# DUPLICATE DETECTION IN MULTI-DOMAIN COMMUNITY QUESTION ANSWERING

K.K.Rasika Kariyawasam

168233K

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2020

# DUPLICATE DETECTION IN MULTI-DOMAIN COMMUNITY QUESTION ANSWERING

K.K.Rasika Kariyawasam

168233K

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science specializing in Data Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2020

# DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part, in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: ……………… Date:……………….

Name: K.K.Rasika Kariyawasam

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the CS6997 MSc Research Project.

Signature of the supervisor: …………………… Date:………………..

Name: Dr. Surangika Ranathunga

## Abstract

Community based question answering forums are very popular these days. People tend to refer community forums for opinions in various fields such as electronics, medical and automobile. It is very easy and useful to find a good opinion freely, but it is hard to choose the correct one when there are thousands of reviews.

There have been several efforts to automate the activities of community-based question answering systems, such as the selection of the most relevant answers to the question (question comment similarity), and identifying the questions already posted that are similar to the new question (question-question similarity). However, there are fewer attempts taken to automate the process of duplicate detection in community question answering systems. At the moment, it is the community itself that manually detects duplicates. The automation attempts are more into individual domains.

The objective of this research is to implement a mechanism that effectively identifies duplicate questions in a data set consisting of question-answer sets from multiple domains. Solution we propose consists of two focus areas such as classification and retrieval. A neural network composed of two parallel LSTM layers (to represent query and candidate question), attention layer and a gradient reversal layer (based on domain) is proposed as the question pair classifier. It’s trained for individual domains (without gradient reversal) and achieved better accuracy than the latest baseline research for this dataset for 9 out of 12 domains. For retrieval the approach was to retrieve 20 candidates using BM25 and re-rank using classifiers trained already. This selects the duplicate into top 10 with better MAP than BM25 does 6 out of 12 domains. Another important observation is that the common model built with all the data combined gained better MAP than the individual models for 7 domains out of 12 in the retrieval case.

**Keywords:** Multi domain data, Siamese neural networks, Domain adaptation, Question pair classification, Duplicate question retrieval

# ACKNOWLEDGEMENTS

First, I'm grateful to Dr. Surangika Ranathunga for giving me the opportunity and further guidance in selecting and conducting this research. Her continuous supervision greatly helped me in keeping the correct phase in research work. I especially appreciate the frequent feedback on the report, which helped me to correct and fine-tune it to this level. Last but not least, my heartfelt gratitude goes to my parents, wife and friends who supported me throughout this effort.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATION

| | |
|---|---|
| NLP | Natural Language Processing |
| QA | Question Answering |
| CQA | Community Question Answering |
| MAP | Mean Average Precision |
| POS | Part of Speech |
| IR | Information Retrieval |
| LTR | Learning to Rank |
| TF | Term Frequency |
| IDF | Inverse Document Frequency |
| CNN | Convolutional Neural Network |