Sentiment and Subjectivity Analysis
An Overview

Alina Andreevskaia
Department of Computer Science
Concordia University
Montreal, Quebec
Canada

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York
USA
Definition

• **Sentiment Analysis**
  - Also called *Opinion Mining*
  - Classify words/senses, texts, documents according to the opinion, emotion, or sentiment they express

• **Applications**
  - Determining critics’ opinions of products
  - Track attitudes toward political candidates
  - etc.
Sub-Tasks

• Determine Subjective-Objective polarity
  - Is the text or language factual or an expression of an opinion?

• Determine Positive-Negative polarity
  - Does the subjective text express a positive or negative opinion of the subject matter?

• Determine the strength of the opinion
  - Is the opinion weakly positive/negative, strongly positive/negative, or neutral?
History

• Current work stems from
  - Content analysis
    • “analysis of the manifest and latent content of a body of communicated material (as a book or film) through classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect” (Webster’s Dictionary of the English Language, 1961)
  • Long history
    - Quantitative newspaper analysis (1890-on)
    - Lasswell, 1941: study of political symbols in editorials and public speeches
    - Gerbner, 1969: establish “violence profiles” for different TV networks; trace trends; see how various groups portrayed
Content Analysis

• Psychology and sociology
  - Analysis of verbal patterns to determine
    • motivational, mental, personal characteristics
      (1940’s)
    • group processes
    • cultural commonalities/differences (Osgood, Suci, and Tannebaum, 1957, semantic differential scales)
  - Major contribution: General Inquirer
    • http://www.wjh.harvard.edu/~inquirer/

• Anthropology
  - Study of myths, riddles, folktales
  - Analysis of kinship terminology (Goodenough, 1972)

• Literary and Rhetorical analysis
  - Stylistic analysis (Sedelow and Sedelow, 1966)
  - Thematic analysis (Smith, 1972; Ide, 1982, 1989)
Other Roots

• **Point of view tracking in narrative** (Banfield, 1982; Uspensky, 1973; Wiebe, 1994)
  - Subjectivity analysis

• **Affective Computing** (Picard, 1997)
  - Develop means to enable computer to detect and appropriately respond to user's emotions

• **Directionality** (Hearst, 1992)
  - Determine if author is positive, neutral, or negative toward some part of a document
History

• Late ‘90’s: first automatic systems implemented for NLP
  – Spertus, 1997
  – Wiebe and Bruce, 1995
  – Wiebe et al., 1999
  – Bruce and Wiebe, 2000

• Now a major research stream
Current work in NLP

- **Sentiment tagging**
  - Assignment of positive, negative, or neutral values/tags to texts and its components
  - Began with focus on binary (positive-negative) classification
  - Recently, include neutrals

- **Little work on other types of affect**
  - Still a focus in much content analysis work in other fields
Other Work

- Assignment of fine-grained affect labels based on various psychological theories (Valitutti et al., 2004; Strapparava and Mihalcea, 2007)
- Detection of
  - opinion holders (Kim and Hovy, 2004; Kim and Hovy, 2005; Kim and Hovy, 2006; Choi et al., 2005; Bethard et al., 2004; Kobayashi et al., 2007)
  - opinion targets (Hurst and Nigam, 2004; Gamon and Aue, 2005; Hu and Liu, 2004; Popescu and Etzioni, 2005; Kim and Hovy, 2006; Kobayashi et al., 2007)
  - perspective (Lin et al., 2006)
  - pros and cons in reviews (Kim and Hovy, 2006a)
  - bloggers’ mood (Mishne and Glance, 2006; Mishne, 2005; Leshed and Kaye, 2006)
  - happiness (Mihalcea and Liu, 2006)
  - politeness (Roman et al., 2005)
- Assignment of ratings to movie reviews (Pang and Lee, 2005)
- Identification of support/opposition in congressional debates (Thomas et al., 2006)
- Prediction of election results (Kim and Hovy, 2007)
Subjectivity

• Focuses on determining subjective words and texts that mark the presence of opinions and evaluations vs. objective words and texts, used to present factual information (Wiebe, 2000; Wiebe et al., 2004; Wiebe and Riloff, 2005)
Terminology

• Many terms
  - Sentiment classification, sentiment analysis
  - Semantic orientation
  - Opinion analysis, opinion mining
  - Valence
  - Polarity
  - Attitude

