

Texts Words Meaning

Elisabeth Bommers

mbr targeting

Ladislaus von Bortkiewicz Chair of Statistics

Humboldt-Universität zu Berlin

<http://lvb.wiwi.hu-berlin.de>



One fits All?

- How to obtain data?
- How to process data?
- How to model data?
- How to interpret results?

Depends on the domain and the question.



Outline

1. Motivation ✓
2. Obtaining Data
3. Processing Data
4. Sentiment Classification
5. Sentiment Classification



Obtaining

- Ready to use data sets
 - ▶ Gold standard corpora
 - ▶ Classification: Reuters Newswire, Twenty Newsgroups, Youtube Spam, ...
 - ▶ Sentiment: reviews from Amazon, TripAdvisor, Rotten Tomatoes, ...
- APIs
 - ▶ Twitter, New York Times, Guardian
 - ▶ Only in real time?
- Crawling / Scraping
 - ▶ Data for (almost) any domain of interest
 - ▶ Legal? Terms of Service, robots.txt



Gold Standard Corpora

Wissler et al. (2014)

- Standard collections for training and evaluation of algorithms
- Manual annotations and more structure than plain text
 - ▶ Syntactical information, lexical knowledge, semantic associations
 - ▶ E.g. entities, grammatical structures
- Multiple experts view the data independently
 - ▶ Inter-annotator agreement
 - ▶ Creation is time-consuming and expensive
- Example: Penn Treebank
 - ▶ 4.5 million English words, GSC for syntactical tagging



New York Times APIs

- Articles
 - ▶ 1851 until today
 - ▶ Headlines, abstracts, first paragraph, meta information
- Book and movie reviews
- Semantics
 - ▶ People, places, organizations
- Geo



Example: NYT API

Crawling and Scraping

□ Crawling

- ▶ Retrieves any information
- ▶ Follows any link
- ▶ General information extraction

□ Scraper

- ▶ Retrieves information from specific page
- ▶ Specific information extraction
- ▶ Easy to obtain high quality data



Legality of Web Scraping

- It is public anyway / Google does it
 - ▶ Added value by search engines
 - ▶ What about log in systems, paywalls, ...?
- Highly context specific
 - ▶ Commerical v non-commercial
 - ▶ Internal v third party use
- Technicalities
 - ▶ Humans don't access thousands of pages, bandwidth usage
 - ▶ Denial-of-service (DoS) attack



European Union

□ Ryanair Ltd v PR Aviation BV (2015)

- ▶ PR Aviation: price comparison of flights
- ▶ Copyright and database right infringement?
- ▶ ToS prohibited data extraction for commercial purposes

□ Decision by Court of Justice of the European Union

- ▶ No infringement of intellectual property, no creative input
- ▶ ToS still apply, liability in terms of breach of contract

□ In contrast NLA v Meltwater (2013)

- ▶ Scraping of news headlines and links to articles
- ▶ Intellectual property is infringed because of creative input



United States

Pro

- Web data is public, should be accessible
- First Amendment protects information gathering
- Unfair market power of Facebook, Google, LinkedIn, ...

Contra

- Copyright infringement
- Breach of contract
- Violation of the Computer Fraud and Abuse Act (CFAA), 1986
forbids unauthorized access to some computers
- Trespass to chattels



LinkedIn v hiQ and vice versa

If you exclude someone from sites like LinkedIn, Facebook and Twitter, you are excluding them from the modern version of the town square.

Laurence Tribe, Harvard law professor

- ▣ hiQ - startup predicting who is when quitting their job
- ▣ LinkedIn: CFAA violation, hiQ: blocked
- ▣ LinkedIn ordered to give access to public profiles



Academia is save, right?



