Introduction to Sentiment Analysis

Text Analytics - Andrea Esuli

What is Sentiment Analysis?

What is Sentiment Analysis?

"Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language."

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

SA works on the **subjective/evaluative/emotive** components of textual information, which have often been ignored in the **objective/factual/topical** analysis usually performed in traditional TA.

Topic vs Sentiment

Topic and sentiment are two main orthogonal dimensions:

- Topic/Fact/Objective information
- Sentiment/Opinion/Subjective information (affective states, emotions. . .)

Topical analysis:

- Discriminating political news from sport news.
- Extracting mention of names of persons in text.

Sentiment analysis:

- Discriminating between favorable and negative attitude toward a subject.
- Identifying the expressions of an emotion and the target of that emotion.

Topic vs Sentiment

Objective information:

*The **4.7-inch** display on the iPhone 6 is arguably its best feature.*

*...concerns have been raised about the relatively low resolution (**1334 x 750 pixels**)*

Source

Topic vs Sentiment

Subjective information:

*The 4.7-inch display on the iPhone 6 is arguably its **best feature**.*

*...concerns have been raised about the **relatively low** resolution (1334 x 750 pixels)*

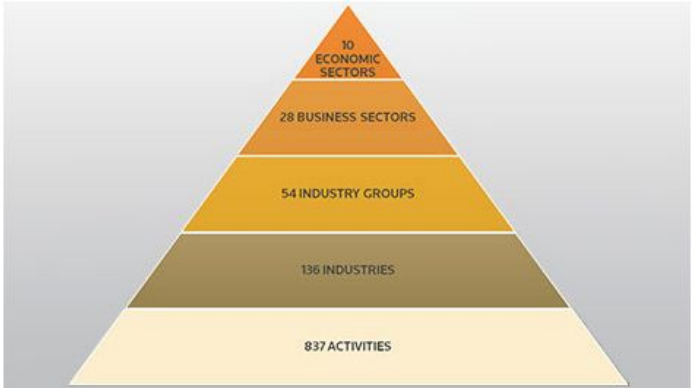
Source

Topic vs Sentiment

Classification of documents:

- with respect to the Thomson Reuters

taxonomy*.



- with respect to the content being a positive, neutral, or a negative evaluation†.

```
{"data": [{"text": "I love Titanic.", "id":1234, "polarity": 4},
          {"text": "I hate Titanic.", "id":4567, "polarity": 0}]}
```

* Source † Source

Topic vs Sentiment

Extraction of information:

- regarding objective properties

  The NBA player Michael Jordan is from the United States of America*

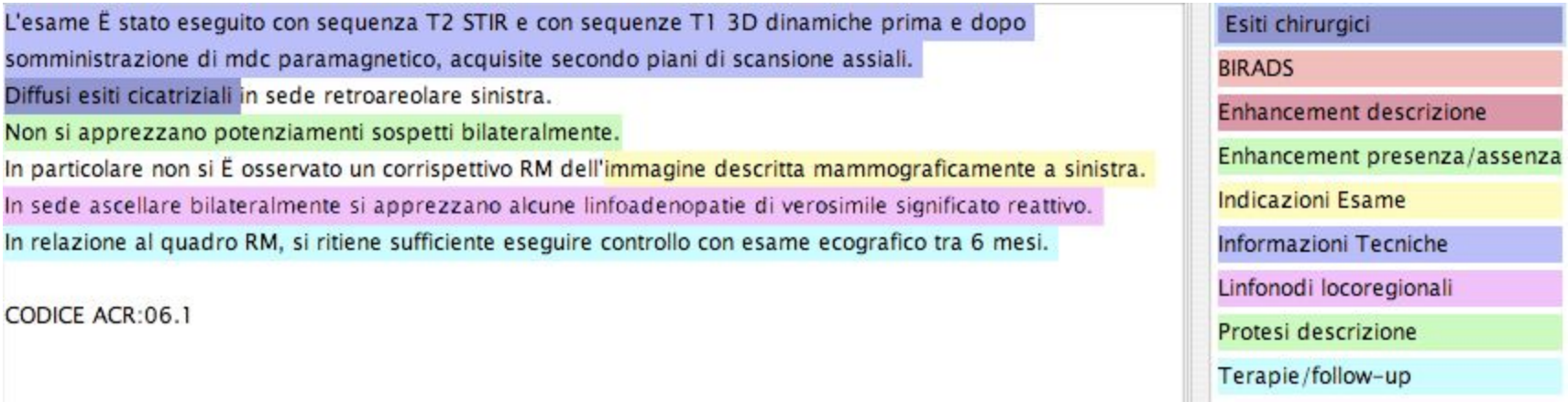
  Organization Person Location

- regarding the expression of opinions.

