



# **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

Copyright © 2018 Waqas Ali

## **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

DECLARATION .....	ii
DEDICATION .....	iii
ACKNOWLEDGEMENT .....	iv
Abstract .....	v
Table of Contents .....	vi
List of Figures .....	ix
List of Tables .....	xi
Chapter 1 .....	1
Introduction.....	1
1.1 News Classification.....	2
1.2 Previous work.....	<b>Error! Bookmark not defined.</b>
1.3 Problem Statement and Objectives .....	<b>Error! Bookmark not defined.</b>
1.4 Research Methodology.....	<b>Error! Bookmark not defined.</b>
1.5 Rest of the content.....	<b>Error! Bookmark not defined.</b>
Chapter 2.....	<b>Error! Bookmark not defined.</b>
Literature Review.....	<b>Error! Bookmark not defined.</b>
2.1 Related Work.....	<b>Error! Bookmark not defined.</b>
2.1.1 Related to pre-processing.....	<b>Error! Bookmark not defined.</b>
1.1.2 Related to news similarity and classification.....	<b>Error! Bookmark not defined.</b>
Chapter 3.....	<b>Error! Bookmark not defined.</b>
Full Text Similarity Approach .....	<b>Error! Bookmark not defined.</b>
3.1 Dataset Collection.....	<b>Error! Bookmark not defined.</b>
3.1.1 Category wise distribution of data .....	<b>Error! Bookmark not defined.</b>
3.2 Pre-Processing.....	<b>Error! Bookmark not defined.</b>
3.2.1 News text tokenization (Step 1).....	<b>Error! Bookmark not defined.</b>
3.2.2 Removal of Diacritic words (Step 2) .....	<b>Error! Bookmark not defined.</b>
3.2.3 Stop words elimination (Step 3) .....	<b>Error! Bookmark not defined.</b>
3.2.4 Stemming Words Removal (Step 4) .....	<b>Error! Bookmark not defined.</b>
3.2.5 News Classification (Step 5).....	<b>Error! Bookmark not defined.</b>
3.2.6 Add Synonyms (Step 6).....	<b>Error! Bookmark not defined.</b>

3.3 Full text similarity.....	<b>Error! Bookmark not defined.</b>
3.3.1 Jaccard Index .....	<b>Error! Bookmark not defined.</b>
3.4 Evaluation and Experimental Results .....	<b>Error! Bookmark not defined.</b>
3.4.1 Overall Results.....	<b>Error! Bookmark not defined.</b>
3.4.2 Confusion Matrix.....	<b>Error! Bookmark not defined.</b>
3.4.3 Results of type accident .....	<b>Error! Bookmark not defined.</b>
3.4.4 Results of type crime .....	<b>Error! Bookmark not defined.</b>
3.4.5 Results of type terrorism.....	<b>Error! Bookmark not defined.</b>
3.4.6 Results of type operations.....	<b>Error! Bookmark not defined.</b>
3.4.7 Results of type natural disaster .....	<b>Error! Bookmark not defined.</b>
Chapter 4.....	<b>Error! Bookmark not defined.</b>
Dataset construction for follow-up .....	<b>Error! Bookmark not defined.</b>
4.1 Types of follow-up.....	<b>Error! Bookmark not defined.</b>
4.1.1 Same day follow-ups.....	<b>Error! Bookmark not defined.</b>
4.1.2 Different day follow-ups.....	<b>Error! Bookmark not defined.</b>
4.1.3 Same text different news.....	<b>Error! Bookmark not defined.</b>
4.2 Overall news distribution.....	<b>Error! Bookmark not defined.</b>
4.3 Follow-up distribution .....	<b>Error! Bookmark not defined.</b>
4.4 Source information of dataset .....	<b>Error! Bookmark not defined.</b>
4.5 Class wise same day source information .....	<b>Error! Bookmark not defined.</b>
4.6 Class wise distinct day source information.....	<b>Error! Bookmark not defined.</b>
Chapter 5.....	<b>Error! Bookmark not defined.</b>
Context Aware Similarity Computational Model.....	<b>Error! Bookmark not defined.</b>
5.1 Context Aware Systems.....	<b>Error! Bookmark not defined.</b>
5.2 Preprocessing Revised .....	<b>Error! Bookmark not defined.</b>
5.3 News Context Information.....	<b>Error! Bookmark not defined.</b>
5.3.1 Date .....	<b>Error! Bookmark not defined.</b>
5.3.2 Day.....	<b>Error! Bookmark not defined.</b>
5.3.3 Class/Category .....	<b>Error! Bookmark not defined.</b>
5.3.4 Location .....	<b>Error! Bookmark not defined.</b>
5.3.5 Source URL link .....	<b>Error! Bookmark not defined.</b>
5.3.6 Number of dead and injured people.....	<b>Error! Bookmark not defined.</b>

5.3.7 Severity .....	<b>Error! Bookmark not defined.</b>
5.3.8 Nouns .....	<b>Error! Bookmark not defined.</b>
5.4 Algorithm .....	<b>Error! Bookmark not defined.</b>
5.4.1 Child vs Parent news.....	<b>Error! Bookmark not defined.</b>
5.4.2 Rules and Heuristics .....	<b>Error! Bookmark not defined.</b>
5.5 Results and Evaluation.....	<b>Error! Bookmark not defined.</b>
5.5.1 Overall results .....	<b>Error! Bookmark not defined.</b>
5.5.2 Confusion Matrix .....	<b>Error! Bookmark not defined.</b>
5.5.3 Results of type accident .....	<b>Error! Bookmark not defined.</b>
5.5.4 Results of type crime .....	<b>Error! Bookmark not defined.</b>
5.5.5 Results of type terrorism.....	<b>Error! Bookmark not defined.</b>
5.5.6 Results of type operation .....	<b>Error! Bookmark not defined.</b>
5.5.7 Results of type natural disaster .....	<b>Error! Bookmark not defined.</b>
Chapter 6.....	<b>Error! Bookmark not defined.</b>
Conclusion .....	<b>Error! Bookmark not defined.</b>
References.....	<b>Error! Bookmark not defined.</b>

