Sentiment Analysis

Oswin Krause
Department of Computer Science

1. Sentiment Analysis: Basics
2. Document Sentiment Classification and Rating
3. Syntex Tree based Sentiment Classification
4. Q&A Project 2

Sentiment Analysis: Basics

Motivation: Product Reviews

Our yard has never looked so good, and we did nothing! The robot is very reliable and solid. Not to mention quiet. I can't say enough good things about it. The app too is a lot of fun. I would like to see the anti-theft expanded. I couldnt hear the alarm from in the house with TV on.

Motivation: Product Reviews

- How do customers perceive a product?
    - What do they like?
    - What are problems?
    - Can we improve the product?
- Reviews are assessed by humans
- Reviews are subjective → many reviews needed
- Automatization: Sentiment Analysis

What is an Opinion?

The blade speed on the robot is low,
which means that it is whisper quiet

- An opinion consists of five parts:
    - An opinion holder (e.g. customer)
    - An entity (e.g. the product)
    - An aspect (a part or attribute of the product)
    - A sentiment(opinion on the aspect)
    - A time (e.g. time of review)
- Parts can be implicit
- Natural language processing is hard:
    - Does “it” refer to the robot or the blade speed?

What is an Opinion? – Formally

Definition

An opinion is a quintuple

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

- where
  - e_i is the ith entity
  - a_{ij} is the jth aspect of e_i
  - h_k is the kth opinion holder
  - t_l is the lth time
  - s_{ijkl} is the sentiment of h_k towards aspect a_{ij} at time t_l
- Goal of Sentiment Analysis:
  - Transform review into set of quintuples

Some Clarifications

- Reviews often have only one entity and opinion holder
- Aspects model abstract part-of relationships
  - The robot has wheels, blades,body…
  - battery-life, reliability, setup-time…
  - GENERAL (the whole entity)
- Several expressions for the same part (roboter, robby)
- Sentiments
  - Categorical: positive, neutral, negative
  - Ratings: (bad) 0–10 (good)
- Sentiments can change over time
- Opinions can be indirect

Our yard has never looked so good!

Exercise: Find the Opinions

- Customer review, 9.3.2015:

  (1) The robot is very reliable and solid. (2) Setting up the fences is complicated. (3) The kids love it.

- Same Customer, 3.6.2015:

  (4) I have never gotten the fences right, thus the robot gets stuck every once in a while. (5) The wheels are often loose and yesterday the engine died. (6) Would not buy again.

Possible Solution

1. (robot, GENERAL, positive, customer, 9.3.2015)
2. (robot, fences, negative, customer, 9.3.2015)
3. (robot, GENERAL, positive, customers kids, 9.3.2015)
4. (robot, fences, neutral, customer, 3.6.2015)
5. (robot, wheels, negative, customer, 3.6.2015)
   (robot, engine, negative, customer, 3.6.2015)
6. (robot, GENERAL, negative, customer, 3.6.2015)

- We do not always agree on a sentiment
- 70% Agreement between humans

Document Sentiment Classification and Rating

Task

Definition

Given an opinion document d about an entity e, determine the overall sentiment s, i.e. the quintuple

$$(\_, \mathit{GENERAL}, s, \_, \_)$$

- For a given document, predict GENERAL sentiment
- Did the reviewer overall like the product?
- Given: Set of documents with known GENERAL sentiment
- Often: Large set of documents with unknown sentiment

Document Sentiment Classification

- Sentiment is categorial e.g. (Positive, Negative)
- Can be seen as classification task
- General approach:
  - Acquire corpus of data
  - Preprocess data to machine readable form
  - Classify using standard algorithms: SVMs, Random Decision Forrests, . . .
- How can we preprocess texts?

Document Sentiment Rating

- Sentiment is on a scale e.g. (bad) 1,2,3,4,5 (good)
- Assumption: close ratings are similar
- Can be seen as regression task
- Or multi-class-classification with confusion matrix
- Regression algorithms:
  - Linear Regression
  - Lasso
  - Random Forrests for regression
  - Neural Networks
- How can we preprocess texts?