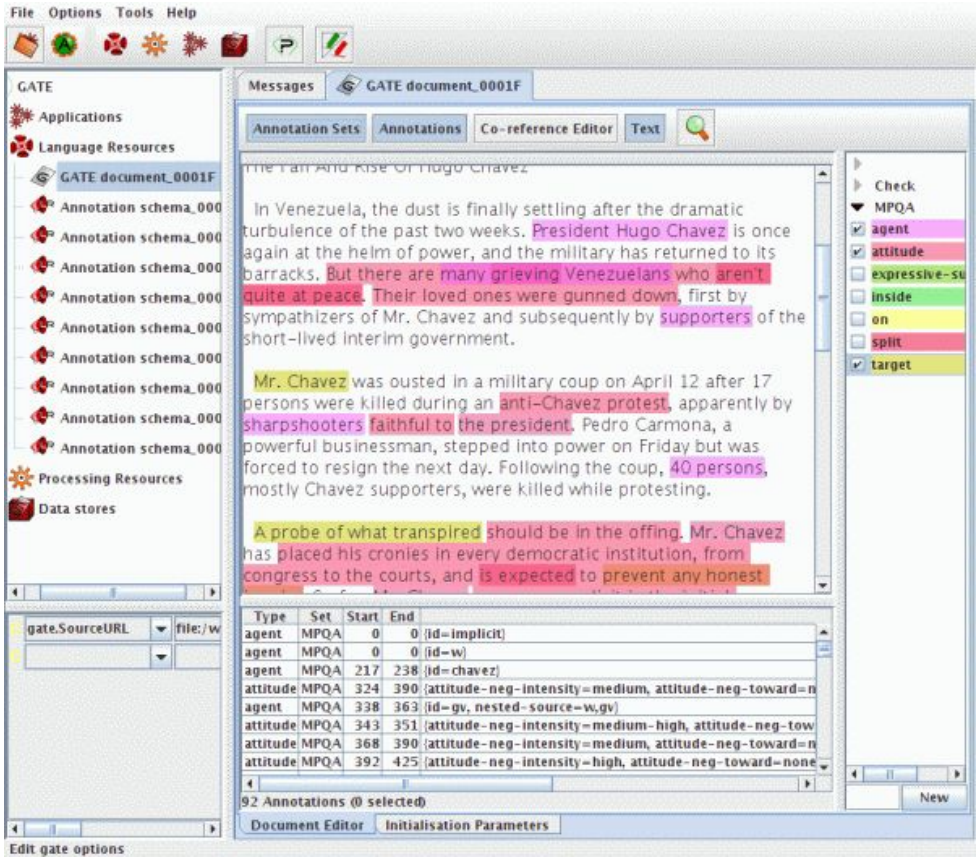
  soldiers with 20 years or more service are generally satisfied with termination packages being offered†

  Agent Attitude Target

* Source † Source



Annotation of radiology reports



Opinion Annotation in GATE

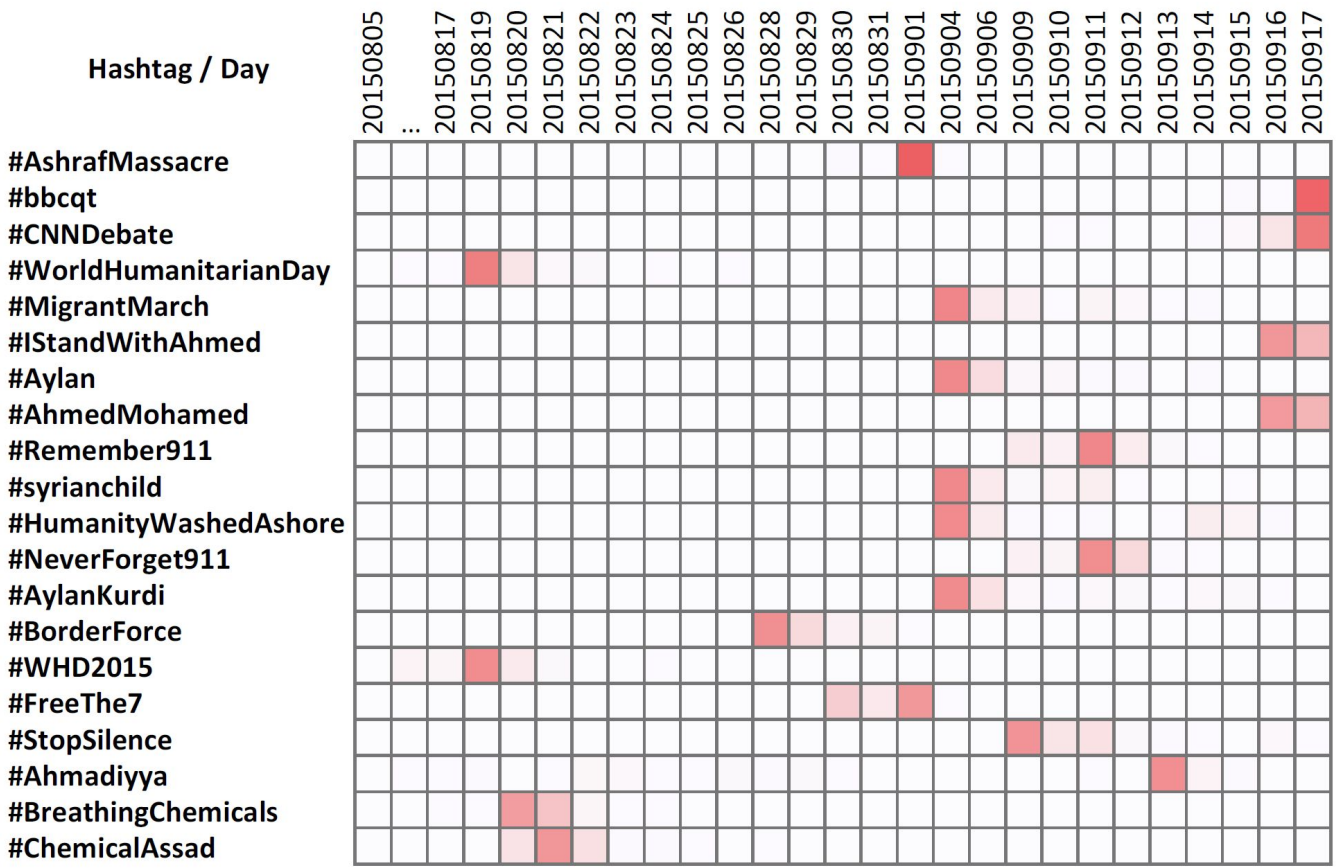
Facts, Sentiments and Big Data

Facts and Big Data

When looking for factual information, the comparison of many sources of information allows to check for its truth, consistency and relevance.

Temporal/spatial anomalies in the use of language, e.g., spikes in the use of words like “earthquake”, “shots”, “explosion”, may allow to recognize events, and gather relevant data about them.

Image source



Event recognition from hashtag use distribution

Sentiment and Big Data

Subjective information is varied by definition.
The more sources are compared, the more the vision of the feelings on the matter is complete.

