TAWI

A MORINGA SCHOOL CHATBOT

MORINGA

Discover · Grow · Transform

Business Understanding

Data Understanding

Data Preparation

EDA

Modelling

Recommendations

TAWI

BUSINESS UNDERSTANDING

Business Problem

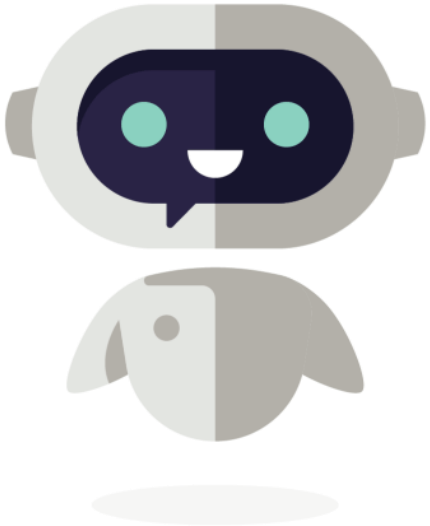
Chatbot Objectives

BUSINESS PROBLEM

A chatbot is a program whose aim is to simulate conversation with human users.

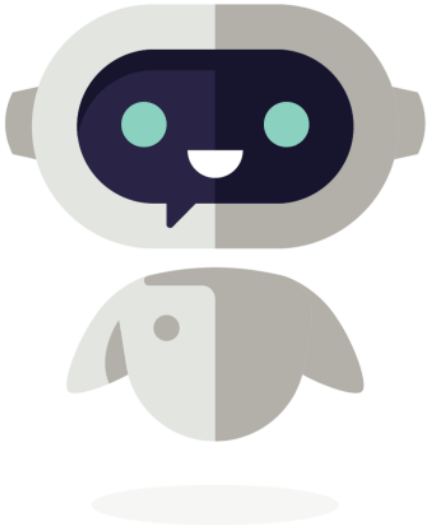
Research suggests that by the end of 2023, more than 75%of customer queries will be resolved by chatbots.

Moringa is one of the best tech schools in Kenya and as it continues to grow, so will the demand for its services. An efficient way to fill this gap would be to employ the use of a chatbot.



OBJECTIVES

- Provide fast, 24/7 Service
- Improve Customer Experience
- Provide Access to Information
- Offer Technical Support
- Data Collection



DATA UNDERSTANDING

01 **link_scraper.py** contains a function designed to go through the Moringa School website and extract hyperlinks present on its pages

02 **web_scraper.py** contains a function that utilizes the links obtained from (01) to scrape text content from the corresponding web pages

03 **scraped_links.json** contains a list of unique URLs obtained during the link scraping process.

04 **moringa_text_corpus.json** stores the text content scraped from the corresponding URLs, organized in a dictionary structure

DATA PREPARATION

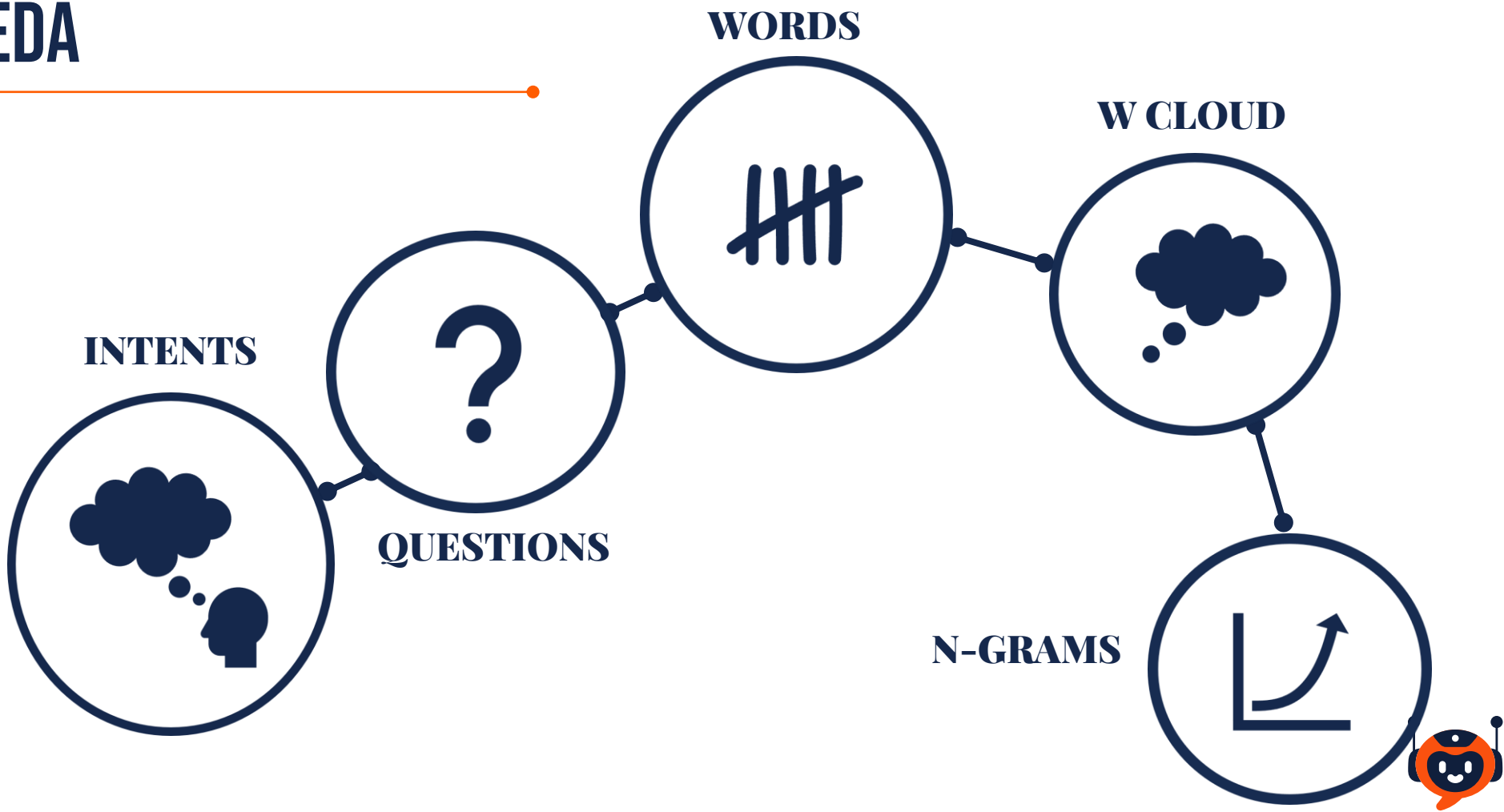
The goal is to create an intent list, which helps the chatbot understand the user's goals. The intent list is stored in “final_intents.json” with the following structure:

“tag": The category of the intent.
Eg
"software_engineering_course_information"