Aaron Swartz

- Harvard research fellow
- Automatic download of articles from JSTOR
- Download via laptop in a closet at MIT
- No civil law suit by MIT and JSTOR
- Federal charges: wire fraud, CFAA violations
- Possible penalty of \$1 million and 35 years in prison

Unclear outcome, suicide on January 11, 2013



Good Manners and Ethics

- Data as a product
- Cost of bandwidth
- Rapid requests slow services for real humans
- robots.txt



Google's robots.txt

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?ptl=true$
Disallow: /imgres
Disallow: /u/
Disallow: /preferences
Disallow: /setprefs
Disallow: /default
Disallow: /m?
Disallow: /m/
Allow: /m/finance
```



Ethical Scraping for Academia

□ Technical

- ▶ Use API if provided
- ▶ Appear as a bot, not as a human
- ▶ Provide user agent string with contact data
- ▶ Decreased rate of requests

□ Usage

- ▶ Strictly non-commercial
- ▶ Restrict further access to academia

□ Ask for permission, not for forgiveness!



Scraping How To



- ▣ Complete framework: Scrapy
- ▣ Fast and easy: BeautifulSoup
- ▣ Low level: lxml



- ▣ Complete framework: RCrawler
- ▣ Fast and easy: rvest
- ▣ Low level: XML



Beeradvocate

- Largest beer rating community
- Founded in 1996, print magazine since 2006
- Mostly “exotic” and craft beer
- Still independent
 - ▶ Anheuser-Busch InBev acquires beer related websites
 - ▶ Examples: RateBeer, The Beer Necessities, ...
- Breweries in the US
 - ▶ Less than 100 in 1980s, more than 5,000 in 2016
 - ▶ Craft beer boom



Individual Review

Sort by: Recent | High | Low | **Top Raters** | Alström Bros

first ← prev | 1-25 | 26-50 | 51-75 | next → last

✓ **Ratings: 5,928** | 📄 **Reviews: 1,915**

4.13/5 rDev -2.1%

look: 4.5 | smell: 4 | taste: 4 | feel: 4 | overall: 4.5

A Great One, a very good IPA with added thickness from chocolate malt. Pours dark and foamy, hoppy, ahh! the taste of citrus C-hop, (there is also Amarillo and Simcoe) and the c-hop aroma. Background roastiness. Full mouthfeel. Delightful as a meal in itself.

📄 259 characters

Sammy, Dec 23, 2007

Source




xPath Inspector

div#rating_fullview | 494.8 × 5179 1,915

4.13/5 rDev -2.1%

look: 4.5 | smell: 4 | taste: 4 | feel: 4 | overall: 4.5

A Great One, a very good IPA with added thickness from chocolate malt. Pours dark and foamy, hoppy, ahh! the taste of citrus C-hop, (there is also Amarillo and Simcoe) and the c-hop aroma. Background roastiness. Full mouthfeel. Delightful as a meal in itself.

 259 characters

Sammy, Dec 23, 2007



Example: Scraper

Processing

- Often depends on theoretical model
 - ▶ Time series models for stock returns
 - ▶ Example 2
- Just a side note in scientific articles
 - ▶ Well established “gold standard” processing for many domains
