Project Synopsis

on

DigiErratum

Submitted as a part of course curriculum for

# Bachelor of Technology

in

# Computer Science



### Submitted by

Govind Singh(2000290120066)

Aman Raj Singh(2100290129003)

Anshul Sharma(2000290120031)

### Under the Supervision of

Prof. Shivani

Assistant Professor

CS Department

# KIET Group of Institutions, Ghaziabad

# Department of Computer Science

**Dr. A. P. J. Abdul Kalam Technical University**

**2021-2022**

# DECLARATION

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person or material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature of Students Name:

Roll No.:

Date:

# CERTIFICATE

This is to certify that Project Report entitled “**DigiErratum**” which is submitted by **Aman Raj Singh, Govind Singh and Anshul Sharma** in partial fulfilment of the requirement for the award of degree B.Tech in Department of Computer Science of Dr.

A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

**Date: Supervisor Signature**

Prof. Shivani Assistant Professor

CS Department

# ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the synopsis of the B.Tech Mini Project undertaken during B.Tech. Third Year. We owe a special debt of gratitude **to Prof. Shivani**, Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, for his/her constant support and guidance throughout the course of our work. His/Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavours have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Dr. P. K Singh, Head of the Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the facultymembersofthedepartmentfortheirkindassistanceandcooperationduringthedevelopmentof our project.

Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Date:

Name:

Roll No:

# ABSTRACT

There are so many sites are available on internet like Grammarly, Trinka AI, Ginger, Typely etc. which provide grammatical error detection ,grammar checking, writing error challenges related to the users. Grammar is essence of language! With people adopting newer slangs to communicate and get closer to the rest of the world, the correct usage Grammar is being affected but our model not only provide grammatical correction to the incorrect sentence speak by the user but it will also perform conversation with the user as a Virtual Assistant, It has flagged way for a new knowledge whenever we can ask queries to machine and can interrelate with IVAs(Intelligent Virtual Assistant) as persons to persons. This new technology involved almost entire world in numerous ways like keen mobiles, laptops, PCs etc. Our model will perform this task by using AI(Artificial intelligence), ML(Machine learning) and NLP(Natural Language Processing) where the model and user communicate with each other but between this communication there are several operations are being performed like Grammatical error detection, Grammar checking and Grammar formalism . Our model helps those people who can not write, who are the initial learner and wants to improve English speaking.

# TABLEOFCONTENTS

|  |  |
| --- | --- |
| CHAPTER1 INTRODUCTION |  |
| 1.1. Introduction……………………................................................... | 1 |
| 1.2 ProblemStatement.……………………........................................ | 1 |
| 1.2. Objective…………………………………………………………. | 2 |
| 1.3. Scope……………………………………………………………… | 2 |
| CHAPTER2 LITERATUREREVIEW……………………………………………...... | 3-8 |
| CHAPTER3 PROPOSEDMETHODOLOGY………………………………….......... | 9-12 |
| 3.1 Flowchart………………………………………………………….. | 13 |
| CHAPTER5 DIAGRAMS…........................................................................................... | 14 |
| CHAPTER6 CONCLUSION…....................................................................................... | 15 |
| REFERENCES….............................................................................................................. | 16 |
|  |  |

**Chapter1.Introduction**

## Introduction to Project

DigiErratum stands for digital erratum (Latin word means error correction) is a software that checks your fluent for grammatical mistakes, appropriate punctuation, misspellings, and issues related to sentence structure as well as it behave as a assistant for the user. It uses different technologies like AI, ML and NPL to do the same.

English as a universal language in the world will get more and more attention, but English is not our mother tongue, and there exist differences in culture and thinking. English grammar is the most difficult problem to solve. There are many English learners, and the number of English teachers is limited, and it is inevitable to use Internet technology to solve the problem of lack of resources so DigiErratum provides us the solution to this by correcting the incorrect sentence speak by the user and the assistant normally ask the user that "Do you want to say this.." as user say yes the assistant continue the conversation.

