Thumbs up? Sentiment Classification using Machine Learning Techniques

Bo Pang and **Lillian Lee**
Department of Computer Science
Cornell University

Shivakumar Vaithyanathan
IBM Almaden Research Center
650 Harry Rd.

Presented By :-

Addepalli Bharat Sai
B19BB002

Overview

- Introduction
- Data Set
- Problem Statement
- Traditional Approach
- Machine Learning Approach
- Results
- Conclusion

Introduction

Huge Amount of Data

Sentiment Analysis

Many different Applications

Data Set

Movie Reviews

For our experiments, we chose to work with movie reviews. Our data source was the Internet Movie Database (IMDb). We selected only reviews where the author rating is expressed either with stars or some numerical value. Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral.

Problem Statement

SENTIMENT BASED CLASSIFICATION

Classification of documents based on the sentiments expressed by them. Intuitions seem to differ as to the difficulty of the sentiment detection problem. On the other hand, it seems that distinguishing positive from negative reviews is relatively easy for humans, especially in comparison to the standard text categorization problem, where topics can be closely related. One might also suspect that there are certain words people tend to use to express strong sentiments, so that it might suffice to simply produce a list of such words by introspection and rely on them alone to classify the texts.

Traditional Approach

Based on Human Interface

Selected Words By the Students

	Proposed word lists	Accuracy	Ties
Human 1	positive: dazzling, brilliant, phenomenal, excellent, fantastic negative: suck, terrible, awful, unwatchable, hideous	58%	75%
Human 2	positive: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting negative: bad, cliched, sucks, boring, stupid, slow	64%	39%

The accuracy — percentage of documents classified correctly — for the human-based classifiers were 58% and 64%, respectively. Here the tie rates — percentage of documents where the two sentiments were rated equally likely — are quite high

New Set of Words From these tests

	Proposed word lists	Accuracy	Ties
Human 3 + stats	positive: love, wonderful, best, great, superb, still, beautiful negative: bad, worst, stupid, waste, boring, ?, !	69%	16%

As we can see that, using these words raised the accuracy to 69%. Also, although this third list is of comparable length to the other two, it has a much lower tie rate of 16%. We further observe that some of the items in this third list, such as “?” or “still”, would probably not have been proposed as possible candidates merely through introspection, although upon result we can see the merit.

Machine Learning Approach

Naive Bayes
Maximum Entropy
Support Vector Machines

To implement these, we used the standard bag of features framework. Let $\{f_1, \ldots, f_m\}$ be a predefined set of 'm' features that can appear in a document. Let $n_i(d)$ be the number of times f_i occurs in document d. Then, each document d is represented by the document vector, $d = (n_1(d), n_2(d), \ldots, n_m(d))$.

Machine Learning Methods

Naive Bayes

Naïve Bayes, for document d and class,
$c^* = \arg\max_c P(c|d)$.
From Bayes rule by assuming all fi's is conditionally independent

$$P_{NB}(c \mid d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)})}{P(d)}$$

Maximum Entropy

Maximum entropy classification, which is better than Bayes as it does not involve any assumption, to estimate P (c | d), we have here Z(d) is the normalization vector and the feature vector

$$F_{i,c} = \begin{cases} 1, & n_i(d) > 0 \; and \; c' = c \\ 0 & otherwise \end{cases}$$

$$P_{ME}(c \mid d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right)$$

Support Vector Machines

Support vector machines (SVM's) in which the basic idea is to find a hyperplane rom the training procedure which is represented by a vector, ω which is also called as margin

$$\vec{\omega} = \sum_j \alpha_i c_j \vec{d_j} \; , \alpha_i \geq 0$$

Results

Based Machine Learning Methods

Results

S.No.	Features	No. of Features	Frequency or Prescence	NB	ME	SVM
(1)	Unigrams	16165	Frequency	78.7	N/A	72.8
(2)	Unigrams	16165	Prescence	81.0	80.4	82.9
(3)	Unigrams + Bigrams	32330	Prescence	80.6	80.8	82.7
(4)	Bigrams	16165	Prescence	77.3	77.4	77.1
(5)	Unigrams + POS	16695	Prescence	81.5	80.4	81.9
(6)	Adjectives	2633	Prescence	77.0	77.7	75.1
(7)	Top 2633 Unigrams	2633	Prescence	80.3	81.0	81.4
(8)	Unigrams + Position	22430	Prescence	81.0	80.1	81.6

Conclusion

Machine learning techniques are quite good in comparison with the human generated baselines. In terms of relative performance Naïve Bayes do the worst and SVMs give the best

Thanks!