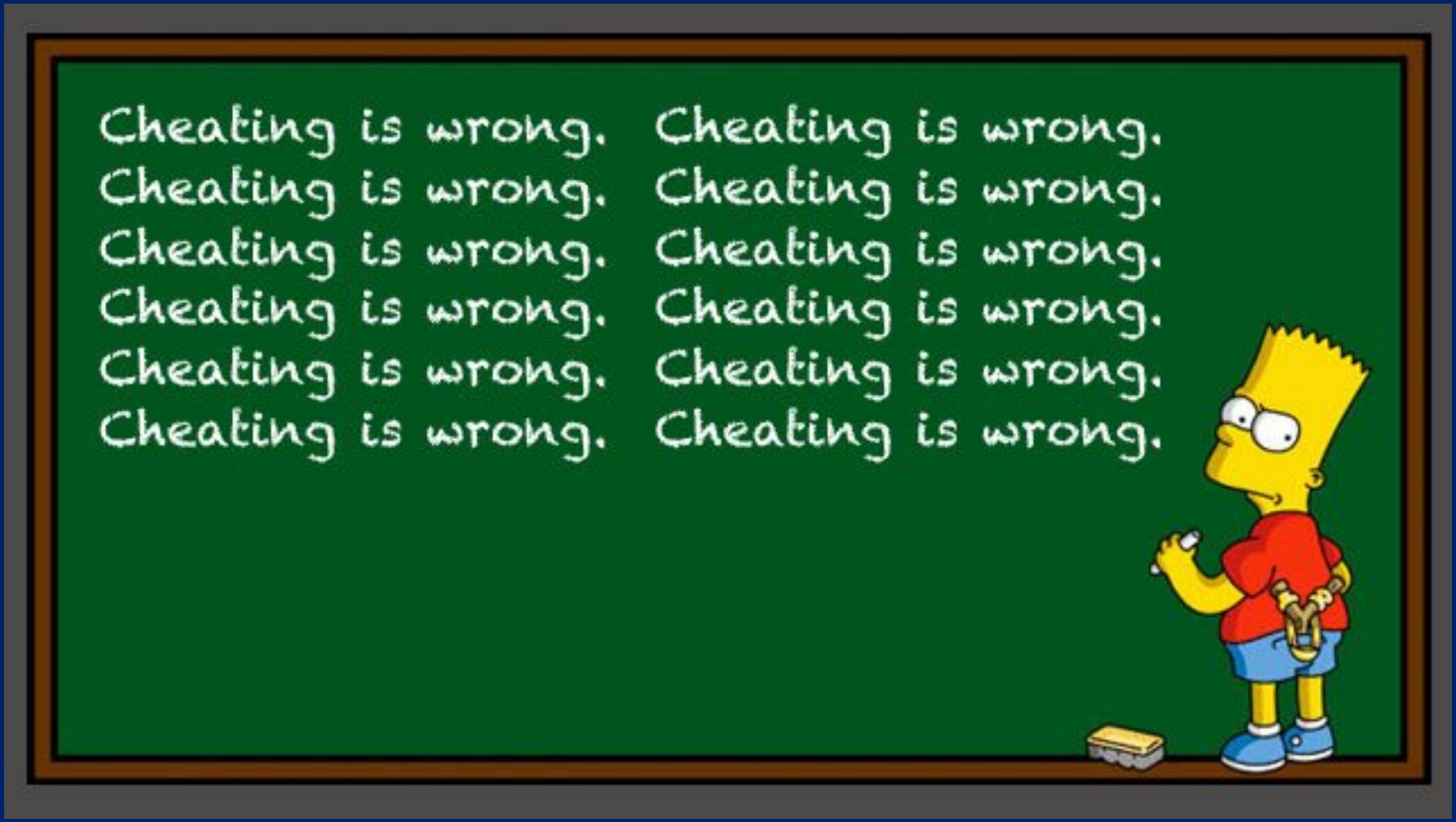
- There is a network effect at play here - once your competitors use these software, there is greater value in you using it too, otherwise you risk falling behind

The “Hack Crash”

- At 1:07 pm on Tuesday, April 23, 2013, a tweet went out over the Associated Press’s Twitter feed: @AP.
- @AP has 2 million followers. The tweet was retweeted 4,000 times in less than five minutes…
- Within minutes, markets crashed: DJIA dropped 140 points; S&P 500 lost $121 billion (1% of its value).
- An AP post on Facebook exposing the hack caused markets to immediately rebound back.

Karppi & Crawford, 2015, Social Media, Financial Algorithms and the Hack Crash. https://doi.org/10.1177/0263276415583139

Gaming the system



Opinions matter… *so it pays to cheat!*

Amazon “Brushing”

1. Seller, S, on Amazon gets name and address of a customer, C
2. S “purchases” an item from their own range and sends it to C, claiming it’s a gift
3. Amazon allows individuals who purchase a gift to leave a review for that item, so seller S leaves a very positive (fake) review
4. Review listed as “verified buyer”. Therefore has more authority, since it’s from somebody who actually bought the product
5. Positive, verified reviews result in higher search listings for S, and greater appeal to customers


Beware of unsolicited packages after Amazon Prime Day — they could be part of a scam

Published: July 22, 2019 8:21 a.m. ET

People across the country have reported receiving dozens of packages or more from Amazon that they never ordered

Received an unexpected gift recently?

