1.1 Sentiment Analysis Applications

Sentiment analysis, also known as opinion mining, is a field that studies people's opinions, sentiments, and emotions towards entities like products, services, organizations, and events. The field has grown significantly with the rise of social media platforms like reviews, forums, blogs, and Twitter, which provide a huge volume of opinionated data in digital form.

- **Business Applications**: Companies use sentiment analysis to understand what consumers and the public think about their products and services. This information is crucial for marketing, public relations, and political campaigns. It's also used to predict sales, box-office revenues for movies, and stock market movements.
- **Social and Political Context**: The opinions expressed on social media can influence public sentiment and impact social and political systems. This has made it a necessity to collect and study opinions from the web.

Sentiment Analysis Research

Sentiment analysis is a challenging and active research area in natural language processing (NLP), data mining, and web mining.

Different Levels of Analysis

Sentiment analysis is typically performed at three main levels of detail:

- **Document Level**: This involves classifying an entire document, like a product review, as either positive or negative. This method works best for documents that focus on a single entity.
- **Sentence Level**: At this level, each sentence is analyzed to determine if it expresses a positive, negative, or neutral opinion. This is closely related to subjectivity classification, which identifies sentences that express opinions versus those that state facts.
- **Entity and Aspect Level**: This is the most detailed level, where the analysis focuses on specific entities and their attributes (aspects). An opinion is seen as a pair of sentiment and a target (the aspect). For example, in the sentence "The iPhone's call quality is good, but its battery life is short," a positive sentiment is found for "call quality" and a negative one for "battery life".

Sentiment Lexicon and its Issues

A **sentiment lexicon** is a list of words and phrases that express positive or negative sentiments. However, using a lexicon alone is not enough because of several complexities:

- **Context-Dependent Words**: A word's sentiment can change depending on the domain. For example, "sucks" is usually negative, but in "This vacuum cleaner really sucks," it can imply a positive sentiment.
- **Neutral Sentences**: Sentences with sentiment words may not express an opinion, such as in questions ("Can you tell me which Sony camera is good?") or conditional sentences.

- **Sarcastic Sentences**: Sarcasm uses positive words to convey a negative meaning, for instance, "What a great car! It stopped working in two days".
- **Implicit Opinions**: Many objective sentences can imply an opinion without using explicit sentiment words. For example, "This washer uses a lot of water" implies a negative opinion because high resource usage is undesirable.

Opinion Spam Detection

Opinion spamming is when people post fake opinions to promote or discredit something. This has become a major issue because it can mislead people and erode trust in online opinions. Opinion spam detection is not just an NLP problem; it also involves analyzing people's posting behaviors.

1.2 Document Sentiment Classification

This task involves classifying an entire document as positive or negative, assuming it discusses a single entity from a single opinion holder. This assumption is often valid for product and service reviews.

Sentiment Classification Using Supervised Learning

Sentiment classification is a text classification problem. It uses machine learning algorithms like Naïve Bayes or Support Vector Machines (SVM) with labeled training data, such as reviews with star ratings. A review with 4 or 5 stars could be labeled as positive, and one with 1 or 2 stars as negative. Features used in this approach include terms and their frequency, parts of speech, sentiment words, and sentiment shifters (e.g., negation).

Sentiment Classification Using Unsupervised Learning (Turney Algorithm)

The Turney algorithm is an unsupervised method that uses sentiment words and phrases to classify documents without training data. It works in three steps:

1. **Extract Phrases**: It extracts two-word phrases that follow specific part-of-speech (POS) patterns.
2. **Estimate Sentiment**: It calculates the "sentiment orientation" (SO) of each phrase by measuring its association with the reference words "excellent" and "poor" using Pointwise Mutual Information (PMI).
3. **Classify Document**: It computes the average SO of all phrases in the document to determine if the overall sentiment is positive or negative.

