Abstract

Email is one of cheapest and fastest availaible tools of communication. Major problem is the increasing number of unsolicited commercial email or called spam. Spam have the impact waste bandwidth of internet connections, reduce data storage, increase computations time and very bother users.

Major approaches to spam detection use the bag of words representation method. But spam content often contain words wrong in grammar and have variation weird punctation like 'f.r.e.e.','f-r-e-e','f r e e'. This conditions affected bag words representation approach not strong. Moreover, this approach need list of stop words, stemming and lemmatizer certain language dependent.

In this final project, approach to spam detection used bag of character n-grams. This approach try to solve problem faced by bag of words representation method. But number of feature produced is still large so used Support Vector Machine(SVM) as classification algorithm able to deal with high dimensional data space.

The result showed that spam detection using character n-grams applicable well. Character n-grams method can avoid using stop-list, stemming and lemmatization language dependent. The result showed that optimal length of character n is n=4 for binary weighting and n=5 for term frequency(TF) weighting.

***Keyword**: spam detection, character n-grams, support vector machine(SVM).*