# Chatbot Implementation Using NLP-Techniques and Machine Learning

Vamsi Krishna Yadav Loya

**Abstract**—This project leverages Natural Language Processing (NLP) techniques and Machine Learning (ML) models aiming to create an efficient question-and-answer chatbot system. Focused on the "SQuAD2.0" Question-Answering dataset, the study combines NLP techniques such as named entity recognition, and n-gram extraction with ML models, particularly k-Nearest Neighbors. The interdisciplinary nature of the project spans data science and machine learning, providing a comprehensive exploration of language data. By integrating diverse methodologies, the project aims to address questions related to the given context. Through systematic data processing, visualization, and predictive modeling, the research objectives are to uncover patterns, evaluate predictive techniques, and contribute to the broader understanding of language data in the context of question-answering systems.

**Keywords**— SQuAD (Stanford Question and Answering Dataset), Natural Language Processing (NLP), Machine Learning (ML), Named Entity Recognition, N-Gram Extraction, K-Nearest Neighbors, BOW (Bag Of Words), TF-IDF (Term Frequency-Inverse Document Frequency).

## 1 Introduction

The usage of question-answering systems is increasing daily, people make frequent use of question-answering systems to find the right answer for different kinds of information. The goal of the question classification process is to accurately assign labels to questions based on the expected answer type. Many question classification techniques have been proposed to help in understanding the actual intent of the user's question, but the abundance of available data has made the process of obtaining relevant information challenging in terms of processing and analyzing it.

In this paper, the focus is on building a question-and-answering chatbot using a KNN model and performing data processing and data transformation using NLP techniques [1] like NER recognition, POS tagging, removing stopwords, and performing Lemmatization. The Experimental results show that our solution leads to accurate identification of answers for different question types.

The rest of the paper is organized as follows: Section 2 outlines previous work on question-and-answer classification. Section 3 provides a detailed description of the Exploratory data analysis. Section 4 reports experimental results. Finally, Section 5 concludes the paper and outlines future work.

### 1.1 Background Domain Knowledge:

There are three major areas in which a person needs to Understand. Natural Language Processing in Python (NLTK) Documentation. This official documentation provides detailed information on the usage of the NLTK library. It clearly explains how to use the library in Python, including Tokenization, Lemmatization, Stemming, and other NLP techniques needed for this project.[1]
Chatbots: An Introduction by Chatbots magazine. This online magazine offers various articles and resources that introduce chatbots, their applications, and how they even work. From the information provided, we can understand the basics of chatbot technology.[2]
Decision Trees: Classification Algorithm.
Knowing decision trees is also a must for this project. This article introduces decision trees, explaining the concept first, construction, and how to implement this topic in the real world.[3]

### 1.2 Sources of Information:

1. Source weblink- NLTK :: Natural Language Toolkit. (n.d.). https://www.nltk.org/
2. Source weblink- Schlicht, M. (2023, April 16). The Complete Beginner's Guide to Chatbots – Chatbots https://chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca
3. Source weblink- Classification Algorithms – Decision. https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_decision_tree.htm

### 1.3 Interesting Questions about data:

1. What are the characteristics of questions that are consistently answered correctly across different models?
2. What is the distribution of question types (who, what, where), and how does it relate to answer types?
3. How does the length of the context passages vary in the dataset? Are there contexts that are consistently longer or shorter?

### 1.4 Research Objectives:

1. To characterize the trends within user queries and interaction patterns in the provided text-based data, using exploratory visualization techniques.
2. To employ a machine learning classification technique (KNN) to predict the response for a particular question given to the chatbot by the user and classify user queries and responses, enhancing the chatbot's ability to interpret and respond effectively.
3. To defend the efficiency and accuracy of the machine learning model KNN for chatbot response classification, ensuring its reliability and defensibility.
4. To evaluate and extract meaningful insights into predictions given by the KNN model thereby enhancing the understanding of user query patterns and responses.

## 2 Related WORKS

The challenges in accurately predicting the answer to a question using ML algorithms still exist in the current era and in the past. Below are some of the papers that have discussed solutions to these challenges and different techniques employed to improve the performance of the Classification model.

### 2.1 Detecting Question Intention using KNN

Question Classification is one of the important applications of information retrieval, as it plays a crucial role in improving the performance of question-answering systems. In This research paper, a new technique is introduced which is question type syntactical patterns for detecting question intention to categorize different question types. In addition, a k-nearest neighbor-based approach has been developed for question classification. Different Experiments show that the approach has a good level of accuracy in identifying different question types [5].