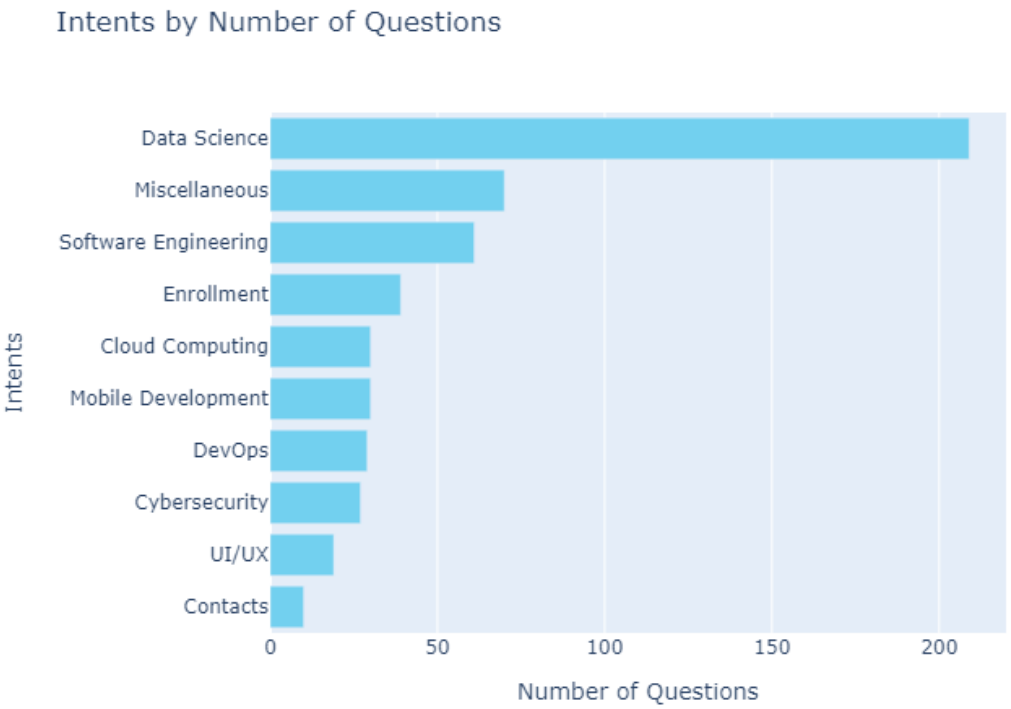
"questions": A list of questions.
Eg
"What topics does the software engineering course cover?“

"responses": A list of responses
Eg
"The software engineering course covers HTML and CSS".

EDA

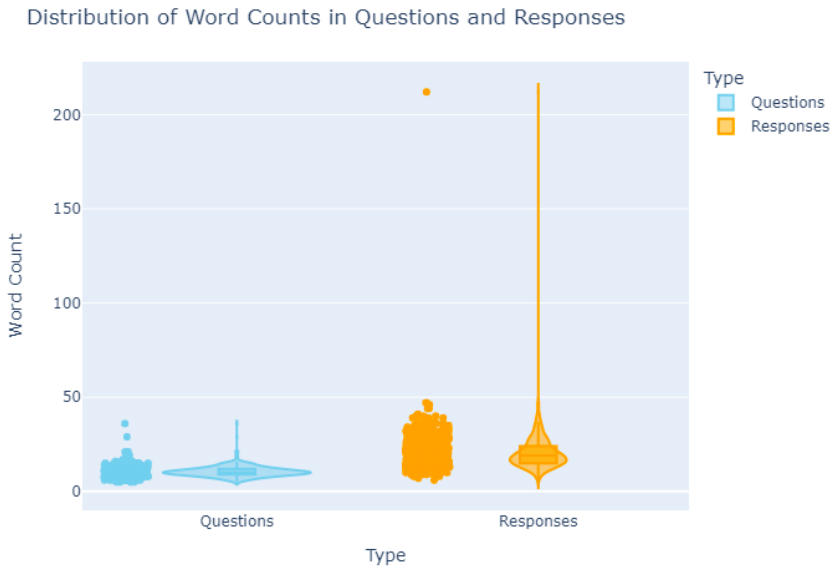


INTENTS

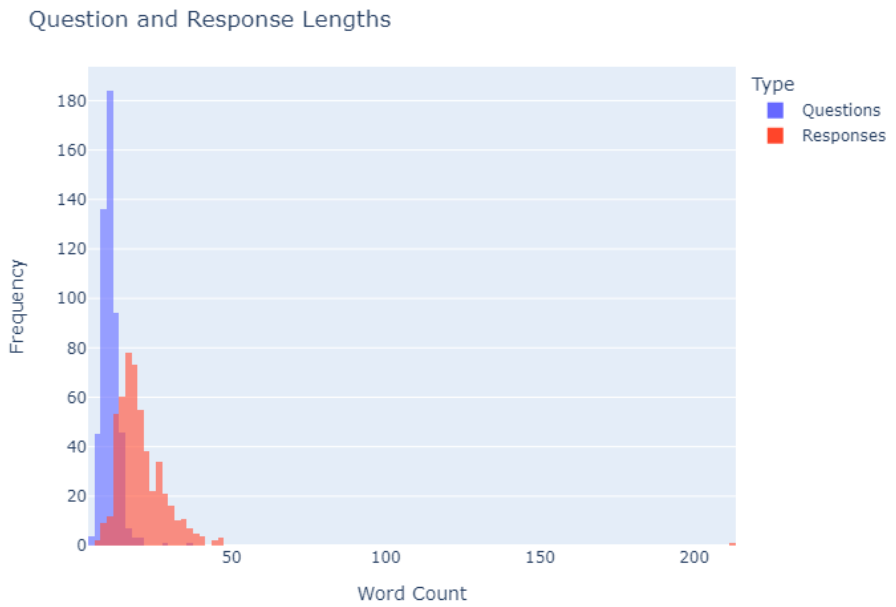


- The dataset comprises 10 unique intents, with an average of approximately 52.40 questions per intent.
- The median number of questions per intent is 30, indicating a balanced distribution.
- Notably, the "Contacts" intent stands out with the fewest questions, totaling 10, suggesting it represents a category with relatively fewer questions compared to others in the dataset.

