Sentiment Analysis and Opinion Mining

Chenghua Lin
Dept. of Computing Science
University of Aberdeen

Introduction to Sentiment Analysis

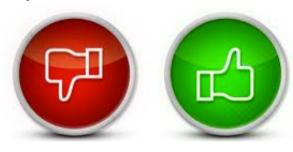
What is Sentiment Analysis

Sentiment

 A thought, view, or attitude, especially one based mainly on emotion instead of reason

Sentiment Analysis

- Computational treatments for discovering opinions and attitudes expressed in text by opinion holders (e.g. positive vs. negative)
- Can be generalised to richer emotion dimensions (e.g., Joy, Sadness, Fear, Anger, etc.)



Definitions

Modelling opinion as a quintuple: (Bing Liu 2012)

(o, a, so, h, t)

Tom thinks this book is great

Opinion

Holder

g weight battery

Opinion

Where

- o is an object or target entity
- a is the aspect or attribute of
- **so** is the sentiment orientation or use οριπίοπ ασουι an object or aspect
- h is an opinion holder
- t is the timestamp when an opinion is expressed.

Different Types of Opinions

- Explicit opinion: opinion or sentiment directly expressed on a target object
 - "My mood is really bad."
 - "His basketball skill is really amazing!"
- Implicit opinion: objective text implying opinion or attitude
 - "The new bike fell apart within two days."
 - "He has learnt a lot from this course."
- <u>Sarcasm</u>: a sharp or bitter remark, usually conveys the opposite of their literal meaning (context-dependent)
 - Context: "The food is totally burned! (very angry)"
 - "You really did a great job!"

Why Sentiment Analysis

- Web 2.0 and the social web
 - has facilitated rich user-generated contents
 - contains valuable information for both business and end-users
- We have a <u>decision support</u> need and an <u>operational</u> need
 - > Marketers and governments
 - Finding out customers' opinions about their products/services
 - Tracking how these opinions evolve over time
 - Accessing public opinion polls on political campaigns
 - **>** Consumers
 - Decision support for purchasing and recommendation
 - What are people saying about X versus Y

Some Practical Examples

Product Review Insights

Customer Reviews Amazon Kindle Keyboard Leather Cover, Black

855 Reviews			Average Customer Review			
5 star:		(594)	**** (855 customer reviews)			
4 star:		(167)	Share your thoughts with other			
3 star:		(47)	customers			
2 star:		(22)	Croate wave aven ravious			
1 star:		(25)	Create your own review			

- What are people's opinions about this product?
- What are the pros and cons?

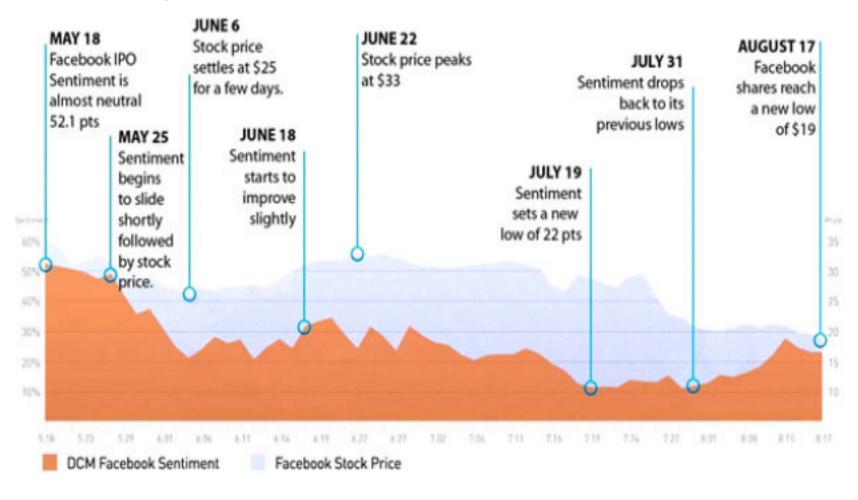
Brand and Consumer Perception

 <u>Music artists analytics</u>: provide aggregated sentiment statistics for artists, songs or albums over all reviews collected online.



Financial Marketing

 Use general Facebook sentiment to predict the company's stock performance. (http://www.dailyfinance.com/)



tweetfeel 4

Percentage of positive sentiment about "iphone 4s",

iphone 4s

Try some Twitter trends: No Basketball Anymore Drake and Big Sean Stop Online Piracy Act Sepp D





Whoa the iphone 4s is great! My dad is thinking about getting one for me for xmas! Thanks dad luv ya!



