

Opinion Mining

Presenter: Mathias Verbeke (KULeuven)

Opponent: Wim Van Eynde (UAntwerpen)

Users Speak Out: Canon PowerShot SX10 IS

BY: Allison J, *DigitalCameraReview.com* Editor

PUBLISHED: 4/13/2009

The [Canon PowerShot SX10 IS](#) has been on store shelves for photographers – since November. User reviews from our site have been in ever since, and we've rounded up a small sample for



Since the release of the SX10, [Canon](#) added the SX11 IS to the line, featuring a CMOS sensor and HD video recording. However, the SX10 remains the most-viewed camera on our website. It sports a 10x optical zoom, a 2.5" LCD, and a 10.0 MP sensor. The whole package is available at [Amazon.com](#) or [B&H Photo](#) shop.

👍 One of the best ultrazooms on the market

Submitted by Kamugin on 1/30/2009

PROS: Huge zoom, hot shoe for external flashgun, tilt LCD, AA batteries equipped, good ergonomic design with useful buttons and controls, plenty of user friendly features and full manual mode, compartment for memory card isn't in the same place as batteries, efficient image stabilization.

CONS: Small LCD, heavy, cheap lens cover and hood, cheap cover for USB connector, hard to open battery hatch, not too good autofocus especially with artificial light, no RAW mode, no battery level indicator.

OVERALL SCORE: 8/10

👎 Slow Lens, Clumsy Controls a Poor Mix

Submitted by Rachel*B on 11/7/2008

PROS: Excellent resolution, reasonably sharp images. Better noise control than Canon's previous "S" models, ISO usable through 400; 800 in a pinch. Image edges, corners are sharper. Lens aberration is less than average for this zoom range. Huge zoom range includes wide angle. Digital zoom images are surprisingly good. iContrast increases dynamic range, means more shadow detail, less highlight clipping. Bright vari-angle LCD with wider viewing.

CONS: Slow lens beyond the 100mm mark – image stabilization can't keep up unless ISO 1600 is employed, which is much too noisy. Need a monopod for much zooming below ISO 800 in less than bright light. Images have an unpleasant blue cast that is hard to stomach. Autofocus is very slow, difficult to achieve indoors and often fails altogether; and while autofocus outdoors in bright light is usually snappy, failure to focus also occurs sometimes even under the most optimal conditions. Control wheel is clumsy and frustrating.

OVERALL SCORE: 5/10

... or how to select a new digital camera?

Overview

1. Introduction
2. Three mining tasks
 1. Sentiment classification
 2. Feature based opinion mining and summarization
 3. Comparative sentence and relation mining
3. Opinion search
4. Opinion spam
5. Conclusion

User generated content

= personal experiences and opinions on almost anything, at review sites, forums, discussion groups, blogs ...

a.k.a. **word-of-mouth behavior**

New opportunity: mine opinions expressed in the user generated content

- > Intellectually very challenging
- > BUT Practically very useful!

Applications

- **Businesses and organizations:** product and service benchmarking, market intelligence.
 - Business spends a huge amount of money to find consumer sentiments and opinions.
 - e.g. Consultants, surveys and focused groups, etc
- **Individuals:** interested in other's opinions when
 - Purchasing a product or using a service,
 - Finding opinions on political topics
- **Ads placements:** Placing ads in the user-generated content
 - Place an ad when one praises a product.
 - Place an ad from a competitor if one criticizes a product.

• ...

Definitions

- **Basic components of an opinion**
 - **Opinion holder:** The person or organization that holds a specific opinion on a particular object.
 - **Object:** on which an opinion is expressed
 - **Opinion:** a view, attitude, or appraisal on an object from an opinion holder.

Definitions (2)

- **Definition (object):** An object O is an entity which can be a product, person, event, organization, or topic. O is represented as
 - a hierarchy of components, sub-components, and so on.
 - Each node represents a component and is associated with a set of attributes of the component.
 - O is the root node (which also has a set of attributes)
- To simplify our discussion, we use **“feature”** to represent both (sub)components and attributes.
- **E.g.:** Canon PowerShot SX10 IS
 - battery
 - * battery life
 - * battery size
 - * ...
 - lens
 - ...



Definitions (3)

- An object O is represented with a finite set of features, $F = \{f_1, f_2, \dots, f_n\}$.
 - Each feature f_i in F can be expressed with a finite set of words or phrases W_i , which are synonyms.

That is to say: we have a set of corresponding synonym sets $W = \{W_1, W_2, \dots, W_n\}$ for the features.