QUESTIONS

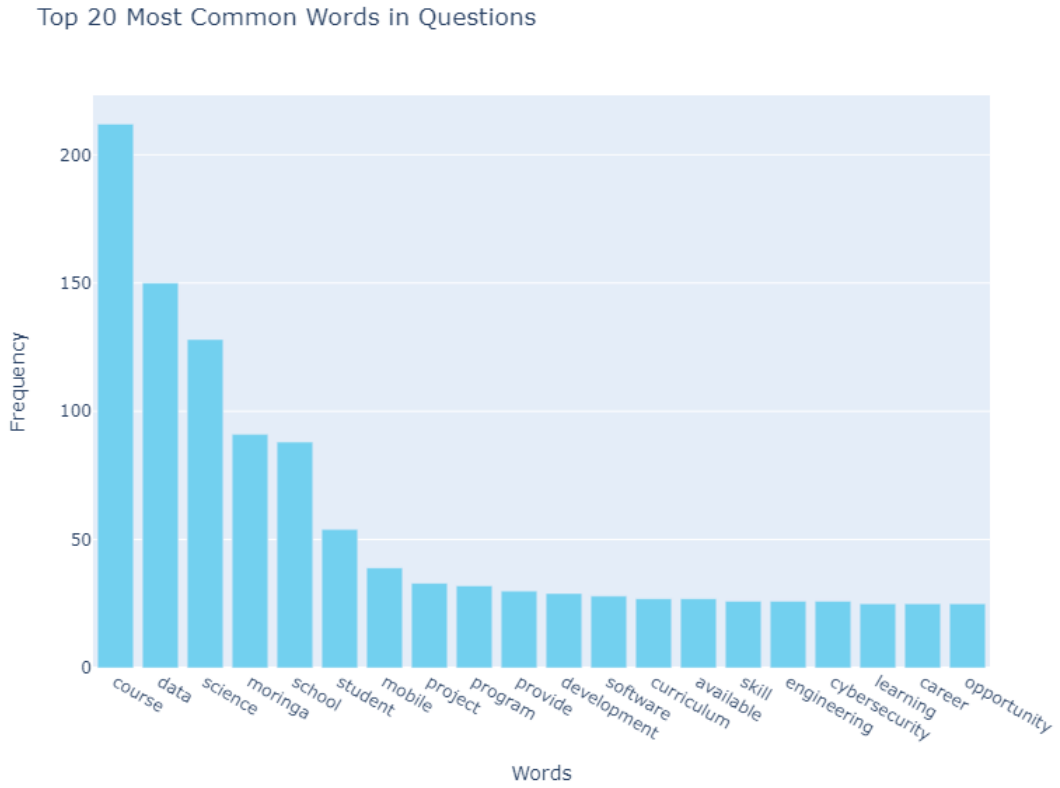


This is a plot showing the distribution of word counts in questions and responses. Questions typically have fewer words, while responses vary more in length.



On average, questions are relatively short (around 10 words), with a moderate level of variability. Responses are longer on average (around 21 words), with greater variability in their lengths.

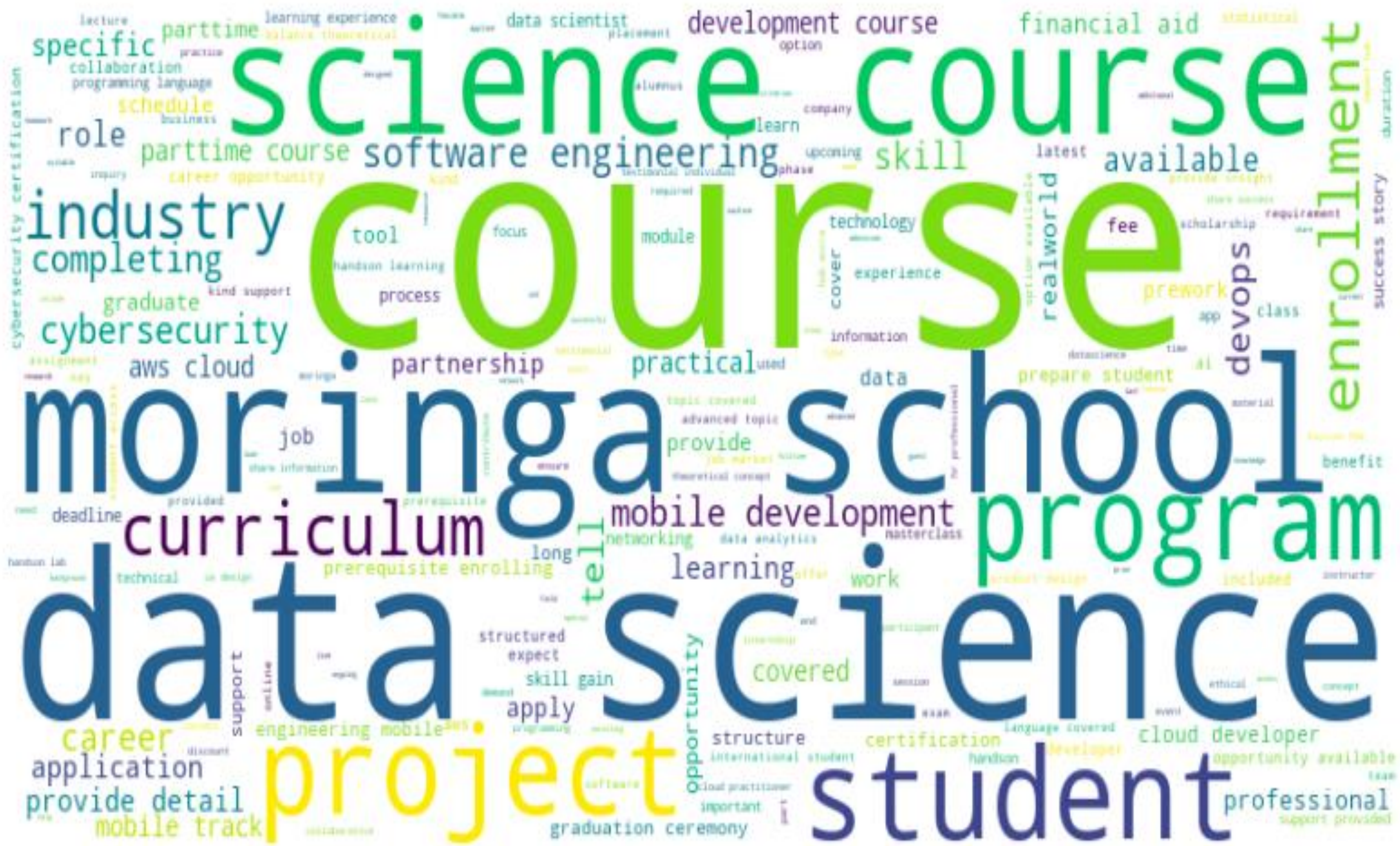
WORDS



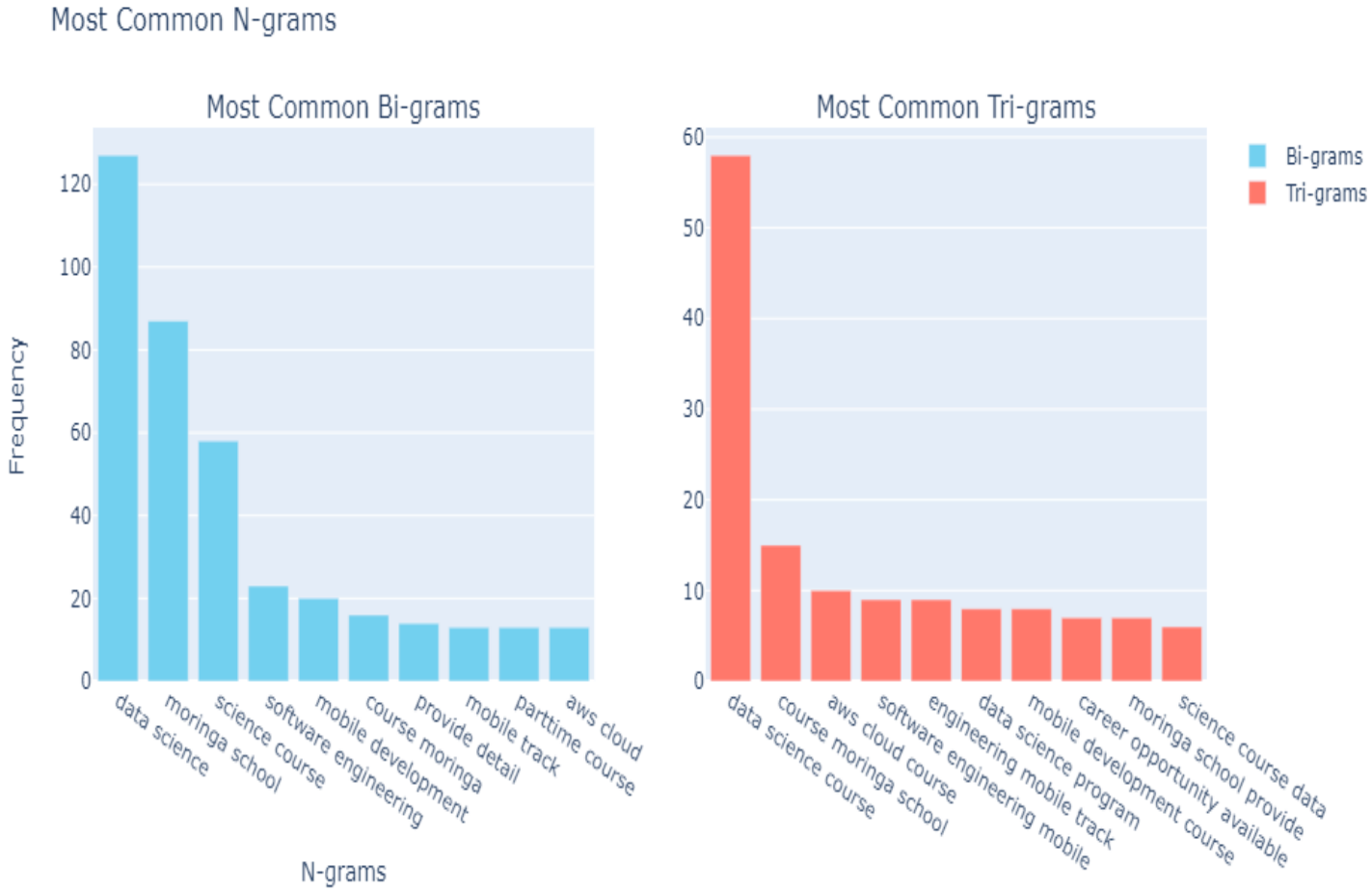
- The analysis highlights key themes in the dataset, emphasizing courses like data science, the school Moringa, and others like education

