CRIME CATEGORIES DETECTION IN URDU NEWS

SUBMITTED BY

HAFIZ MUHAMMAD FAYYAZ (FA20-MSCS-0027)

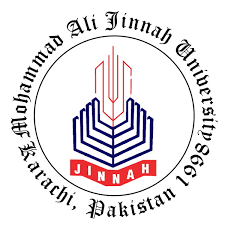
SUPERVISED BY

Dr. TEERATH DAS



THESIS SUBMITTED TO THE FACULTY OF COMPUTING, MOHAMMAD ALI JINNAH UNIVERSITY, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE/ SOFTWARE ENGINEERING

FALL 2021



**THESIS CERTIFICATE**

*This is to certify that the thesis titled, “Crime Categories Detection in Urdu News”, is submitted to the Department of Computer Science, Spring 2021, by Muhammad Fayyaz for the award of the degree of Master of Science in the discipline of Computer Science. The thesis has been carried out under my supervision. I certify that the work submitted is original and not plagiarized from any other source, except as specified in the references. Neither the thesis nor the work contained therein has been previously submitted to any other institution for a degree.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor’s name and signature

**MOHAMMAD ALI JINNAH UNIVERSITY**

# ORIGINAL LITERARY WORK DECLARATION

| **PROGRAM** | MS (Computer Science) | |
| --- | --- | --- |
|  |  | |
| Student Name: Muhammad Fayyaz | | Reg. No: fa20-mscs-0027 |
| Email: [fa20mscs0027@maju.edu.pk](mailto:fa20mscs0002@maju.edu.pk) | | Mobile No: 0311-2174362 |

Research Title:

| \_\_CRIME CATEGORIES DETECTION IN URDU NEWS\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  I do solemnly and sincerely declare that:  1. I am the author of this work;  2. This work is original;  3. Use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work; I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;  4. I hereby assign all and every rights in the copyright to this work to Mohammad Ali Jinnah University (MAJU), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of MAJU having been first had and obtained;  5. I am fully aware that if in the course of making this work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by MAJU. |
| --- |

Student’s Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Place: Karachi-Pakistan.

Acknowledgments

All praises are due to only Almighty Allah.

I would like to thank my supervisor Dr. Teerath Das, for his kind guidance and time given. department of computer science for being so supportive. Finally yet importantly. I would like to thank my family and friends for their support.

Executive Summary

In the last few years, the Internet is the most significant and useful source of getting up any kind of data. With rapid growth in use of online media sites such as urdu news sites, and international Urdu news media has been increased greatly. Collection of the data is one of the major challenges and appealing activities in the discipline of machine learning. The two primary reasons that has made the collection of data such a significant concern; Firstly, as machine learning is being more commonly used, new implementations are being seen that do not typically have ample branded data. Second, deep learning techniques create features automatically, unlike conventional machine learning, which saves practical engineering costs, but can require greater volumes of labelled data in return. In our proposed system, First we have to create a specialized process dataset in huge amounts of the urdu crime news, which will be generated after scrapping and then we have to label that dataset to identify the crime type. The labelled data will be used in machine learning to train the system which gives the best result based on the given input. Integration of Big Data and Artificial Intelligence (Al) and offers numerous resources for new science.

*Keywords*: [Urdu Language, Crime Categorization, Machine Learning Classifiers]

# Table of Contents

[ORIGINAL LITERARY WORK DECLARATION 3](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.gjdgxs)

[CHAPTER 1 7](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.30j0zll)

[INTRODUCTION 7](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.1fob9te)

[CHAPTER 2](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.3znysh7) 8

[LITERATURE REVIEW](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.2et92p0) 8

[CHAPTER 3](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.tyjcwt) 9

[METHODOLOGY](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.3dy6vkm) 9

[References](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.2s8eyo1) 11

[PLAGIARISM VERIFICATION (for Final Submission)](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.17dp8vu) 12

[PLAGIARISM REPORT](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.3rdcrjn) 13

# List of Tables

[Table 1: Introduction ---------------------------------------------------------------------------------](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.26in1rg) 07

[Table 2: Related Work -------------------------------------------------------------------------------](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.26in1rg) 08