• Here, use the term sentiment
Theories of Emotion and Affect

• Osgood's *semantic differential* (Osgood, Suci, and Tannenbaum, 1957)
  - Three recurring attitudes that people use to evaluate words and phrases
    • Evaluation (good-bad)
    • Potency (strong-weak)
    • Activity (active-passive)

• Ortony's *saliency-imbalance theory* (Ortony, 1979)
  - defines metaphors in terms of particular relationships between topic and vehicle
Theories of Emotion and Affect

• Martin’s Appraisal Framework
  - http://www.grammatics.com/appraisal/

• Three sub-types of attitude
  - Affect (emotion)
    • evaluation of emotional disposition
  - Judgment (ethics)
    • normative assessments of human behavior
  - Appreciation (aesthetics)
    • assessments of form, appearance, composition, impact, significance etc of human artefacts and individuals
### Theories of Emotion and Affect

- **Elliot’s Affective Reasoner**
  - [http://condor.depaul.edu/~elliott/ar.html](http://condor.depaul.edu/~elliott/ar.html)

<table>
<thead>
<tr>
<th>GROUP</th>
<th>SPECIFICATION</th>
<th>CATEGORY LABEL AND EMOTION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well-Being</td>
<td>appraisal of a situation as an event</td>
<td>joy, distress</td>
</tr>
<tr>
<td>Fortune’s-of-</td>
<td>presumed value of a situation as an event affecting another</td>
<td>happy, gloating, resentment, jealousy, envy, sorry-for</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospect-based</td>
<td>appraisal of a situation as a prospective event</td>
<td>hope, fear</td>
</tr>
<tr>
<td>Confirmation</td>
<td>appraisal of a situation as confirming or disconfirming an expectation</td>
<td>satisfaction, relief, fears-confirmed, disappointment</td>
</tr>
<tr>
<td>Attribution</td>
<td>appraisal of a situation as an act of some agent accountable</td>
<td>pride, admiration, shame, reproach</td>
</tr>
<tr>
<td>Attraction</td>
<td>appraisal of a situation as containing an attractive or unattractive object</td>
<td>liking, disliking</td>
</tr>
<tr>
<td>Well-being/</td>
<td>compound emotions</td>
<td>gratitude, anger, gratification, remorse</td>
</tr>
<tr>
<td>Attribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attraction/Attri-</td>
<td>compound emotion extensions</td>
<td>love, hate</td>
</tr>
<tr>
<td>bution</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Theories of Emotion and Affect

• Ekman’s basic emotions
  - Ekman’s work revealed that facial expressions of emotion are not culturally determined, but universal to human culture and thus biological in origin
  - Found expressions of anger, disgust, fear, joy, sadness, and surprise to be universal
  • Some evidence for contempt
Resource Development

• Semantic properties of individual words are good predictors of semantic characteristics of a phrase or a text that contain them
• Requires development of lists of words indicative of sentiment
Manually-created Lists

• General Inquirer (GI)
  - Best known extensive list of words categorized for various categories
  - Developed as part of content-analysis project (Stone et al., 1966; Stone et al., 1997)
  - Three main word lists:
    • Harvard IV-4 dictionary of content-analysis categories (includes Osgood’s three dimensions of value, power and activity)
    • Lasswell’s dictionary
      - eight basic value categories (WEALTH, POWER, RESPECT, RECTITUDE, SKILL, ENLIGHTENMENT, AFFECTION, WELLBEING) plus other info
    • Five categories based on social cognition work of Semin and Fiedler (1988)
      - Verb and adjective types
  - Some words tagged for sense
  - Recognized as a gold standard for evaluation of automatically produced lists
WordNetAffect

- Developed semi-automatically by Strapparava and colleagues
  - assigned affect labels to words in WordNet
  - expanded the lists using WordNet relations such as synonymy, antonymy, entailment, hyponymy
- Includes
  - semantic labels based on psychological and social science theories (Ortony, Elliot, Ekman)
  - valence (positive or negative)
  - arousal (strength of emotion)
- 2004 version covers 1314 synsets, 3340 words
- Part of WordNet Domains
  - http://wndomains.itc.it/
Others