4.7 out of 5 stars

5 star	82%
4 star	12%
3 star	2%
2 star	1%
1 star	2%

See all 1,771 reviews

"*It takes great photos and it is a very good quality camera.*"
[redacted] | 478 reviewers made a similar statement

"*Very easy to use as a point and shoot camera, as I will be taking some classes to really learn how to use more features.*"
[redacted] | 362 reviewers made a similar statement

"*It is the best thing I have spent money on, and I don't think anyone will regret buying this camera.*"
[redacted] | 151 reviewers made a similar statement

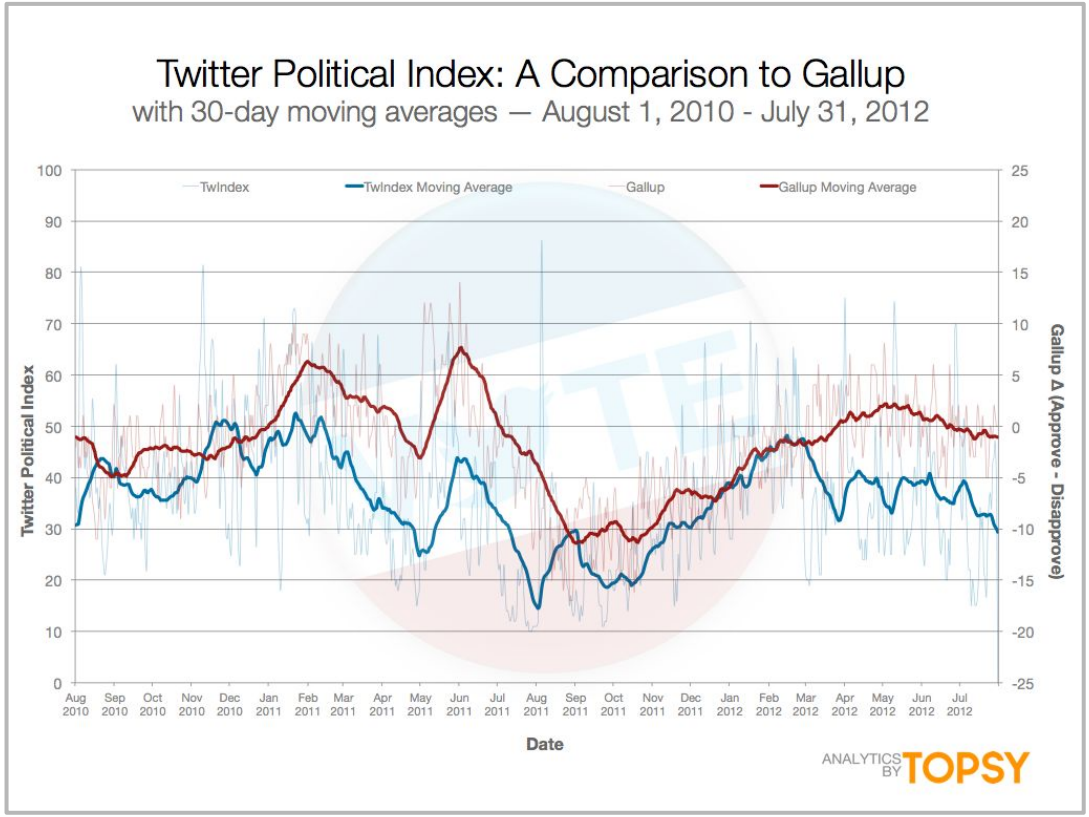
Broke right out of the box

By [redacted] on January 22, 2015

Style: w/ 18-55mm lens | Package Type: Standard Packaging

I have been begging for this camera for ever so I bought it right here off amazon as I got the the box I opened it and the hole box fell apart the camera fell and the lens focus ring broke as well as the lcd screen got scratch. I wouldn't be complaining but the camera didn't come with a warranty so I wasted 500 dollars on a broken camera.

Sentiment and Big Data



[Twindex]

Why Sentiment Analysis?

(Is it of practical use?)

Why Sentiment Analysis?

When we have to take a decision we look for the opinion of the others.

The textual user-generated content that is

- shared on the Web/social networks,
- written in open-ended questions in questionnaires,
- sent to companies as feedback, . . .

contains

- **voluntarily produced,**
- **unconstrained,**
- **first-hand/personal,**
- **fresh,**

evaluative information about our topic of interest.

Why Sentiment Analysis?

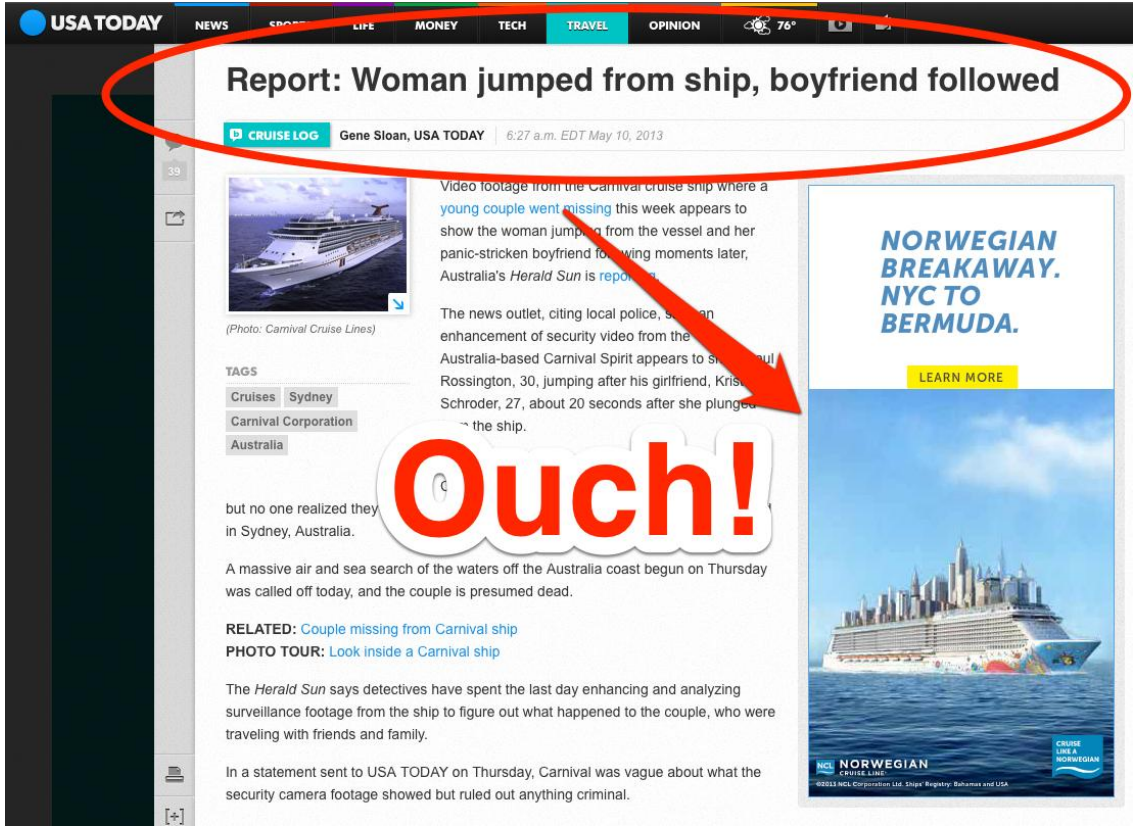
Practical example: customers satisfaction questionnaires.

