Faculty of Electrical Engineering & Computer Science
Master Degree Course Electronics Engineering, MSc EE

HSB
Hochschule Bremen
City University of Applied Sciences

Master 2021

Ms. Ravali Mummadi

Analysis and Prediction of Content Popularity on Twitter using Rate-limited Data.

Aim of this thesis is to analyze the backdrop of popularity of Twitter trends with the rate limited real-time data, subsequently predict/model the popularity using machine learning techniques to be used in assessing networking technologies. Network traffic has significance in the planning & design of a network all through the improvement in networking technologies. As popularity of content drives the network traffic, understanding the dynamics of popularity helps build realistic network traffic models to fill-in for the actual endless network traffic which cannot be retained, for the performance evaluation as well as enhancement purposes. Being able to predict the upcoming popularity of the current data is essential to regulate and plan resources and enhancements to meet the changing requirements for performance optimization. As there are no unlimited sources of streaming data, the purpose of this thesis is to analyze the behavioral characteristics on the rate-limited data offered to developers by Twitter.

The idea is to pre-process the data collected through the Twitter API, perform exploratory data analysis, then experiment on various machine learning classification and regression algorithms, to obtain a model that yields the highest accuracy for the best prediction and determine if the limited data is adequate for at least near-accurate prediction, as many unanticipated events cannot to be trained into the model.

Existing literature have been conducted on diffusion, sentiment and event-specific analyses. Much of the research has tended to focus on the accuracy but not on the crucial aspect of choosing features fed into the model. A key limitation of a recent research by Devi P.S et al., 2020 is not considering the importance of features fed into the model. Although their model has achieved an accuracy of 94.4%, I would argue that excluding Feature selection procedure only leads to curse of dimensionality, overfitting, reduced accuracy and speed. The significance of my thesis is to extend the analysis to emphasize on features and subsequently, model the prediction on wide number of trends, but rate-limited data offered by Twitter.