[Table 3: Methodology -------------------------------------------------------------------------------](https://docs.google.com/document/d/1DdGgHLmL8Dz5o_pV3VCn5PKNK_Ww4O4hf89rD8Cj8ps/edit#heading=h.26in1rg) 09

# List of Figures

Fig.1 Steps involved in analyzing the crime categories09

Crime Categories Detection in Urdu News

Muhammad Fayyaz

Mohammad Ali Jinnah University

Author Note

Department of Computer Science, 22-E, Block-6, P.E.C.H.S., Lal Kothi Stop, Main Shahrah-e-Faisal, Karachi, 75400, Sindh, Pakistan. Email ID: [fa20mscs0027@maju.edu.pk](mailto:fa20mscs0027@maju.edu.pk)

## 

## 

## CHAPTER 1

## INTRODUCTION

In the last few years, the Internet is the most significant and useful source of getting up any kind of data. With rapid growth in use of online media sites such as urdu news sites, and international Urdu news media has been increased greatly. Collection of the data is one of the major challenges and appealing activities in the discipline of machine learning. The two primary reasons that has made the collection of data such a significant concern; Firstly, as machine learning is being more commonly used, new implementations are being seen that do not typically have ample branded data. Second, deep learning techniques create features automatically, unlike conventional machine learning, which saves practical engineering costs, but can require greater volumes of labelled data in return. In our proposed system, First we have to create a specialized process dataset in huge amounts of the urdu crime news, which will be generated after scrapping and then we have to label that dataset to identify the crime type. The labelled data will be used in machine learning to train the system which gives the best result based on the given input. Integration of Big Data and Artificial Intelligence (Al) and offers numerous resources for new science.

Text mining assumes a critical part in extracting significant data in numeric structure utilizing different strategies, methods, calculations, and so on. Text mining changes the unstructured information over to organized information and gives more valuable data as it will assist us with tracking down wrongdoing classes recognition. Text mining is an extremely immense and un-investigated point and it is generally another region in the vertical of wrongdoing classes identification.

The principle purpose is to focus on the crime categories areas using various text mining techniques and algorithms.

# CHAPTER 2

# LITERATURE REVIEW

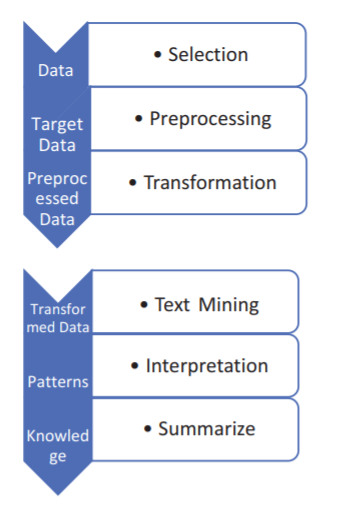
Before, research has been done to dissect opinions on Twitter information. Nonetheless,

the majority of the examination has been done in English, Chinese, Arabic and other languages. Various works have been done on Urdu language using supervised machine learning approach. This part explains the work done in Urdu.

In the related work surveyed, several notable works on text classification of different samples in other languages is discussed here. Young and Jeong [1] applied a new Feature Scaling technique using Naïve Bayes. This method was tested on large collection of news article and readily it became so successful that it outperformed all the other existing ranking schemes such as Information Gain. Wang [2] established an optimal text classification algorithm based on SVM classification technique and found it to be quite efficient enough in text classification of large amount of news samples compared to other two known classifiers namely Decision Tree and KNN algorithms. On the other hand, Kashif [3] choses SVM methodology for text classification of Urdu headlines. The documents under investigation were first normalized, stemmed and stop words were removed from the collection for frequency measurement. The term frequency of each word in the vocabulary was measured along with the inverse document frequency of the chosen corpus. Wahab [4] discussed an extensive assessment of different state-of-the-art machine learning methods for text categorization. Ali and Ijaz [5] presents a comparative study of Urdu text classification using Naive Bayes and SVM classifiers on 26,067 documents. In their experimental finding, they observed SVM to produce higher accuracy than Naive Bayes classifier. The collection includes six different categories namely Finance, News, Culture, Consumer Information, Sports and Personal Communication. Additionally, they found that stemming is not very effective but stop words removal improves the overall classifier accuracy.