- `Are you happy with us? yes/no`
- `How much are you happy on a scale from 0 to 10?`
- `Your vote is determined by our: ☐ rates ☐ service ☐ other`
- `Write here any other feedback: _____`

The first three answers can be directly **automatically processed** to extract **statistical information**.

The last answer to an **open-ended question** is the only potential source of **unexpected information**.

Why Sentiment Analysis?



Why Sentiment Analysis?

Home » World » Article

Cambodian plane crashes, 20 feared dead

June 25, 2007 - 7:01PM

Tools sponsored by Credit Cards

A chartered plane flying today between two popular tourist destinations in Cambodia has crashed, with at least 20 people on board feared killed, an aviation official said.

The plane, a Russian-made AN-24, was flying from Siem Reap - where the famous Angkor Wat temple complex is located - to Sihanoukville, a coastal city with access to beaches, said Him Sarun, Cabinet chief for the Secretariat of Civil Aviation.

An official at Siem Reap airport said 13 of the passengers were from South Korea, three were Czech, one was Russian and five were Cambodian.

He said the plane carried a crew of five Cambodians.

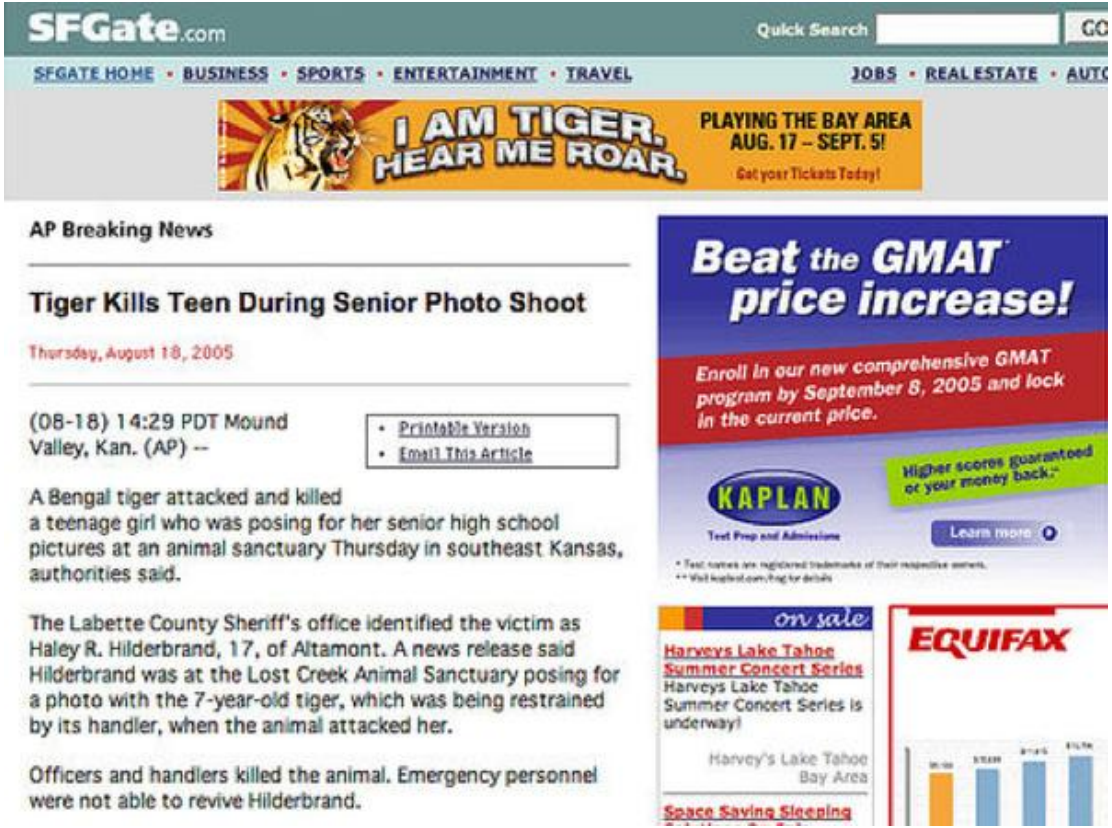
The plane belonged to a small Cambodian

Advertisement

You fly.
We pay.

Until June 30
we pay on-road costs

Why Sentiment Analysis?



Why Sentiment Analysis?

Why Sentiment Analysis?