• Whissell’s *Dictionary of Affect in Language* (DAL) (Sweeney and Whissell, 1984; Whissell, 1989; Whissell and Charuk, 1985)

• Affective Norms for English Words (ANEW) (Bradley and Lang, 1999)

• Sentiment-bearing adjectives by Hatzivassilioglou and McKeown (1997)

• Sentiment and subjectivity clues from work by Wiebe
Caution

• **Limitations of manually annotated lists**
  - limited coverage
  - low inter-annotator agreement
  - diversity of the tags used
Automatically-created Lists

Corpus-based methods

- Hatzivassiloglou and McKeown (1997) (HM)
  - builds on the observation that some linguistic constructs, such as conjunctions, impose constraints on the semantic orientation of their constituents
  - clustered adjectives from the Wall Street Journal in a graph into positive and negative sets based on the type of conjunction between them
  - cluster with higher average frequency was deemed to contain positive adjectives, lower average frequency meant negative sentiment

- Limitations
  - algorithm limited to adjectives (also adverbs -- Turney and Littman, 2002)
  - requires large amounts of hand-labeled data to produce accurate results
Web As Corpus

- Peter Turney (Turney, 2002; Turney and Littman, 2002; Turney and Littman, 2003)
  - More general method, does not require previously annotated data for training
  - Induce sentiment of a word from the strength of its association with 14 seed words with known positive or negative semantic orientation
  - Two methods for association:
    - Point-wise mutual information
    - Semantic latency
  - Used web as data
    - ran 14 queries on AltaVista using NEAR operator to acquire co-occurrence statistics with the 14 seed words
Turney, con’t

• Results evaluated against GI on a variety of test settings gave up to 97.11% accuracy for top 25% of words
• Size of the corpus had a considerable effect
  – 10-million word corpus instead of the full Web content reduced accuracy to 61.26–68.74%
• LSA performed relatively better than PMI on 10m word corpus
  – LSA more complex, harder to implement
End of An Era

• Due to its simplicity, high accuracy and domain independence, the PMI method became popular.

• In 2005, AltaVista discontinued support for NEAR operator, upon which the method relied.

• Attempts to substitute NEAR with AND led to considerable deterioration in system performance.
Other Approaches

• Bethard et al. (2004)
  - Used two different methods to acquire opinion words from corpora
    • calculated frequency of co-occurrence with seed words taken from Hatzivassiloglou and McKeown, computed the log-likelihood ratio
    • computed relative frequencies of words in subjective and objective documents
  - First method produced better results for adverbs and nouns, gave higher precision but lower recall for adjectives
  - Second method worked best for verbs
Other Approaches

• Kim and Hovy (2005)
  - separated opinion words from non-opinion words by computing their relative frequency in subjective (editorial) and objective (non-editorial) texts from TREC data

• Riloff et al. (2003), Grefenstette et al. (2006)
  - Used syntactic patterns
  - Learn lexico-syntactic expressions characteristic for subjective nouns
Automatically-created Lists

Dictionary-Based Methods

- Addresses some of the limitations of corpus-based methods
- Use semantic resources such as WordNet, thesauri
- Two approaches:
  - Rely on thesaural relations between words (synonymy, antonymy, hyponymy, hyperonymy) to find similarity between seed words and other words
  - Exploit information contained in definitions and glosses
Use of WordNet

• Kim and Hovy (2004, 2005)
  – Extended word lists by using WordNet synsets
  – Ranked lists based on sentiment polarity assigned to each word in synset based on WordNet distance from positive and negative seed words

• Similar approach used by Hu and Liu (2004)
SentiWordNet

• Developed by Esuli and Sebastiani
  - Trained several classifiers to give rating for positive, negative, objective for each synset in WordNet 2.0
    • Scores from 0 - 1
  - Freely available

• http://sentiwordnet.isti.cnr.it
Beyond Synsets

• Kamps et al. (2004)
  - Tagged words with Osgood’s three semantic dimensions
  - Computed shortest path through WordNet relations connecting a word to words representative of the three categories (e.g., good and bad for evaluation)

• Esuli and Sebastiani (2005)
  - Classified words in WordNet into positive and negative based on synsets, glosses and examples
Further Beyond Synsets