### 2.2 Question Classification using SVM

To build a question and answering chatbot, the Question classification technique is very important. This paper presents research work on automatic question classification through machine learning approaches. It presents experiments with five machine learning algorithms: Nearest Neighbors (NN), Naive Bayes (NB), Decision Tree (DT), Sparse Network of Winnows (SNoW), and Support Vector Machines (SVM) using two kinds of features: bag-of-words and bag-of-n-grams [6].

### 2.3 Systematic Comparison of Vectorization Methods in Classification Context

The goal of vector space modeling is to project words in a language corpus into a vector space in such a way that words that are similar in meaning are close to each other. The paper presents a comparison of different existing text vectorization methods in natural language processing. The first focuses on creating word vectors considering the entire linguistic context, while the second focuses on creating document vectors in the context of the linguistic corpus of the analyzed texts. The comparison of these text vectorization methods is done by checking the accuracy of classification; NBC and k-NN were used for the classification to avoid the influence of the choice of the method itself on the result. The conducted experiments provide a basis for further research for better automatic text analysis [7].

### 2.4 Document classification using KNN with fuzzy bags of word representation

Text classification is used to classify the documents depending on the words, phrases, and word combinations according to the declared syntax. Many applications use text classification such as Question and Answering Chatbot.

In the proposed paper, keywords are extracted from documents using TF-IDF and Word Net. The TF-IDF algorithm is mainly used to select the important words by which a document can be classified. Word Net is mainly used to find similarities between these candidate words. The words that have the maximum similarity are considered as Topics(keywords).[8]

### 2.5 Investigating the impact of data scaling on the k-nearest neighbor algorithm

This study investigates the impact of data scaling techniques on the performance of the k-nearest neighbor (KNN) algorithm using ten different datasets from various domains. Three commonly used data scaling techniques, min-max normalization, Z-score, and decimal scaling, are evaluated based on the KNN algorithm's performance in terms of accuracy, precision, recall, F1-score, runtime, and memory usage. The results show that data scaling significantly affects the performance of the KNN algorithm, and the choice of scaling method can have significant implications for practical applications [9].

### 2.6 Natural Language Processing: State of The Art, Current Trends and Challenges

Natural language processing (NLP) has recently gained much attention for representing and analyzing human language computationally. It has spread its applications in various fields such as machine translation, email spam detection, information extraction, summarization, medical, and question answering etc. This paper distinguishes by discussing different levels of NLP and components of Natural Language Generation (NLG) followed by presenting the history and evolution of NLP, state of the art presenting the various applications of NLP and current trends and challenges [10].

## 3 Exploratory Data Analysis

The dataset used for this project is SQUAD(2.0) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage. The dataset has more than 1 million rows out of which 20,000 rows were used to perform experimentations for this project. The download file is a .json file so we converted this to a csv file and extracted the top 20,000 rows. Even the unnecessary columns like row id and title are removed. Therefore, this data can be used for analysis and model training as needed.

Each entry in the dataset encapsulates a question paired with its corresponding answer. These entries represent the fundamental units of analysis, portraying diverse linguistic styles, topics, and complexities inherent in real-world interactions.

Our project has three quantitative features which are character count, word count and unique word count. The EDA is performed on this by using different type of plots like box plots, scatter plots and histograms. Box plots provide insights into question length variations. They help identify outliers, showcase data spread, and highlight central tendencies, offering a visual representation of how questions vary in length.

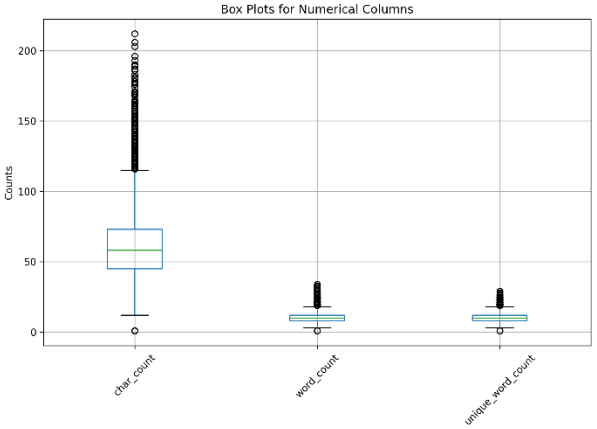


Figure 1: Box plot for char_count vs word_count vs unique_word_count

Scatter plots were created for word count vs unique word count, word count vs character count, and unique word count vs character count. These plots reveal positive correlations, indicating relationships between different linguistic metrics. For example, as word count increases, unique word count and character count also tend to increase.