Fake Reviews: Opinion Spam

- Fake reviews are common.
- In 2013:
  - Yelp estimated that they filter out 25% of posted reviews as suspected fake.
  - Samsung fined in Taiwan for paying people to praise Samsung products and criticise HTC products in online forums.
  - NY state fined 19 companies, many of them ‘reputation management’ firms, for posting thousands of fake reviews. The NY Times reported that some of these reputation management firms then “went on review sites that criticised their fake review operations and wrote fake reviews denying they wrote fake reviews”!

Sockpuppetry

- A sockpuppet is an online identity used for purposes of deception.
- The sockpuppet poses as an independent third-party unaffiliated with the main account operator
- Used to praise, defend, or support a person or organisation, or to manipulate public opinion
- Use of many fake personas by one person or organisation can give a misleading impression of an opinion being held widely and diversely

An army of sockpuppets

Sockpuppetry

- Sockpuppets work because people are strongly influenced by the opinions of others, especially when many others agree. Variants of sockpupperty include:
- Ballot Stuffing: Submit multiple votes in online poll in favour of the puppeteer
- Sybil Attack: Multiple false identities used to create influence within a peer-to-peer online network
- Stealth Marketing: Multiple sockpuppets, each claiming to be an enthusiastic supporter of product/service/ideology
- Strawman Sockpuppet: To make a particular point of view look foolish in order to generate negative sentiment against it.; or to advance arguments that their puppeteers can easily refute
- Astroturfing: Use sockpuppets to make a message (e.g., political, advertising, religious or public relations) appear as though it originates from, and is supported by, grassroots participants. Name is derived from AstroTurf, a brand of synthetic carpeting designed to resemble natural grass (play on word "grassroots").

Detecting Sockpuppets

Sockpuppets can be difficult to spot, except when used clumsily…

- American Petroleum Institute (probably) set up 25 accounts of ‘regular guys/gals’ to tweet in favour of the Keystone XL Tar Sands Pipeline.
- New accounts, identical simultaneous tweets; all removed when highlighted in press.

Persona Management Software for managing sockpuppets:

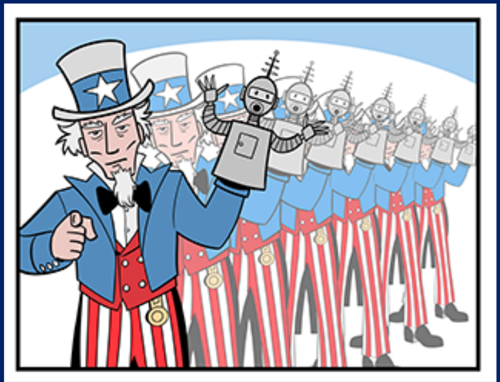
- Creates diverse plausible and geographically consistent online personas, with static IP addresses assigned to each
- Puppeteers have randomly selected (geographically diverse) IP addresses which change each day.
- Traffic is blended with users outside the organisation to provide ‘excellent cover and powerful deniability.’
- Commissioned by US military in 2011, probably for use in the Middle East as part of ‘Operation Earnest Voice’ to spread pro-US propaganda.
- An example of US government astroturfing.


Fake Twitter accounts used to promote tar sands pipeline

Oil companies believed to be behind fake accounts by 'SarahMama2', 'droidude7816' and 'JennyJohnson10' in support of planned Keystone XL tar sands pipeline from Alberta to Texas

● Read all our coverage of tar sands

▲ A protester against the Keystone XL tar sands pipeline. Several fake Twitter accounts appear to have been created in support of the pipeline. Photograph: Nati Harnik/AP



Opinion Spam Detection: Duplicate Reviews

- Spotting opinion spam is difficult and requires effort. Can we automate?
- Jindal and Liu (2008) used a supervised learning approach.
  - They identified duplicate reviews on Amazon (same review, different product or reviewer) as likely fakes. This was used as the training corpus.
- Features used:
  - Review: e.g., word n-grams, brand-name count;
  - Reviewer: e.g., mean and SD of rating given, % of reviews which were first or only reviews;
  - Product: price, sales rank.
- Tentative observations:
  - Negative outlier reviews are heavily spammed; positive outliers less so.
  - Singleton reviews are often fake.
  - Top-ranked reviewers are more likely to post fake reviews.
- Results tentative because it assumes duplicate equals fake.
  - E.g., it may be that reviews of top-ranked reviewers are more likely to be plagiarised by others, and so more likely to be duplicated even though the original was not fake.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (WSDM '08). ACM, New York, NY, USA, 219-230. https://doi.org/10.1145/1341531.1341560

Opinion Spam Detection: Atypical Behaviours

- Lim et. al. (2010) use data mining with models of atypical behaviour to assign a spam score to reviewers.
- Suspect behaviours include:
  - Promoting or victimising a few target products.
  - Targeting a group of products in a short period of time.
  - Tending to always give very high or low scores.
  - Giving ratings which deviate from those of other reviewers of a given product.