- Andreevskaia and Bergler (2006)
  - Take advantage of the semantic similarity between glosses and head words
  - Start with a list of manually annotated words, expand with synonyms and antonyms, search WordNet glosses for occurrences of seed words
  - If a gloss contains a word with known sentiment, head word deemed to have same sentiment
  - Suggest that overlap measure reflects the centrality of the word in the sentiment category
Role of Neutrals

• Most work cited so far classifies words as positive or negative

• Results vary between 60–80% agreement with GI as gold standard

• Adding neutrals severely reduces accuracy by 10–20%, depending on part of speech
Problems

• Words without strong positive or negative connotations are difficult to categorize accurately
  - Strength of +/- affinity can be used as a measure
  - Highest accuracy for words on extremes of the +/- poles

• Many words have both sentiment-bearing and neutral senses
  - E.g., great typically tagged as positive, but according to statistics in WordNet, used neutrally 75% of occurrences
  - Solve this by using sense-tagged word lists
Sense-tagged lists

- The need for sense-level sentiment annotation has recently attracted considerable attention.
- Development of methods to devise sense-tagged word lists.
Sense-tagged Word Lists

• Andreevskaiia and Bergler (2006)
  - applied gloss-based sentiment tagging to sense level
  - extended their system by adding a word sense disambiguation module (Senti-Sense)
    - used syntactic patterns to disambiguate between sentiment-bearing and neutral senses
  - learned generalized adjective-noun patterns for sentiment-bearing adjectives from unambiguous data
  - abstracted learned patterns to higher levels of hypernym hierarchies using predetermined propagation rules
  - applied learned patterns to disambiguate adjectives with multiple senses in order to locate senses that bear sentiment
Sense-tagged Word Lists

• Esuli and Sebastiani (2007)
  - Application of random walk PageRanking algorithm to sentiment tagging of synsets
  - Takes advantage of the graph-like structure of the WordNet hierarchy
Sense-tagged Word Lists

- Wiebe and Mihalcea (2006)
  - Sense-level tagging for subjectivity (i.e., neutrals vs. sentiment-bearing words)
  - Automatic method for sense-level sentiment tagging based on Lin’s (1998) similarity measure
  - Acquire a list of top-ranking distributionally similar words
  - Compute WordNet-based measure of semantic similarity for each word in the list
Beyond Positive and Negative

- **Ide (2006)**
  - Bootstrapped *sense-tagged word lists for semantic categories* based on
    - FrameNet frames (e.g., commitment, reasoning, etc.)
    - GI categories (*hostile/friendly weak/strong, submit/dominate, power/cooperation*)
  - Treated lexical units associated with a given frame as a “bag of words”
  - Used *WordNet::Similarity* to compute similarity among senses of the words
    - Relation set consisted of different pairwise combinations of synsets, glosses, examples, hypernyms, and hyponyms
  - Based on results, compute “suresenses” (strongest association) and retain; iterate
  - Augment with “suresenses” for synsets, hypernyms, hyponyms
- Resulting lists very high on precision (98%), lower on recall
- Used hierarchical clustering to group senses in a given category into positive and negative and finer-grained distinctions

<table>
<thead>
<tr>
<th>acclaim</th>
<th>denigrate</th>
<th>belittle</th>
<th>condemn</th>
<th>charge</th>
<th>accuse</th>
<th>ridicule</th>
</tr>
</thead>
<tbody>
<tr>
<td>acclaim1 extol1 laud1 commend4 commend1 praise1 cite2</td>
<td>denigrate1 deprecate2 execrate2</td>
<td>belittle2 disparage1 reprehend1 censure1 denounce1 remonstrate3 blame2 castigate1</td>
<td>condemn1 decry1 excoriate1 deprecate1</td>
<td>accuse2 charge2 recriminate1</td>
<td>accuse1 denigrate2</td>
<td>deride1 ridicule1 gibe2 scoff1 mock1 scoff2 remonstrate2</td>
</tr>
</tbody>
</table>
Manually Annotated Corpora

