Incidents Management System (IMS) – A software for management of Human Loss News Reports

By

ANSAR ABBAS CHISHTY

15026050006

Thesis presented in partial fulfillment of the requirements for the degree of

MS (COMPUTER SCIENCE)

2015 – 2017

Approved by:

_____

Dr. ADNAN ABID, Ph.D. ASSOCIATE PROFESSOR, SST
UMT, LAHORE, PAKISTAN

_____

UZMA FAROOQ ASSISTANT PROFESSOR, SST
UMT, LAHORE, PAKISTAN

ACCEPTED AND APPROVED ON BEHALF OF THE UNIVERSITY

_____

Dr. SHAUKAT IQBAL KAMBOH, Ph.D. DEAN, SST
UMT, LAHORE, PAKISTAN

SCHOOL OF SYSTEMS AND TECHNOLOGY

UNIVERSITY OF MANAGEMENT AND TECHNOLOGY, LAHORE, PAKISTAN

ABSTRACT

In today's world, with the exponential rise of human population and due to emergence of modern technology, every society is facing such challenges that they have never faced before during entire course of history. One of the main challenges among many is to secure and protect lives of subjects from any kind of emergency deaths and injuries. Human lives are lost or damaged due to accidents, disasters and various criminal activities and during law and enforcement agencies operations at regular pace in every society. This loss is priceless and very sensitive matter for every society. One way of protecting people from such situations is to take preemptive actions so that such incidents can be made little to happen. For this purpose every nation sets up departments those plan different strategies to prevent from such situations. These departments and their personals require different tools and technologies to assist them in their planning's and completion of relevant operational tasks.

We are living in the age of modern science and a unique miracle of modern age science is information technology. We are in the realm of information technology. Back bone of informational applications is data. Different kind of data is generated every day in every society. Media is considered fourth pillar of society and its role in different walks of daily life is above doubts and non-debatable. In every society lot of accidents, crimes and other incidents happen that involve loss of human lives. These incidents do not go unnoticed. These are reported by different segments of society and regular media is one of them besides many social media platforms. Newspaper agencies are one of regular media outlets. There reports are reliable and trust worthy. Newspaper websites are modern form of newspapers. These are easily accessible and updated at regular intervals as compared to old classical styled printed newspapers. Archives of these news websites are easily and widely accessible at zero cost.

Pakistan is a nation of worth 200 million inhabitants in the north western region of South Asia. It is passing through a very crucial period of its history. Besides regular incidents of human loss like every other nation, it is also bleeding by the hands of many terrorist organizations operating in Pakistan and many other neighboring countries of the region. It is been more than a decade that Pakistani nation is fighting against these terrorists. Almost every day terrorists target different segments of this nation. It is need of the day to curb such activities, data about such incidents must be captured, collected in an organized way, analyzed in a systematic way to find different useful information, share this information among different departments dealing with these events in a joint and organized manner. So that relevant authorities can deal with these situations proactively and effectively.

In this thesis, we proposed a state of the art framework known as Incidents Management System or IMS, to tackle the challenges described above. The proposed framework can be helpful in addressing and avoiding human loss for authorities, as this system can help them incollecting, processing and analyzing the real-time data about human loss reports. The historical data of all such reports would also be stored in an efficient

repository for analysis. We developed a prototype of many of the key components of the proposed framework crawler, classifier and reporter. We evaluated various classifiers for this problem against a middle sized data of 1600 plus reports. A number of performance metrics are used to measure the performance of these classifiers such as MNB, SVM, KNN and RF.The prototype of this system demonstrated the successful collection and analysis.

DEDICATION

I dedicate my research work to my parents, family, friends and teachers who encouraged, guided and helped me throughout my studies.