- The prevalence of words like 'course,' 'data,' 'science,' 'moringa,' 'student,' 'develop,' and 'learn' indicates a focus on data science education with a learner-centric approach.

WORD CLOUD

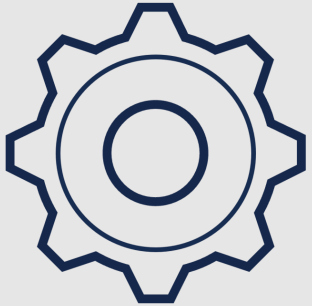

course
science course
moringa school
data science
project
student
program
curriculum
enrollment
industry
software engineering
development course
financial aid
specific
parttime
data scientist
learning experience
lecture
collaboration
programming language
schedule
business
role
parttime course
career opportunity
completing
graduate
cybersecurity
cybersecurity certification
aws cloud
partnership
practical
skill
available
latest
technology
fee
scholarship
requirement
success story
devops
realworld
cover
experience
information
module
focus
tool
handson learning
process
kind support
prework
class
app
data
prepare student
topic covered
advanced topic
provide
job
deadline
mobile development
networking
data analytics
masterclass
long
tell
learning
work
prerequisite enrolling
technical
benefit
structured
expect
skill gain
apply
covered
opportunity
support
online
career
engineering mobile
structure
certification
cloud developer
opportunity available
professional
support provided
application
provide detail
mobile track
graduation ceremony
important
international student
cloud practitioner
handson lab
learn
company
option
placement
alumnus
upcoming
phase
used
duration
included
instructor

N-GRAMS

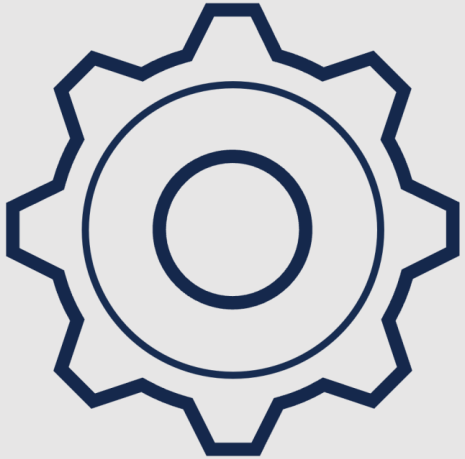

Most Common N-grams
Most Common Bi-grams
Frequency
120
100
80
60
40
20
0
data science
moringa school
science course
software engineering
mobile development
course moringa
provide detail
mobile track
parttime course
aws cloud
N-grams
Most Common Tri-grams
60
50
40
30
20
10
0
data science course
course moringa school
aws cloud course
software engineering mobile
engineering mobile track
data science program
mobile development course
career opportunity available
moringa school provide
science course data
Bi-grams
Tri-grams