- Scarcity of manually annotated resources for system training and evaluation
- Some work uses user-created rankings in online product, book, or movie reviews
  - ranking scale (good-bad, liked-disliked, etc.)
  - easily available, fast to collect
  - Drawbacks
    - contain significant amount of noise
      - erroneous ratings
      - misspellings
      - phrases in different languages
    - exacerbated when researchers automatically break reviews into sentences or snippets
## Available Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Level of annotation</th>
<th>Annotation type(s)</th>
<th>Corpus size</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPQA</td>
<td>Phrases and sentences</td>
<td>Private states</td>
<td>535 documents (10 657 sentences)</td>
<td><a href="http://www.cs.pitt.edu/wiebe/mpqa">http://www.cs.pitt.edu/wiebe/mpqa</a></td>
</tr>
<tr>
<td>Opinion corpus</td>
<td>Expressions and sentences</td>
<td>Subjectivity and objectivity</td>
<td>2 sets of documents, 500 sentences each</td>
<td><a href="http://www.cs.pitt.edu/wiebe/pub4.html">http://www.cs.pitt.edu/wiebe/pub4.html</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>from WSJ Treebank</td>
<td></td>
</tr>
<tr>
<td>Product reviews</td>
<td>Product features</td>
<td>Sentiment features</td>
<td>500 reviews</td>
<td><a href="http://www.cs.uic.edu/liub/FBS/FBS.html">http://www.cs.uic.edu/liub/FBS/FBS.html</a></td>
</tr>
<tr>
<td>SemEval-07 Task 14 dataset</td>
<td>Headlines</td>
<td>Sentiment -100 to +100 polarity scale, 6 basic emotions</td>
<td>2225 news headlines</td>
<td><a href="http://www.cse.unt.edu/~rada/affectivetext/">http://www.cse.unt.edu/~rada/affectivetext/</a></td>
</tr>
</tbody>
</table>
Inter-annotator Agreement

- Rate of agreement among human annotators may reveal important insights about the task, provide a critical **baseline** for system evaluation
- Few inter-annotator agreement studies conducted to date
- Inter-annotator agreement on popular sentiment genres such as movie reviews, blogs, so far unexplored
- Agreement depends on
  - unit annotated (sentence or text)
  - annotation type (subjectivity or sentiment)
  - domain and genre
IA Studies

• Sentence subjectivity labels (Wiebe et al., 1999)
  - annotators classified sentence as subjective if it contained any significant expression of subjectivity
  - multiple rounds of training and annotation instructions adjustment
  - pairwise Kappa over WSJ test set ranged from 0.59 to 0.76
Variable Results

• Kim and Hovy (2004)
  - Relatively high agreement ($\kappa=0.91$) between two annotators who assigned positive, negative, and n/a labels to 100 newspaper sentences

• Gamon and Aue (2005)
  - Similar study using car reviews produced pairwise Kappa of 0.70 - 0.80
Granularity

• Strapparava and Mihalcea (2007)
  - Study suggests inter-annotator agreement substantially lower on fine-grained types of annotation
  - Six annotators assigned sentiment score and Ekman’s six basic emotions scores to news headlines
  - Agreement (Pearson correlation) 78.01 for sentiment
  - Agreement for emotion labels ranged from 36.07 (surprise) to 68.19 (sadness)
IA for texts

• Wiebe et al., 2001
  - Two genres
    • flames (hostile, inflammatory messages)
      - κ=0.78
    • opinion pieces from WSJ
      - κ=0.94–0.95
  - Domain and genre may affect level of inter-annotator agreement
Sentiment Analysis

- Ultimate goal of sentiment and subjectivity annotation is the analysis of clauses, sentences, and texts

- Resources:
  - Lists
  - Annotated (validated) corpora
    - Subjectivity analysis uses MPQA corpus
      - reliable gold standard for training and evaluation of newspaper texts
    - Sentiment analysis must rely on small manually annotated test sets
      - created ad hoc
      - seldom made publicly available
      - too small for machine learning methods
Sentiment Analysis

• Single text often includes both positive and negative sentences

• Sentences and clauses regarded as the most natural units for sentiment/subjectivity annotation
  - Sentiment usually more homogenous in sentence than whole text
  - But harder to identify due to a limited number of clues
  - Work on improving system performance by performing analysis simultaneously at different levels
    • Sentence-level analysis has high precision
    • Text-level analysis has high recall
Words vs. Other Features

