Web services such as online banking, gaming, and social networking have rapidly evolved as has the reliance upon them by people to perform everyday tasks. As a result, a large amount of information is uploaded on a daily basis to the web. The openness of the web exposes opportunities for criminals to upload malicious content. Despite extensive research, email based spam filtering techniques are unable to protect other web services. Therefore, a counter measure must be taken that generalizes across web services to protect the user from malicious hosts. This paper describes an approach that classifies URLs automatically based on their lexical and host-based features. The usability of Mahout is demonstrated for such scalable machine learning problems and online learning is considered over batch learning due to its useful properties. The classifier achieves 93-97% accuracy by detecting a large number of malicious hosts, with a modest false positive rate.

Mohammed Nazim Feroz is a graduate student majoring in Computer Science at Texas Tech University. He is currently working on his thesis with Dr. Susan Mengel in Computer Security. His primary interests lie in the field of Artificial Intelligence, Computer security, and Web application development. Mohammed completed his Bachelor’s degree in Information Technology, at Anna University, India, 2011. During his undergraduate study, he has done his research on Healthcare IT and Artificial Intelligence areas and submitted three papers on international conferences.