With the increasing update of computers and the Internet, tens of thousands of users tend to write and communicate in English in their daily work. For users whose native language is not English, writing in English is a major obstacle for them. Grammar checking technology originated from the application of natural language understanding. The standard of grammar directly affects the fluency of sentences. The grammar checking system introduced in this article can efficiently check out grammatical errors in sentences and automatically generate correct sentences to replace the wrong one.

# 1.2 Problem Statement

During these days most of the people wants to learn and speak English but they are self-effacing when they try to communicate, so using DigiErratum they can easily practice English speaking without any hesitation and improve their English speaking and feel confident while talking to someone.

Sometime we need to write a big paragraph on some grammatical error correction site to check weather our paragraph is correct or not. It takes so much time but using DigiErratum we can do the same, by only speaking the whole paragraph. It is for any kind of age people and who are physically challenged person they can communicate with the assistant easily and improve their English speaking. It is basically works as virtual assistant so there will be no need of writing. It will save our time.

1

# Objective

With the help of this Project following objectives can be achieved:

* A user friendly assistant system.
* A model that will help every age group people to improve/learn English.
* Learning via communicating with the device.
* Self-effacing person can practice on DigiErratum and improve their English as well as enhance his/her confidence and become a good public speaker.

# 1.4 Scope

The overall purpose of DigiErratum is to Improve English language free of cost so that everyone can access it easily and it is very user-friendly model where user can select his/her topic of interest to communicate with the assistant which makes it an interactive platform where user enjoy learning. By providing all these features DigiErratum is a futuristic model and in future it can be one of the best platform to learn English.

2

**Chapter2: Literature Review**

### (1)

**Analysis of community question-answering issues via machine learning and deep learning : State-of-the-art review**

### AUTHORS– Pradeep Kumar Roy, Sunil Saumya, Jyoti Prakash Singh, Snehasish Banerjee.

Community question-answering sites (CQAs) continue to serve as useful avenues for Internet users to exchange knowledge. Over the last couple of decades, community question-answering sites (CQAs) have been a topic of much academic interest. Scholars have often leveraged traditional machine learning (ML) and deep learning (DL) to explore the ever-growing volume of content that CQAs engender. To clarify the current state of the CQA literature that has used ML and DL, this paper reports a systematic literature review. The goal is to summarise and synthesise the major themes of CQA research related to (i) questions, (ii) answers and (iii) users. The final review included 133 articles. Dominant research themes include question quality, answer quality, and expert identification. In terms of dataset, some of the most widely studied platforms include Yahoo! Answers, Stack Exchange and Stack Overflow. The scope of most articles was confined to just one platform with few cross-platform investigations. Articles with ML out number those with DL.

3

### (2)

**Machine learning for the educational sciences**

### AUTHORS-Sven Hilbert, Stefan Coors, Elisabeth Kraus

Machine learning (ML) provides a powerful framework for the analysis of high-dimensional datasets by modelling complex relationships, often encountered in modern data with many variables, cases and potentially non-linear effects. The impact of ML methods on research and practical applications in the educational sciences is still limited, but continuously grows, as larger and more complex datasets become available through massive open online courses (MOOCs) and large-scale investigations. The educational sciences are at a crucial pivot point, because of the anticipated impact ML methods hold for the field. To provide educational researchers with an elaborate introduction to the topic, it provide an instructional summary of the opportunities and challenges of ML forth educational sciences, show how a look at related disciplines can help learning from their experiences, and argue for a philosophical shift in model evaluation. We demonstrate how the overall quality of data analysis in educational research can benefit from these methods and show how ML can play a decisive role in the validation of empirical models. Specifically, it(1)provide an overview of the types of data suitable for ML and(2) give practical advice for the application of ML methods.