Table of Contents

ABSTRACT 2

DEDICATION 4

1. Introduction 10

1.1 CONTRIBUTION OF THIS THESIS ..... **ERROR! BOOKMARK NOT DEFINED.**
1.2 THESIS OVERVIEW ..... **ERROR! BOOKMARK NOT DEFINED.**

2. Related Work ..... **Error! Bookmark not defined.**

3. CORPUS ..... **Error! Bookmark not defined.**

3.1 NEWS WEBSITES ..... **ERROR! BOOKMARK NOT DEFINED.**
3.2 CATEGORIES ..... **ERROR! BOOKMARK NOT DEFINED.**
3.3 CRAWLER ..... **ERROR! BOOKMARK NOT DEFINED.**
3.4 STORY EXTRACTION ..... **ERROR! BOOKMARK NOT DEFINED.**
3.5 PREPROCESSING ..... **ERROR! BOOKMARK NOT DEFINED.**
3.6 CORPUS STATISTICS ..... **ERROR! BOOKMARK NOT DEFINED.**

4. NEWS REPORTS CATEGORIZATION ..... **Error! Bookmark not defined.**

4.1 APPROACHES ..... **ERROR! BOOKMARK NOT DEFINED.**
4.1.1 Multinomial Naïve Bayes ..... ***Error! Bookmark not defined.***
4.1.2 Support Vector Machines ..... ***Error! Bookmark not defined.***
4.1.3 K Nearest Neighbor ..... ***Error! Bookmark not defined.***
4.1.4 Random Forest ..... ***Error! Bookmark not defined.***
4.2 MORPHOLOGICAL ANALTSIS ..... **ERROR! BOOKMARK NOT DEFINED.**

5. EXPERIMENTAL EVALUATION AND ANALYSIS .. **Error! Bookmark not defined.**

5.1 DATA SET ..... **ERROR! BOOKMARK NOT DEFINED.**
5.2 EVALUATION MEASURES ..... **ERROR! BOOKMARK NOT DEFINED.**
5.2.1 Accuracy: ..... ***Error! Bookmark not defined.***
5.2.2 Confusion Matrix: ..... ***Error! Bookmark not defined.***
5.2.3 Time: ..... ***Error! Bookmark not defined.***
5.2.4 ROC curve: ..... ***Error! Bookmark not defined.***
5.2.5 AUROC: ..... ***Error! Bookmark not defined.***
5.3 WEKA: ..... **ERROR! BOOKMARK NOT DEFINED.**
5.4 SYSTEM SPECIFICATIONS: ..... **ERROR! BOOKMARK NOT DEFINED.**
5.5 EXPERIMENTS AND ANALYSIS: ..... **ERROR! BOOKMARK NOT DEFINED.**
5.5.1 MNB ..... ***Error! Bookmark not defined.***
5.5.2 MNB^{S} ..... ***Error! Bookmark not defined.***
5.5.3 MNB^{T} ..... ***Error! Bookmark not defined.***
5.5.4 MNB^{TS} ..... ***Error! Bookmark not defined.***
5.5.5 Best of MNBs ..... ***Error! Bookmark not defined.***
5.5.6 $SVM^{(RBF)}$ ..... ***Error! Bookmark not defined.***
5.5.7 $SVM^{S(RBF)}$ ..... ***Error! Bookmark not defined.***
5.5.8 $SVM^{T(RBF)}$ ..... ***Error! Bookmark not defined.***
5.5.9 $SVM^{TS(RBF)}$ ..... ***Error! Bookmark not defined.***
5.5.10 $SVM^{(L)}$ ..... ***Error! Bookmark not defined.***