Preprocessing I

- We must find words that are pointing towards a sentiment
    - Good, bad, poor, wonderful, love, hate,…
    - But phrases can be negated: “not bad”
- Typical stop-words can be important for sentiment classification
    - but, doesn’t, nor, not,…
- Possible techniques using Bag-of-Words approach
    - n-grams
    - TF.IDF
    - Sentiment Orientation
    - …
- Bag of words can not learn sentence level semantic

Preprocessing II: Pointwise Mutual Information

- Pointwise Mutual Information

$$\text{PMI}(\text{term}_1, \text{term}_2) = \log\left(\frac{P(\text{term}_1, \text{term}_2)}{P(\text{term}_1)P(\text{term}_2)}\right)$$

- Rememeber $P(\text{term}_1|\text{term}_2) = \frac{P(\text{term}_1,\text{term}_2)}{P(\text{term}_2)}$

$$\text{PMI}(\text{term}_1, \text{term}_2) = \log\left(\frac{P(\text{term}_1|\text{term}_2)}{P(\text{term}_1)}\right)$$

- Is term_1 more probable in sentences in which term_2 occurs than on average?

Preprocessing II: Sentiment Orientation

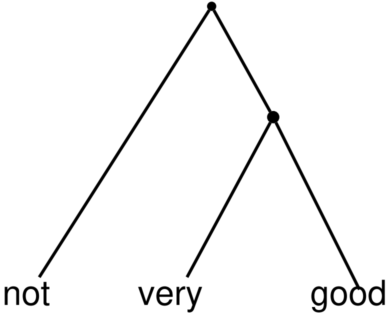
- Sentiment Orientation

$$SO(\text{term}) = \text{PMI}(\text{term}, \text{“good”}) - \text{PMI}(\text{term}, \text{“bad’})$$

- Measures whether a given term occurs more often together with “good” than “bad” in a document
- Average over several choices of word-pairs
- Pick set of words with largest average $SO(\text{term})$
- Does not require labels

Syntex Tree based Sentiment Classification

Binarized Abstract Syntax Trees

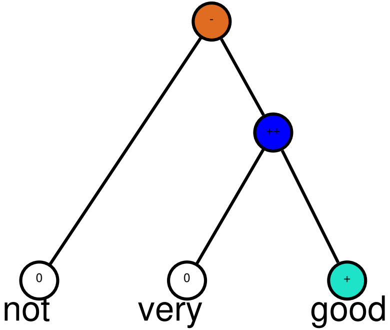


- ASTs encode grammatic structure of phrases

→ Computer readable form of grammar

- Requires Natural language parser for target language

Labeling ASTs

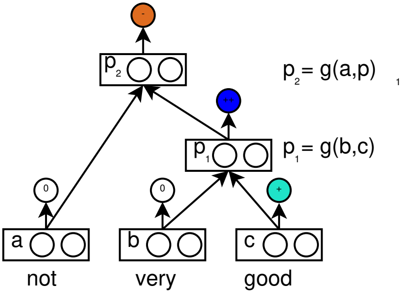


- For every word and node we need a sentiment label

→ Much more fine grained information

- Helps with words like: “but”, “not”, . . .

Neural Network from Syntax tree



- We assign a vector to every word and phrase
- Vector is used for sentiment classification
- Vectors of words are learned
- g is a function that assigns vectors, e.g. trained neural network

Training the Neural Network

- Pick $g(a,b)$
- Pick sentence and create AST
- Look up vectors of words
- Assign vectors to inner nodes using g
- Assign sentiment labels using vectors and classifier
- Update $g(a,b)$, word vectors and classifier using gradient descent
- Repeat until convergence

Summary

- Sentiment Analysis is about finding the optinions in a document
- Opinions are a quintuple involving entity, aspect, sentiment, opinion holder and time
- Sentiment Analysis performed on sentence or document level
- Simple: Bag of words and classification/regression methods
- Preprocessing to find words important for sentiment
- Preprocessing can be done without labels (Sentiment Orientation)
- Learning proper semantic is problematic (not good, but,. . .)
- Methods exploiting syntax information (AST) are sometimes beneficial

Q&A Project 2