 **SIMATS SCHOOL OF ENGINEERING**

**SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL**

# SCIENCES CHENNAI-602105

**Automatic Question Tagging is a critical application of Natural Language Processing**

**A CAPSTONE PROJECT REPORT**

*Submitted in the partial fulfillment for the award of the degree of*

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

**Submitted by**

**N. RISHITHA (192210240)**

**Under the Supervision of**

**Dr. Kanimozhi**

**January 2025**

# DECLARATION

We,  **N. Rishitha,** students of **Bachelor of Engineering in Computer Science**, Department of Computer Science and Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, hereby declare that the work presented in this Capstone Project Work entitled **Automatic Question Tagging is a critical application of Natural Language Processing**

is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics.

## N. RISHITHA (192210240)

Date: 5/1/2025

Place: Chennai

# CERTIFICATE

This is to certify that the project entitled **“Automatic Question Tagging is a critical application of Natural Language Processing”** submitted by

**N. Rishitha,** has been carried out under my supervision. The project has been submitted as per the requirements in the current semester of B. Tech Information Technology.

Teacher-in-charge

Dr. Kanimozhi

# Table of Contents

|  |  |
| --- | --- |
| **S.NO** | **TOPICS** |
| 1 | **Abstract** |
| 2 | **Introduction** |
| 3 | **Project Description** |
| 4 | **Problem Description** |
| 5 | **Architecture / UML Diagrams** |
| 6 | **Approach / Algorithm Description** |
| 7 | **Implementation** Coding |
| 8 | **Output**  Output with Screenshots |
| 9 | **Conclusion** |
| 10 | Future Enhancement  **References** |

## ABSTRACT

## Automatic Question Tagging is a critical application of Natural Language Processing (NLP) that automates the process of assigning relevant tags or categories to questions. This project focuses on developing an efficient tagging system that utilizes advanced text classification techniques to improve the organization and accessibility of content on educational platforms, forums, and customer support systems.

## The system involves comprehensive data preprocessing to clean and prepare text, followed by feature engineering to extract meaningful representations of questions using techniques like Bag-of-Words, TF-IDF, and embeddings. A labeled and annotated dataset is used to train and evaluate supervised machine learning models, ensuring high accuracy and relevance in tagging.

## This project provides a hands-on learning experience in implementing NLP workflows, from data preprocessing to model deployment. It equips students with essential skills for solving real-world text classification problems and highlights the practical impact of NLP in automating repetitive tasks, enhancing user experience, and improving information retrieval systems.

## INTRODUCTION

In the digital era, questions serve as the foundation of knowledge exchange across educational platforms, online forums, and customer support systems. The exponential growth of user-generated content has created a pressing need for efficient question categorization and management. Manual tagging of questions is time-consuming, error-prone, and inefficient, especially at scale. This calls for an automated solution capable of streamlining the process while maintaining accuracy and relevance.

The Automatic Question Tagging System leverages advancements in Natural Language Processing (NLP) to address this challenge. By employing sophisticated text classification techniques, the system can automatically assign relevant tags or categories to questions based on their content. This improves content organization, enhances search functionality, and provides a seamless user experience.

Key features of this system include:

1. **Data Preprocessing:** Cleaning and normalizing text to prepare it for classification.
2. **Feature Engineering:** Extracting meaningful representations using methods like Bag-of-Words, TF-IDF, or embeddings.
3. **Supervised Learning:** Training classification models using labeled datasets to ensure high accuracy.
4. **Scalability:** Adapting to growing datasets and varied domains, making it suitable for diverse applications.

This paper discusses the design, implementation, and impact of the Automatic Question Tagging System, emphasizing its role in automating and optimizing content categorization. It highlights the potential of NLP-driven solutions in enhancing knowledge management across interconnected platforms.

## Project Description

The **Automatic Question Tagging System** is an innovative solution that automates the process of categorizing and tagging questions, enhancing the organization and accessibility of content on educational platforms, forums, and customer support systems. Leveraging cutting-edge Natural Language Processing (NLP) techniques, this project addresses the challenges of manual tagging by providing a scalable, accurate, and efficient alternative.

The system processes raw text data through comprehensive **data preprocessing** methods such as text normalization, tokenization, and stop-word removal. It employs advanced **feature engineering** techniques like Bag-of-Words, TF-IDF, and embeddings to extract meaningful representations of the questions. These features are used to train supervised machine learning models, enabling accurate tag predictions based on labeled datasets.

The project includes:

* **Text Classification Models:** Using algorithms like Support Vector Machines (SVM), Random Forest, or neural networks to assign relevant tags.
* **Integration Capabilities:** Seamlessly integrating with existing platforms for real-time tagging and content management.
* **Scalability:** Adapting to large datasets and diverse question formats to support growing user bases.

The **Automatic Question Tagging System** bridges the gap between manual and automated categorization, providing a user-friendly and efficient solution. Its real-world applications demonstrate the transformative potential of NLP in content management and knowledge organization, offering significant benefits to both individual users and organizations.

The increasing frequency and sophistication of cyberattacks have brought digital identity protection to the forefront of modern cybersecurity challenges. Traditional password management methods, such as using weak or repetitive passwords, storing credentials in insecure locations, or relying on memory, expose both individuals and organizations to significant risks. Cybercriminals exploit these vulnerabilities through methods likephishing, malware attacks, and brute force attempts, leading to widespread breaches and identity theft.