Sentiment Rating Prediction

This task predicts a review's numerical rating (e.g., 1-5 stars) rather than a simple positive/negative class. This is typically treated as a regression problem. Researchers have used methods like SVM regression and graph-based semi-supervised learning for this purpose.

Cross-Domain Sentiment Classification

Classifiers trained on one domain often fail when applied to another because of differences in language and vocabulary. This problem is called domain adaptation or transfer learning.

- **Challenges**: The main challenges are that words can have different sentiments in different domains, and there is often a lack of labeled training data in the new domain.
- **Approaches**:
    - **Domain-independent features**: Using features that are less sensitive to domain changes, such as N-grams.
    - **Transfer learning methods**: Techniques like **Structural Correspondence Learning (SCL)**, which finds a set of "pivot features" that are common in both the source and target domains to build a more adaptable classifier.

Cross-Language Sentiment Classification

This task involves analyzing sentiment in multiple languages.

- **Approaches**:
    - **Lexicon-based approach**: Translating a sentiment lexicon from a resource-rich language (like English) to a target language to build a classifier.
    - **Co-training method**: This method uses a labeled corpus in one language and an unlabeled corpus in another. The documents are translated to create two views of the data (e.g., English and Chinese) and a co-training algorithm is used to learn a robust classifier for both languages simultaneously.

2.1 Sentence Subjectivity and Sentiment Classification

This approach aims to classify the sentiment in individual sentences, which is a more fine-grained analysis than document-level classification.

Subjectivity Classification

Subjectivity classification distinguishes between sentences that express personal opinions (subjective) and those that state facts (objective). Objective sentences can still imply opinions, so it's a challenging task.

Here are some approaches:

1. **Supervised Learning**: Using classifiers like Naïve Bayes and features such as pronouns, adjectives, and adverbs.
2. **Unsupervised Methods**: Using the presence of subjective expressions to determine if a sentence is subjective.
3. **Sentence Similarity**: Assuming that sentences with similar topics and tone are more likely to have the same classification.
4. **Bootstrapping**: An iterative process that uses a small set of highly accurate subjective and objective sentences to automatically learn patterns and classify more sentences.

5. **Mincut-based Algorithm**: A graph-based method that classifies sentences by considering the relationships between them and assuming that nearby sentences are likely to have the same label.
6. **Social Media-Specific Features**: Using platform-specific clues like hashtags and emoticons to classify the subjectivity of tweets.
7. **Multi-class Classification**: A more complete approach that classifies sentences into four categories: subjective and evaluative, opinion-implied objective, non-opinion objective, and subjective but non-evaluative.

Sentence Sentiment Classification

This task determines if a subjective sentence is positive or negative. A key challenge is that a single sentence can contain multiple opinions.

Here are some approaches:

1. **Lexicon-based Methods**: Using a sentiment lexicon to assign scores to words in a sentence and then aggregating those scores to determine the overall sentiment.
2. **Aggregation of Scores**: Methods that either sum or multiply the scores of sentiment words in a sentence, while also accounting for negation and other contextual factors.
3. **Semi-supervised Learning**: Using a combination of labeled and unlabeled data to train a classifier that can handle three classes: positive, negative, and "other" (for no or mixed opinions).
4. **Hierarchical Models**: Learning sentiment at both the sentence and document levels simultaneously to improve the accuracy of both classifications.
5. **Social Media Features**: Using hashtags, smileys, and punctuation from platforms like Twitter to help classify the sentiment of a sentence.

Dealing with Conditional Sentences

Conditional sentences describe hypothetical situations and pose challenges for sentiment analysis because a sentiment word may not necessarily express a direct opinion. For example, the sentence "If I can find a good camera in the shop, I will buy it" contains the word "good" but expresses no opinion on a specific camera.

To address this, a "divide-and-conquer" approach is needed, where different types of sentences are handled with specialized techniques. Narayanan, Liu, and Choudhary's work focused on conditional sentences and used a supervised learning approach with linguistic features like tense patterns and conditional connectives to accurately determine their sentiment.