Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

Text Analytics in the Wild
Trends and Technologies

Mark Cieliebak

20.11.2013

About Us

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

ZHAW

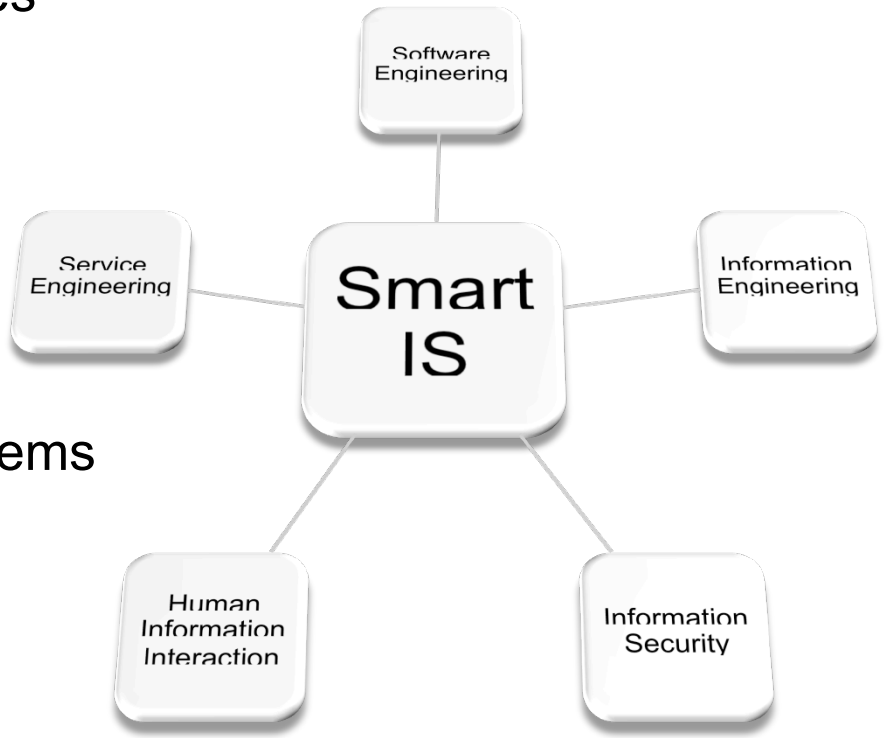
Zurich University of Applied Sciences

- more than 10'000 students
- ~230 Professors (FTE)

InIT

Institute of Applied Information Systems

- since 2005
- ~35 professors
- Labs on Cloud Computing, Data Science, Visual Computing etc.



What is "Text Analytics"?

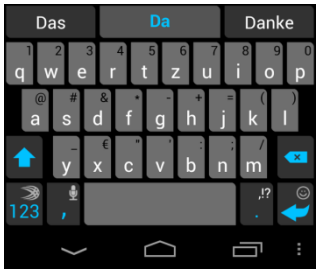
Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

Synonyms: Text Mining, Natural Language Processing (NLP)

[1]


[6]

Goal: Turn text into information



[2]

