

**A project Report On**

AUTOMATIC QUESTIONS TAGGING SYSTEM

SUBMITTED TO

**SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES**

In partial fulfillment of the award of the course of

THEORY OF COMPUTATION-CSA1337

**SUBMITTED By**

# T.Rechan Reddy

**192211812**

**SUPERVISOR**

Dr.C.Anitha



**SAVEETHA SCHOOL OF ENGINEERING,**

**SIMATS CHENNAI-602105**

**MARCH-2024**

**BONAFIDE CERTIFICATE**

Certified that this project report AUTOMATIC QUESTIONS TAGGING SYSTEM**,** is the bonafide work **T.Rechan Reddy.** who carried out the project work. Certified further, that to the best of my knowledge the work reported herein does not form any other project report .

|  |  |
| --- | --- |
| **Project Supervisor** | **Head of the Department** |
| **Date :** | **Date:** |

**TABLE OF CONTENTS:**

**TABLE OF CONTENTS:**

|  |  |  |
| --- | --- | --- |
| **SNO** | **CONTENT** | **PAGE NO:** |
| **1)** | INTRODUCTION | 3 |
| **2)** | Problem Definition and Algorithm: | 3-4 |
| **3)** | Algorithm Definition: | 4 |
| **4)** | Pseudocode**:** | 5 |
| **5)** | Experimental Evaluation: | 6 |
| **6)** | Conclusion | 7 |
| **7)** | Bilbiography. | 8 |

**Introduction:**

The Automatic Questions Tagging System harnesses the power of Natural Language Processing (NLP) to revolutionize the way questions are categorized and analyzed. By leveraging advanced linguistic algorithms, this system automatically assigns appropriate tags to questions, facilitating efficient information retrieval and enhancing user experience. Through NLP techniques such as semantic analysis, syntactic parsing, and machine learning models, the system can accurately identify the intent and context of questions, ensuring precise tagging for improved searchability and organization. This innovative approach not only streamlines question handling processes but also paves the way for more sophisticated applications in knowledge management and information retrieval domains.

**Problem Definition and Algorithm:**

**Task Definition:**

The task is to develop an Automatic Questions Tagging System using Natural Language Processing (NLP) to accurately assign tags to input questions, improving information retrieval and organization, which is crucial for efficient knowledge management and user experience enhancement.

**Algorithm Definition:**

**Algorithm used:**

**Preprocessing:**

* Tokenization: Split the input question into individual tokens (words).
* Stopword Removal: Remove common stopwords (e.g., "the," "is," "are") that don't contribute significantly to the meaning of the question.
* Lemmatization/Stemming: Convert words to their base forms to reduce variations (e.g., "running" to "run").

Feature Extraction:

* N-grams Generation: Create n-grams (sequences of tokens) to capture contextual information (e.g., bi-grams, tri-grams).
* Part-of-Speech Tagging: Label each token with its part of speech (e.g., noun, verb, adjective).
* Named Entity Recognition (optional): Identify named entities like people, organizations, locations, etc., if applicable.

**Vectorization:**

* Convert the preprocessed text into numerical vectors using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec, GloVe).
* Optionally, combine different feature vectors (e.g., TF-IDF vectors, POS tag vectors) into a unified representation.

Machine Learning Model:

* Train a classification model (e.g., Support Vector Machine, Random Forest, Neural Network) using a labeled dataset of questions and their corresponding tags.
* Use the trained model to predict tags for new, unseen questions based on their vector representations.

Postprocessing:

* Filter out irrelevant or noisy tags generated by the model.

**Pseudocode:**

function automaticQuestionTagging(question):

preprocessed\_text = preprocess(question)

feature\_vector = extractFeatures(preprocessed\_text)

predicted\_tags = machineLearningModel.predict(feature\_vector)

refined\_tags = postprocess(predicted\_tags)

return refined\_tags

function preprocess(text):

tokens = tokenize(text)

tokens = removeStopwords(tokens)

tokens = lemmatize(tokens)

return tokens

function extractFeatures(preprocessed\_text):

ngrams = generateNGrams(preprocessed\_text)

pos\_tags = partOfSpeechTagging(preprocessed\_text)

named\_entities = recognizeNamedEntities(preprocessed\_text)

feature\_vector = vectorizeFeatures(ngrams, pos\_tags, named\_entities)

return feature\_vector

function machineLearningModel.train(training\_data):

# Train a classification model using training\_data

model = trainModel(training\_data)

return model

function machineLearningModel.predict(feature\_vector):

# Use the trained model to predict tags for the input feature\_vector

predicted\_tags = model.predict(feature\_vector)

return predicted\_tags

function postprocess(predicted\_tags):

refined\_tags = filterIrrelevantTags(predicted\_tags)

refined\_tags = applyRules(refined\_tags)

return refined\_tags

**Experimental Evaluation:**

**3.1 Methodology**

Methodology refers to the systematic approach used to conduct research or solve problems. It involves defining the research questions, selecting appropriate data collection methods, and analyzing the data to draw conclusions. A well-defined methodology enhances the credibility and validity of research findings

**Results**

The Automatic Questions Tagging System in NLP has shown promising results in accurately categorizing question types and improving question answering systems' performance. This technology enhances understanding and processing of queries, leading to more efficient information retrieval and interaction with users.

**Discussion**

An Automatic Questions Tagging System in NLP streamlines question classification by assigning relevant tags to queries, enhancing information retrieval and analysis. This technology leverages machine learning algorithms to accurately categorize questions based on their semantic and syntactic features.

**Related Work**

* Existing research on automatic question tagging systems in NLP includes methods utilizing deep learning models such as recurrent neural networks (RNNs) and transformers for semantic understanding and tag prediction.
* Prior work has also explored rule-based approaches, hybrid models combining machine learning and linguistic rules, and ensemble techniques to enhance the accuracy and efficiency of question tagging systems.

**Future Work**

* Developing hybrid models that integrate both deep learning techniques and linguistic rules to improve the accuracy and interpretability of question tagging.
* Exploring multi-task learning frameworks to simultaneously address related NLP tasks such as question classification, information retrieval, and sentiment analysis within the context of question tagging.
* Investigating the application of pre-trained language models and transfer learning techniques to leverage large-scale labeled datasets and improve the generalization of question tagging systems across diverse domains and languages.

**Conclusion**

In conclusion, advancements in Automatic Question Tagging Systems within NLP have shown promise through the integration of deep learning models, linguistic rules, and hybrid approaches. Future research efforts should focus on enhancing model interpretability, leveraging multi-task learning frameworks, and evaluating practical applications for improved user experiences.

**Bilbiography**.

Huang, X., & Yates, A. (2018). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Li, H., & Roth, D. (2002). Learning question classifiers. Proceedings of the 19th international conference on Computational linguistics-Volume 1, 1, 1-7.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

Seo, M., & Kwak, S. (2020). A Survey on Deep Learning Techniques for Question Classification. Applied Sciences, 10(6), 2015.

Yang, Z., Yang, D., & Dyer, C. (2016). Hierarchical attention networks for document classification. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 1480-1489.