

[Text Processing] Sentiment Analysis

Loïc Barrault

l.barrault@sheffield.ac.uk
Department of Computer Science

November 12, 2019

Text Processing: Overview

- Introduction [RG]
- Text encoding [RG]
- Information Retrieval [RG]
- Text Compression [RG]

Text Processing: Overview

- Introduction [RG]
- Text encoding [RG]
- Information Retrieval [RG]
- Text Compression [RG]
- Sentiment Analysis [LB]
- Information Extraction [LB]

Text Processing: SA and IE Assessment

COM3110: 10 Credit module:

- Exam [75%]
- Assignment [25%]

COM4115: 15 Credit module:

- Exam [50%]
- Assignment 1 [25%]
- Assignment 2 [25%]

Assignment 2:

- Handout on Tuesday 19 November (next week)
- Deadline: Friday 13 December

[Sentiment Analysis]

Motivations, definitions and more

The world is digital!



[Source: <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>]

The world is digital!

People express their **emotions**, **sentiments** or **opinions**

- comments on products (Amazon, Rakuten)
- comments on movies (Rotten Tomatoes, IMDB, Youtube)
- experience in restaurants (Yelp!, Trip Advisors)
- community websites (Facebook, Twitter, Instagram, LinkedIn, Reddit, Flickr)

General goal of SA

Extract **emotions**, **sentiments** or **opinions**

- expressed by humans in **texts**
- Use that information for business or intelligence purposes

→ Answer questions like:

- Do people liked this movie?
- Do people agree with this law?
- Customers: should you buy this washing machine?
- Prediction: who will win next election?

Sentiment analysis = opinion mining

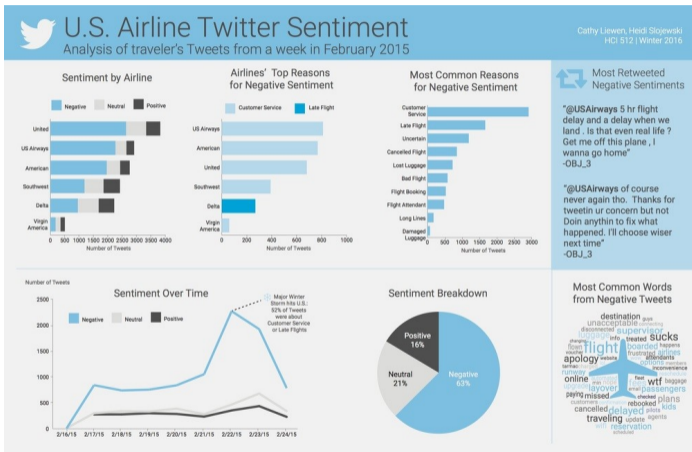
- considered as equivalent in our case...
- ...although a sentiment does not necessarily express an opinion!

Importance of SA

To make decisions!

- Politics: can replace surveys, polls, etc.
 - Market: complete/verify/correct vendor's advice
 - Television: decide whether to continue or stop a TV show
 - at a larger scale!
- people express themselves on the web on any subject
- provides an estimate of the global opinion

Example: social media monitoring



[Source: <https://www.heidisiojewski.com/blog/2016/3/11/m301zbt9orke84cuhcw37j1wwumn6n>]

An example + concepts

An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

An example + concepts

An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

- "❤️ you" is a positive **opinion**

An example + concepts

An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

- "❤️ you" is a positive **opinion**
- "I normally ❤️ you" is a less positive **opinion**

An example + concepts

An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

- "❤️ you" is a positive **opinion**
- "I normally ❤️ you" is a less positive **opinion**
- "was the worst." is a negative **opinion**

An example + concepts

An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

- "❤️ you" is a positive **opinion**
- "I normally ❤️ you" is a less positive **opinion**
- "was the worst." is a negative **opinion**
- "@JetBlue" and "Late Flight flight experience": the **target of opinions**

An example + concepts

An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

- "❤️ you" is a positive **opinion**
- "I normally ❤️ you" is a less positive **opinion**
- "was the worst." is a negative **opinion**
- "@JetBlue" and "Late Flight flight experience": the **target of opinions**
- "2 hours on runway, no wifi & tv not working properly" is the **objective** part