## List of Figures

- Figure 3.1. Representation of Dataset Distribution .....**Error! Bookmark not defined.**
- Figure 3.2. Pre-processing on raw news data .....**Error! Bookmark not defined.**
- Figure 3.3. Full text similarity model architecture .....**Error! Bookmark not defined.**
- Figure 3.4. Overall results by full text similarity approach in the form of bars showing correctness and incorrectness.....**Error! Bookmark not defined.**
- Figure 3.5. Overall results of the full text similarity approach through confusion matrix .... **Error! Bookmark not defined.**
- Figure 3.6. Accidents related data in the form of bars showing correctness and incorrectness of follow-up and distinct news respectively.....**Error! Bookmark not defined.**
- Figure 3.7. Crimes related data in the form of bars showing correctness and incorrectness of follow-up and distinct news respectively.....**Error! Bookmark not defined.**
- Figure 3.8. Terrorism related data in the form of bars showing correctness and incorrectness of follow-up and distinct news respectively.....**Error! Bookmark not defined.**
- Figure 3.9. Operations related data in the form of bars showing correctness and incorrectness of follow-up and distinct news respectively.....**Error! Bookmark not defined.**
- Figure 3.10. Disasters related data in the form of bars showing correctness and incorrectness of follow-up and distinct news respectively.....**Error! Bookmark not defined.**
- Figure 4.1. First screenshot of news related to natural disaster, same day follow-up ..... **Error! Bookmark not defined.**
- Figure 4.2. Second screenshot of news related to natural disaster, same day follow-up..... **Error! Bookmark not defined.**
- Figure 4.3. News related to terrorism crawled from Dawn on 19 Aug 2011**Error! Bookmark not defined.**
- Figure 4.4. Follow-up news of blast incident happened on 19 Aug 2011, crawled from CNN .....**Error! Bookmark not defined.**

Figure 4.5. Follow-up news of blast incident happened on 19 Aug 2011, crawled from Tribune Express .....**Error! Bookmark not defined.**

Figure 4.6. News of natural disaster related to flood in Gilgit Baltistan taken from maverickpakistanis.com on 14 august 2018 .....**Error! Bookmark not defined.**

Figure 4.7. News of natural disaster related to flood in Gilgit Baltistan**Error! Bookmark not defined.**

Figure 5.1. Showing preprocessing for context processing .....**Error! Bookmark not defined.**

Figure 5.2. Hierarchy of features to build context of a news.....**Error! Bookmark not defined.**

Figure 5.3. Follow Up News Identification System Using Context Information**Error! Bookmark not defined.**

Figure 5.4. Overall results by context aware similarity approach in the form of bars showing correctness and incorrectness of the system .....**Error! Bookmark not defined.**

Figure 5.5. Overall results of context aware similarity approach through confusion matrix **Error! Bookmark not defined.**

## List of Tables

- Table 3.1. Shows the Source name and Source Links .....**Error! Bookmark not defined.**
- Table 3.2. Overall News Distribution into five categories .....**Error! Bookmark not defined.**
- Table 3.3. Overall Follow-ups Source Information .....**Error! Bookmark not defined.**
- Table 3.4. Category wise follow ups source information .....**Error! Bookmark not defined.**
- Table 3.5. Overall results of follow-ups prediction by full text similarity model ..... **Error! Bookmark not defined.**
- Table 3.6. Class Accident result of follow-ups prediction by full text similarity model..... **Error! Bookmark not defined.**
- Table 3.7. Class Crime result of follow-ups prediction by full text similarity model ..... **Error! Bookmark not defined.**
- Table 3.8. Class Terrorism result of follow-ups prediction by full text similarity model ..... **Error! Bookmark not defined.**
- Table 3.9. Class Operation results of follow-ups prediction by full text similarity model.... **Error! Bookmark not defined.**
- Table 3.10. Class Natural Disaster result of follow-ups prediction by full text similarity model .....**Error! Bookmark not defined.**
- Table 4.1. Overall News Distribution into five categories of distinct and follow-up news .. **Error! Bookmark not defined.**
- Table 4.2. Overall Follow-ups Distribution of five categories into same**Error! Bookmark not defined.**

Table 4.3. Source information about same day and different day follow-ups respectively... **Error! Bookmark not defined.**

Table 4.4. Category wise same day follow ups source information **Error! Bookmark not defined.**

Table 4.5. Category wise different day follow ups source information **Error! Bookmark not defined.**

Table 5.1. Overall results of same day follow-ups and different day follow-ups prediction by using context aware similarity model..... **Error! Bookmark not defined.**

Table 5.2. Class Accident result of same day follow-ups and different day follow-ups prediction by context aware similarity model..... **Error! Bookmark not defined.**

Table 5.3. Class Crime result of same day follow-ups and different day follow-ups prediction by context aware similarity model..... **Error! Bookmark not defined.**

Table 5.4. Class Terrorism result of same day follow-ups and different day follow-ups prediction by context aware similarity model..... **Error! Bookmark not defined.**

Table 5.5. Class Operation result of same day follow-ups and different day follow-ups prediction by context aware similarity model..... **Error! Bookmark not defined.**

Table 5.6. Class Natural Disaster result of same day follow-ups and different day follow-ups prediction by context aware similarity model..... **Error! Bookmark not defined.**



## **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].