Abstract

With the help of Social Networking sites many news providers used to share their news headlines on the micro blogging sites such as twitter. We are proposing a system to classify tweets into different groups and labels so that the user can identify the particular tweet from particular category. We will use 120 character tweets for our analysis purpose. Various active and verified twitter accounts would be chosen to extract the tweets. Each tweet is to be classified into 2 category-spam and non-spam. Then further spam group is classified as advertisement, malicious and URL links. The non-spam tweets are classified into 6 labels. These classified tweets then are used to train the various machine learning techniques. Words of each tweet considered as features and a feature vector was created using bag-of-words approach in order to create the instances. The data will be trained using SVM (Support Vector Machine), Naive Bayes and K neighbor machine learning techniques and their efficiency will be compared.
References

1. Ersin Yar, Lemi Baruh, Syleyman S. Kozat 2016 Online Text Classification for Real Life Tweet Analysis
7. Support vector mechanism by David Meyer: The interface to libsvm in package e1071
9. Machine learning course by Andrew Ng: https://www.coursera.org/learn/machine-learning
10. Basic text mining in R: https://rstudio-pubs-staic.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html

Index Terms

Computer Science  Information Sciences

Keywords

SVM -Support Vector Mechanism NLP -Natural Language Processing NB-Naïve Bayes KNN-K Nearest Neighbor