@ssaudiaaK fuck iphone 4s





I really Love iphone 4s



@Ahmedmob8 thank u for listening they are

I hate iphone 4s please come u



re and Kids Luda



I hate iphone 4s please come up with a jailbreak before I sell this shit lol @Gojohnnyboi



@andyrmyers tell her it would be easier to transfer it to your 3gs and then suggest that you could use the shiny new iphone 4s...! WIN

Sentiment Analysis Tasks

The Big Picture

Sentiment classification

Is a document/sentence positive or negative?

Subjectivity detection

Whether given text expresses opinions (subjective) or reports facts (objective)

Opinion holder/target identification

Who express a specific opinion? What features of the iPhone 5 do customers like?

Opinion summarisation

Summarise opinions over multiple review documents towards a certain product

Opinion retrieval

How do people think of iPhone5?

Sentiment dynamics prediction and tracking

How does people's views on Mac change over time?

Opinion spam detection

Opinion spam detection: Identify fake/ untruthful reviews.

Sentiment Classification

Sentiment Classification

- Goal: classify the overall sentiment orientation expressed in a given text, i.e. positive, negative or (possibly) natural.
- Classification can involve different levels of granularity
 - Document-level
 - Sentence-level
 - Word/phrases-level
- Traditional sentiment classification techniques
 - Lexicon-based approaches
 - Corpus-based approaches

Sentiment Classification Techniques



- Lexicon-based approaches
 - Use sentiment lexicons as prior knowledge
 - Unsupervised/weakly supervised learning



- Corpus-based approaches
 - Annotated corpus with class labels available (e.g. positive or negative)
 - Supervised learning, e.g., Naïve Bayes (NB),
 SVMs, Maximum Entropy Model (MaxEnt),
 etc.

Preprocessing

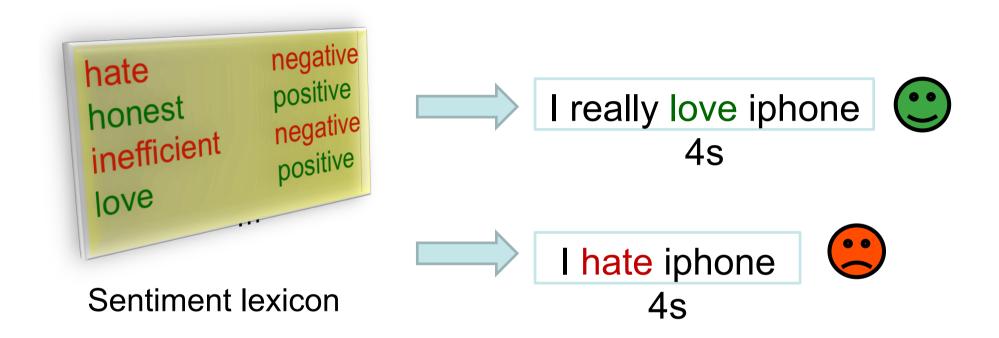
- Text preprocessing
 - An essential part of any NLP system
 - Segment text into appropriate unit (e.g. characters, words, sentences, etc.) and prepare in appropriate forms (e.g. canonical form) before passing for further processing
 - Also called text tokenization and normalization
- Easier in some languages (e.g. English) and highly nontrivial for the languages without space between words (e.g. Chinese).
- Have a great impact on NLP system performance, e.g. speed, accuracy, etc.

Preprocessing (cont.)

Typical steps of preprocessing in English:

- Clean up unwanted stuff, e.g. HTML tags. But sometimes can be helpful, e.g. tags preserving document structures like headings, sections.
- Text unit segmentation: normally white space and punctuations as word boundaries. Instances like 'e-mail', 'aren't', can be problematic.
- Stopword removal: (1) remove the most frequent words that do not carry much meaning. E.g., 'the', 'and', 'a', etc. (2) can greatly reduce corpus size.
- Stemming: convert the inflected words to their stem, e.g.,
 'stocking', 'stocks', 'stocked' → 'stock'.
 - Porter stemmer
 - Reduce vocabulary size
 - May collapse words with different meaning into the same stem 'pass', 'passe' → 'pass'

