Improved Feature Selection for Better Classification in Twitter

Abstract

Social networks are widely used as a communication tools by millions of people and their friends. In today's era everybody is online and use social network for interaction, to gain knowledge, for business purpose, politics and many more. But along with positive approach of using these tools some infect many negative approaches are also applied on these tools for executing malwares and spam messages. Spam on twitter has become one of the most trending topics of research in recent years. And many researchers have done work on it but make some very complex structure to detect spam but still cannot achieve that level of accuracy in detection. So to gain the greater level of accuracy and to reduce the complexity of structure this work proposes a simplified model to detect the spam tweets which are spread by unauthorised users or by spammers. And this is analysed by feature extraction and applying classifiers. The text and content attribute features are extracted by pre-processing and forming a feature vector matrix. Moreover K-nearest neighbour (KNN) and decision tree two classifier algorithms are applied to show the comparative results. The results are evaluated with False positive rate (FPR), F- measures, True positive rate (TPR) and accuracy with improved detection results.
References

- Saini Jacob Soman, Dr. S. Murugappan, "Detecting Malicious Tweets in Trending Topics using Clustering and Classification", 2014 International Conference on Recent Trends in Information Technology, IEEE
- Mashable. Twitter now has more than 200 million monthly active users. [Online]. Available: http://mashable.com/2012/12/18/twitter-200-million-active-users/
- Saini Jacob Soman, Dr. S. Murugappan, "Bayesian Probabilistic Tensor Factorization for Malicious Tweets in Trending Topics", 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), IEEE

Index Terms
Computer Science  Information Sciences

**Keywords**
- K-nearest neighbour
- Decision tree classifier algorithm
- Pre-processing
- Social network
- Spam detection