A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts

Thierry Goeckel

Department of Computational Linguistics

November 25, 2010
Outline

1. Introduction

2. Method
   - Architecture
   - Context and Subjectivity Detection
   - Cut-based classification

3. Evaluation Framework

4. Experimental Results
   - Basic Subjectivity extraction
   - Incorporating context information
Why Sentiment Analysis

- Opinion rather than facts
- Useful for companies and recommender systems to create subjective summaries
Why Sentiment Analysis

- Opinion rather than facts
- Useful for companies and recommender systems to create subjective summaries
Document Polarity Classification

- Polarity in movie review: “thumbs up” or “thumbs down”

- Previous approaches focused on selecting indicative lexical features as in “good” or “bad”
• Polarity in movie review: “thumbs up” or “thumbs down”

• Previous approaches focused on selecting indicative lexical features as in “good” or “bad”
Subjectivity Approach

1. Labeling of sentences as either subjective or objective
2. Apply a standard machine-learning classifier to the extract
Subjectivity Approach

- Irrelevant or even potentially misleading text is not considered.
- Subjectivity extracts accurately represent the sentiment in a compact form.
- Results show statistically significant improvement or maintain the same level with a lot less data.
- Minimum cut formulation provides an efficient, intuitive and effective means for integrating inter-sentence-level contextual information with traditional bag-of-words features.
Subjectivity Detector

- Determines whether sentence is subjective or not
- Combination of sentence-level subjectivity detection with document-level sentiment polarity
Subjectivity Detector

n-sentence review

subjective sentence?

m-sentence extract (m<=n)

positive or negative review?

subjectivity detector

default polarity classifier

+/-

subjectivity extraction

Architecture
Context and Subjectivity Detection
Cut-based classification
Outline

1. Introduction
2. Method
   - Architecture
   - Context and Subjectivity Detection
   - Cut-based classification
3. Evaluation Framework
4. Experimental Results
   - Basic Subjectivity extraction
   - Incorporating context information
Context and Subjectivity Detection

- *Coherence*: sentences not classified in isolation
- Achieved by an efficient and intuitive algorithm relying on finding *minimum cuts*
**Concepts**

- **Classes** $C_1$ and $C_2$: positive and negative classes
- **Individual scores** $\text{ind}_j(x_i)$: non-negative estimate of each item $x_i$’s preference for being in $C_j$
- **Association scores** $\text{assoc}(x_i, x_k)$: non-negative estimate of how important it is that $x_i$ and $x_k$ be in the same class.
**Definition**

A cut \((S, T)\) of G is a partition of its nodes into sets \(S = \{s\} \cup S'\) and \(T = \{t\} \cup T'\), where \(s \notin S'\), \(t \notin T'\). Its cost \(\text{cost}(S, T)\) is the sum of the weights of all the edges crossing from \(S\) to \(T\). A minimum cut of \(G\) is one of minimum cost.
Practical Advantages

- Co-ordination of algorithms deriving the individual scores and methods assigning the association scores.
- Use of maximum-flow algorithms with polynomial asymptotic running times
Test data contains 1000 positive and 1000 negative reviews of movies released pre-2002

Default popularity classifiers

- Tested: support vector machines (SVMs)
- Tested: naive Bayes (NB)

Used: Unigram-presence features
Test data contains 1000 positive and 1000 negative reviews of movies released pre-2002

Default popularity classifiers

- Tested: support vector machines (SVMs)
- Tested: naive Bayes (NB)

Used: Unigram-presence features
Subjectivity Dataset

- Web mining on rottentomatoes.com and imdb.com
  - Subjective data: 5000 “snippets” from rottentomatoes.com
  - Objective data: 5000 sentences of plot summaries from imdb.com
  - “Snippets” and sentences at least ten words long and from movies released after 2001 (to prevent overlap with polarity dataset)
Subjectivity Detectors

- Polarity classifiers as “basic” sentence-level subjectivity detectors
  → to produce extracts of original reviews
- Creation of a family of cut-based subjectivity detectors
  → determine the subjectivity status of all the sentences
  per-item and pairwise relationship information
  - Using either Naive Bayes or SVMs
  - Association scores set either to zero
  - or to degree of proximity controlled by three parameters:
    - Threshold $T$: maximum distance of two sentences to be still “proximal”
    - Non-increasing function $f(d)$: decay of influence of proximal sentences with respect to distance $d$
    - The constant $c$: controls the relative influence of the association scores
Experiments

First, the basic subjectivity extraction algorithms (on sentence-level) are examined.

Second, the more sophisticated form of subjectivity extraction with context information is evaluated (using min-cut paradigm):

→ Subjectivity extracts improve polarity classification, otherwise at least same accuracy as with the full review.
→ The created extracts are both smaller and more effective as input to a default polarity classifier than the original text.
Results

Accuracy for 5-sentence abstracts (def + MB)

Accuracy for 10-sentence abstracts (def + SWM)

Accuracy for subjective abstracts (def + MB)

Accuracy for subjective abstracts (def + SWM)

Indicates statistically significant improvement in accuracy
1 Introduction

2 Method
   - Architecture
   - Context and Subjectivity Detection
   - Cut-based classification

3 Evaluation Framework

4 Experimental Results
   - Basic Subjectivity extraction
   - Incorporating context information
Basic Subjectivity Extraction

- NB produces better results than SVM method
- Achieves 86.4% (vs. 82.8% without extraction)
  → Extracts preserve (and even clarify) the sentiment underlying the original document
- Subjectivity extracts are much more compact than original reviews
  → about 60% of the source’s words
- Tests with varying lengths and positions in the text
  → Extracts containing as few as 5 to 15 sentences are almost as informative as the full review
Outline

1 Introduction

2 Method
   • Architecture
   • Context and Subjectivity Detection
   • Cut-based classification

3 Evaluation Framework

4 Experimental Results
   • Basic Subjectivity extraction
   • Incorporating context information

Thierry Goeckel  A Sentimental Education
Incorporating context information

- Context-aware graph-based subjectivity detectors create more informative extracts
- But: extracts are longer
- Further tests: reduce the association scores for sentences in different paragraphs
  → Graph-cut formulation produces better results than standard classifiers
Relation between subjectivity detection and polarity classification shows that the former can compress reviews and still retain polarity information.

Using contextual information via the minimum-cut framework leads to an improvement in polarity-classification accuracy.
Pang, Bo and Lillian Lee
A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.