E.g. battery size, battery dimensions, battery magnitude, ...
- **Model of a review:** An opinion holder j comments on a subset of the features $S_j \subseteq F$ of object O .
 - For each feature $f_k \in S_j$ that j comments on, he/she
 - chooses a word or phrase from W_k to describe the feature, and
 - expresses a positive, negative or neutral opinion on f_k .

Mining Tasks: overview

- Sentiment classification
 - Classifies an evaluative text as being positive or negative, no details are discovered about what people liked or disliked
 - Usually at the document level
- Featured-based opinion mining and summarization
 - Identifies product features that have been commented on by reviewers and determines whether the comments are positive or negative
 - At the sentence level
 - E.g. « the battery life is too short » => negative comment on « battery life »
- Comparative sentence and relation mining
 - Compares one object against one or more other similar objects
 - E.g. « the battery life of camera A is much shorter than that of camera B » => extract comparative relations

Document-level sentiment classification

- Classify documents (e.g., reviews) based on the overall sentiments expressed by opinion holders (authors)
 - Positive, negative, and (possibly) neutral
 - Since in our model an object O itself is also a feature, then sentiment classification essentially determines the opinion expressed on O in each document (e.g., review).
- Similar but different from **topic-based text classification**.
 - In topic-based text classification, topic words are important.
 - In sentiment classification, sentiment words are more important, e.g., great, excellent, horrible, bad, worst, etc.
- Mainly at the document-level, but also extendable to the sentence-level.

1. Introduction
2. **Three mining tasks**
3. Opinion Search
4. Opinion Spam
5. Conclusion

- Overview
- Document-level sentiment classification
 - Sentence-level sentiment classification
 - Comparative sentence and relation extraction
 - Application

Document-level sentiment classification: example

Canon's PowerShot SX10 IS is a 10 Megapixel super-zoom camera with a 20x optically-stabilised lens and a 2.5in flip-out screen. It has an **excellent** resolution. Digital zoom images are surprisingly **good**.

Announced in September 2008 alongside the higher-end SX1 IS, it's the successor to the best-selling PowerShot S5 IS and retains its main body shape, articulated screen, AA battery power and movies with stereo sound, but within the camera there's been some major changes. Now, it has **better** noise control than Canon's previous "S" models. However, autofocus is **very slow**.

POSITIVE DOCUMENT!

Document-level sentiment classification

Three approaches

1. Classification based on sentiment phrases
2. Classification using text classification methods
3. Classification using a score function

Document-level sentiment classification

Disadvantages

- It does not give details on what people liked or disliked
 - > Specific features of an object that the author likes or dislikes cannot be identified
- It is not easily applicable to non-reviews, e.g. forum and blog postings
 - > Main focus may not be evaluation or review, but still contain a few opinion sentences.

=> **Sentence-level sentiment classification**

Sentence-level sentiment classification: example

Canon's PowerShot SX10 IS is a 10 Megapixel super-zoom camera with a 20x optically-stabilised lens and a 2.5in flip-out screen. It has an excellent resolution. **Digital zoom images are surprisingly good.**

Announced in September 2008 alongside the higher-end SX1 IS, it's the successor to the best-selling PowerShot S5 IS and retains its main body shape, articulated screen, AA battery power and movies with stereo sound, but within the camera there's been some major changes. Now, it has better noise control than Canon's previous "S" models. However, autofocus is very slow.

POSITIVE SENTENCE!

Sentence-level sentiment classification

Three tasks

1. Identify and extract features of the product that the reviewers have expressed their opinions on, called product features.
E.g.: "the picture quality of this camera is amazing"
-> Product feature: "picture quality"
2. Determining whether the opinions on the features are positive, negative or neutral
3. Group feature synonyms
Produce a summary

1. Introduction
2. **Three mining tasks**
3. Opinion Search
4. Opinion Spam
5. Conclusion

- Overview
Document-level sentiment classification
Sentence-level sentiment classification
Comparative sentence and relation extraction
Application

Different review formats

Format 1

My SLR is on the shelf

by [camerapun4](#). Aug 09 '04

Pros: Great photos, easy to use, very small

Cons: Battery usage; included memory is stingy.

I had never used a digital camera prior to purchasing
have always used a SLR ... [Read the full review](#)

Format 3

GREAT Camera., Jun 3, 2004

Reviewer: [jprice174](#) from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The **pictures** coming out of this camera are amazing. The **'auto'** feature takes great pictures most of the time. And with digital,

Format 2



out of 10

"It is a great digitbal still camera for this century"

September 1, 2004

Pros:

It's small in size, and the rotatable lens is great. It's very easy to use, and has fast response from the shutter. The LCD has increased from 1.5 in to 1.8, which gives bigger view. It has lots of modes to choose from in order to take better pictures.