 - ▶ No scientific news value



Part of Speech Tagging

- More than nouns, verbs, adjectives
 - ▶ Penn Treebank: 45 tags
 - ▶ Brown Corpus: 87 tags
- Useful for
 - ▶ Information retrieval
 - ▶ Word-sense disambiguation
 - ▶ Shallow parsing of names and other named entities
- Math Tagging, [Kristianto et al \(2012\)](#)
 - ▶ Extract mathematical definitions
 - ▶ Pattern matching and machine learning



Example: Beer Review

An OK lager. Light, crisp, but nothing special. Stacked against the great pilsners of the world or similar offerings, this beer is mediocre. But still solid enough to put it head and shoulders above any macro.



Nouns

An OK **lager**. **Light**, crisp, but **nothing** special. Stacked against the great **pilsners** of the **world** or similar **offerings**, this **beer** is mediocre. But still solid enough to put it **head** and **shoulders** above any **macro**.

NOUN

- ▣ Which topic?
- ▣ Which beer category? Lager, pilsner



POS Tagging is not perfect

An OK **lager**. **Light**, crisp, but **nothing** special. Stacked against the great **pilsners** of the **world** or similar **offerings**, this **beer** is mediocre. But still solid enough to put it **head** and **shoulders** above any **macro**.

NOUN

- Which topic? Beer
- Which beer category? Lager, pilsner



Verbs

An OK **lager**. **Light**, crisp, but **nothing** special. **Stacked** against the great **pilsners** of the **world** or similar **offerings**, this **beer is** mediocre. But still solid enough to **put** it **head** and **shoulders** above any **macro**.

NOUN, VERB

- ▣ Relationships between nouns
- ▣ Here: not very informative

Dependency Tree



Adjectives

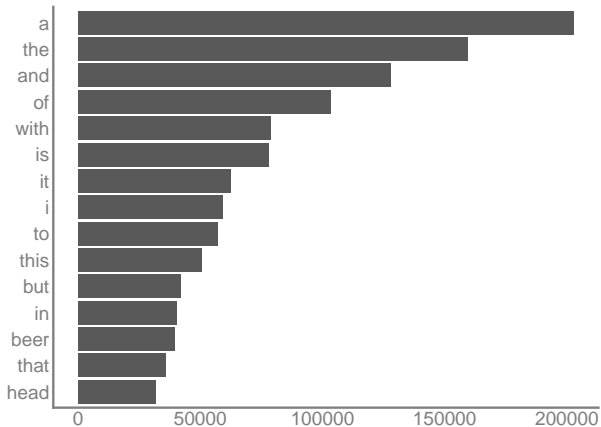
An **OK** **lager**. **Light**, **crisp**, but **nothing** **special**. **Stacked** against the **great** **pilsners** of the **world** or **similar** **offerings**, this **beer** **is** **mediocre**. But still **solid** enough to **put** it **head** and **shoulders** above any **macro**.

NOUN, **VERB**, **ADJ**

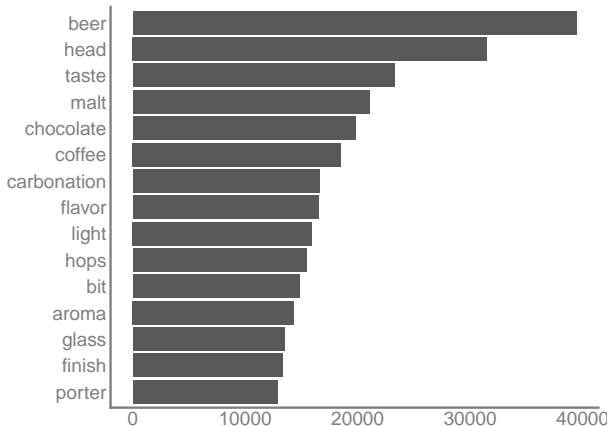
- What is the opinion? OK, solid, (nothing) special, mediocre
- How is something? Crisp



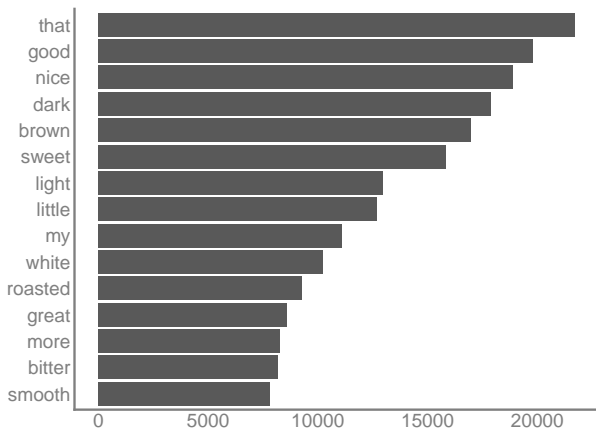
All Words



Nouns



Adjectives



Text Normalization: Basics

[(an, ok lager), (light, crisp, but, nothing, special), (stacked, against, the, great, pilsners, of, the, world, or, similar, offerings, this, beer, is, mediocre), (but, still, solid, enough, to, put, it, head, and, shoulders, above, any, macro)]

- ▣ Lowercase
- ▣ Removal of non alphabetic characters
- ▣ Word and sentence lemmatization



Text Normalization: Lemmatization

(stacked, stack), (pilsners, pilsner)