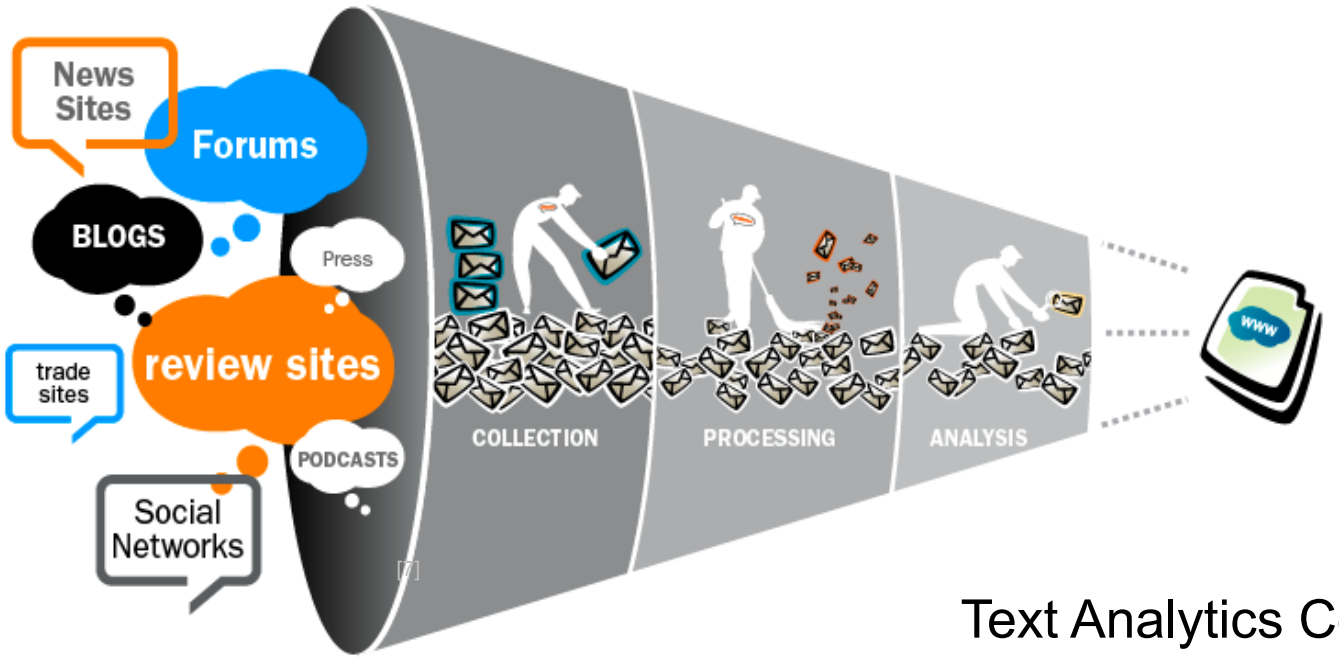
[5]

[3]

[4]

Sample Application: Social Media Monitoring

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



Text Analytics Components:

- Find relevant documents
- Hot topic detection
- Sentiment detection

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

"Netgear's WiFi Analytics is a free Android app that I find very handy when it comes to troubleshooting and monitoring a home network.

It's simple, but that simplicity is its strength and gives home users insight into a home wireless network, without need to know a lot about networking." [8]

Simple Sentiment Detection

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

Idea: Count number of positive and negative words

"This analysis is good[+1]."	1 (pos)
"I find it beautiful[+1] and good[+1]."	2 (pos)
"It looks terrible[-1]."	-1 (neg)
"This car has a blue color."	0 (neu)

Use Sentiment-Dictionary:

POSITIVE:	NEUTRAL:	NEGATIVE:
good	hello	bad
love	see	hate
nice	I	ugly
…	…	…

Improvements

Sample Text

- "This analysis is good."
  "This analysis is excellent."
- "The car really very expensive. "
- "This car has an appealing design and comfortable seats, but it is expensive."

Solution

- *Fine-Grained Dictionaries:*
  "This analysis is good [+2]. "
  "This analysis is excellent [+3].
- *Detect Booster Words:*
  "This car really very expensive[-1 -1 -2] ."
- *New Category "Mixed":*
  "This car has an appealing[+1] design and comfortable[+1] seats, but it is expensive[-1]. "
  → MIX(+2/-1)

Improvements 2

Sample Text

- This analysis is not good.
- The car is appealing and I do not find it expensive.
- I do not find the car expensive and it is appealing.

Solution

- *Invert all Scores:*
  "This analysis is not[*-1] good[+3]."
- *Invert only score of words occuring after the negation:*
  "The car is appealing[+3] and I do not[*-1] find it expensive[-2]"

→ Need to “understand” the sentence
→ Need linguistic analysis!