MODELLING

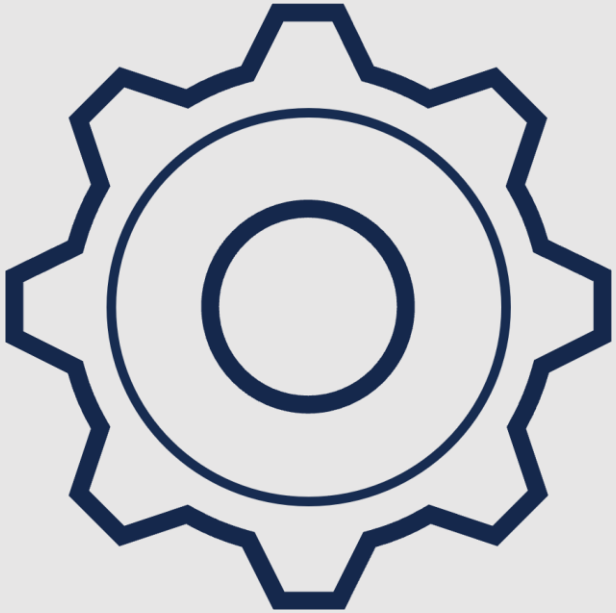
ARCHITECTURE



MACHINE LEARNING MODELS



DEEP LEARNING MODELS



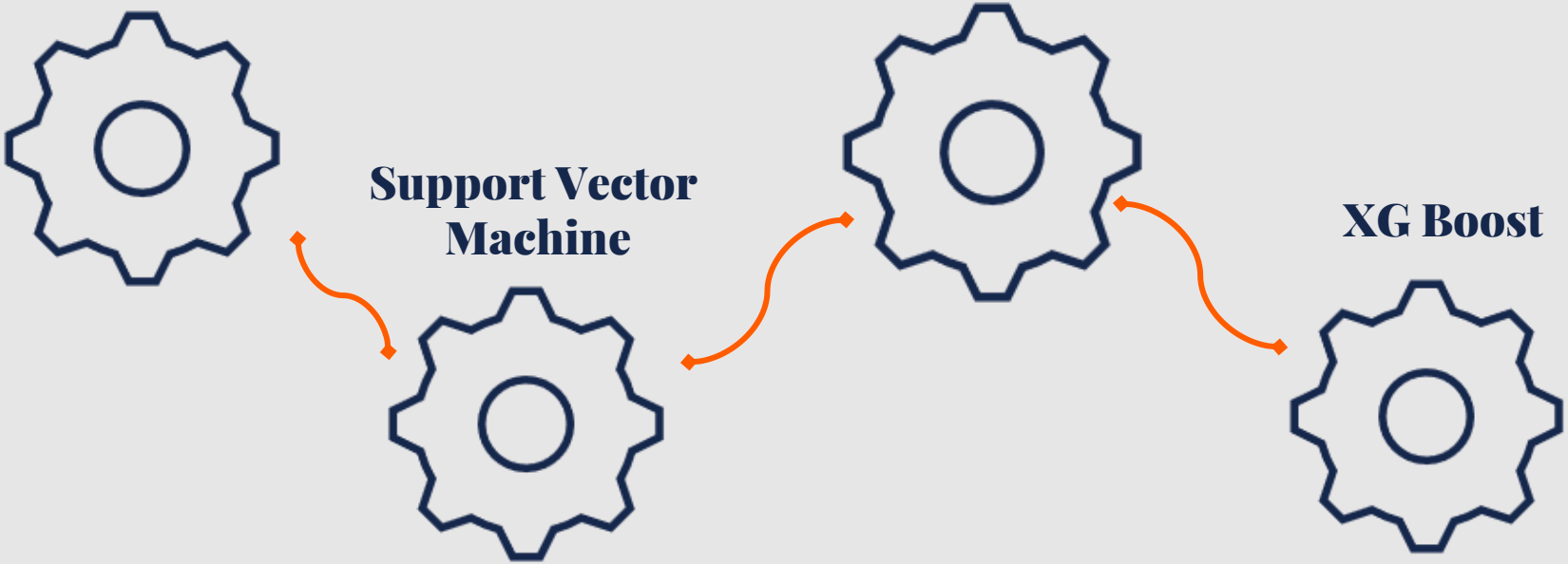
MACHINE LEARNING MODELS

Multinomial Naïve Bayes

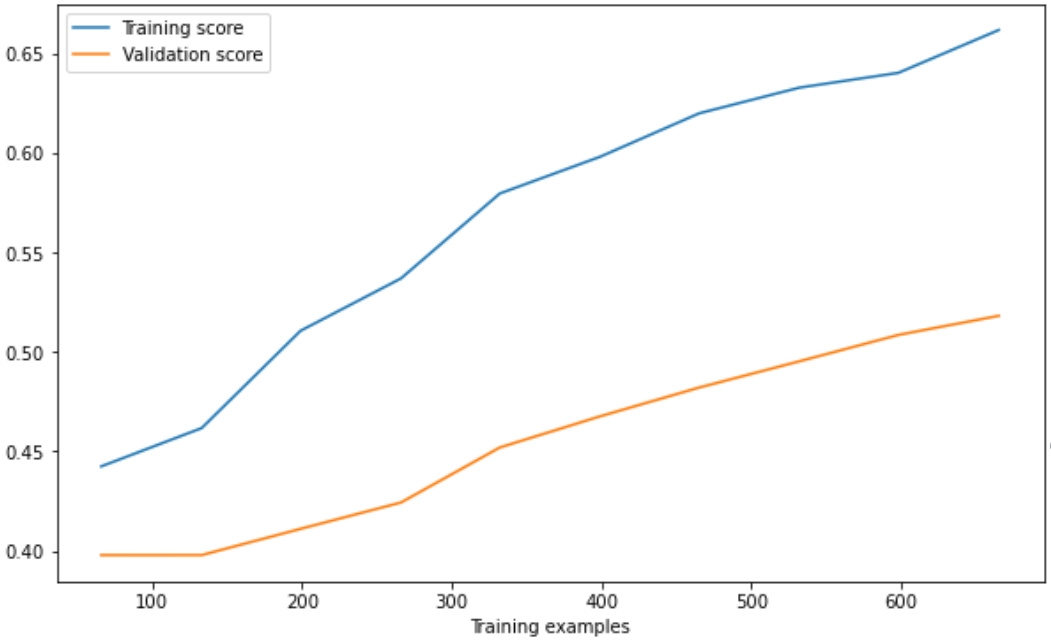
Support Vector Machine

Random Forest

XG Boost



MULTINOMIAL NAÏVE BAYES



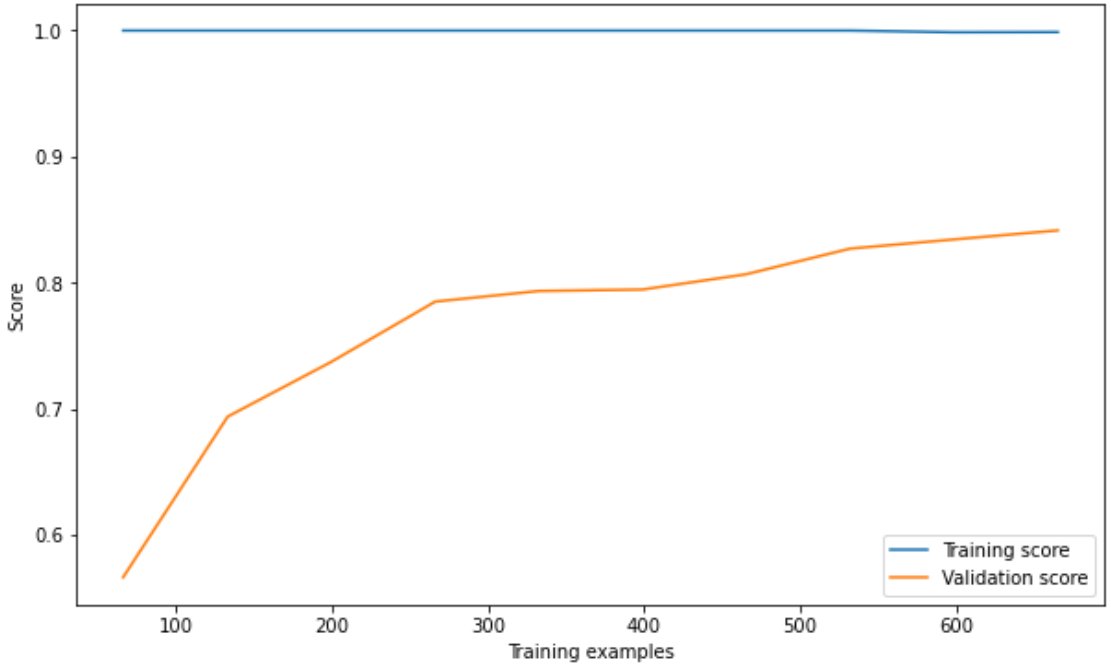
1.Multinomial Naive Bayes:

1. Accuracy: 78%
2. Notable variations in precision and recall among classes.
3. Adjustments suggested for classes with low precision or recall.