5.5.11 SVM$^{S(L)}$ ..... ***Error! Bookmark not defined.***
5.5.12 SVM$^{T(L)}$ ..... ***Error! Bookmark not defined.***
5.5.13 SVM$^{TS(L)}$ ..... ***Error! Bookmark not defined.***
5.5.14 Best of SVMs ..... ***Error! Bookmark not defined.***
5.5.15 Selection of K for KNN: ..... ***Error! Bookmark not defined.***
5.5.16 KNN$^{(ED)}$ ..... ***Error! Bookmark not defined.***
5.5.17 KNN$^{S(ED)}$ ..... ***Error! Bookmark not defined.***
5.5.18 KNN$^{T(ED)}$ ..... ***Error! Bookmark not defined.***
5.5.19 KNN$^{TS(ED)}$ ..... ***Error! Bookmark not defined.***
5.5.20 KNN$^{(CS)}$ ..... ***Error! Bookmark not defined.***
5.5.21 KNN$^{S(CS)}$ ..... ***Error! Bookmark not defined.***
5.5.22 KNN$^{T(CS)}$ ..... ***Error! Bookmark not defined.***
5.5.23 KNN$^{TS(CS)}$ ..... ***Error! Bookmark not defined.***
5.5.24 Best of KNNs ..... ***Error! Bookmark not defined.***
5.5.25 RF ..... ***Error! Bookmark not defined.***
5.5.26 RF^{S} ..... ***Error! Bookmark not defined.***
5.5.27 RF^{T} ..... ***Error! Bookmark not defined.***
5.5.28 RF^{TS} ..... ***Error! Bookmark not defined.***
5.5.29 Best of RFs ..... ***Error! Bookmark not defined.***
5.5.30 COMPARISON OF BEST OF ALL APPROACHES ..... ***Error! Bookmark not defined.***
6. APPLICATION ..... **Error! Bookmark not defined.**
6.1 INCIDENT ENTRY MODULE: ..... **ERROR! BOOKMARK NOT DEFINED.**
6.1.1 Stats Extraction ..... ***Error! Bookmark not defined.***
6.2 REPORTING MODULE ..... **ERROR! BOOKMARK NOT DEFINED.**
7. CONSLUSION AND FUTURE DIRECTIONS ..... **Error! Bookmark not defined.**
7.1 SUMMARY OF OUR WORK ..... **ERROR! BOOKMARK NOT DEFINED.**
7.2 FUTURE WORK ..... **ERROR! BOOKMARK NOT DEFINED.**
Bibliography ..... **Error! Bookmark not defined.**

List of Tables

Table 3-1: Top 5 English news websites of Pakistan .....**Error! Bookmark not defined.**
Table 3-2: Categories and their abbreviations .....**Error! Bookmark not defined.**
Table 3-3: Split of news stories .....**Error! Bookmark not defined.**
Table 4-1: Approaches and their variants, ^{T} is for TF-IDF and ^{S} is for Stemming.....**Error! Bookmark not defined.**
Table 5-1: Split of news stories .....**Error! Bookmark not defined.**
Table 5-2: Confusion Matrix .....**Error! Bookmark not defined.**
Table 5-3: Evaluation Measures and their notations.....**Error! Bookmark not defined.**
Table 5-4: α of MNB .....**Error! Bookmark not defined.**
Table 5-5: CM of MNB .....**Error! Bookmark not defined.**
Table 5-6: α of MNB^{S} .....**Error! Bookmark not defined.**
Table 5-7: CM of MNB^{S} .....**Error! Bookmark not defined.**
Table 5-8: α of MNB^{T} .....**Error! Bookmark not defined.**
Table 5-9: CM of MNB^{T} .....**Error! Bookmark not defined.**
Table 5-10: α of MNB^{TS} .....**Error! Bookmark not defined.**
Table 5-11: CM of MNB^{TS} .....**Error! Bookmark not defined.**
Table 5-12: α of $SVM^{(RBF)}$ .....**Error! Bookmark not defined.**
Table 5-13: CM of $SVM^{(RBF)}$ .....**Error! Bookmark not defined.**
Table 5-14: α of $SVM^{S(RBF)}$.....**Error! Bookmark not defined.**
Table 5-15: CM of $SVM^{S(RBF)}$ .....**Error! Bookmark not defined.**
Table 5-16: α of $SVM^{T(RBF)}$ .....**Error! Bookmark not defined.**
Table 5-17: CM of $SVM^{T(RBF)}$ .....**Error! Bookmark not defined.**
Table 5-18: α of $SVM^{TS(RBF)}$.....**Error! Bookmark not defined.**
Table 5-19: CM of $SVM^{TS(RBF)}$.....**Error! Bookmark not defined.**
Table 5-20: α of $SVM^{(L)}$ .....**Error! Bookmark not defined.**
Table 5-21: CM of $SVM^{(L)}$ .....**Error! Bookmark not defined.**
Table 5-22: α of $SVM^{S(L)}$ .....**Error! Bookmark not defined.**
Table 5-23: CM of $SVM^{S(L)}$ .....**Error! Bookmark not defined.**
Table 5-24: α of $SVM^{T(L)}$.....**Error! Bookmark not defined.**
Table 5-25: CM of $SVM^{T(L)}$.....**Error! Bookmark not defined.**
Table 5-26: α of $SVM^{TS(L)}$ .....**Error! Bookmark not defined.**
Table 5-27: CM of $SVM^{TS(L)}$ .....**Error! Bookmark not defined.**
Table 5-28: α of $KNN^{(ED)}$.....**Error! Bookmark not defined.**
Table 5-29: CM of $KNN^{(ED)}$ .....**Error! Bookmark not defined.**
Table 5-30: α of $KNN^{S(ED)}$ .....**Error! Bookmark not defined.**
Table 5-31: CM of $KNN^{S(ED)}$.....**Error! Bookmark not defined.**
Table 5-32: α f $KNN^{T(ED)}$ .....**Error! Bookmark not defined.**
Table 5-33: CM of $KNN^{T(ED)}$.....**Error! Bookmark not defined.**
Table 5-34: α of $KNN^{TS(ED)}$ .....**Error! Bookmark not defined.**
Table 5-35: CM of $KNN^{TS(ED)}$ .....**Error! Bookmark not defined.**
Table 5-36: α of $KNN^{(CS)}$.....**Error! Bookmark not defined.**

