MSc Project Report

AUTOMATIC SUMMARIZATION OF PRIVACY POLICIES USING DEEP LEARNING

A dissertation by

Devmie Wannakuwatte

RGU No: 1613578

Supervised By

Mr. Buddhika Kasthuriarachchy

August 2018

A report submitted as part of the requirements for the degree of MSc in Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

ABSTRACT

With the increase in web users and websites, the safety of the users' privacy online is a major concern. When accessing a website or using a computer software, all users are presented with a Privacy Statement, that details, what and how any information is collected from the users. Most of the users do not take time to read this policy, even if they do read, most of the time the language is hard to grasp for an average user.

To assist in this task, this research aims to create a tool that summarizes the content of a privacy policy, to identify what information is being captured via a privacy policy. The proposed application will take a full text privacy policy as an input and will generate a visual representation of what information is being collected.

The model used for this information retrieval is a GUR based Recurrent Neural Network for the text classifier which can identify 8 different categories of information that is collected. Based on the testing carried out, this model yielded an accuracy of 0.81 on test data. The average wait time to get an output was about 497 ms, for a policy of average 2000 words. Based on the results, the research can be considered as a successful attempt in summarizing privacy policies, which may give the end user an advantage in understanding the content of privacy policies in a much easier manner.

Keywords:

Natural Language Processing, Text Classification, Recurrent Neural Network