2.GridSearchCV Multinomial Naive Bayes:

1. Overall accuracy: 73.33%
2. Notable variations in precision and recall, emphasizing the need for adjustments.

SUPPORT VECTOR MACHINE



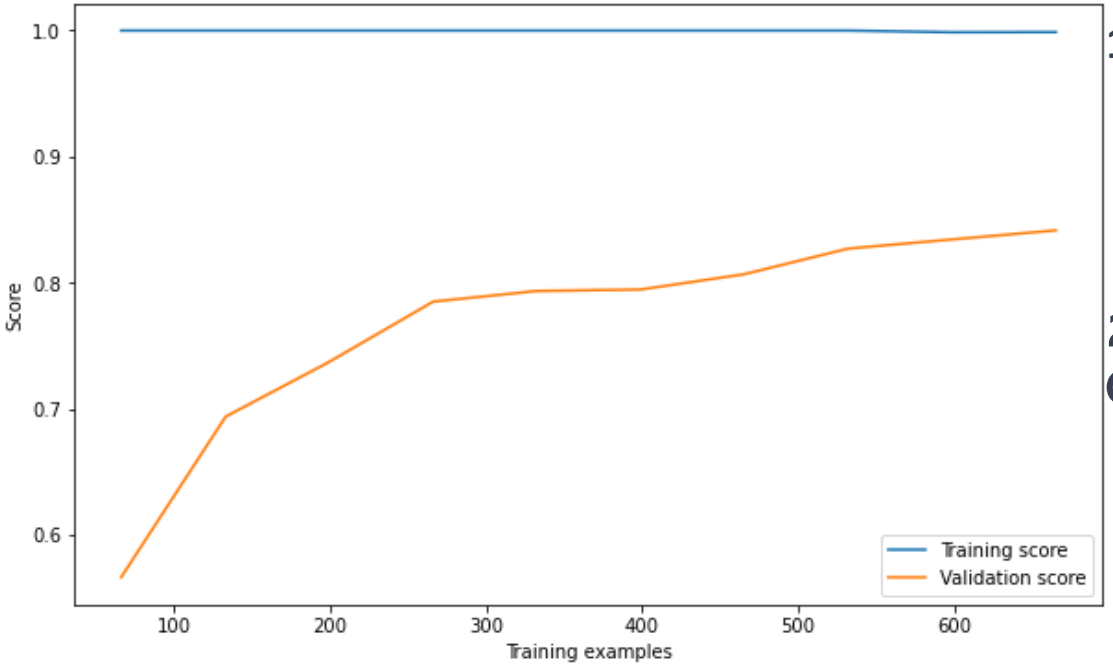
1.TF-IDF Vectorizer and SVM:

1. Accuracy: 81.9%
2. Varied precision and recall values across classes.
3. Consideration of hyperparameter tuning for enhanced performance.

2.Cross-validation with TF-IDF + SVM:

1. Average accuracy: 80.2%
2. Challenges observed in precision and recall for certain classes.

RANDOM FOREST



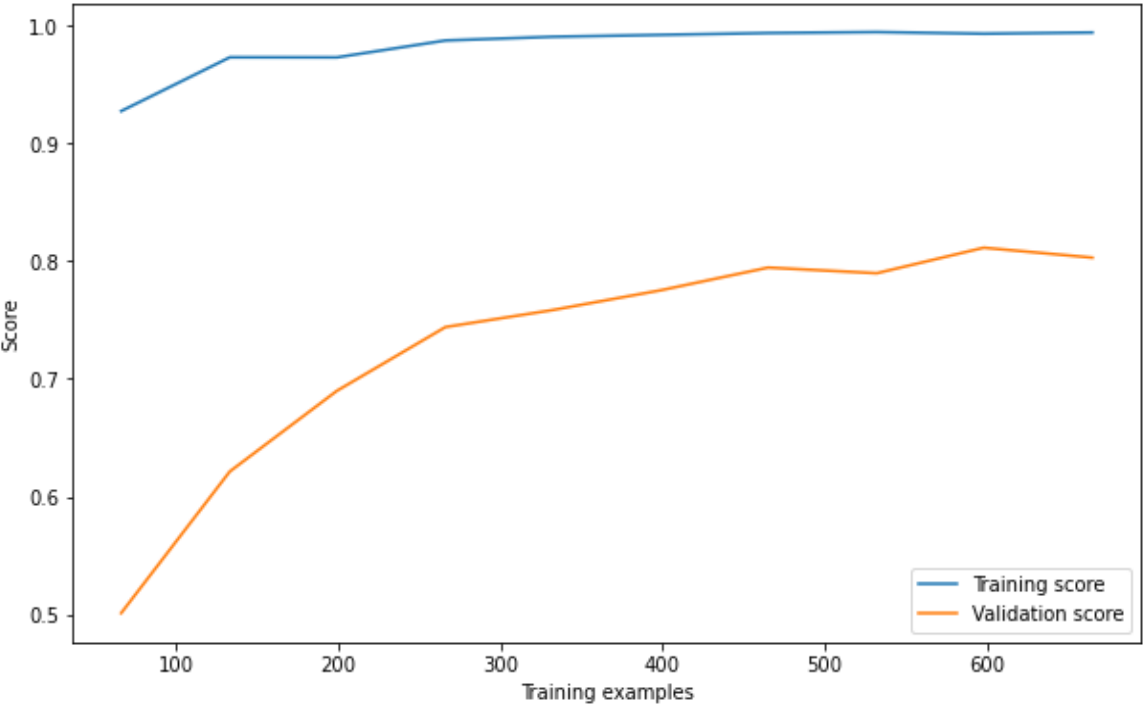
1. Random Forest Classifier:

1. Accuracy: 81%
2. Challenges in precision and recall for specific classes.

2.GridSearchCV for Random Forest Classifier:

