Abstract

Sentiment analysis or opinion mining is a process of categorizing and identifying the sentiment expressed in a particular text. The need of automatic sentiment retrieval of the text is quite high as amount of reviews obtained from the Internet are huge in number. Reviews on various 'E-commerce websites', 'social networks', and 'movie review websites' come up huge in number regularly. These reviews on popular products help in determining the public opinion towards the product. An averaged histogram model is proposed in the process that deals with text classification in continuous variable approach. After data cleaning and feature extraction from the reviews, average histograms are constructed for every class, containing a generalized feature representation in that particular class. Histograms of every test elements are then matched with the averaged histograms of every class using k-Nearest Neighbor and Naïve Bayesian Classifier. Results showed on 3000 reviews a steady classification accuracy of 79-80% with the Naïve Bayesian Classifier with very little cost of computation, and increase in the number of training dataset k-Nearest Neighbor can give up to a high accuracy of 85%. This work proposed here is language independent, neither include any dictionary nor depend on the
meaning of any word.

References

8. imdb: Maas et. al., 2011 'Learning word vectors for sentiment analysis'
9. amazon: McAuley et. al., 2013 'Hidden factors and hidden topics: Understanding rating dimensions with review text'


Index Terms

Computer Science Information Sciences

Keywords

Averaged Histogram