Building a Classifier to analyze the user sentiment about a Product

Chaithanya Kumar (A20327683)

Hina Garg (A20342375)

PROBLEM

- To predict if a tweet posted by a user holds positive or negative sentiment towards a product.

APPROACH

- Collect data (user tweets) from Twitter
- Tokenize the data to form a feature vector
- Build a text classifier using Logistic Regression and Naive Bayes classification methods
- Perform cross validation technique on training data to assess accuracy of the classifier.
- Implement below experiments to check if classifier's accuracy can be improved
  - Build different tokenizers
  - Vary different parameters to create a feature vector.
- Predict document labels in testing data and compute classifier's accuracy
- Analyze the top errors in testing data

DATA

- Tweet collection - Related to a product(iPhone) posted by different Twitter users
- Collected documents and labelled - Training data set (398) and Testing data set (392)
- Used "Twitter Search API" to search relevant tweets using search parameter (query, count, lang)
- Tweets sentiment category and statistics (shown below)

Labels/Category	Train document counts	Test document counts
positive	200	193
negative	198	199

RESULTS

- Classification methods used:
  - Logistic Regression
  - Naive Bayes
- Test Accuracy Results
  - Logistic Regression: 0.8189
  - Naive Bayes: 0.7984

- Used efficient feature vector construction to improve classifier accuracy in Logistic regression

- Both methods failed in identifying true labels of documents for the test data:
  - Logistic Regression: 40 false positive 31 false negative
  - Naive Bayes: 48 false positive 31 false negative

CONCLUSIONS

- This project presents a method to collect a corpus that can be used to train a sentiment classifier
- Analyses on how sentiment predictions on test data can be affected by the presence of certain negative words like: 'not, 'don't, 'won't, 'can't.. etc.
- Built both unigrams and bigrams feature vectors to improve the accuracy of classifier
- Observation: Classifier's accuracy reduces as the size of testing data increases while keeping the same size of training data