

Using Lexical and Machine Learning in Sentiment Mining

One of the biggest challenges in sentiment mining is language expressions. Thwarted and negated expressions can be difficult to analyze for detecting the proper sentiments. Thwarted sentences can contain many words that may have a polarity opposite to the polarity of the expression itself. Similarly a negated expression can be difficult to detect for an automated entity or program because there are negative words followed by positive verbs, adverbs, nouns or adjectives.

There are two main types of approaches used by automated sentiment analysis systems to find out the appropriate sentiments, machine learning and lexicon based approach.

Lexical Based Approach

Lexicon approaches focuses on building successful and efficient dictionaries through which the words in the sentences are compared. If a certain word being looked for is found its polarity value is added to the total polarity score of the text. For example while analyzing a blog if a match of a certain word is found to be 'excellent', being a positive word itself will improve the overall polarity score of the blog.

The classification of a statement in lexical approaches depends on the scoring it receives. This calls for a lot of work going into the process of determining which lexical information works best. For some analysts the subjectivity of a sentence could be determined through hand tagged lexicons. This has provided about 80% accuracy rates in single phrases. Many researches in sentiment mining have opted various techniques to determine the correct polarity of words. The working of lexical approaches for analyzing blogs can be outlined as such:

1. Preprocessing the blog by removing punctuations and stripping HTML tags.
2. Initializing the blog polarity score.
3. Tokenizing the blog posts by determining the features and comparing it to similar list of words from dictionary.
4. For thwarted expressions the words in the dictionary are assigned weights for calculating the minimum path distance from the original word used to describe a feature.
5. Comparing the words and allotting a score to the feature to classify the overall score of the blog.

Machine Learning Approach

In machine learning approaches a series of feature vectors are chosen. The selection of features is crucial in this process for the successful rate of classification. Generally a variety of unigrams or n-grams are chosen for feature vectors where unigrams are single words and n-grams are two or more words in sequential order. This classification to determine the polarity of a document can be done using the algorithms of Super Vector Machines and the Naïve Bayes.

For machine learning methods the creation of feature vectors in a blog for sentiment mining can be done by the following method.

1. Firstly part of speech tagger is applied to each blog post.
2. Collecting the adjectives and adverbs is necessary to make a word set compound.
3. The word set compound will consist of a certain number adjectives and adverbs for comparison.
4. Through the blog post the number of positive words, negative words and negating words are identified.
5. The frequency of each word is then checked with the compound set of popular words.

The machine learning approach is more accurate in the sense that each of the classifier is trained on a collection of representative data called the corpus. The series of feature vectors and collection of tagged corpora are for the training purpose and this can be applied to an untagged corpus of text. It is also called supervised learning for sentiment mining because of this aspect. The only con to this approach is that the corpus will require retraining if it is applied elsewhere.