Q&A Here's why experts say all kids ages 6 and up should be screened for obesity

A government panel recommends that all children ages 6 and up be screened for obesity and referred for treatment if necessary. (Dreamstime)

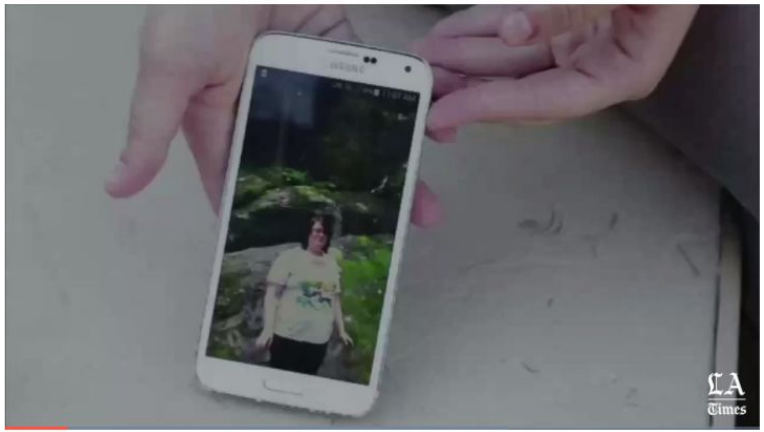
By Karen Kaplan
Contact Reporter

JUNE 20, 2017, 3:55 PM

ADVERTISEMENT

Why Sentiment Analysis?

As obesity keeps rising, more Americans are just giving up



New research shows if you know you're overweight or obese, and you know your extra pounds are unhealthy, that you may not be making a stab at losing weight anymore.

By Melissa Healy · Contact Reporter

MARCH 7, 2017, 3:35 PM

It stands to reason that if you know you're overweight or obese, and you know your extra pounds are unhealthy, that you've made a stab at losing weight. Right?

ADVERTISEMENT

Why Sentiment Analysis?

Why Sentiment Analysis?

Health News

Share this article: Print

Study: 75 percent of Americans will be overweight by 2020

Associated Press
Posted on September 23, 2010 at 12:41 PM

Recommend

PARIS – Citizens of the world's richest countries are getting fatter and fatter and the United States is leading the charge, an organization of leading economies said Thursday in its first ever obesity forecast.

Three out of four Americans will be overweight or obese by 2020, and disease rates and health care spending will balloon, unless governments, individuals and industry cooperate on a comprehensive strategy to combat the epidemic, the study by the Organization for Economic Cooperation and Development said.

Why Sentiment Analysis?

11-year-old charged with driving drunk

REUTERS

NEWS ALERTS

Get an alert when there are new stories about:

- Orange Beach, Alabama
- Perdido Key, Florida
- Chevrolet Monte Carlo

Add Selected Alerts

» More Alerts

ELSEWHERE ON THE WEB

CNN.COM
Explosion rocks besieged mosque in Pakistani capital

ABC NEWS
'Out of the Blue': Do Aliens Exist?

THE CHRISTIAN SCIENCE MONITOR
William Bratton: Lauded chief of troubled LAPD

Fri Jul 6, 3:23 PM ET

MIAMI (Reuters) - An 11-year-old girl was charged with drunken driving after leading police on a chase at speeds of up to 100 mph that ended when she flipped the car in an Alabama beach town.

A video camera in the police car captured the look of surprise on the officer's face when he approached the wrecked car and got a look at the motorist.

The Mobile Press-Register newspaper said the patrolman saw the Chevrolet Monte Carlo speeding and flashed his lights to signal the driver to stop. Instead, the car sped faster, traveling at up to 100 mph (160 kph) before sideswiping another vehicle and flipping over in the Gulf Coast town of Orange Beach, Alabama, on Tuesday night.

ADVERTISEMENT

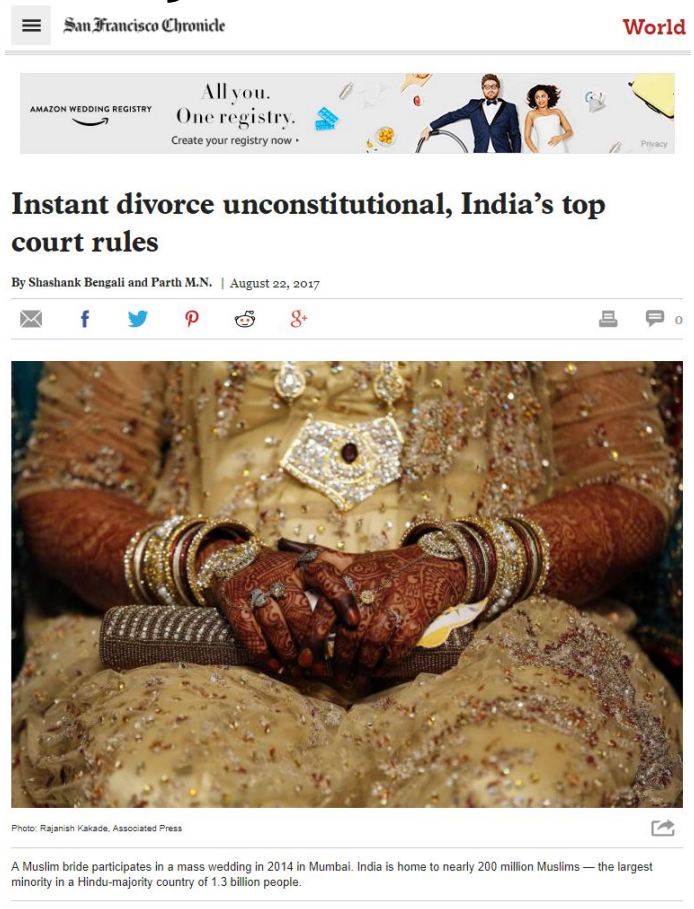
The young driver, who lived nearby in Perdido Key, Florida, was treated at a hospital for scrapes and bruises and released to relatives. Police also charged her with speeding, leaving the scene of an accident and reckless endangerment.

The car belonged to a relative and police were still trying to find out where she got the alcohol. There was none in the vehicle but her blood alcohol level was over the limit for adult motorists, police told the newspaper.

Email Story IM Story Printable View

Why Sentiment Analysis?

Why Sentiment Analysis?