4

### (3)

**A Comprehensive Survey of Grammar Error Correction**

### AUTHORS - Emma Strubell , Ananya Ganesh and Andrew McCallum

Grammar error correction (GEC) is an important application aspect of natural language processing techniques. The past decade has witnessed significant progress achieved in GEC for the sake of increasing popularity of machine learning and deep learning, especially in late 2010s when near human level GEC systems are available. However, there is no prior work focusing on the whole recapitulation of the progress. We present the first survey in GEC for a comprehensive retrospect of the literature in this area. We first give the introduction of five public datasets, data annotation schema, two important shared tasks and four standard evaluation metrics. More importantly, we discuss four kinds of basic approaches, including statistical machine translation based approach, neural machine translation based approach, classification based approach and language model based approach, six commonly applied performance boosting techniques for GEC systems and two data augmentation methods. Since GEC is typically viewed as a sister task of machine translation, many GEC systems are based on neural machine translation (NMT) approaches, where the neural sequence-to-sequence model is applied. Similarly, some performance boosting techniques are adapted from machine translation and are successfully combined with GEC systems for enhancement on the final performance. Furthermore, we conduct an analysis in level of basic approaches, performance boosting techniques and integrated GEC systems based on their experiment results respectively for more clear patterns and conclusions. Finally, we discuss five prospective directions for future GEC researches. Be sure that you adhere to these limits; otherwise, you will need to edit your abstract accordingly. The abstract must be written as one paragraph, and should not contain displayed mathematical equations or tabular material. The abstract should include three or four different keywords or phrases, as this will help readers to find it. It is important to avoid over-repetition of such phrases as this can result in a page being rejected by search engines. Ensure that your abstract reads well and is grammatically correct. We present the first survey in grammar error correction (GEC) for a comprehensive retrospect of existing progress. We first give definition of the task and introduction of public datasets, annotation schema, and two important shared tasks. Then, four dominant basic approaches and their development are explained in detail. After that, we classify the numerous performance boosting techniques into six branches and describe their application and development in GEC, and two data augmentation methods are separately introduced due to their importance. More importantly, in Section 6, after the introduction of the standard evaluation metrics, we give in-depth analysis based on empirical results in aspects of approaches, techniques and integrated GEC systems for a more clear pattern of existing works. Finally, we present five prospective directions based on existing progress in GEC. We hope our effort could provide assistance for future researches in the community.

5

### (4)

**Intelligent Virtual Assistant knows your life**

### AUTHORS- Yu Wang, Yuelin Wang, JieLiuɨ , and Zhuo Liu

In the IoT world, intelligent virtual assistant (IVA) is a popular service to interact with users based on voice command. For optimal performance and efficient data management, famous IVAs like Amazon Alexa and Google Assistant usually operate based on the cloud computing architecture. In this process, a large amount of behavioral traces that include user’s voice activity history with detailed descriptions can be stored in the remote servers within an IVA ecosystem. If those data (as also known as IVA cloud-native data) are leaked by attacks, malicious person may be able to not only harvest detailed usage history of IVA services, but also reveals additional user-related information through various data analysis techniques. In this paper, we firstly show and categorize types of IVA-related data that can be collected from popular IVA, Amazon Alexa. We then analyze an experimental dataset covering three months with Alexa service, and characterize the properties of user’s lifestyle and life patterns. Our results show that it is possible to uncover new insights on personal information such as user interests, IVA usage patterns and sleeping/wake-up patterns. The results presented in this paper provide important implications for and privacy threats to IVA vendors and users as well.

In recent days, cloud-based IoT devices are evolving rapidly and spreading widely in our lives. Many people are becoming accustomed to interacting with various IoT consumer products, such as intelligent virtual assistants. In these circumstances, lots of data are being produced in real time in response to user behaviors. Interestingly, a large amount of behavioral traces that include user’s voice activity history with detailed descriptions can be stored in the emote servers. Until now, there has been little research reported on analysis of intelligent virtual assistant related data collected from cloud servers. In this paper, we showed and categorized types of IVA-related data that can be collected from a popular IVA, Amazon Alexa. We then analyzed an experimental dataset from Amazon Alexa, and characterized several properties of a user’s lifestyle and life patterns. Our results showed that it is possible to uncover new insights on personal information such as IVA usage patterns, user’s interests and sleeping/wake-up patterns. The results presented in this work provide important implications for security and privacy threats to IVA vendors and users as well.