Lexicon-Based Approaches



- Use sentiment words as reference features for polarity detection
- Does not rely on labelled data for training

Corpus-based Approaches

- Basic idea: treat sentiment classification as a binary classification problem with two topics, i.e. 'positive' and 'negative'.
- Supervised classification algorithms
 - Naïve Bayes (NB)
 - Support Vector Machines (SVM)
 - Maximum Entropy (MaxEnt), etc.
- Commonly used benchmark dataset
 - Movie reviews (http://www.cs.cornell.edu/people/pabo/movie-review-data/)
 - Product reviews (http://www.cs.jhu.edu/~mdredze/datasets/sentiment/)

Corpus-based methods (cont.)

**** Some flaws, but overall, GREAT, 25 Oct 2011

**** The best? Maybe so....., 26 Oct 2011

A limited device, 29 Oct 2011

By <u>A reviewer</u> (United Kingdom) - <u>See all my reviews</u>

This review is from: Apple iPhone 4S 16GB Black (Electronics)

I'm not "an Apple fanatic with the ethos 'if it aint Apple don't bother'", so you will get something balanced here, but I will say that I purchased an iPhone 4S with a strong desire to like it. I really tried my best and intended to use it exclusively, but due to me having already experienced Android, it had to go back to the shop.

I don't care who makes a product or what their marketing is like, I care about how versatile and useful the product is and in this respect I just couldn't avoid the obvious conclusion that this device is deficient. Shock, horror, Apple?! Yes, they don't walk on water, they just have slick marketing.

What were the problems? I'll just list those I discovered in the few days using the phone. Some of these I suppose are going to be subjective but I'll just tell you how I found it:

Training set

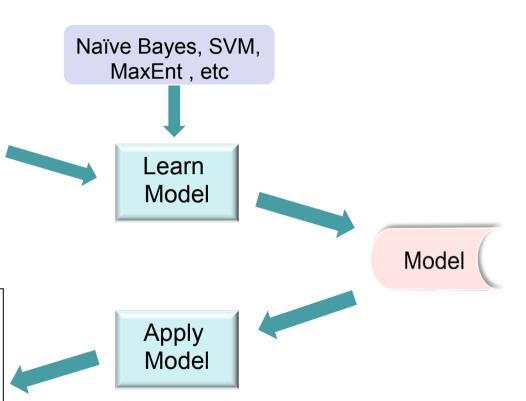
By M. Bond (London) - See all my reviews

By Dr. W. E. Allen "wallen200" (Belfast, UK) - See all my reviews

This review is from: Apple iPhone 4S 16GB Black (Electronics)

The first thing I need to say is that the Apple iPhone 4S is the best smart phone in the market at present, and unless something radical happens will probably be the best smart phone until the iPhone 5 is released. I am not going to labour all the features, these have been well covered in the description and the previous reviews. However I will say that this phone is definitely not worth upgrading to from the iPhone 4 and even if you have an iPhone 3GS I would say it would be better to wait until the next generation iPhone comes out. The reason I say this is that this phone has really only two differences from the iPhone 4 - Siri and a higher resolution camera. I will discuss these first.

Test set



Rely on syntactic or co-occurrence patterns in large text corpora

Example: Supervised Learning

Supervised machine learning algorithms for sentiment classification (Pang et al, 2002, 2004)

 Goal: predict the sentiment orientation of a document as positive or negatives

Algorithms

- Navie Bayes (NB)
- Support Vector Machines (SVMs)
- Maximum entropy (MaxEnt)

Data and setting

- 700 positive (4-5 stars) and 700 negative (1-2 stars) reviews
- 3-fold cross-validation
- No stemming or stopword removal

Feature Engineering

- Features used for building classifier
 - Unigrams [I, like, the, new, iPad]
 - Bigrams [I_like, like_the, the_new, new_iPad]
 - POS [I(p), like(v), the(d), new(a), iPad(n)]
 - Negation [not_good]
 - Punctuation [!!]
 - Position
 - Frequency vs. presence

Performance

- SVMs performs best
 - With 82.9% accuracy on the Movie review dataset
 - Based on unigram (presence) features

	Features	# of	frequency or	NB	ME	SVM
		features	presence?			
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

What you should know

- What is opinion mining
- What is sentiment classification
- Text preprocessing
- Feature engineering
- How to perform corpus-based sentiment classification