San Francisco Chronicle World

AMAZON WEDDING REGISTRY
All you. One registry.
Create your registry now ›
Privacy

Instant divorce unconstitutional, India's top court rules

By Shashank Bengali and Parth M.N. | August 22, 2017

0

Photo: Rajanish Kakade, Associated Press

A Muslim bride participates in a mass wedding in 2014 in Mumbai. India is home to nearly 200 million Muslims — the largest minority in a Hindu-majority country of 1.3 billion people.

Sentiment Analysis tasks

Sentiment Analysis tasks

Most of SA research and applications are focused on the simple **positive vs negative** dichotomy (or a graded scale among this two opposites).

Most common SA tasks:

- Subjectivity/polarity classification
- Regression
- Opinion extraction
- Quantification

There is also research on **emotions**, **attitude** and **humor** in human language.

Classification

Classification: determining the attitude of the author of a document toward the document subject matter.

By **subjectivity**: determining if the text contains or not subjective evaluations.

"The movie is set in WW2" → Objective

"The plot is confusing" → Subjective

By **polarity**: determining if the subjective evaluations are positive or negative with respect to its topic.

"This movie is a masterpiece" → Positive

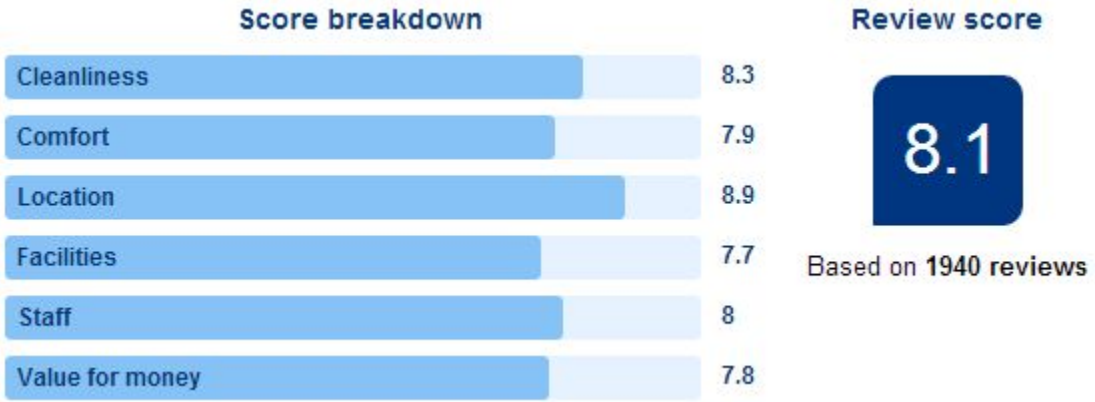
Regression

Regression: extending the polarity classification problem to a ordinal scale.

Typical scenario: “Star rating” of product reviews.

“This phone is not worth its price” ➡ 

Regression can produce a global evaluation or be focused on specific aspects.



Extraction

Extraction: identifying the expressions of an opinion, its properties, and the target of that opinion.

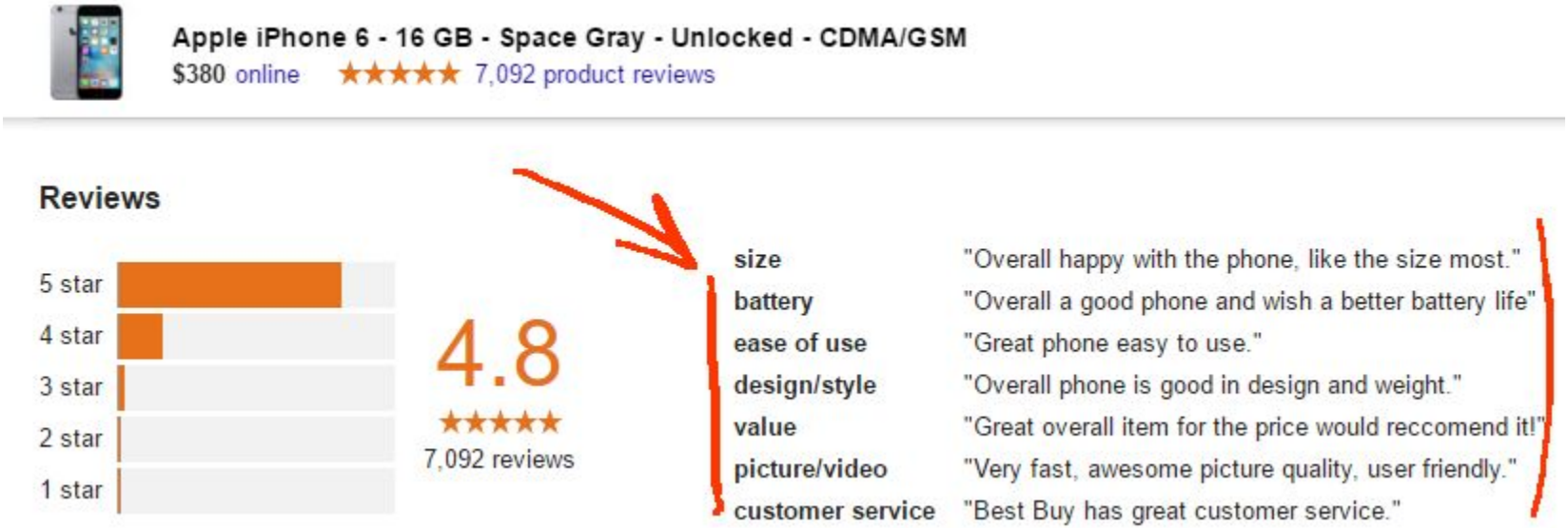
> *"The phone has a great display but it is killed by the small battery"*

`(display: great, positive), (battery: small, negative)`

Extraction is often modeled as a **classification** problem **at the word level**.

The output of extraction contribute to build a **knowledge base**, which can be then **queried** by traditional methods from Information Retrieval and **Data Mining**.