6

### (5)

**Virtual Assistant for Enhancing English Speaking Skills**

### AUTHORS- Ayushi Desai, Yash Gandhi, Jaynil Gaglani, Nikahat Mulla

Over the years with the advent of social media and messaging apps, people have been using jargon, abbreviated words, and casual language while chatting with other people. This leads to a lack of conversational skills during interviews, job meetings, or even daily conversations. Poorly spoken English has been a prime factor due to which students are unsuccessful in clearing the interviews for a job. There are many studies that indicate that an overwhelming percentage of engineers in the country cannot speak English fluently which is required for high end consulting jobs. Present-day institutions provide solutions for improving English speaking but are expensive. Hence, there is a need for an instantly available conversing partner to hone communication skills. We propose a virtual assistant that can communicate with the user in an attempt to improve English speaking skills. The system consists of SynQG model for question generation, RoBERTa Grammar Error Correction model and praat-parselmouth for speech analysis. The user practices English speaking by answering the questions generated by the system. A thorough speech analysis report is provided to the user based on these answers highlighting mistakes as well as strengths in areas like grammar and pronunciation.

7

### (6)

**Personality Research and Assessment in the Era of Machine Learning**

### AUTHORS - Clemens Stachl, Florian Pargent, Sven Hilbert, Markus Buhner

The increasing availability of high-dimensional, fine-grained data about human behaviour, gathered from mobiles ensing studies and in the form of digital footprints, is poised to drastically alter the way personality psychologists perform research and undertake personality assessment. These new kinds and quantities of data raise important question sabouthowtoanalysethedataandinterprettheresultsappropriately.Machinelearningmodelsarewellsuitedtothesekindsofdata,allowingresearcherstomodelhighlycomplexrelationshipsandtoevaluatethegeneralizabilityandrobustnessoftheir results using resampling methods. The correct usage of machine learning models requires specialized methodological training that considers issues specific to this type of modelling. Here, we first provide a brief overview of past studies using machine learning in personality psychology. Second, we illustrate the main challenges that researchers face when building, interpreting, and validating machine learning models. Third, we discuss the evaluation of personality scales, derived using machine learning methods. Fourth, we highlight some key issues that arise from the use of latent variables in them modelling process. We conclude with an outlook on the future role of machine learning models in personality research and assessment.

8

# CHAPTER3: PROPOSED METHODOLOGY

Over the years with the advent of social media and messaging apps, people have been using jargon, abbreviated words, and casual language while chatting with other people. This leads to a lack of conversational skills during interviews, job meetings, or even daily conversations. Poorly spoken English has been a prime factor due to which students are unsuccessful in clearing the interviews for a job. There are many studies that indicate that an overwhelming percentage of engineers in the country cannot speak English fluently which is required for high end consulting jobs [1].

Grammar error correction (GEC) is an important application aspect of natural language processing techniques. The past decade has witnessed significant progress achieved in GEC for the sake of increasing popularity of machine learning and deep learning, especially in late 2010s when near human level GEC systems are available. . Since GEC is typically viewed as a sister task of machine translation, many GEC systems are based on neural machine translation (NMT) approaches, where the neural sequence-to-sequence model is applied. Ensure that your abstract reads well and is grammatically correct. We present the first survey in grammar error correction (GEC) for a comprehensive retrospect of existing progress[2]. Machine learning models are well suited to these kinds of data, allowing researchers to model highly complex relationships and to evaluate the generalizability and robustness of their results using resampling methods. The correct usage of machine learning models requires specialized methodological training that considers issues specific to this type of modelling.[3].