An example + concepts

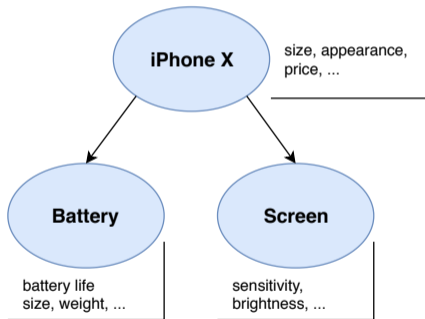
An example

@JetBlue , I normally ❤️ you , but this Late Flight flight experience was the worst.
2 hours on runway, no wifi & tv not working properly

- "❤️ you" is a positive **opinion**
- "I normally ❤️ you" is a less positive **opinion**
- "was the worst." is a negative **opinion**
- "@JetBlue" and "Late Flight flight experience": the **target of opinions**
- "2 hours on runway, no wifi & tv not working properly" is the **objective** part
- "I" is the **holder of opinion**

Target of opinions

- **Target of opinions** is a **person**, **event**, **organisation**, or **topic**
- Represented as a hierarchy of **components** having a set of **attributes**
→ let's call them both **features**
- An **opinion** is expressed on any **feature**



Factual vs. subjective data

Facts

- ***something that is known to have happened or to exist, for which proof or information exists***
- can be expressed using *keywords*
- less/not subject to interpretation
- might be difficult to identify though!

Opinions / subjective data

- ***a thought, belief, judgment about something or someone***
- no proof required
- can be hard to express with keywords

Factual vs. subjective data: examples

What do people think of

- @SouthwestAir just did, thank you
- @united which is why my next flights to Miami will be on another airline.
- @JetBlue please provide me your direct email for me to explain.
- @united worst service ever 🙄
- It costed 500 dollars.
- It costed 500 dollars!!!

Subjectivity analysis

Subjectivity classification is often the first step for sentiment analysis.

Aims: decide whether a text is **objective** or **subjective**

- **objective**: It costed 500 dollars.
- **subjective**: this Late Flight flight experience was the worst.

However, it is not always as simple as that...

- **objective** sentences can express opinion **indirectly**:
 - *My phone broke in the second day.*
- **subjective** sentences do not always express positive or negative opinions:
 - *I think they will refund us.*

Bing Liu's model for Sentiment Analysis

An **opinion** is a quintuple $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$, where:

- o_j is the targeted **object** of opinion. Also called **entity**.
- f_{jk} is a **feature** of the object o_j . Also called **aspect**.
- so_{ijkl} is the **sentiment value**
- h_i is the **sentiment holder**
- t_l is the **time**

so_{ijkl} can take several forms:

- positive, negative, neutral
- 1 to 5 stars, as in movie reviews
- more granular ratings.



Example

On 25th Oct. 2018, John Doe wrote:

“I bought the new iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life is not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop.”

Example

On 25th Oct. 2018 , John Doe wrote:

"I bought the new iPhone a few days ago. It was such a nice phone . The touch screen was really cool . The voice quality was clear too. Although the battery life is not long , that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive , and wanted me to return it to the shop."

- o_j : iPhone
- f_{jk} : phone, touch screen, voice quality, battery life, phone
- so_{ijkl} : positive , positive , positive , negative , negative
- opinion holder h_i : I, I, I, I, mother
- time t_j : 25/10/2018

Opinion mining

- Discover all quintuples $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$ in a document
- **Structure the unstructured**
 - Use in data visualisation tools
 - **Quantitative** and **qualitative** analysis
 - To answer questions like:
 - Do people like the new iPhone?
 - Among them, what are the features that they most dislike?
 - Among people that don't like the iPhone, what feature is missing?

Sentiment Analysis: not just ONE problem

Discover all quintuples $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$ in a document

- o_j – the target **object** or **entity**:
 - **Named Entity Recognition.**
 - **Coreference resolution.**
- f_{jk} – a **feature** or **aspect**: **Information Extraction** (more about that later...).
- so_{ijkl} – **sentiment value**: **Sentiment Identification**
- h_i – **sentiment holder**: **Information/Data Extraction**
- t_l – **time**: **Information/Data Extraction**

Sentiment Analysis: granularity: document level

- **Document** level

- Assumption: document focuses on a **single object** from a **single opinion holder**.
- Goal: discover ($_, _, so, _, _$)
→ ignore o_j, f_{jk}, h_i, t_l