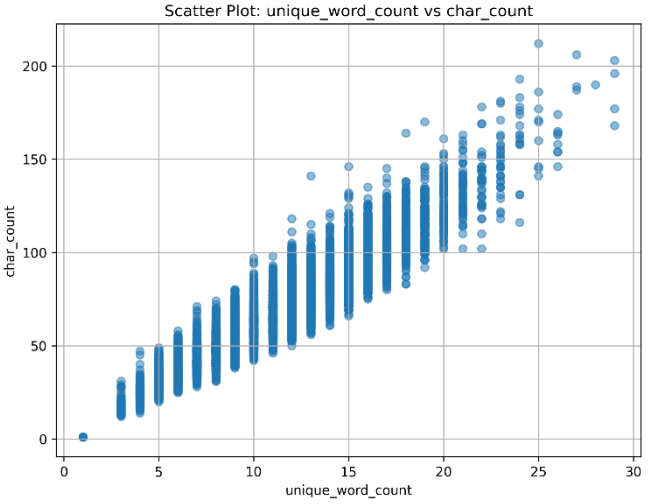


Figure 2: Scatter plot for unique words vs character count

Two qualitative features that are used in the project are questions and answers. The EDA is performed on the qualitative features as well as the quantitative features by calculating the summary statistics. It tells us the most frequent question question and answers along with least frequent questions and answers along with the max, min, median of the quantitative features.

Summary Statistics for Quantitative Features:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| char_count | 20000.0 | 60.85265 | 21.868984 | 1.0 | 45.0 | 58.0 | 73.0 | 212.0 |
| word_count | 20000.0 | 10.38255 | 3.648461 | 1.0 | 8.0 | 10.0 | 12.0 | 34.0 |
| unique_word_count | 20000.0 | 10.05950 | 3.310318 | 1.0 | 8.0 | 10.0 | 12.0 | 29.0 |

Figure 3: Summary statistics for Quantitative features

Summary Statistics for Qualitative Features:

| Feature | Number of Categories | Most Frequent Category | Least Frequent Category |
|---|---|---|---|
| question | 19932 | [Which Caribbean nation is in the top quartile of HDI (but missing IHDI)?, Who won this season of Idol?] To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France? | |
| answers | 14771 | [three] | Saint Bernadette Soubirous |

Figure 4: Summary statistics for qualitative features

## 4 Methodology

The chosen predictive technique is a KNN (K-Nearest Neighbors) model, specifically tailored for question-answering classification. KNN is one of the simplest of all machine learning algorithms. KNN is used in pattern

$$d(x, y) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}$$

recognition for classification and regression. For both classification and regression, the input consists of the k nearest training examples in sample space and the output depends on classification or regression. KNN has been used in statistical estimation and pattern recognition. There are different measures for distance calculation like Euclidean, Euclidean Squared, City-block and Chebyshev. Among all these Euclidean is most popular choice to measure the distance between to the two points. Euclidean distance between two points x and y of M dimensions is given by:
For the question and answering dataset KNN is used as a classification algorithm. This model predicts the most relevant answers based on the input questions and contextual information. The KNN model is applied to the transformed dataset, leveraging various features like question content and context. The model utilizes TF-IDF vectorization for textual representation, considering the relevance of words in questions and answers. During testing, the model predicts the top sentences related to the given context and question using a chosen k value.
**Best Model:** The KNN model with Bag of Words Vectorizer and a chosen k value of 5 emerges as the best-performing model. It exhibits high accuracy, precision, and recall compared to alternative models, making it the optimal choice for the question-answering classification problem.

## 5 Results and Discussion

To evaluate the experimental results precision, recall and accuracy metrics are used. The experimentation and analysis conducted on the question-answering system using various data transformations, visualization techniques, and predictive models have yielded valuable insights. The focus of the discussion will be on the performance of the predictive technique, the influence of different factors on its decisions, and the feasibility of achieving the research objectives. The KNN model with Bag of Words (BoW) vectorization and a k value of 5, **demonstrated commendable performance**. The results are show in below Figure 5.

```
KNN_BOW_K_equals_to_five - Testing Accuracy: 0.80

KNN_BOW_K_equals_to_five - Precision: 0.80

KNN_BOW_K_equals_to_five - Recall: 0.79
```

Figure 5: Results of KNN model with K value is 5

The 80% accuracy indicates out of 100 questions 80 questions were correctly answered as accuracy gauges the overall correctness of the model by measuring the ratio of correctly predicted instances to the total instances Accuracy gauges the overall correctness of the model by measuring the ratio of correctly predicted instances to the total instances.
A precision of 80% indicates the relevance of the predicted context to the given question, enhancing the understanding of the model's accuracy in predicting relevant information. Recall of 79% reflects the model's sensitivity in identifying actual question contexts, providing insights into its ability to recognize a significant portion of relevant instances.