In the IoT world, intelligent virtual assistant (IVA) is a popular service to interact with users based on voice command. For optimal performance and efficient data management, famous IVAs like Amazon Alexa and Google Assistant usually operate based on the cloud computing architecture. In this process, a large amount of behavioral traces that include user’s voice activity history with detailed descriptions can be stored in the remote servers within an IVA ecosystem. If those data (as also known as IVA cloud-native data) are leaked by attacks, malicious person may be able to not only harvest detailed usage history of IVA services, but also reveals additional user-related information through various data analysis techniques. In this paper, we firstly show and categorize types of IVA-related data that can be collected from popular IVA, Amazon Alexa. We then analyze an experimental dataset covering three months with Alexa service, and characterize the properties of user’s lifestyle and life patterns.

9

Our results show that it is possible to uncover new insights on personal information such as user interests, IVA usage patterns and sleeping/wake-up patterns. The results presented in this paper provide important implications for and privacy threats to IVA vendors and users as well[4].

Machine learning (ML) provides a powerful framework for the analysis of high-dimensional datasets by modelling complex relationships, often encountered in modern data with many variables, cases and potentially non-linear effects. The impact of ML methods on research and practical applications in the educational sciences is still limited, but continuously grows, as larger and more complex datasets become available through massive open online courses (MOOCs) and large-scale investigations. The educational sciences are at a crucial pivot point, because of the anticipated impact ML methods hold for the field. To provide educational researchers with an elaborate introduction to the topic, it provide an instructional summary of the opportunities and challenges of ML for the educational sciences, show how a look at related disciplines can help learning from their experiences, and argue for a philosophical shift in model evaluation. We demonstrate how the overall quality of data analysis in educational research can benefit from these methods and show how ML can play a decisive role in the validation of empirical models. Specifically, it(1) provide an overview of the types of data suitable for ML and(2)give practical advice for the application of ML methods[5].

10

**3.1 Flowchart**

**Diagram

Description automatically generated**

### Fig:1 Virtual Assistant interaction with user

**A picture containing diagram

Description automatically generated**

**Fig:2 WorkingModel**

**11**

Providing voice command as input using English language

DigiErratum speech recognizer will response after receiving the voice command

Converting the voice command into text and comparing it with the text inside smart assistant

The system allows user with a voice notification in English to insure right command