- **Reviews** usually satisfy the assumption

- Positive = 4 or 5 ☆, Neutral = 3 ☆, Negative = 1 or 2 ☆

Sentiment Analysis: granularity: sentence level

- **Sentence** level SA is similar to document level
 - Assumption: sentence contains a **single opinion** from a **single opinion holder**
 - Only an **intermediate** step
 - Consists of two steps:
 - ① **Subjectivity** classification \Rightarrow **detect** if the sentence expresses an opinion
 - ② **Sentiment** classification \Rightarrow identifies the sentence **polarity**

Sentiment Analysis: granularity: sentence level

- **Sentence** level SA is similar to document level
 - Assumption: sentence contains a **single opinion** from a **single opinion holder**
 - Only an **intermediate** step
 - Consists of two steps:
 - ① **Subjectivity** classification \Rightarrow **detect** if the sentence expresses an opinion
 - ② **Sentiment** classification \Rightarrow identifies the sentence **polarity**
- ① **Subjectivity** classification
 - Bootstrapping method [1]
 - Sentence similarity [2]
 - Naive Bayes

Sentiment Analysis: granularity: sentence level

- **Sentence** level SA is similar to document level
 - Assumption: sentence contains a **single opinion** from a **single opinion holder**
 - Only an **intermediate** step
 - Consists of two steps:
 - ① **Subjectivity** classification \Rightarrow **detect** if the sentence expresses an opinion
 - ② **Sentiment** classification \Rightarrow identifies the sentence **polarity**
- ① **Subjectivity** classification
 - Bootstrapping method [1]
 - Sentence similarity [2]
 - Naive Bayes
- ② **Sentiment** classification
 - Rule-based
 - Corpus -based

\rightarrow detailed later... stay tuned!

Sentiment Analysis: granularity: feature level 1/2

Motivation

- Document and sentences may contain **mixed** opinions
 - **Targets** of opinions (**features**) are ignored
 - SA at this level provides a **general** opinion on the **object**
 - Does not mean that **opinion holder** likes/dislikes everything about it.
-
- **Feature** level aims to provide a more fine-grained analysis
 - **which** component people like/dislike
 - more informative analysis
 - consists of 5 steps

Sentiment Analysis: granularity: feature level 2/2

① Identify **entities** or **objects**

- Given a set Q of **seed entities** of class C , and a set D of **candidate entities**, determine which of the entities in D belong to C .
→ binary classification problem

Sentiment Analysis: granularity: feature level 2/2

① Identify **entities** or **objects**

- Given a set Q of **seed entities** of class C , and a set D of **candidate entities**, determine which of the entities in D belong to C .

→ binary classification problem

② Extract object **features** that have been commented on by the **opinion holder**

- Frequency-based methods [3]
- Challenge: infrequent feature extraction

Sentiment Analysis: granularity: feature level 2/2

① Identify **entities** or **objects**

- Given a set Q of **seed entities** of class C , and a set D of **candidate entities**, determine which of the entities in D belong to C .

→ binary classification problem

② Extract object **features** that have been commented on by the **opinion holder**

- Frequency-based methods [3]
- Challenge: infrequent feature extraction

③ Group similar **features**

- e.g. *screen* and *touch screen*
- e.g. *power usage* and *battery life*

Sentiment Analysis: granularity: feature level 2/2

① Identify **entities** or **objects**

- Given a set Q of **seed entities** of class C , and a set D of **candidate entities**, determine which of the entities in D belong to C .

→ binary classification problem

② Extract object **features** that have been commented on by the **opinion holder**

- Frequency-based methods [3]
- Challenge: infrequent feature extraction

③ Group similar **features**

- e.g. *screen* and *touch screen*
- e.g. *power usage* and *battery life*

④ Classify the opinions as **positive**, **neutral** or **negative**

⇒ detailed later

Sentiment Analysis: granularity: feature level 2/2

- 1 Identify **entities** or **objects**
 - Given a set Q of **seed entities** of class C , and a set D of **candidate entities**, determine which of the entities in D belong to C .
 - binary classification problem
- 2 Extract object **features** that have been commented on by the **opinion holder**
 - Frequency-based methods [3]
 - Challenge: infrequent feature extraction
- 3 Group similar **features**
 - e.g. *screen* and *touch screen*
 - e.g. *power usage* and *battery life*
- 4 Classify the opinions as **positive**, **neutral** or **negative**
 - ⇒ detailed later
- 5 [optional] Produce a summary of all feature-based opinions