Software-based password managers, while offering convenience, often come with their own set of risks. These systems are susceptible to malware infections, phishing scams, and system vulnerabilities. Additionally, the dependence on master passwords introduces a single point of failure, making all stored credentials vulnerable if the master password is compromised. Cloud-based solutions, though popular, further increase risks by transmitting and storing sensitive data online, exposing it to potential interception and unauthorized access.

The lack of secure, user-friendly solutions often discourages users from adopting best practices, leaving them at heightened risk. This underscores the urgent need for a robust, comprehensive password management system that not only protects digital identities but also simplifies security practices for widespread adoption.

## Architecture



## Algorithm Description

## Key Features:

## Hardware Enclave: Isolates sensitive data and operations.

## Biometric Authentication: Primary access control.

## End-to-End Encryption: Protects data in transit and at rest.

## Strong Password Generation: Creates secure credentials.

## Multi-Factor Authentication: Enhances security for high-risk accounts.

## Offline Backup: Enables data recovery.

## Benefits:

## Enhanced Security: Protects against data breaches and attacks.

## Improved Usability: Simplifies password management.

## Increased Convenience: Streamlines user experience.

## Enhanced Privacy: Minimizes data exposure.

## This system provides a robust and user-friendly solution for managing digital credentials securely.

## CODE

## #include <stdio.h>

## #include <string.h>

## #include <ctype.h>

## #define MAX\_QUESTIONS 100

## #define MAX\_LENGTH 256

## // Function prototypes

## void to\_lowercase(char \*str);

## void preprocess\_text(char \*text);

## void classify\_question(const char \*question);

## int main() {

## int n;

## char questions[MAX\_QUESTIONS][MAX\_LENGTH];

## printf("Enter the number of questions: ");

## scanf("%d", &n);

## getchar(); // Consume newline character after the number input

## printf("Enter the questions:\n");

## for (int i = 0; i < n; i++) {

## fgets(questions[i], MAX\_LENGTH, stdin);

## questions[i][strcspn(questions[i], "\n")] = '\0'; // Remove newline character

## }

## printf("\nClassifying questions:\n");

## for (int i = 0; i < n; i++) {

## preprocess\_text(questions[i]);

## classify\_question(questions[i]);

## }

## return 0;

## }

## // Convert string to lowercase

## void to\_lowercase(char \*str) {

## for (int i = 0; str[i]; i++) {

## str[i] = tolower(str[i]);

## }

## }

## // Preprocess text (convert to lowercase, remove special characters)

## void preprocess\_text(char \*text) {

## to\_lowercase(text);

## for (int i = 0, j = 0; text[i]; i++) {

## if (isalnum(text[i]) || isspace(text[i])) {

## text[j++] = text[i];

## }

## }

## }

## // Classify question based on keywords

## void classify\_question(const char \*question) {

## if (strstr(question, "math") || strstr(question, "algebra") || strstr(question, "geometry")) {

## printf("Tag: Math - %s\n", question);

## } else if (strstr(question, "science") || strstr(question, "physics") || strstr(question, "biology")) {

## printf("Tag: Science - %s\n", question);

## } else {

## printf("Tag: General - %s\n", question);

## }

## }

## Output

## 

## Conclusion

In conclusion, the **"Automatic Questions Tagging System"** provides a comprehensive and innovative approach to managing and organizing textual data through Natural Language Processing (NLP). By leveraging advanced text classification techniques and robust data preprocessing methods, this system offers a structured solution for automatically categorizing questions into relevant tags, making it highly valuable in educational, corporate, and technical domains.

The implementation of effective preprocessing steps—such as text normalization, tokenization, and cleaning—ensures that the input data is optimized for accurate analysis and classification. Moreover, the project highlights the significance of feature engineering techniques, including bag-of-words and term frequency-inverse document frequency (TF-IDF), which transform raw textual data into meaningful representations suitable for machine learning models.

The use of labeled and annotated datasets as a foundation for training and testing underscores the importance of high-quality data in achieving reliable results. This not only facilitates the development of accurate classification models but also emphasizes the real-world challenges of managing and curating datasets.

By combining these elements, the "Automatic Questions Tagging System" demonstrates how NLP can be applied to simplify complex tasks, such as managing large volumes of questions in educational or customer support scenarios. Its ability to handle diverse question types and classify them efficiently makes it an essential tool for modern applications.

As the demand for intelligent systems capable of understanding and processing natural language continues to grow, the "Automatic Questions Tagging System" stands out as a forward-thinking solution. This project not only equips students with essential NLP skills but also highlights the importance of integrating data preprocessing, feature engineering, and classification techniques to tackle real-world challenges. Ultimately, this system contributes to the broader goal of enhancing human-computer interaction and simplifying information management in today's digital landscape.

## Future Enhancements

## Integration of Deep Learning: Incorporate advanced models like BERT or GPT for improved semantic understanding and context-aware tagging.

## Multilingual Support: Add support for multiple languages to make the system accessible globally.

## Real-Time Processing: Optimize the system for real-time tagging of streaming questions for faster responses.

## References

1. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing*. Pearson.  
   (Covers foundational concepts in NLP, including text classification and preprocessing techniques.)
2. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space."  
   (Discusses word embeddings that can enhance text classification systems.)
3. TensorFlow NLP: https://www.tensorflow.org/text  
   (Resources for implementing text classification models using machine learning and deep learning.)
4. "Understanding Text Classification in NLP" by Towards Data Science <https://towardsdatascience.com>.
5. The Stanford Question Answering Dataset (SQuAD): <https://rajpurkar.github.io/SQuAD-explorer>.