### 5.1 Performance of Predictive Technique (MLA):

Three Alternate models were built based on different vectorizers and by choosing different K values. As a part of Experiment one, the chosen predictive technique is a KNN model with TF-IDF vectorization and a k value of 5, demonstrated **good performance**. The key metrics evaluated include accuracy, precision, and recall. The model achieved an accuracy of 74%, precision of 73%, and recall of 72% on the test data.
As a part of Experiment two, the chosen predictive technique is a KNN model with Bag of Words (BoW) vectorization and a k value of 1, demonstrated **bad performance**. The model achieved an accuracy of 63%, precision of 64%, and recall of 62% on the test data.
As a part of Experiment three, the chosen predictive technique is a KNN model with count vectorization and a k value of 1, demonstrated **poor performance**. The model achieved an accuracy of 58%, precision of 59%, and recall of 57% on the test data.

### 5.2 Factors influencing the Decision(MLA):

During experimentation, various factors were identified that significantly influenced the model's decisions:
**Quality of Questions:** The model's performance was highly dependent on the quality and informativeness of the input questions. Well-constructed questions with clear context and relevance to the dataset resulted in more accurate predictions.
**Context Understanding:** The context in which a question is asked played a crucial role. The model's ability to identify the relevant context and subsequently predict accurate answers was evident. Ambiguous or contextually challenging questions led to less precise predictions.
**Data Transformations:** The applied data transformations, including Bi-Grams, Part-of-Speech Tags, Named Entity Recognition (NER), and Sentiment Analysis, significantly improved the dataset's quality. These transformations enriched the data by preserving context, enhancing grammatical understanding, and adding emotional context.

### 5.3 Predictive techniques to Answer Research Objectives:

1. Through exploratory visualization techniques, we analyze trends within user queries and interaction patterns in the provided data. Our focus is on understanding the dynamics of language data through systematic processing and visualization.
2. Employing the KNN classification technique, we predict responses to user questions. The chatbot's ability to

interpret and respond effectively is enhanced through the classification of user queries and responses.
3. The efficiency and accuracy of the KNN model are rigorously defended, ensuring its reliability in predicting chatbot responses. We evaluate the model's performance to establish its defensibility in real-world scenarios.
4. By evaluating and extracting meaningful insights from predictions made by the KNN model, we aim to enhance our understanding of user query patterns and responses. This contributes to the broader comprehension of language data in question-answering systems.

## 6 Conclusion

In conclusion, the experimentation results indicate a successful implementation of the predictive technique, with high accuracy and precision. Factors such as question quality, context understanding, data transformations, and visualization insights played pivotal roles in influencing the model's decisions. The feasibility of the predictive technique for answering research objectives is evident, and continuous refinement of the model and dataset can further enhance its performance. The comprehensive approach involving both natural language processing techniques and machine learning models provides a solid foundation for the development of effective question-answering systems.

## References


[1] Chatbots Development Using Natural Language Processing: A Review. (2022, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10017592

[2] Schlicht, M. (2023, April 16). The Complete Beginner's Guide To Chatbots - Chatbots Magazine. Medium. https://chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca

[3] Chatbots Development Using Natural Language Processing: A Review. (2022, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10017592.

[4] Holotescu, C., & Holotescu, V. (2016, September 1). MOOCBuddy: a chatbot for personalized learning with MOOCs. ResearchGate. https://www.researchgate.net/publication/304037510_MOOCBuddy_a_chatbot_for_personalized_learning_with_MOOCs

[5] Mohasseb, Alaa & Bader-El-Den, Mohamed & Haig, Ella. (2018). Detecting Question Intention Using a K-Nearest Neighbor Based Approach. 10.1007/978-3-319-92016-0_10.

[6] Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 26–32. ACM (2003)

[7] Gumińska, Urszula & Poniszewska-Maranda, Aneta & Ochelska-Mierzejewska, Joanna. (2022). Systematic Comparison of Vectorization Methods in Classification Context. Applied Sciences. 12. 5119. 10.3390/app12105119.

[8] Prasanna, P.L. & Manogni, S. & Tejaswini, P. & Kumar, K.T. & Kesharaju, Manasa. (2019). Document classification using KNN with fuzzy bags of word representation. International Journal of Recent Technology and Engineering. 7. 631-634.

[9] Pagan, Muasir & Zarlis, Muhammad & Candra, Ade. (2023). Investigating the impact of data scaling on the k-nearest neighbor algorithm. Computer Science and Information Technologies. 4. 135-142. 10.11591/csit.v4i2.p135-142.

[10] Khurana, Diksha & Koli, Aditya & Khatter, Kiran & Singh, Sukhdev. (2022). Natural Language Processing: State of The Art, Current Trends and Challenges. Multimedia Tools and Applications. 82. 10.1007/s11042-022-13428-4.