Valid Text

Error Correction will correct the incorrect text

Executed the command by Smart Assistant

Performing the corresponding tasks associated with the command

**Fig:3 A Flowchart of Voice command recognizer**

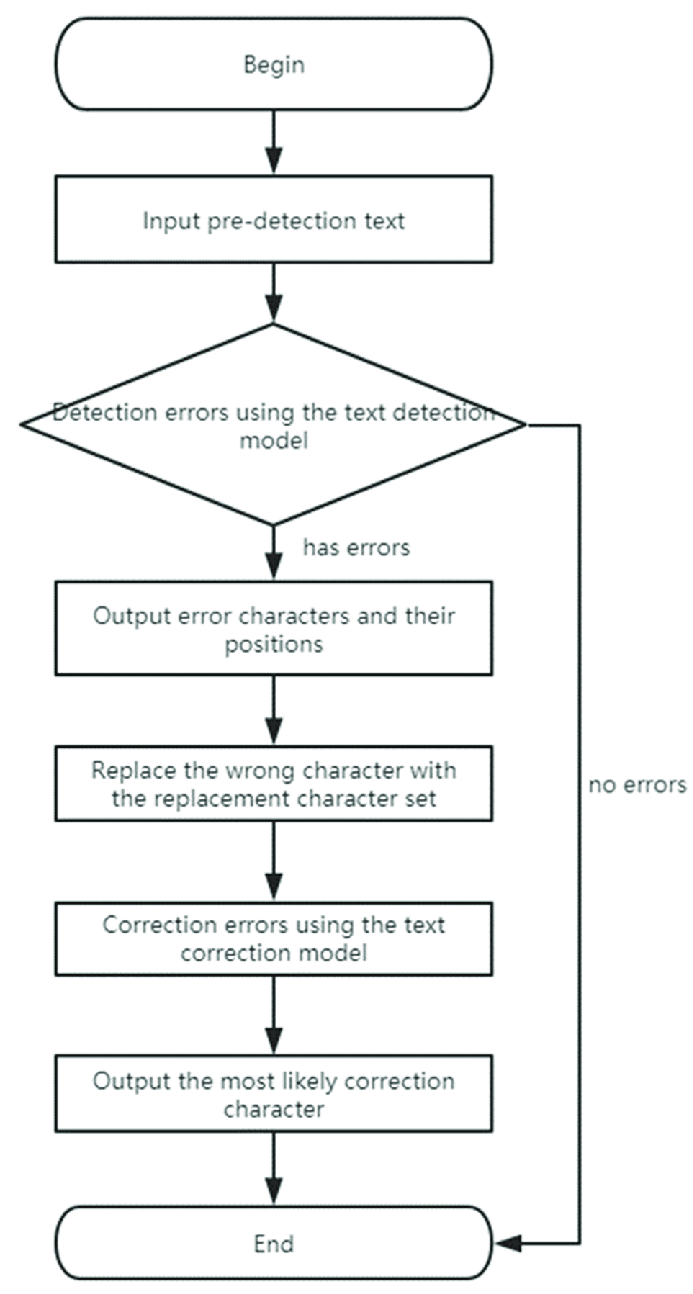
**12**

**Diagram

Description automatically generated**

### Fig:4 Working Model of Grammar Correction

13

****

**Fig:4 Working model of Error Detection**

**14**

# CHAPTER 6 CONCLUSION

We have shown that it is possible to create a model that can communicate and also correct the incorrect sentences with the help of its various features we have go through various technologies to make this model like AI (Artificial Intelligence) , ML (Machine learning) and NLP (Natural language processing) and we have obtained encouraging results.

One of the most important learnings from this project is that **ASSISTANT IS KING** without assistant the conversation with a machine is not possible and we would not have been able to build a robust enough model.

From a personal perspective, we have tremendously enjoyed this project, the hardest so far,as it enabled me to gain more practical experience of Machine learning, AI and NLP and more other concept as well.

15

**REFERENCES**

1. O. Omelianchuk, "GECToR - Grammatical Error Correction: Tag Not Rewrite", Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 163-170, 2020.
2. D. Naber, “A rule-based style and grammar checker,” 01 2003. [2] M. Gamon, C. Leacock, C. Brockett, W . Dolan, J . Gao, D. Belenko, and A. Klementiev, “Using statistical techniques and web search to correct esl errors,” CALICO J ournal, vol. 26, pp. 491–511, 01 2013. [3] D. Dahlmeier, H. T. Ng, and S. M. W u, “Building a large annotated corpus of learner English: The NUS corpus of learner English,” in Proceedings of the Eighth W orkshop on Innovative Use of NLP for Building Educational Applications. Atlanta, Georgia: Association for Computational Linguistics, J un. 2013, pp. 22–31.
3. Aggarwal, C. C. (2016). Recommender systems. [electronic book]: The textbook. Cham: Springer.
4. Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models.
5. IDC. (2015). Explosive Internet of Things Spending to Reach $1.7 Trillion in 2020 According to IDC. http://www.businesswire.com/news/home/20150602005329/en/ Gatner. Gartner Says Worldwide Spending on VPA-Enabled Wireless Speakers Will Top $2 Billion by 2020. https://www. gartner.com/newsroom/id/3464317
6. Y. Sun, J. Zhang and Y. Xiong, “Data Security and Privacy in Cloud Computing”, International Journal of Distributed Sensor Networks, Jan 1, 2014
7. Rhonda Ascierto. 2018. Uptime Institute Global Data Center Survey. Technical report, Uptime Institute. DzmitryBahdanau, Kyunghyun Cho, and YoshuaBengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference for Learning Representations (ICLR), San Diego, California, USA. James Bergstra and YoshuaBengio. 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb):281–305. James S Bergstra, R´emiBardenet, YoshuaBengio, and Bal´azsK´egl. 2011. Algorithms for hyper-parameter optimization. In Advances in neural information processing systems, pages 2546–2554.

16