Cons:

It almost has no cons, it would be better if the LCD is bigger and it's going to be best if the model is designed to a smaller size.

Feature extraction – Format 1

- **Label sequential rules (LSR)** are a special kind of sequential patterns, discovered from sequences. It is used in a supervised method.
- The training data set is a set of sequences, e.g.,
“Included memory is stingy”
is turned into a sequence with POS tags.
 $\langle \{ \text{included, VB} \} \{ \text{memory, NN} \} \{ \text{is, VB} \} \{ \text{stingy, JJ} \} \rangle$
then turned into
 $\langle \{ \text{included, VB} \} \{ \$\text{feature, NN} \} \{ \text{is, VB} \} \{ \text{stingy, JJ} \} \rangle$

(where VB = verb, NN = noun, JJ = adjective)

Feature extraction – Format 1

- Based on a set of training sequences, we can mine label sequential rules, e.g.,
 $\langle \{ \text{easy, JJ} \} \{ \text{to} \} \{ *, \text{VB} \} \rangle \rightarrow \langle \{ \text{easy, JJ} \} \{ \text{to} \} \{ \$\text{feature, VB} \} \rangle$
[sup = 10%, conf = 95%]
E.g.: “easy to use” -> “use” identified as a feature
- **Feature Extraction**
 - Only the right hand side of each rule is needed.
 - The word in the sentence segment of a new review that matches \$feature is extracted.
 - Conflict resolution (if multiple rules are applicable)

Feature extraction – Formats 2 & 3

- Reviews of these formats are usually complete sentences
 - Explicit feature: picture
 - e.g., “the pictures are very clear.”
 - Implicit feature: size
 - e.g. “It is small enough to fit easily in a coat pocket or purse.”

- Extraction: **Frequency based approach**
 - Frequent features
 - Infrequent features -> Sequential pattern mining based on sentiment words
 - E.g.: “The pictures are absolutely amazing.”
 - > Picture is a frequent feature and we know that “amazing” is a positive opinion word
 - “The software is amazing”
 - => “software” is identified as an infrequent feature

Feature extraction – Formats 2 & 3

Improvement of precision

=> Remove noun phrases that may not be product features

Evaluate noun phrases by computing the Pairwise Mutual Information between the phrase and **meronymy discriminators** associated with the product class.

E.g. camera class => “of camera”, “camera has”, “camera comes with”,...

$$\text{PMI}(f, d) = \frac{\text{hits}(f^d)}{\text{hits}(f)\text{hits}(d)}$$

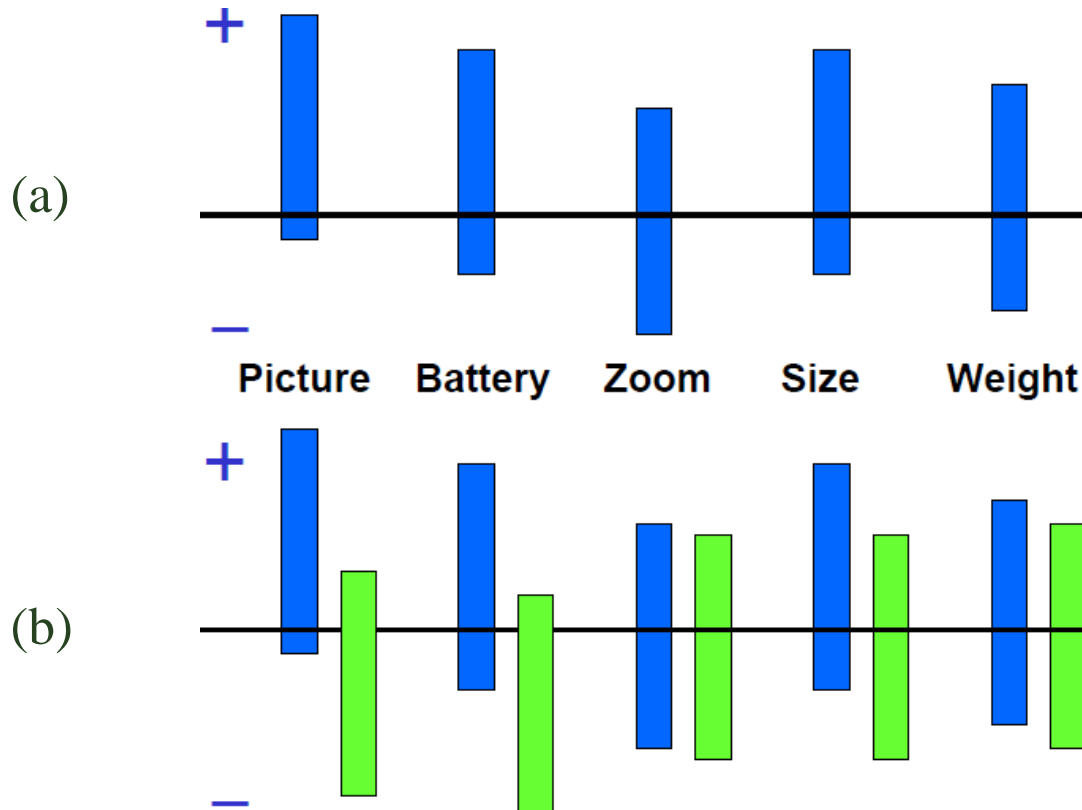
where f = candidate feature (step 1) & d = discriminator