Extraction

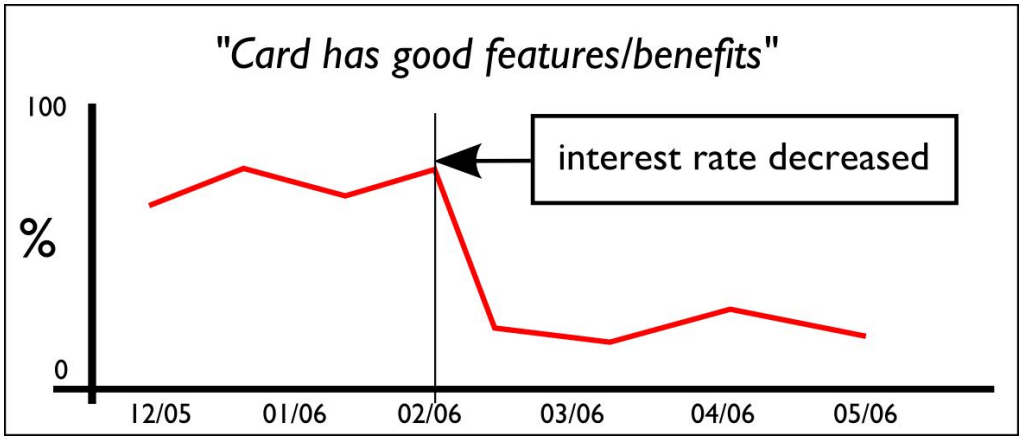


Example of extraction of aspect-related relevant evaluations, Google Shopping

Quantification

Quantification is an aggregate analysis problem: a set of documents is processed as single entity in order to determine some properties of the whole set.

- Determining the proportion, and its trend over time, of positive reviews about a product.



Sentiment Analysis methods

There is no one-stop solution for Sentiment Analysis.

Sentiment Analysis is not a single problem.
Sentiment Analysis is not a dataset.
Sentiment Analysis is not a lexicon.
Sentiment Analysis is not an algorithm.

Sentiment Analysis is a special scenario for text analysis problems.

A “standard” method produces 70-90% of the result.

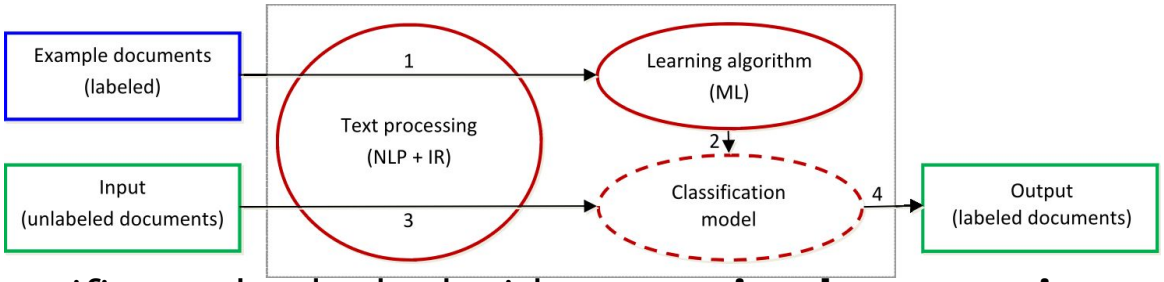
Exploiting the characteristic that are specific of a given Sentiment Analysis problem produces that 10-30% improvement that separates an average solution from a good one.

Sentiment Analysis methods

Multidisciplinary approach:

- Natural Language Processing
- Information Retrieval
- Machine Learning

The template solution to a sentiment analysis problem is the same of a “generic” one, e.g.:



Most of sentiment-specific methods deal with **capturing how sentiment are expressed in natural language**.

The language of opinions

The language of opinions

The language we use to express our subjective evaluations is **one of the most complex parts of language**.

There are many components in the language of opinions:

- Global/Domain-specific lexicon.
- Valence shifters/Comparative expressions.
- Irony, sarcasm, common knowledge.
- . . .

The main aim of NLP/IR/ML applied to Sentiment Analysis is to **recognize** sentiment expressions and to **model** them into **semantic abstractions**.

The language of opinions

Some **words** have a **globally** recognized **sentiment valence** in any context of use, e.g.: *“good”*, *“poor”*, *“perfect”*, *“ugly”*

“A good tool that works perfectly”

“I had an horrible experience”

General purpose lexical resources list these words associating sentiment labels to them, e.g.:

- The General Inquirer lexicon
- WordNet affect
- SentiWordNet

The language of opinions

Domain/aspect-specific expressions: words that have a sentiment valence only when used in the context of a **specific domain**, or when they are associated with a **specific aspect**.

“The phone is made of cheap plastic”

“The carrier offers cheap rates”

“We have got a warm welcome”

“We have got a warm beer”

A collection of text from the domain can be used to build a **domain lexicon**.

The language of opinions

Negation and **valence shifters**: they do not determine sentiment directly but have **influence** on it.

It is difficult to determine their **scope** and **combined effect**.

"This is a very good car" (increment)

"This car is not very good" (flip, decrement)

"I don't like the design of the new Nokia but it contains some intriguing functions"

"Not only is this phone expensive but it is also heavy and difficult to use"

Workshop on Negation and Speculation in NLP

The language of opinions

Punctuation, emoticons, emoji:

“7AM battery 100% - 9AM 30% :(“

Irony, sarcasm:

“Light as a bulldozer”

“The most useful idea since the DVD rewinder”

Common knowledge:

“Windows Vista: the new Windows ME”

“Windows 7: the new Windows XP”

5,448 of 5,551 people found the following review helpful

Good for clensing

By Davie Blossom on August 17, 2009

Format: Paperback

My wife and I were actually quite impressed with this item. Between the two of us it took 9 days to finish and I must say I found it hard to get used to during the first few sittings, but gradualy became more comfortable as time went on.

Although neither of us paid too much attention to the content, the sheer volume and quality of the paper contained within was pleasantly suprizing. The paper feels crisp, heavy and tough, yet the pages seperate with just the slightest tug of a thumb and forefinger.

The only real controversy this book created for my wife and I was mostly my fault. You see after I finished the last page I was too lazy to replace the book with another and I left the empty cover sitting on top of the sistern. Let me tell you I copped a nagging!