Table 5-37: CM of $KNN^{(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-38: α of $KNN^{S(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-39: CM of $KNN^{S(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-40: α of $KNN^{T(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-41: CM of $KNN^{T(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-42: α of $KNN^{TS(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-43: CM of $KNN^{TS(CS)}$ ..... **Error! Bookmark not defined.**
Table 5-44: α of RF ..... **Error! Bookmark not defined.**
Table 5-45: CM of RF ..... **Error! Bookmark not defined.**
Table 5-46: α f RF^{S} ..... **Error! Bookmark not defined.**
Table 5-47: CM of RF^{S} ..... **Error! Bookmark not defined.**
Table 5-48: α of RF^{T} ..... **Error! Bookmark not defined.**
Table 5-49: CM of RF^{T} ..... **Error! Bookmark not defined.**
Table 5-50: α of RF^{TS} ..... **Error! Bookmark not defined.**
Table 5-51: CM of RF^{TS} ..... **Error! Bookmark not defined.**
Table 5-52: α of all best ..... **Error! Bookmark not defined.**
Table 5-53: T of all best ..... **Error! Bookmark not defined.**
Table 5-54: Consolidated CM of all best ..... **Error! Bookmark not defined.**
Table 5-55: Area under ROC of all best ..... **Error! Bookmark not defined.**
Table 5-56: Class-wise Area under ROC of all best ..... **Error! Bookmark not defined.**

List of Figures

Fig 1.1: A snippet of a crime and order related news report..... 10
Fig 3.1: Corpus generation process.....**Error! Bookmark not defined.**
Fig 3.2: Snippet of a news web page showing news story and noise .... **Error! Bookmark not defined.**
Fig 3.3: News web page as a DOM tree .....**Error! Bookmark not defined.**
Fig 3.4: Preprocessing steps.....**Error! Bookmark not defined.**
Fig 4.1: News categorization process .....**Error! Bookmark not defined.**
Fig 4.2: Stemming process.....**Error! Bookmark not defined.**
Fig 4.3: TF-IDF of a word or term n in a news report d.....**Error! Bookmark not defined.**
Fig 4.4: An extract from a news report.....**Error! Bookmark not defined.**
Fig 4.5: Common English language stop words .....**Error! Bookmark not defined.**
Fig 5.1: Classifiers and their variants .....**Error! Bookmark not defined.**
Fig 5.2: Overall accuracy of all variants of MNB .....**Error! Bookmark not defined.**
Fig 5.3: Class-wise accuracy of all variants of MNB .....**Error! Bookmark not defined.**
Fig 5.4: Overall accuracy of all variants of SVM.....**Error! Bookmark not defined.**
Fig 5.5: Class-wise accuracy of all variants of SVM.....**Error! Bookmark not defined.**
Fig 5.6: α for different values of K .....**Error! Bookmark not defined.**
Fig 5.7: Overall accuracy of all variants of KNN.....**Error! Bookmark not defined.**
Fig 5.8: Class-wise accuracy of all variants of KNN.....**Error! Bookmark not defined.**
Fig 5.9: Overall accuracy of all variants of RF.....**Error! Bookmark not defined.**
Fig 5.10: Class-wise accuracy of all variants of RF .....**Error! Bookmark not defined.**
Fig 5.11: Class-wise accuracy of all best.....**Error! Bookmark not defined.**
Fig 5.12: Class-wise ROC curves of all best .....**Error! Bookmark not defined.**
Fig 6.1: Incident entry web form .....**Error! Bookmark not defined.**
Fig 6.2: Pie chart showing % of injured persons per province against accident category .....**Error! Bookmark not defined.**
Fig 6.3: Incident Layered Map.....**Error! Bookmark not defined.**
Fig 6.6: Incident Map View.....**Error! Bookmark not defined.**
Fig 6.7: Year comparison report .....**Error! Bookmark not defined.**
Fig 6.8: Dead/Injured summary report.....**Error! Bookmark not defined.**