1. Average cross-validation accuracy: 81.87%
2. Class-specific insights provided for refinement.

XG BOOST



1. Accuracy: 82.86%

2. Varied performance across classes, suggesting further tuning or additional data.

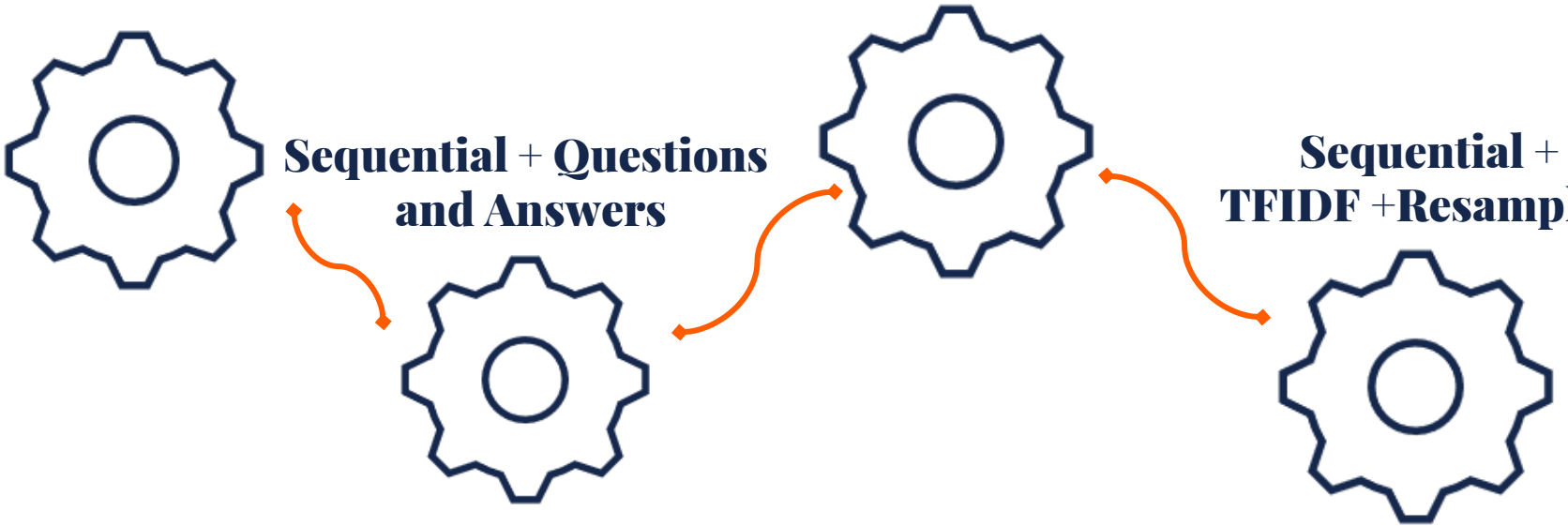
DEEP LEARNING MODELS

Base – Sequential Model

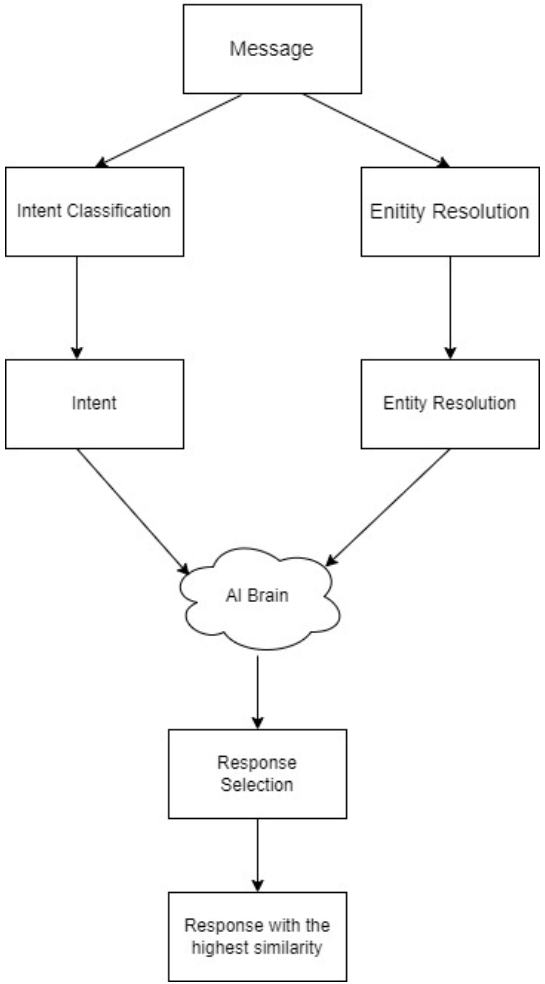
Sequential + Questions and Answers

Sequential + Regularization

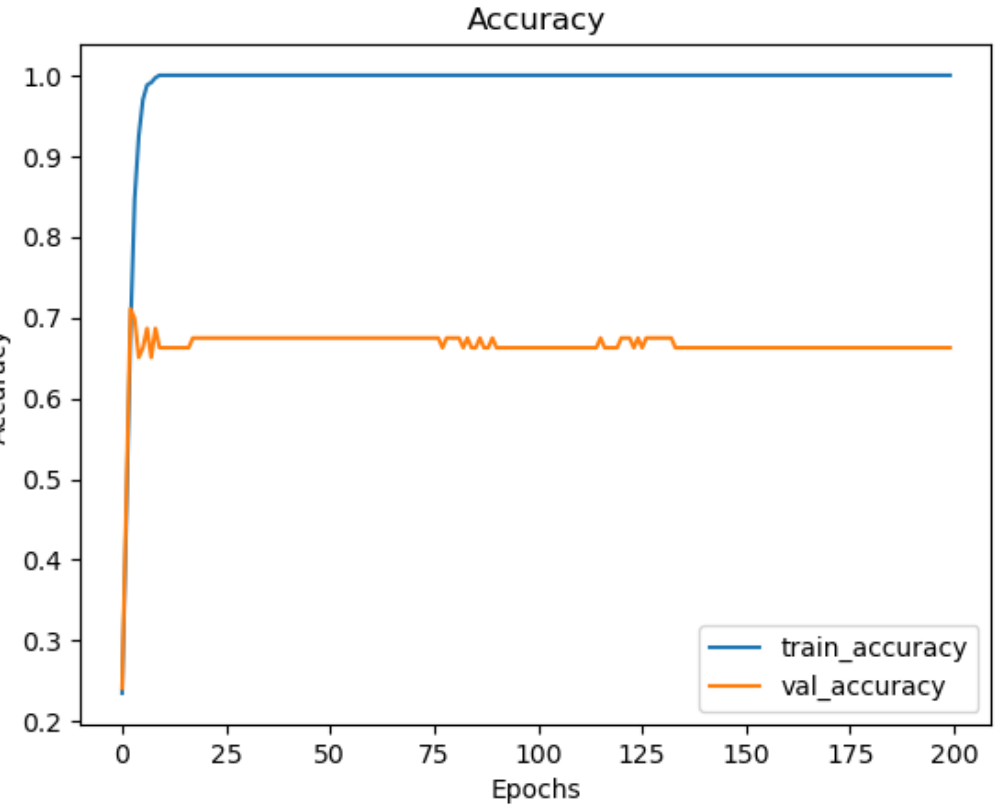
Sequential + TFIDF +Resampling



ARCHITECTURE



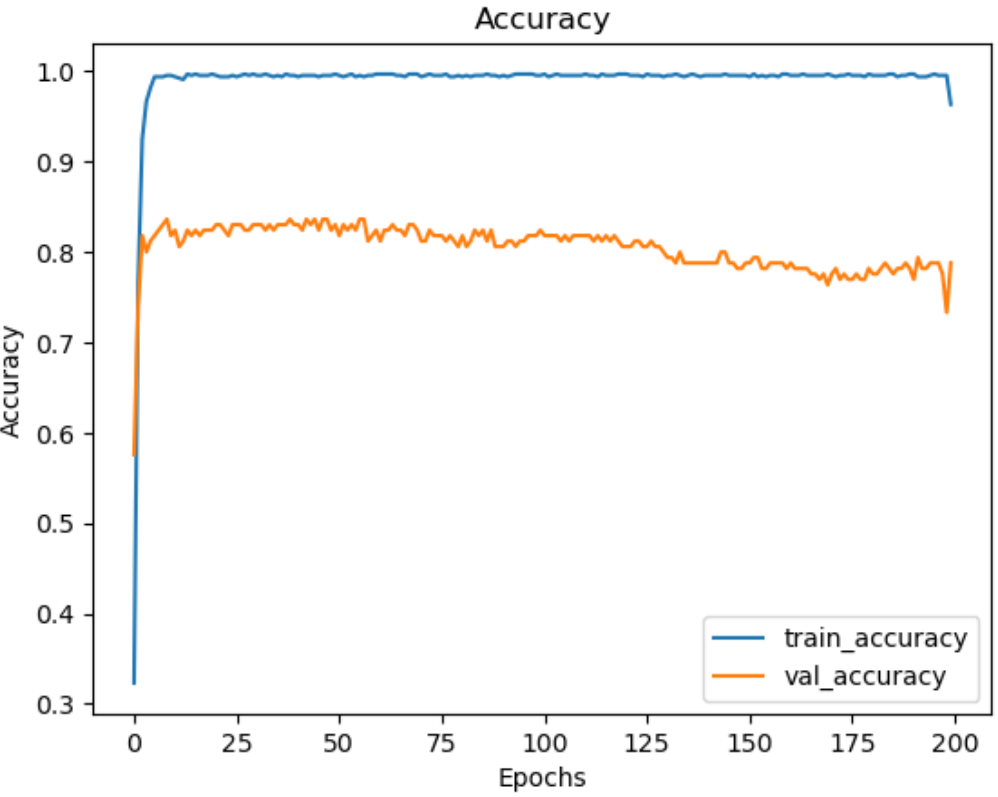
BASE - SEQUENTIAL



The base model demonstrates a high training accuracy of approximately 98%, suggesting effective learning from the training data.