Linguistic Analysis

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

<table>
<tr><td colspan="10">Sentence</td></tr>
<tr><td colspan="7">Sentence</td><td rowspan="3">Conj.</td><td colspan="3" rowspan="2">Sentence</td></tr>
<tr><td rowspan="2">Noun Phrase</td><td colspan="6">Verb Phrase</td></tr>
<tr><td rowspan="2">Verb</td><td rowspan="2">Adverb</td><td rowspan="2">Verb</td><td colspan="2">Noun Phrase</td><td rowspan="2">Adj.</td><td>Noun Phrase</td><td colspan="2">Verb Phrase</td></tr>
<tr><td>Det.</td><td>Det</td><td>Noun</td><td>Det.</td><td>Verb</td><td>Participle</td></tr>
<tr><td>I</td><td>do</td><td>not</td><td>find</td><td>the</td><td>car</td><td>expensive</td><td>and</td><td>it</td><td>is</td><td>appealing</td></tr>
</table>

-> Invert scores of words being in the same phrases as negation:

“I do not find the car expensive[+2]
and it is appealing[+3].” → +5 (pos)

Rule-based Sentiment Detection

"This movie was like a horror event."

<table>
<tr><td colspan="2">Noun Phrase</td><td colspan="5">Verb Phrase</td></tr>
<tr><td rowspan="3">Determiner</td><td rowspan="3">Noun</td><td rowspan="3">Verb</td><td colspan="4">Prepositional Phrase</td></tr>
<tr><td rowspan="2">Preposition</td><td colspan="3">Noun Phrase</td></tr>
<tr><td>Determiner</td><td>Noun</td><td>Noun</td></tr>
<tr><td>This</td><td>movie</td><td>was</td><td>like</td><td>a</td><td>horror</td><td>event</td></tr>
</table>

Sentiment Dictionary: *like* (Verb): +2, *like* (PP): 0

⇒ "This movie was like[0] a horror[-2] event." → -2 (neg)

Technology Stack

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

for Sentiment Detection on Web Documents using Linguistic Analysis

Linguistic Framework

- Gate: text analysis framework with experimentation GUI; includes sentence splitter, tokenizer, chunker, pos-tagger, etc; plugin-mechanism for new libraries; very widely used. http://gate.ac.uk/
- OpenNLP: Java-based, easy to use; includes tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, etc; by Apache Foundation. http://opennlp.apache.org/

Web Boilerplate Extraction

- Boilerpipe: removes ads, menus, HTML-tags etc. http://boilerpipe-web.appspot.com/

Language Detection

- Nutch language identifier plugin: 14 languages preimplemented, „easy“ to add more. http://wiki.apache.org/nutch/LanguageIdentifierPlugin

Sentence Splitting (various programming languages)

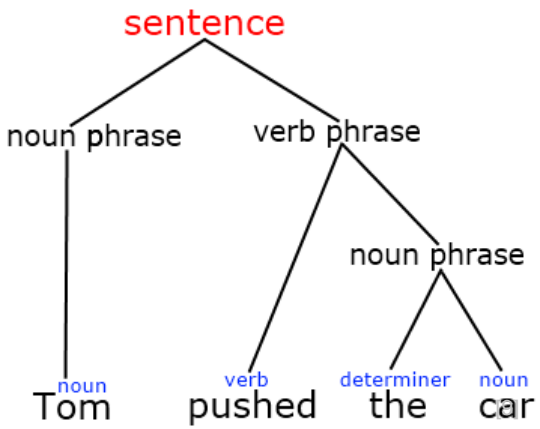
- OpenNLP S 2.2 http://opennlp.apache.org/
- CoreNLP R 18.7 http://nlp.stanford.edu/software/corenlp.shtml
- GATE R 60.6 http://gate.ac.uk
- LingPipe R 1.7 http://alias-i.com/lingpipe/
- MxTerminator S 2.8 ftp://ftp.cis.upenn.edu/pub/adwait/jmx/
- Punkt U 7.1 http://nltk.org/api/nltk.tokenize.html
- RASP R 0.3 http://ilexir.co.uk/applications/rasp/
- Splitta S 31.5 http://code.google.com/p/splitta
- tokenizer R 0.3 http://www.cis.uni-muenchen.de/~wastl/misc/
- opennlp ist java und easy to use.

