****

**Project Report**

**On**

**Author Identification in Online Hindi Text**

*Submitted in partial fulfilment of the requirements*

*for the award of the degree of*

**BACHELOR OF TECHONOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

***Submitted by***

**Kishan Kumar Mishra**

**Enroll. No. CSE/517/14**

**Mayank Kumar Gupta**

**Enroll. No. IT/515/14**

***Under the supervision of***

**Mr. Muhammad Ahsan Chishti**

**Ms. Azra Nazir**

**Department of Computer Science**

**National Institute of Technology, Srinagar**

**J&k June 2018**



**CERTIFICATE**

****

This is to certify that the project titled **AUTHOR IDENTIFICATION IN ONLINE HINDI TEXT** has been completed by **Kishan Kumar (CSE/517/14) and Mayank Kumar (CSE/515/14)** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science**.



**STUDENT’S DECLARATION**

We, hereby declare that the work, which is being presented in the project entitled **AUTHOR IDENTIFICATION IN ONLINE HINDI TEXT** in partial fulfilment of the requirements for the award of thedegree of **Bachelor of Technology** in **Computer Science** in the session 2018, is an authentic record of our own work carried out under the supervision of **Mr. Muhammad Ahsan Chisti** and **Ms. Azra Nazir**, Department of Computer Science, National Institute of Technology,Srinagar.

Dated: 7th June 2018

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (i) Name |  |  | Kishan Kumar |  | |
| Signature | \_\_\_\_\_\_\_\_\_\_\_\_\_ | | | | |
| (ii) Name |  | Mayank Kumar | | |  |
| Signature | \_\_\_\_\_\_\_\_\_\_\_\_\_ | | | |  |

**ACKNOWLEDGEMENT**

The work would not have been possible to come to present shape without the able guidance, support and encouragement of a number of people.  We would like to show our gratitude to our guide Muhammad Ahsan Chisti and Co-guide Azra Nazir with whom support we were able to complete our final year project.

**ABSTRACT**

Authorship Identification is the task of identifying the author of a given document. From machine learning perspective it is a single label multi class classification problem. It has important applications in cybersecurity where it can be used to identify authors of documents. Online text can be data from blogs, emails, discussion, forums, etc. The writing style of any author is unique. As fingerprints can identify an individual, the writing style too work like the same and identify the author of unknown document. Extracting features from online text is difficult since online text is very small as compared to literary work, books. Author identification is done by first extracting the features which characterize the writing style of the author and then feeding these features to some learning machine algorithm. Features used in author identification are called Stylometric features and comprised of Lexical features, Syntactical features, Structural features and content specific features. Model is generated by training the machine learning algorithm with dataset of documents and then this model is used to identify the author of unknown text document. Author Identification is language dependent. Our work is related to author identification pertaining to documents written in Hindi. To our knowledge no work has been done in this field when it comes to Hindi Language. So in this project, we will do author identification of Online text like blogs written in Hindi language.

**TABLE OF CONTENTS**

Students Declaration……………………………………………………………….........ii

Certificate……………………………………………………………………….………iii

Acknowledgement……………………………………………………………….……...iv

Abstract……………………………………………………………………….…………v

**Chapter 1: Introduction**

* 1. Context……………………………………………………………....….01
  2. Characteristics of Online Text…………………………….……………01
  3. Motivation……………..…………………………………………….….02

**Chapter 2: Literature Survey**

**Chapter 3: Design**  06

3.1 General Procedure………………………………..…………………….…06

3.2 Features………………………..…………………………………………..07

3.2.1 Lexical…………………………………………………………....07

3.2.2 Syntactical………………………………………………………..08

3.2.3 Structural…………………………………………………………09

3.2.4 Content-Specific………………………………………………….09

3.3 Machine Learning Algorithms……………………………………………...09

3.3.1 Logistic Regression……………………………………………….09

3.3.2 Naïve Bayes……………………………………………………….10

3.3.3 SVM……………………………………………………………….11

3.3.4 Random Forest Classifier………………………………………….11

3.3.5 K Nearest Neighbors……………………………………………....12

**Chapter 4: Implementation**

4.1 Data………………………………………..………………………………...13

4.2 Formation of Feature vector……………………..…………………………..13

4.2.1 Bag of words Model………………………………………………..….13

4.2.2 Character n grams……………………………………………………...14

4.2.3 Function Words……………………...…………………………………14

4.2.4 Part of Speech Tagging………………………………………………….16

4.2.5 Secondary Features……………………………………………………...16

4.2.6 Combination of Features………………………………………………...17

4.3 Text Normalisation..............................................................................................17

4.3.1 TF Normalisation………………………………………………………...17

4.3.2 IDF Normalisation………………………………………………………..17

4.4 Feature Normalisation…………………………………………………………..18

4.5 Model Training and Evaluation…………………………………………………18

4.5.1 Choice of Evaluation Metric……………………………………………...18

4.5.2 Approach………………………………………………………………….18

4.5.3 Cross Validation Strategy…………………………………………………19

4.5.4 Learning Models used……………………………………………………..19

**Chapter 5:Results**

**Chapter 6: Conclusions and Future Work**

6.1 Conclusions………………………………..…………..…………………….…14

4.2 Future Work………………………..…………………………………………..15

**References…………………………………………………………………………………...24**

# LIST OF TABLES

Table 1:Finding optimum threshold value for function words……………………………….15

**LIST OF FIGURES**

Figure 1: General Procedure for Author Identification………………………………………7