- **Words/unigrams** provide good accuracy in sentence-level sentiment tagging
  - Yu and Hatzivassiloglou (2003)
    - no significant improvement when bigrams or trigrams added to feature set
    - presence of sentiment-bearing words works better than more sophisticated scoring methods
  - Riloff et al. (2006)
    - similar results for unigrams vs. unigrams plus bigrams, extraction patterns
    - subsumption hierarchy and feature selection brought less than 1% improved accuracy
Other Features vs. Words

- Some studies show gains with use of features
  - Wilson et al. (2005), Andreevskaia et al. (2007)
    • syntactic properties of the phrase (role in sentence structure, presence of modifiers and valence shifters) yield statistically significant increases in system performance sentiment and subjectivity analysis
  - Gains in subjectivity analysis:
    • presence of complex adjectival phrases (Bethard et al., 2004)
    • similarity scores (Yu and Hatzivassiloglou, 2003)
    • position in the paragraph (Wiebe et al., 1999)
  - Gains in sentiment analysis
    • syntactic patterns and negation (Hurst and Nigam, 2004; Mulder et al., 2004; Andreevskaia et al., 2007)
    • knowledge about the opinion holder (Kim and Hovy, 2004)
    • target of the sentiment (Hu and Liu, 2004)
Multiple Features

• Some experiments suggest improvement gained when multiple feature sets combined
  - Hatzivassiloglou and Wiebe, 2000
    • combination of lists of adjectives tagged with dynamic, polarity, and gradability labels best predictor of sentence subjectivity
  - Riloff et al. (2003)
    • accuracy of subjectivity tagging results improved with addition of each new feature (25 in all)
  - Gamon and Aue (2005)
    • sentiment tagging using larger number of features increased average precision and recall
SemEval-2007 Affective Text Task

• Opportunity to compare systems for sentiment analysis run on the same dataset of 2000 manually annotated news headlines

• Machine learning methods had highest recall at the cost of low precision
  - naïve Bayes-based CLaC-NB (Andreevskaia and Bergler, 2007)
  - word-space model-based SICS (Sahlgren et al., 2007)

• Knowledge-based unsupervised approaches had highest precision, but low recall because few sentiment clues per headline
  - CLaC (Andreevskaia and Bergler, 2007)
  - UPAR7 (Chaumartin, 2007)
General Conclusion (so far)

• Subjectivity classification
  - statistical approaches that use naïve Bayes or SVM significantly outperform non-statistical techniques

• Binary (positive-negative) sentiment classification
  - non-statistical methods that rely on presence of sentiment markers in a sentence, or on strength of sentiment associated with these markers, yield as good or better accuracy than statistical approaches

• BUT: both methods show some strengths in both tasks

• Need LOTS more research
Text-level Analysis

• **Features**
  - Choice of features used by a sentiment annotation system is a critical factor
  - Wide range of features used:
    • lists of words
    • lemmas or unigrams
    • Bigrams
    • higher order n-grams
    • part-of-speech
    • syntactic properties of surrounding context
    • etc.
Use of Features

- Use of multiple models can improve the performance of both sentiment and subjectivity classifiers
  - Use of different features within the same classifier or in a community of several classifiers improves system performance
Features

• Studies show improvement when using ngrams vs. unigrams (Dave et al., 2003; Cui et al., 2006; Aue and Gamon, 2005; Wiebe et al., 2001; Riloff et al., 2006)

• Improvement when n-gram models / word lists augmented with
  - context information (Wiebe et al., 2001b; Andreevskaia et al., 2007)
  - features associated with syntactic structure (Gamon, 2004)
  - combination of words annotated for semantic categories related to sentiment or subjectivity (Whitelaw et al., 2005; Fletcher and Patrick, 2005; Mullen and Collier, 2004)
Feature Selection

• Costly or computationally unfeasible to use all features
  - Aue and Gamon (2005)
    • n-grams selected based on Expectation Maximization algorithm
  - Riloff et al. (2006)
    • Use a subsumption hierarchy for n-grams (unigrams and bigrams) and extraction patterns
      - if a feature’s words and dependencies are a superset of a more general ancestor in the hierarchy, discard
      - only features with higher information gain were allowed to subsume less informative ones
  • for both subjective/objective and positive/negative text-level classifications, a combined use of subsumption and traditional feature selection improves performance of subjective/objective and positive/negative text-level classification
Part of Speech