Lim, Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady W. Lauw. *Detecting Product Review Spammers using Rating Behaviors*. in *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2010)*. 2010. https://doi.org/10.1145/1871437.1871557

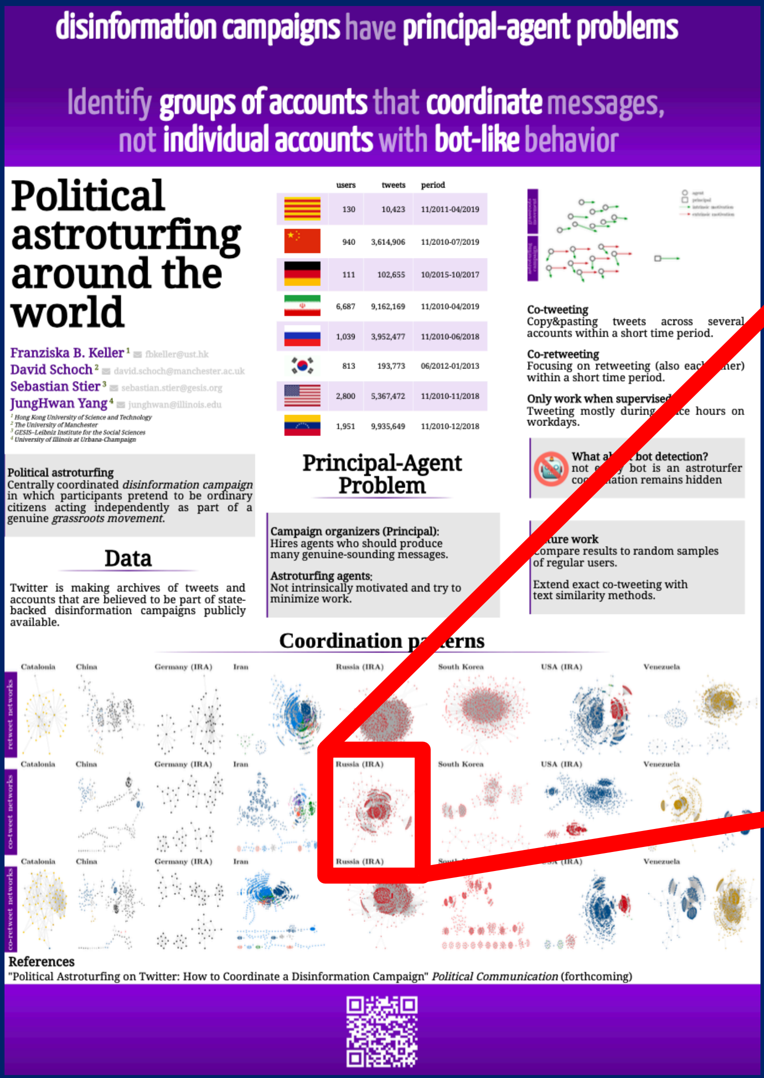
Sock Puppet Detection (Bu et. al. 2013)

- Create a network of links between online personas based on each time they express a similar view in a given discussion chain.
- Identify clusters within this network.
- Analyse writing styles within clusters and identify those with similar characteristics.
- Use heuristics on these to identify likely sock puppets;
  - Similar IP addresses
  - Similar online names and/or avatars
  - Similar registration times
  - Similar login patterns

[Note that Persona Management will beat this approach…]

Zhan Bu, Zhengyou Xia, Jiandong Wang, A sock puppet detection algorithm on virtual spaces, Knowledge-Based Systems, Volume 37, January 2013, Pages 366-377. https://doi.org/10.1016/j.knosys.2012.08.016

Detecting Political Astroturfing:
Network Analysis (2019)



Identifying political astroturfing through co-tweet networks

Summary

- Online opinions (e.g., reviews, tweets) have great economic value
  - Influencing behavior of consumers, policy decisions, financial markets
- Sentiment analysis enables mining and aggregation of the many opinions expressed on the internet.
  - Document, sentence, and aspect level analysis
  - Supervised and unsupervised approaches
- But as opinions have economic and political impact, ‘opinion spam’ is now a considerable problem.
  - Fake reviews, sockpuppets, astroturfing
  - Malicious / unethical economics of opinions
- Machine learning, behavioural analysis, and network analysis can help us to detect and combat opinion spam

…and remember, whatever you do, *don’t annoy Kylie Jenner*

Example Questions

January 2019

A.10: What is "opinion spamming", and why might it be used?

[3 marks]

January 2017

A.11 In the context of sentiment analysis, if a pair of terms has a low pointwise mutual information, this means (a) the words are too rare, (b) the words are irrelevant, (c) the words do not co-occur, (d) the words are too frequent

[1 mark]

A.12 Why might identifying "sentiment shifters" be important in aspect classification?

[1 mark]

A.13 How could a restaurant owner benefit from the use of a sock puppet?

[1 mark]