1. Introduction

News is not only about gathering and reporting facts and figures; it is in fact information that is collected from the society and at the same time it impacts our society in many different ways. For instance, news effect subjects of a society in the ways they perform their duties and make their choices and decisions.

DAWN

At least 70 dead as bomb rips through Lal Shahbaz shrine in Sehwan, Sindh

DON'T JUST VISIT, LIVE IT.

Soldiers standing after the suicide attack at shrine of Lal Shahbaz Qalandar. —APP

At least 70 people were killed and more than 150 injured in a suicide attack on the shrine of Lal Shahbaz Qalandar in Sehwan on Thursday evening.

"So far 70 people have been killed and more than 150 have been wounded," Inspector General Police Sindh A.D. Khawaja said.

"Many of the wounded are in critical condition and they will be shifted to Karachi as soon as navy helicopters and the C-130 plane reach the nearest airport."

Medical Superintendant Dr Moinuddin Siddiqui of Sehwan Taluka Hospital confirmed that 61 bodies were received by the hospital.

"Almost all the bodies were brought here. We have handed over 26 bodies to the heirs of the deceased, while those that remain unidentified are at the hospital," said Siddiqui

Deputy Commissioner Munawar Maheesar confirmed the dead included at least four children and 12 women.

The Assistant Superintendent of Police in Sehwan said a suicide bomber entered the shrine through its Golden gate. The attacker blew himself up after throwing a grenade, which failed to explode, he added.

Read: Military on Sehwan attack: 'Pakistan will respond to terrorist attacks from hostile powers'

The explosion took place in the area where the *dhamaal* (a Sufi ritual) was being performed after evening prayers.

A large number of devotees, from different faiths and from across the country, frequent the shrine on Thursdays — a day of spiritual significance in Pakistan's shrine culture.

Security for shrines was tightened across the province following the attack.

Surge of attacks in Pakistan

Sehwan is located 193km to the north east of Karachi.

Meet the change makers who brought 229,000 students to schools across Pakistan

Quarter of a million children were enrolled in 65 districts under the ILMPOSSIBLE programme, 40% students are girls.

Fig 1.1: A snippet of a crime and order related news report

Among all the news, crime and order news are of great importance for a civilized society, no matter how big or small these are. Avoiding human loss due to crime or disaster is a big challenge for any society and it cannot be tolerated at any level. Timely reporting and analysis for different patterns of such events and then forward out this information to take decisions is very crucial to the concerned field formations so that they can take appropriate measures to avoid such incidents in future. This information must be updated and readily available on daily basis so that ill-fated decisions and measures could be avoided.

Organization and management of vast volumes of electronic text information is a great challenge. Text classification could be used as an essential technique to handle this issue. Text classification is assigning predefined categories to textual data. There are lot of applications of text classification in natural language processing like prediction of user

preferences, news filtering and email filtering and many more. A number of machine learning techniques have been used to classify texts like rule induction, Naïve Bayes (NB), decision tree induction, K-Nearest Neighbours (KNN), Rocchio and Support Vector Machine (SVM).

[1]