University of Victoria | Graduate Studies

Notice of the Final Oral Examination
for the Degree of Master of Applied Science

of

HADEER SAAD AHMED

BEng (Ahram Canadian University, 2012)

"Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity"

Department of Electrical and Computer Engineering

Tuesday, October 24, 2017
10:00 A.M.
Engineering Office Wing
Room 430

Supervisory Committee:
Dr. Issa Traore, Department of Electrical and Computer Engineering, University of Victoria (Supervisor)
Dr. Lin Cai, Department of Electrical and Computer Engineering, UVic (Member)

External Examiner:
Dr. Alex Thomo, Department of Computer Science, UVic

Chair of Oral Examination:
Dr. Laurel Bowman, Department of Greek and Roman Studies, UVic

Dr. David Capson, Dean, Faculty of Graduate Studies

Abstract

In recent years, deceptive contents such as fake news and fake reviews, also known as opinion spams, have increasingly become a dangerous prospect, for online users. Fake reviews affect consumers and stores alike. Furthermore, the problem of fake news has gained attention in 2016, especially in the aftermath of the last US presidential election. Fake reviews and fake news are a closely related phenomenon as both consist of writing and spreading false information or beliefs. The opinion spam problem was formulated for the first time a few years ago, but it has quickly become a growing research area due to the abundance of user-generated content. It is now easy for anyone to either write fake reviews or write fake news on the web.

The biggest challenge is the lack of an efficient way to tell the difference between a real review or a fake one; even humans are often unable to tell the difference. In this thesis, we have developed an n-gram model to detect automatically fake contents with a focus on fake reviews and fake news. We studied and compared two different features extraction techniques and six machine learning classification techniques. Furthermore, we investigated the impact of keystroke features on the accuracy of the n-gram model. We also applied semantic similarity metrics to detect near-duplicated content. Experimental evaluation of the proposed using existing public datasets and a newly introduced fake news data set introduced indicate improved performances compared to state of the art.