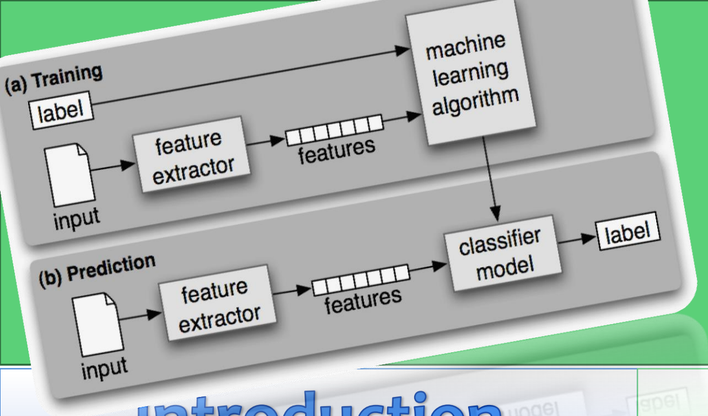
Twitter Sentiment Analysis

Ajay Singh 12056
Dept of Computer Science & Engg,IIT Kanpur
Mentor – Prof Amitabha Mukherjee



Introduction

Sentiment analyis refers to the use of natural language processing, text analysis and computational linguisticsto identify and extract subjective information in source materials.

Approach

- Preprocessing of tweets
- Filtering for Feature Vector size reduction
- Machine learning classifiers
  1. Naïve Bayes Classifier
  2. Maximum Entropy Classifier
- Unigrams used as features

Dataset

http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

- Consists of 20,000 positive and 20,000 negative tweets
- Tweets are collected from all topics and issues

Previous Work

- Sentiment analysis using larger pieces of text has already been done on a large scale.
- Twitter sentiment analysis has been attempted through machine learning as well as keyword matching.

Results

Using unigrams as features,
Accuracy of:

- Naïve Bayes Classifier – 76%
- Maximum Entropy classifier – 75.4%

Future Work

- Neutral tweets need to be classified as a lot of the tweets are factual or unbiased news
- Semantics of tweet need to considered as sentiment depends on the perspective
- Bigrams need to used in feature extraction along with unigrams

Challenges

- Presence of usernames
- Links and URLS along with tinyurls
- Repeated letters in a word to stress an emotion
- Hashtags
- Punctuations and additional spaces

Conclusion

- Even though unigram feature extractor is the simplest, it fails to identify negations. Using bigrams will help a lot in increasing the accuracy of the classifier
- Presence of neutral tweets too causes a dip in the accuracy

References

- sentiment140, Go, Bhayani and Huang, Stanford University.
- Twitter sentiment classier using Python and NLTK Laurent Luce.
- Naive Bayes Classier Jacob Perkins.
- Twitter Sentiment Niek Sanders