# CHAPTER 3

# METHODOLOGY



**Fig.1 Steps involved in analyzing the crime categories**

**A. COLLECTION OF DATA**

This is the initial step of text mining. It includes collection of data (in the form of text). The data set of urdu crime news between 2021-2016 is used. The data set is scraped from the news site. Now the information (data set) is ready to use for Machine Learning.

**B. DATA PRE-PROCESSING**

This step includes removal of redundancies, errors, spelling mistakes and the data is processed under the following steps:

**a).Tokenization:** In the textual data the sentences are broken into small pieces such as words, symbols, keywords and others called tokens. The sentences become free from commas, spaces, etc.

**b) Removal of stop word:** Under this step the sentences become free from unwanted words It also involves removal of HTML, Xml tags.

**c) Stemming:** Stemming can be described as a technique mainly utilized for finding the stem of a particular word. By implementing it mainly for the grammatical reasons, the textual data utilizes words like organize, organizes, and organizing. After stemming only organize word is left and the other two words are neglected. The principle focus is to remove modulation of the words.

**C. TEXT TRANSFORMATION**

This is the transformation of textual data into two main notations named as vector space model and bag of words.

**D. FEATURE SELECTION (ATTRIBUTE SELECTION)**

In feature selection, the relevant features are selected and the irrelevant features are removed using the model notations. This process is also known as variable selection. The features selected are further used in model creation. In this process the database required less space and less computations.

**E. DATA MINING/ PATTERN SELECTION**

In this process the text mining process merges with the data mining. Here the classical data mining methods and techniques are used on the data obtained by previous processes.

a) Classification with k-Nearest Neighbor

b) K-Means Clustering

# References

1. E. Youn and M. K. Jeong, “Class dependent feature scaling method using naive Bayes classifier for text datamining,” Pattern Recognit. Lett., vol. 30, no. 5, pp. 477–485, 2009.
2. Z.-Q. Wang, X. Sun, D.-X. Zhang, and X. Li, “An optimal SVM-based text classification algorithm,” in Machine Learning and Cybernetics, 2006 International Conference on, 2006, pp. 1378–1381.
3. K. Ahmed, M. Ali, S. Khalid, and M. Kamran, “Framework for Urdu News Headline Classification,” 2016.
4. W. Khan, A. Daud, J. A. Nasir, and T. Amjad, “A survey on the state-of-the-art machine learning models in the context of NLP,” Kuwait J. Sci., vol. 43, no. 4, pp. 95–113, 2016.
5. A. R. Ali and M. Ijaz, “Urdu text classification,” in Proceedings of the 7th international conference on frontiers of information technology, 2009, p. 21.
6. T. Zia, M. P. Akhter, Q. Abbas, C. Study, F. S. Approaches, and U. T. Categorization, “Comparative study of feature selection approaches for Urdu text categorization,” Malaysian J. Comput. Sci., vol. 28, no. 2, pp. 93–109, 2015.
7. Conference: 17th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications (SERA 2019)
8. Conference: 2019 International Conference on contemporary Computing and Informatics (IC3I)
9. Urdu sentiment analysis using supervised machine learning approach." International Journal of Pattern Recognition and Artificial Intelligence 32.02 (2018): 1851001.
10. Amjad, Kamran, et al. "Exploring Twitter news biases using urdu-based sentiment lexicon." 2017 International Conference on Open Source Systems & Technologies (ICOSST). IEEE, 2017.

# PLAGIARISM VERIFICATION (for Final Submission)

(Note: *This form is filled at the final submission of Thesis/Research Project Reports for graduation. Library MUST submit his/her/their work in turnitin repository*)

Title of Research Project: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Total Pages \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Registration ID \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Supervisor:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Program: Faculty: Faculty of Computing

This is to report that the above document was scanned for similarity detection and has been submitted in turnitin repository. Process and outcome is given below:

Software used **turnitin** Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Similarity Index\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Total word count\_\_\_\_\_\_\_\_\_\_\_\_\_\_

File Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Digital Receipt No \_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Checked by Library Staff Head Librarian

(Name & Signature) (Signature & Stamp)

*\*Please attach a plagiarism report mentioning similarity index duly signed and stamped by the library at the end of the thesis/ Research Project.*

# PLAGIARISM REPORT

Attach the title page and summary of the plagiarism report provided by the coordinator (checked by the library). Plagiarism report from any source other than the library is not acceptable.