Opinion Orientation Classification

Two main techniques

1. Using sentiment words and phrases
 1. Identify sentiment words and phrases
 2. Assign scores (positive: +1, negative: -1)
 3. Sum up scores
 - If positive: sentence is positive
 - If negative: sentence is negative
 - Negation near sentence word => opinion reversed
2. Methods for sentiment classification (see previous)

Summarization



Visualization of feature-based opinion summary and comparison
(a) Feature-based summary of opinions on a digital camera
(b) Opinion comparison of two digital cameras

Comparative sentence and relation extraction

- **Two types of evaluation**
 - Direct opinions: "This digital camera is bad"
 - Comparisons: "Camera X is not as good as camera Y"

=> different language constructs.

- Direct expression of sentiments are good
But... comparison may be better.
Since... good or bad, compared to what?

Comparative sentence and relation extraction

Three tasks

Given a collection of evaluative texts

1. Identify comparative sentences.
2. Categorize different types of comparative sentences.
3. Extract comparative relations from the sentences.

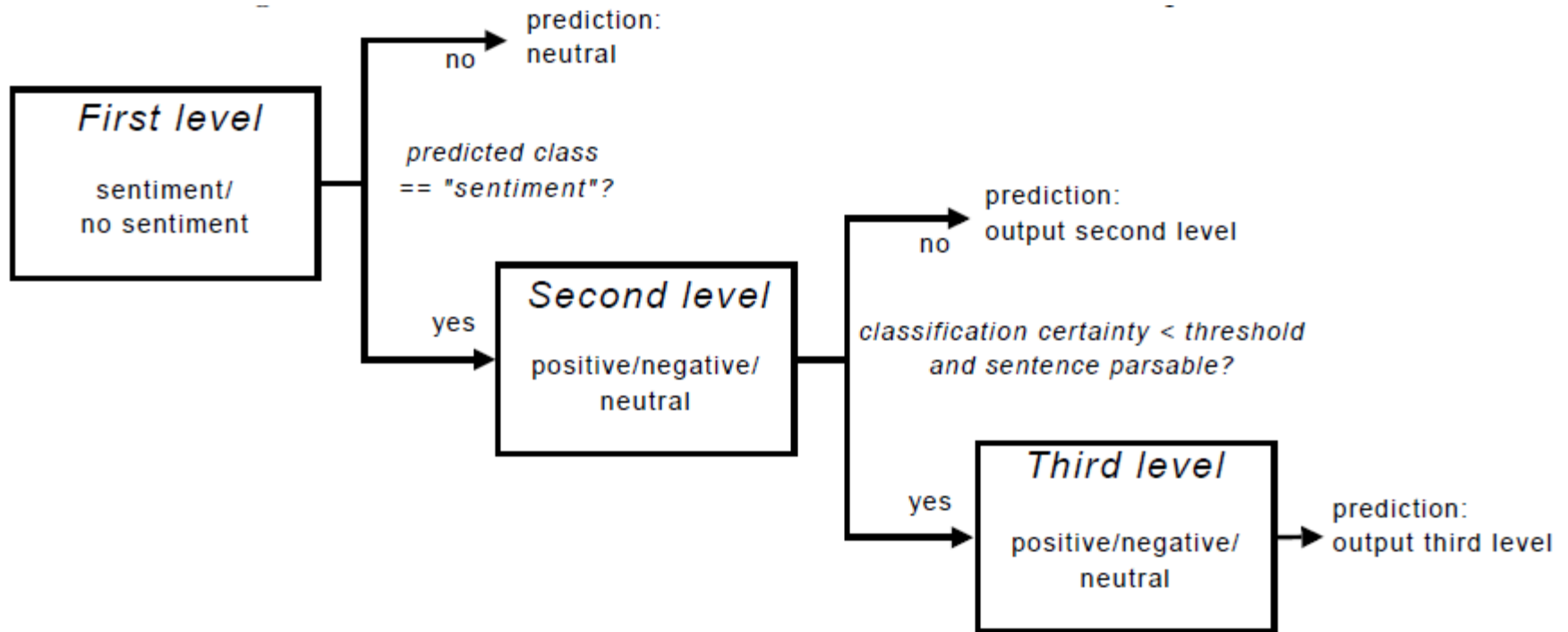
Application: sentiment analysis in multilingual web texts