Part-of-Speech Tagger

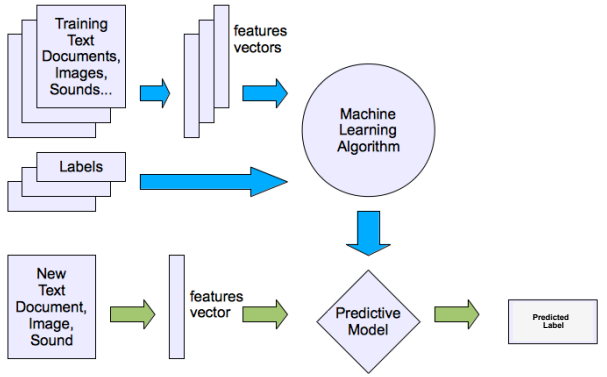
- TreeTagger: language-independent, several languages already implemented; also for lemmatization. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- TweetNLP: POS-tagger just for Twitter. http://www.ark.cs.cmu.edu/TweetNLP/

Approaches to Sentiment Detection

Rule-Based



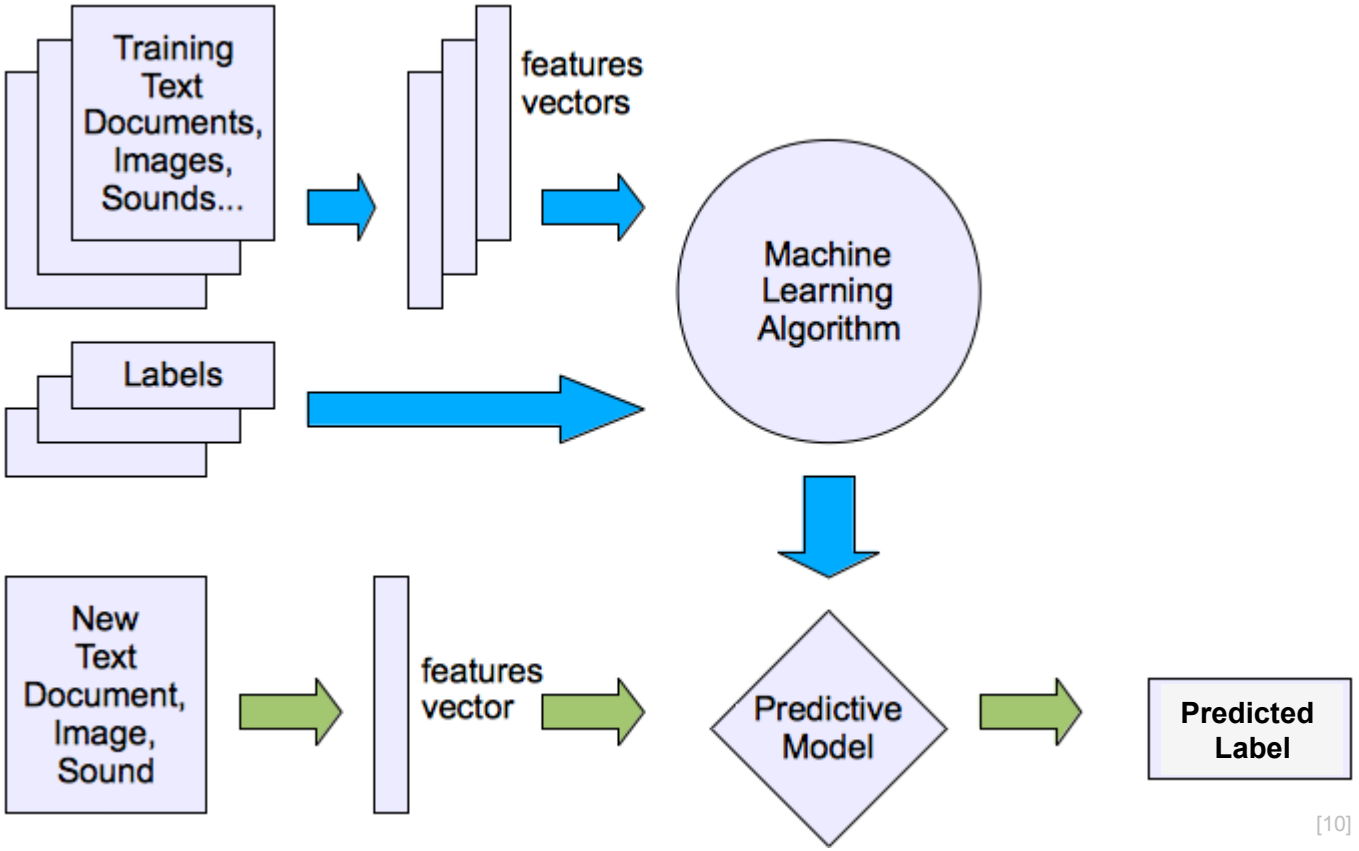
Corpus-Based



[10]

Corpus-Based Sentiment Detection

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



[10]

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

Annotated Corpus

Sentence	Polarity
This analysis is good.	Pos
It looks awful.	Neg
This car has a blue color.	Neu
This car has an appealing design, comfortable seats, but it is expensive.	Mix