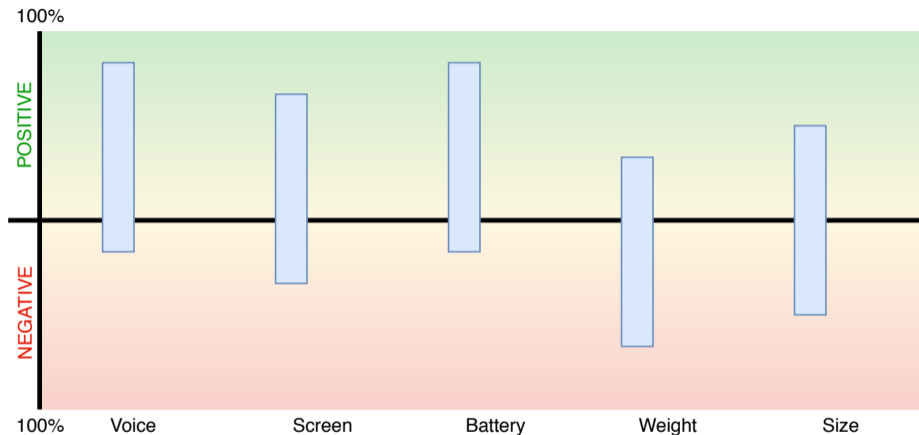
Sentiment Analysis: result

On 25th Oct. 2018 , John Doe wrote:

"I bought the new iPhone a few days ago. It was such a nice phone . The touch screen was really cool . The voice quality was clear too. Although the battery life is not long , that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive , and wanted me to return it to the shop."

Component	Polarity	Comment
Feature1		The touch screen was really cool.
<i>screen</i>	Positive	The touch screen was so easy to use and can do amazing things.
<i>touch screen</i>		The screen is so fluid!
	Negative	The screen is easily scratched.
		It's very difficult to remove finger marks from the touch screen.
Feature2		The battery stood for 2 days.
<i>battery life</i>	Positive	The battery lasts very long.
	Negative	I had to charge it after 2 hours...
		...

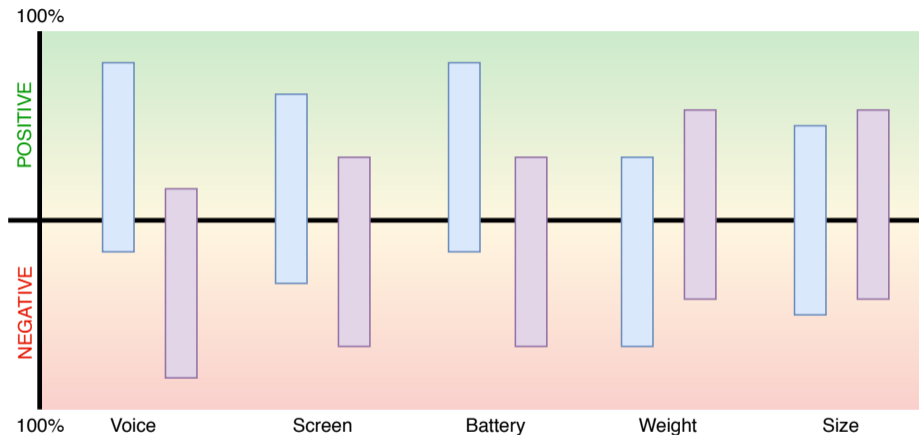
Sentiment Analysis: opinion observer



Summary of reviews for smartphone 1

[4]

Sentiment Analysis: opinion observer

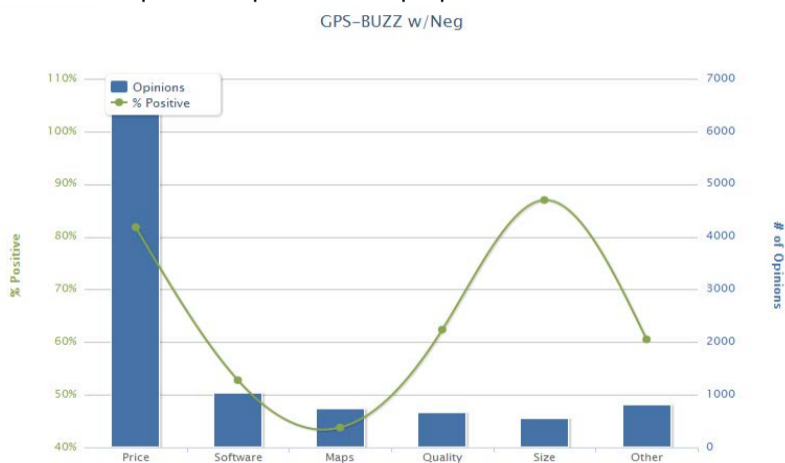


Comparison of summaries of reviews for two smartphones

[4]