Overall the item was an interesting change, but we found it a bit expensive and not as suitable as the regular stuff available at the supermarket. Together my wife and I have decided just to stick with normal sorbent two-ply rolls from now on, and would suggest others do the same.


AutoExec

A model for Sentiment Analysis

A model for Sentiment Analysis

Referring to Bing Liu's model, an opinion, in the context of a sentiment analysis problem, can be defined as a **quintuple**:

$$<e_i, a_{ij}, s_{ijkl}, h_k, t_l>$$

where

- e_i is the **entity** that is the **target** of the opinion
- a_{ij} is the **aspect** of the entity e_i that is the target of the opinion
- s_{ijkl} is the **sentiment** toward a_{ij} expressed by h_k at time t_l
- h_k is the **holder** of the opinion, i.e., who expresses the opinion
- t_l is the **time** the when opinion has been expressed

A model for Sentiment Analysis

The entity-aspect pair identify the subject of the opinion expression, which can be refer to a main object, a sub-part, or an aspect of a sub-part.

“iPhone is great”

`<e=iPhone, a=GENERAL,...>`

GENERAL indicates that the entity as a whole is the target of opinion.

“iPhone battery sucks”

`<e=iPhone, a=battery,...>`

A model for Sentiment Analysis

Sub-parts/aspects can be arranged in a hierarchy.

“iPhone display has a good resolution, but colors are washed out”

<e=iPhone, a=display,...>

<e=iPhone, a=display,...>

<e=iPhone, a=display/resolution,...>

<e=iPhone, a=display/color,...>

A model for Sentiment Analysis

Sentiment can be defined as binary “positive” vs “negative” labeling, include also a “neutral” label, or use a graded scale.

“iPhone display has a good resolution, but colors are washed out”

`<e=iPhone, a=display/resolution, s=positive,...>`

`<e=iPhone, a=display/color, s=negative,...>`

“iPhone display has an amazing resolution, but colors are bit washed out”

`<e=iPhone, a=display/resolution, s=5/5,...>`

`<e=iPhone, a=display/color, s=2/5,...>`

A model for Sentiment Analysis

The opinion holder may be the **writer** of the text, or the text may **report** someone else's opinion:

"I love my new bicycle"

`<e=bicycle, a=GENERAL, s=positive, h=WRITER,...>`

"My friend hates my new bicycle"

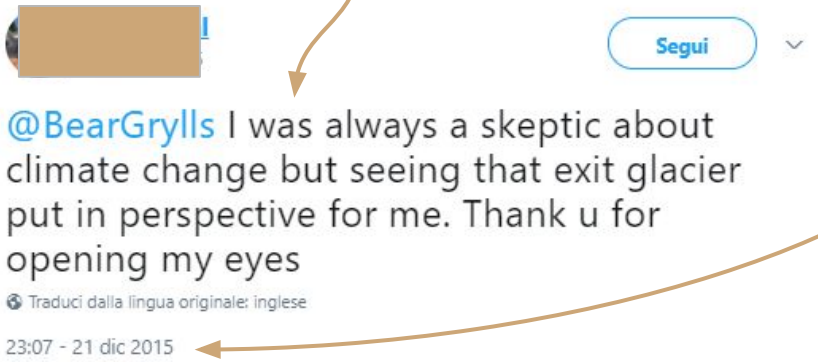
`<e=bicycle, a=GENERAL, s=positive, h=EnviousFriend,...>`

Tracking opinion holders is useful, e.g., in social debates analysis and recurring market research activities.

A model for Sentiment Analysis

Time is a relevant dimension whenever the analysis process is **recurrent** or it is focused on an evolving situation, e.g., elections, social reaction to relevant events.

In many cases time can be tracked from **metadata**.
A dedicated analysis can improve dating accuracy.



A model for Sentiment Analysis

Liu's model is a simple model for direct, non-contextualized, and non-comparative opinions.

“A boring story if you expect to see an action movie.”
“The role and the actor don’t fit together”
“Both X and Y are good, but X is better than Y”

Yet, it covers most of the applications, which can be seen as more or less simplified instances of the model.

It’s an example of a framework to translate the **unstructured information** contained in text into a **structured knowledge base**, on which traditional **data mining** methods can be applied.

Sentiments, Emotions, Humor

Affective computing

Modern Sentiment Analysis applications are mainly **data mining oriented** and focused on the evaluations expressed toward the subject matter of the text.

There is also active research on the topic of **affective computing**, more related to **psychology** and **cognitive sciences**.

In affective computing the focus is on the **human computer interaction**, aiming at identifying the **emotions** and **feelings conveyed** by the text **to the reader**.

Affective computing

Recognizing the expression of six basic emotions: anger, disgust, fear, joy, sadness and surprise:

"He looked at his father lying drunk on the floor" (disgust)

"She was leaving and she would never see him again" (sadness)

"She turned and suddenly disappeared from their view" (surprise)

"They celebrated their achievement with an epic party" (joy)

Strapparava and Mihalcea. *Learning to Identify Emotions in Text*. SAC 2008

Computational humor

Generating and recognizing humor: jokes, puns, wordplay.

“Beauty is in the eye of the beholder”
“Beauty is in the eye of the beer holder”

Generation is usually based on templates, recognition is mainly based on stylistic features.

An example of application is building a language playground for people with complex communication needs.

Ritchie et al. *A practical application of computational humour*. ICCC 2007.
Mihalcea and Strapparava. *Learning to Laugh (Automatically): Computational Models for Humor Recognition*. Computational Intelligence, 2006.

Irony and sarcasm

Irony and sarcasm are pervasive on social media.

Both are linguistic phenomena that rely on context and common knowledge.

Irony and sarcasm

Research on computational recognition of irony is at an early stage, mainly focusing on syntactic features.
Data is often collected from tweets with *#ironic* or *#sarcasm* hashtag.

Wallace, "Computational irony: A survey and new perspectives" AIR 2015
Hernández & Rosso "Irony, Sarcasm, and Sentiment Analysis" Chapter 7 in "Sentiment Analysis in Social Networks" Liu, Messina, Fersini, Pozzi