This car has a very appealing design, comfortable seats, but it is really expensive.	Mix
This analysis is not good.	Neg
This car has an appealing design, comfortable seats and it is not expensive.	Mix
This movie was like a horror event.	Neg
This car is appealing and is not expensive.	Mix
...	...

Challenges in Sentiment Detection

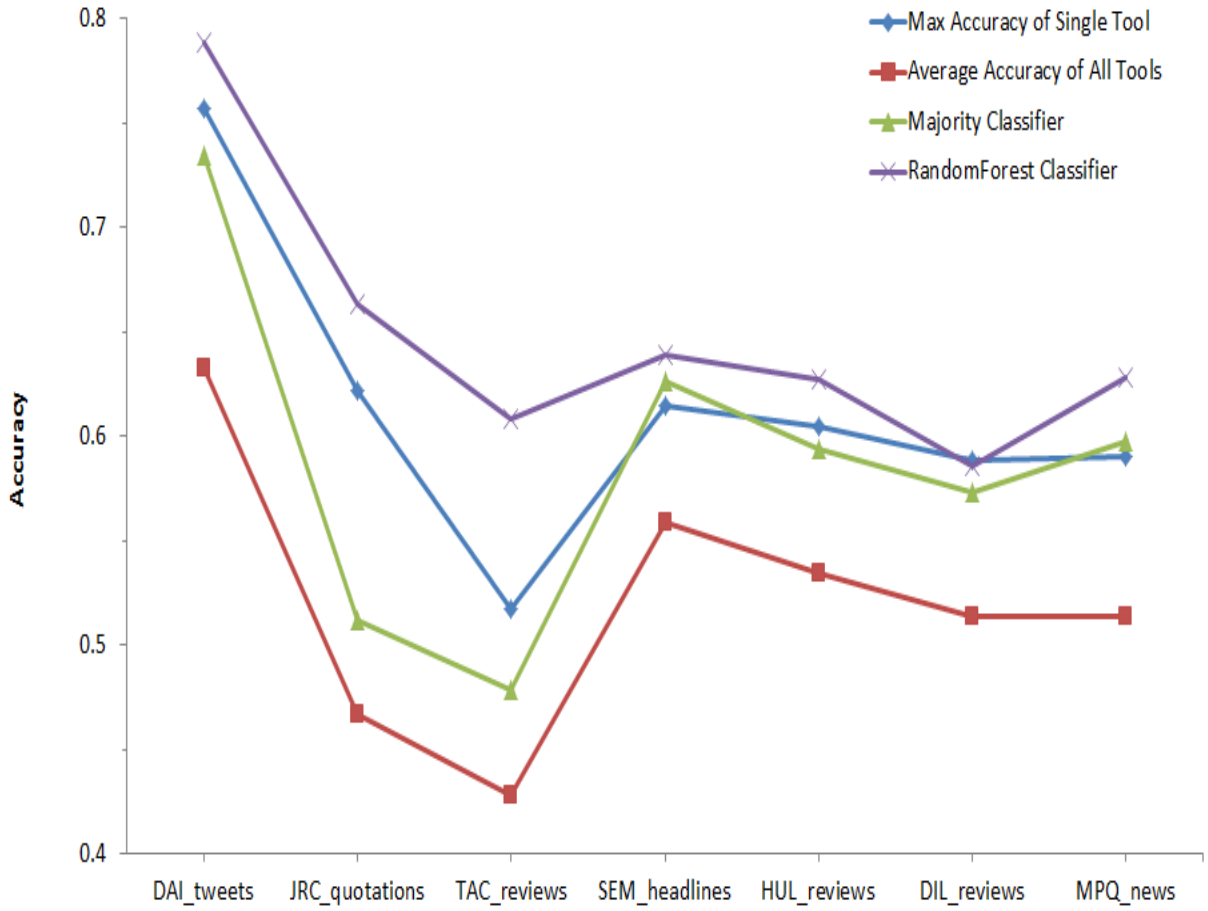
- **Sarcastic Statements**
  "What a great car, it stopped working in the second day."
- **Conditional Statements**
  "If I can find a good camera, I will buy it."
  "If you are looking for a good phone, buy Nokia"
- **Context Dependencies**
  "Salaries for software engineers are extremely high."