Sentiment Analysis: indicators

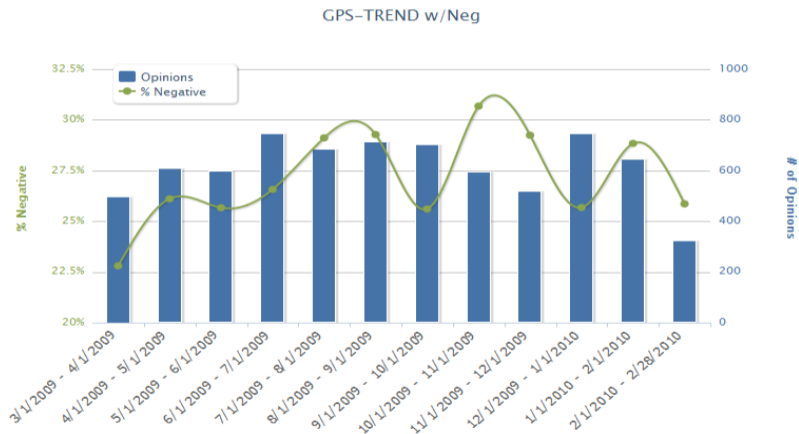
- absolute number of expressed opinions and proportion



[5]

Sentiment Analysis: indicators

- evolution of opinions across time



[5]

Sentiment Analysis: challenges

On 4th Nov. 2018 , John Doe wrote:

This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone with Bluetooth . We called each other when we got home.

The voice on my phone was not so clear , worse than my previous phone.

The battery life was short too.

My girlfriend was quite happy with her phone.

I wanted a phone with good sound quality . So my purchase was a real disappointment . I returned the phone yesterday.

Sentiment Analysis: challenges

On 4th Nov. 2018 , John Doe wrote:

This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone with Bluetooth . We called each other when we got home.

The voice on my phone was not so clear , worse than my previous phone.

The battery life was short too.

My girlfriend was quite happy with her phone.

I wanted a phone with good sound quality . So my purchase was a real disappointment . I returned the phone yesterday.

→ Identify all the components of an opinion

Sentiment Analysis: challenges

On 4th Nov. 2018 , John Doe wrote:

This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone with Bluetooth . We called each other when we got home.

The voice on my phone was not so clear , worse than my previous phone.

The battery life was short too.

My girlfriend was quite happy with her phone.

I wanted a phone with good sound quality . So my purchase was a real disappointment .

I returned the phone yesterday.

→ Identify all the components of an opinion

→ Identify the relations

Sentiment Analysis: Challenges

Discover all quintuples $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$ in a document

- o_j – the target **object** or **entity**:
 - **Named Entity Recognition.**
 - **Coreference resolution.**
- f_{jk} – a **feature** or **aspect**: **Information Extraction** (more about that later...).
- so_{ijkl} – **sentiment value**: **Sentiment Identification**
- h_i – **sentiment holder**: **Information/Meta-Data Extraction**
- t_l – **time**: **Information/Meta-Data Extraction**

In addition:

- **Relation Extraction**
- **Synonym match**: e.g. "voice" == "sound quality"

References I

E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” pp. 105–112, 2003. [Online]. Available: <https://www.aclweb.org/anthology/W03-1014/>

H. Yu and V. Hatzivassiloglou, “Towards answering opinion questions,” in *Proceedings of the 2003 conference on Empirical methods in natural language processing -*, vol. 10. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 129–136. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1119355.1119372>

M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. New York, New York, USA: ACM Press, 2004, p. 168. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1014052.1014073>

References II

B. Liu, M. Hu, and J. Cheng, “Opinion observer: analyzing and comparing opinions on the Web,” *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1060745.1060797>

B. Liu, “Sentiment Analysis and Opinion Mining,” in *Proceedings of AAAI*, San Francisco, California, USA, 2011. [Online]. Available: http://www.cs.uic.edu/~lzhang3/paper/opinion_survey.pdf