(offerings, offering), (is, be)

(it, -PRON-), (shoulders, shoulder)

- Map a word to its canonical form
- Alternative to stemming



Text Normalization: Special Cases

□ Dates

- ▶ Canonical form: 22.11.2017 → 11/22/2017
- ▶ Relative to absolute: Yesterday → 11/22/2017

□ Abbreviations

- ▶ Common: United States of America → USA
- ▶ Specific: ordinary least squares (OLS)

□ Numbers and units

- ▶ Hundred → 100
- ▶ \$100 → 100_dollar → 10000_cent

What is the most important aspect of the domain?



Text Filtering: Noun Chunks

(An, OK, lager), (Light), (the, great pilsners)

(the, world), (similar, offering), (this, beer)

(it), (any, macro)

- ▣ Parse dependencies of nouns
- ▣ Additional filtering for noun, adjective, verb combinations



Evaluation Processing

Pred True \	Unprocessed		Processed	
	-1	1	-1	1
-1	677	723	725	675
1	498	902	307	1,093
Accuracy	0.56		0.65	
Recall	0.58	0.56	0.70	0.62
Precision	0.48	0.64	0.52	0.78



Evaluation Pipeline

Pred True \	Unprocessed		Processed	
	-1	1	-1	1
-1	1,304	96	1,274	126
1	70	1,330	94	1,306
Accuracy	0.94		0.92	
Precision	0.93	0.95	0.91	0.93
Recall	0.95	0.93	0.93	0.91



Sentiment Distribution

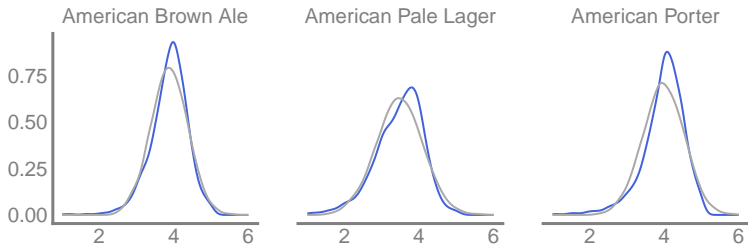


Figure: Estimated Densities of [Beer Ratings](#) and Simulated Normal Distribution



Texts Words Meaning

Elisabeth Bommes

mbr targeting

Ladislaus von Bortkiewicz Chair of Statistics

Humboldt-Universität zu Berlin

<http://lvb.wiwi.hu-berlin.de>



Bibliography



Wissler, L. and Almasraee, M. and Díaz, D. M. and Paschke, A.
The Gold Standard in Corpus Annotation
IEEE GSC, 2014



Kristianto, G Y, Ngien, M Q, Matsubayashi, Y and Aizawa, A
Extracting definitions of mathematical expressions in scientific papers
Proc. 26th JSAI, 2012



Härdle, W. K. and Lee, Y. J. and Schäfer D. and Yeh Y. R.
Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies
J. Forecast., 2009





Hu, M. and Liu, B.

Mining and Summarizing Customer Reviews

10th ACM SIGKDD, 2004



Loughran, T. and McDonald, B.

When is a liability not a liability?

J. Financ., 2011



Malo, Pekka and Sinha, Ankur and Korhonen, Pekka and Wallenius, Jyrki and Takala, Pyry

Good debt or bad debt

Journal of the Association for Information Science and Technology,
2014



Wilson, T. and Wiebe, J. and Hoffmann, P.

Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis

HLT-EMNLP, 2005





Zhang, J., Chen C. Y., Härdle, W. K. and Bommers, E.
Distillation of News into Analysis of Stock Reactions
JBES, 2016

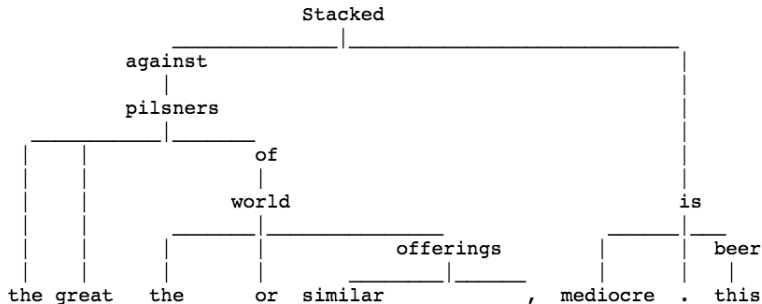


Zhang, X., Yichao, W., Wang, L. and Runze, L.
A Consistent Information Criterion for Support Vector Machines in Diverging Model Spaces
J. Mach. Learn. Res., 2016



Appendix

Grammar based Dependency Tree

[Back](#)