- **User Expectations vs. Reality**
  one "sentScore" vs. ambivalent docs
- **Opinion Spam Detection**

- **Huge Effort for New Languages**

Sentiment Detection: State of the Art

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



Evaluation of 7 commercial tools on 30'000 documents [11]

Main Results:

- Best tool has 60% accuracy on all documents
- "Easiest" document type is Tweets: up to 78% accuracy
- Meta-classifier beats each single tool

Contact

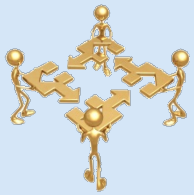
Mark Cieliebak

Institute of Applied Information Technology (InIT)

ZHAW, Winterthur

Email: ciel@zhaw.ch, Website: www.zhaw.ch/~ciel

Research Interests

- Open Data 
- Social Network Analysis 
- Automated Test Generation 
- Sentiment Detection 

References

[1] http://www.teleportmyjob.com/blog/wp-content/uploads/2013/05/find_job.jpg
[2] http://screenshots.de.sftcdn.net/de/scrn/3340000/3340669/swiftkey-23-630x535.png
[3] https://lh4.ggpht.com/Fs4HLaMCgmY6O7BWGKZB8LdOM8GmiRAbKUcdJysNp0k74xmPMTSGVHigwvaLy_Tw=w300
[4] http://revaxmedia.co.uk/wp-content/uploads/2012/09/Siri-What-Can-I-Help-You-With.jpg
[5] http://www.betadaily.com/wp-content/uploads/2011/02/BusinessPlanExecutiveSummary.jpg
[6] http://everycook.org/cms/en/hardware-en/pictures-en#
[7] http://cdn.fulltraffic.net/images/blog/social-media-monitoring-process.png
[8] http://www.pcmag.com/article2/0,2817,2426416,00.asp
[9] http://i1059.photobucket.com/albums/t424/shmi11y/UoS%20LingSite/tompushedthecar.png
[10] http://bigsnarf.files.wordpress.com/2013/04/supervised.png
[11] Cieliebak et al., ESSEM 2013.

Sentiment Detection for Product Reviews

"McAfee Antivirus & Security for Android … is free for All Access users.

Even if you don't have All Access, McAfee is certainly attractive with its well-known name, … and enormous slew of security features.

But I was disappointed that the app didn't live up to its incredible potential, and would recommend looking at lower-priced alternatives if you're not an All Access user."