However, a significant performance gap is observed, with the validation accuracy plateauing at 30-40%, raising concerns about overfitting

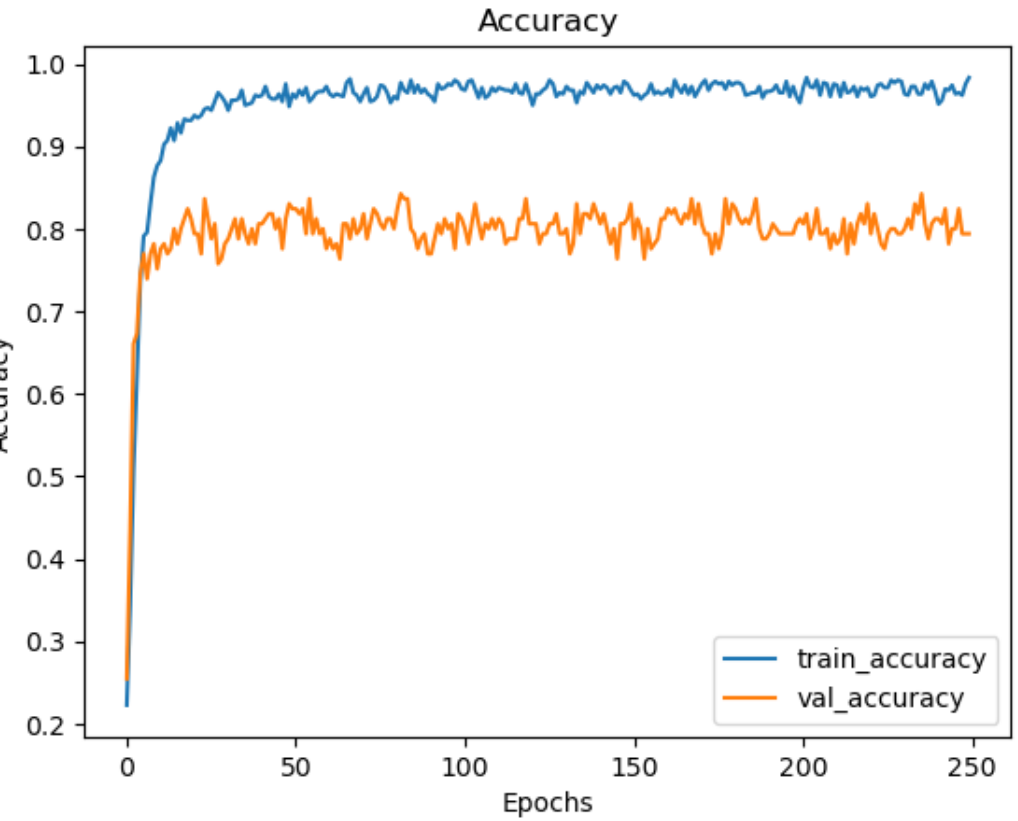
SEQUENTIAL + QNA



The model exhibits notable improvement, with training accuracy at approximately 99.42% and validation accuracy at around 58.78%.

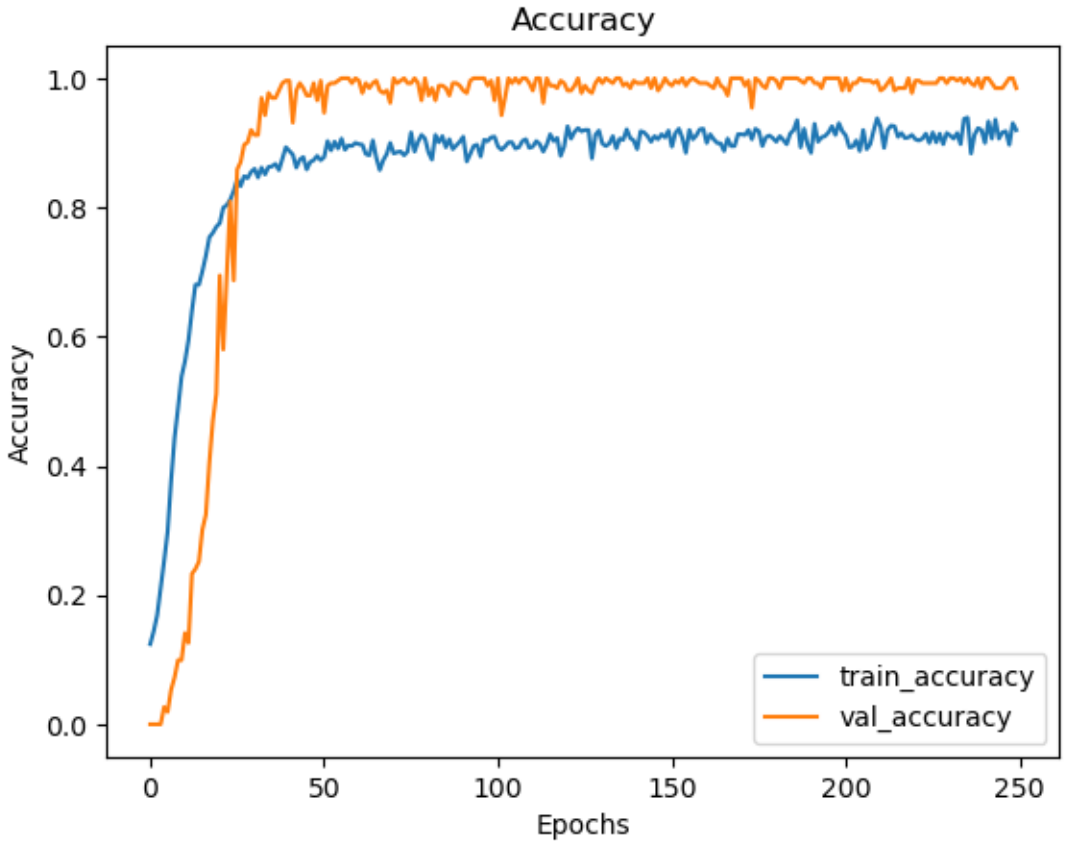
Despite this progress, there is potential for further enhancement in the validation accuracy.

SEQUENTIAL + REGULARIZATION



The gap between training and validation scores is not drastic, indicating that the regularization and dropout are having a positive effect.

SEQUENTIAL + TFIDF, RESAMPLING



The model achieved high accuracy on the training set, with the training accuracy steadily increasing and reaching nearly 1.0

Further steps may include using more advanced models

CONCLUSION & RECOMMENDATIONS

1.Intent Classification:
The intent classification model achieved good performance, accurately categorizing user queries into predefined intent categories. This model can effectively route user queries to appropriate response handlers or workflows.

2. Response Generation:
The response generation system, utilizing cosine similarity with TF-IDF vectors, showed promising results in mapping user queries to appropriate responses. However, there is room for improvement, particularly in handling context and generating diverse responses.

3. Continuous Learning and Adaptation:
Implement mechanisms for continuous learning and adaptation, allowing the chatbot to evolve over time based on user feedback, new data, and changing requirements.
Monitor performance metrics and user satisfaction to measure the effectiveness

Interact with TAWI at:
https://tawi-chatbot.onrender.com/

THANKYOU!