Context Aware News Similarity for the identification of follow-ups among human loss related news articles

By

Waqas Ali (15007114012)

A Dissertation Submitted For the Degree of
Master in Software Engineering in the Department of Computer Science,
Faculty of Science, University of Management and Technology, Lahore

Month year

August 2018

Supervised by:

Dr. Adnan Abid

DECLARATION

I declare that this dissertation is my own original work. Where collaborations with other researchers are involved, or materials generated by other researchers are included, the parties and/or materials are acknowledged or are explicitly referenced as appropriate.

This work is being submitted for the degree of Master in Software Engineering at the University of Management and Technology. This thesis has not been submitted to any other university or institution for any other degree or examination.

Date

Signature

Waqas Ali

DEDICATION

This thesis is dedicated to my family and friends.

A special feeling of gratitude to my loving parents, their proficient

Guidance and splendid inspiration had

Positively influenced my life and made me capable

Of attaining this degree

Successfully.

ACKNOWLEDGEMENT

I am highly grateful to Allah Almighty who blessed me with strength, wisdom and proper light in order to complete my research work.

I am thankful to my kind supervisor, Dr. Adnan Abid for his precious guidance, which paved the way to steer me a right. He not only guided me throughout this work but also provided me with his unprecedented and continuous support, encouragement and motivation. This work never has been completed without his enthusiastic supervision.

The prayer of my lovely parents, sisters and teachers always remained with me in this difficult work. Thanks for their selflessness support. My friends and my fellows, who remembered me in their prayers and extended their moral support to me are also worthy for my complements.

Abstract

Text similarity plays an important role in document clustering, plagiarism detection, automatic student answer grading, information retrieval and language translation systems. Many researchers have studied on string, corpus and knowledge based approaches to resolve the problem of document similarity. In this paper research has been made on full text similarity and context information of a news. Time complexity and follow-ups distribution are main challenges for full text similarity which is computed by using jaccard index. The proposed system in this paper uses context information of news to resolve these issue and to classify the news into three categories i.e. same day follow-ups, different day follow-ups and distinct news. Context is built by using well known Stanford parser which further enriched with factor of severity to reduce the target dataset for similarity computation. Results showed that context aware similarity approach is better than traditional full text similarity approach in efficiency and accuracy.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 News Classification

News classification is the task of classifying a news into predefined categories on the basis of training set. Simply, if ‘ni’ is a news article of the entire dataset of news articles ‘N’ and {c1, c2, c3, …., cn} is set of all the classes, then news classification assigns one class cj to a news article ni. Depending upon the characteristics of news text. News may have one or more than one labels. If a news text belong to one class, then it is single-class-label and if it belongs to more than 1 classes, then it is called as multi-class-label [1].