- Recognition as a supervised classification task
 - Support Vector Machines
 - Multinomial Naïve Bayes
 - Maximum Entropy
- Cascaded and aggregated learners
 - At deep level: expensive feature extraction techniques (e.g. sentence parsing for difficult cases)
 - At certain level, features become language dependent (e.g. treatment of negation, sentence parsing)
- Active learning
 - Reduces the workload in annotation
 - Examples to annotate are automatically selected from the set of unlabeled examples

1. Introduction
2. **Three mining tasks**
3. Opinion Search
4. Opinion Spam
5. Conclusion

Overview
Document-level sentiment classification
Sentence-level sentiment classification
Comparative sentence and relation extraction
→ Application

Cascaded learners



[Moens, Boiy, 2008]

Opinion Search

Goal: Whenever you need to make a decision, you may want some opinions from others.

=> Instead of searching for facts, search for opinions

- Queries such as
 - Opinions: "Motorola cell phones"
 - Comparisons: "Motorola vs. Nokia"
- Cannot be done yet
 - Accuracy
 - Opinion Spam (see further)

Typical opinion search queries

- Find the opinion of a person or organization (opinion holder) on a particular object or a feature of the object.
 - E.g., what is Barack Obama's opinion on abortion?
- Find positive and/or negative opinions on a particular object (or some features of the object), e.g.,
 - customer opinions on a digital camera.
 - public opinions on a political topic.
- Find how opinions on an object change over time.
- How object A compares with object B?
 - Gmail vs. Hotmail

Opinion Spam

- = Human activities (e.g. write spam reviews) that try to deliberately mislead readers or automated opinion mining systems by
- giving undeserving positive opinions to some target objects in order to promote the objects (**hype spam**) and/or
 - giving unjust or false negative opinions on some other objects in order to damage their reputation (**defaming spam**).

Types of spam and spammers:

- Manual and automated spam
- Individual spammers and group spammers

Hiding techniques

An individual spammer

1. Build up a reputation
2. Register multiple times at a site using different user-ids
3. Give reasonably high rating but write critical review
4. Write either only positive reviews on own products or only negative reviews on the products of competitors, but not both

A group of spammers

1. Every member reviews same product to lower rating deviation
2. Write reviews when product is launched to take control of the product
3. Write reviews at random or irregular intervals
4. Divide group in sub-groups so that each sub-group can spam at different web sites

Spam detection

- **Review centric spam detection**
 - Compare content similarity
 - Detect rating and content outliers
 - Compare average ratings form multiple sites
 - Detect rating spikes
- **Reviewer centric spam detection**
 - Watch early reviews
 - Detect early remedial actions
 - Compare review ratings of the same reviewer on products form different brands
 - Compare review times
- **Server centric spam detection**

=> Important, because without effective detection, opinions on the Web may become useless.

Conclusion

- Very useful applications (opinion search) that exploit the opportunities of user-generated content.
- But:
 - Challenging
 - Threat of opinion spam

=> Rather new research domain, so lot of research to be done!

Questions?

Bibliography

This presentation is based on (and inspired by):

- B. Liu (2007). Chapter 11 of *Web Data Mining* (Springer 2007)
- Boiy, E. & Moens, M.-F. (2008). A machine learning approach to sentiment analysis in multilingual Web texts.
- M. Hu and B. Liu(2004). Mining and Summarizing Customer Reviews. In *Proc. of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'04)*, pp. 168–177, 2004.
- A.-M. Popescu, and O. Etzioni. Extracting Product Features and Opinions from Reviews. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP'05)*, 2005.