• **Subjectivity**
  - *Adjectives* best predictors of subjectivity (Hatzivassiloglou and Wiebe, 2000)
  - *Modals, pronouns, adverbs, cardinal numbers* also used as subjectivity clues (Wiebe et al., 1999; Bruce and Wiebe, 2000)

• **Sentiment**
  - *Combined use* of words from all parts-of-speech produced more accurate tags (Blair et al., 2004; Salvetti et al., 2004)
Feature Generation

• Most sentiment classifiers use standard machine learning techniques to learn and select features from labeled corpora
  - Works well when large labeled corpora available for training and validation (e.g., movie reviews)
  - Falls short when training data is
    • scarce
    • different domain or topic
    • different time period
• Led to increased interest in unsupervised and semi-supervised approaches to feature generation
Some Promising Research

• Systems trained on a small number of labeled examples and large quantities of unlabelled in-domain data perform relatively well (Aue and Gamon, 2005)

• Structural correspondence learning applied to small number of labeled examples sufficient to adapt to new domain (Blitzer et al., 2007)

• So far performance of these methods inferior to supervised approaches and knowledge-based methods

• Availability of word lists and clues makes knowledge-based approaches an attractive alternative to supervised machine learning when labeled data is scarce
Domain Effects

• Variety of domains used in sentiment analysis
  - movie, music, book and other entertainment reviews
  - product reviews
  - Blogs
  - dream corpus
  - Etc.

• Choice of domain can have a major impact on results
Movie Reviews

• Popular domain in sentiment research

• Positive and negative words/expressions do not necessarily convey the opinion holder's attitude
  - E.g., "evil" used in movie reviews when referring to characters or plot, does not convey sentiment toward the movie itself (Turney, 2002)

• Simple counting of positive and negative clues in movie review texts insufficient

• Clues acquired from out-of-domain sources often fail
Product Reviews

• Sentiment towards whole product sum of the sentiment towards its parts, components, and attributes (Turney, 2002)

• General word lists perform better on product reviews (Turney, 2002; Kennedy and Inkpen, 2006)
Attitude Influence

- Texts with positive sentiment easier to classify than negative ones
  - Kennedy and Inkpen, 2006; Hurst and Nigam, 2004; Dave et al., 2003; Koppel and Schler, 2006; Chaovalit and Zhou, 2005

- Possible explanations:
  - positive documents more uniform (Dave et al., 2003)
  - positive clues have higher discriminant value (Koppel and Schler, 2006)
  - negative texts characterized by extensive use of negations and other valence shifters that reverse the sentiment conveyed by individual words (e.g., not bad) (Pang and Lee, 2004)
Attitude Influence

• Improvement in accuracy when valence shifters taken into account (Kennedy and Inkpen, 2006; Andreevskaia et al., 2007)
  - but negative impact reported when negation included in feature set (Dave et al., 2003)

• Use of balanced evaluation sets with equal number of positive and negative documents has become a standard in sentiment research
Classification Algorithms

• Wide variety of classification approaches used:
  - simple keyword counting methods, with or without scoring
  - rule-based methods
  - content analytical methods (statistical)
  - SVM, naïve Bayes and other statistical classifiers used alone, sequentially, or as a community

• Comparison of results does not provide a definite answer as to which of these methods is the best for sentiment or subjectivity tagging
  - choice of features and training domain have a more impact on accuracy than choice of classification algorithm
  - comparison of performance of systems evaluated on different domains or different feature sets not conclusive
Summary

• Sentiment and subjectivity analysis has evolved into a strong research stream in NLP

• State-of-the-art systems can reach up to 90% accuracy on certain domains

  – But need a generally applicable method
Research Directions

• Development of semi-supervised machine-learning approaches that will maximize the usefulness of the available resources and ensure domain adaptation with limited in-domain data

• Creation of reliable and extensive resources such as lists of words and expressions, syntactic patterns, combinatorial rules, and annotated corpora

• Creation of uniform ways to denote and represent sentiment and subjectivity annotation
Bibliography


Bibliography


Bibliography

- Dave, K., S. Lawrence, and D. M. Pennock: 2003, ‘Mining the Peanut gallery: opinion extraction and semantic classification of product reviews’. In: (WWW03). Budapest, Hungary, pp. 519-528.
Bibliography


Bibliography


Bibliography

Bibliography


Bibliography


Bibliography


Bibliography


