Improving University Customer Support with Learning to Rank Chatbot

Subject: BIA 660-A

Group D:

Han Chau – 20012654

Manvendra Shrimal - 20011277

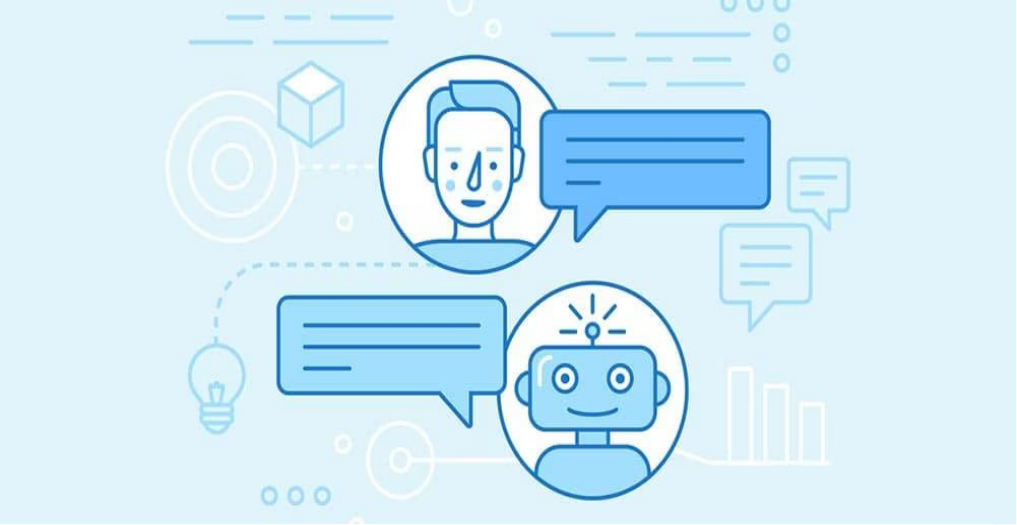
Yashita Vajpayee - 20011237

Content

- Problem and research question
- Data collection and pre-processing
- Method and results
  - Model 1 – SVM
  - Model 2 – RF & Bidirectional LSTM neural network architecture
- Insights and analysis

Problem and research question

- How can we effectively classify questions into different categories?
- How can we identify the most suitable answer for a given question?

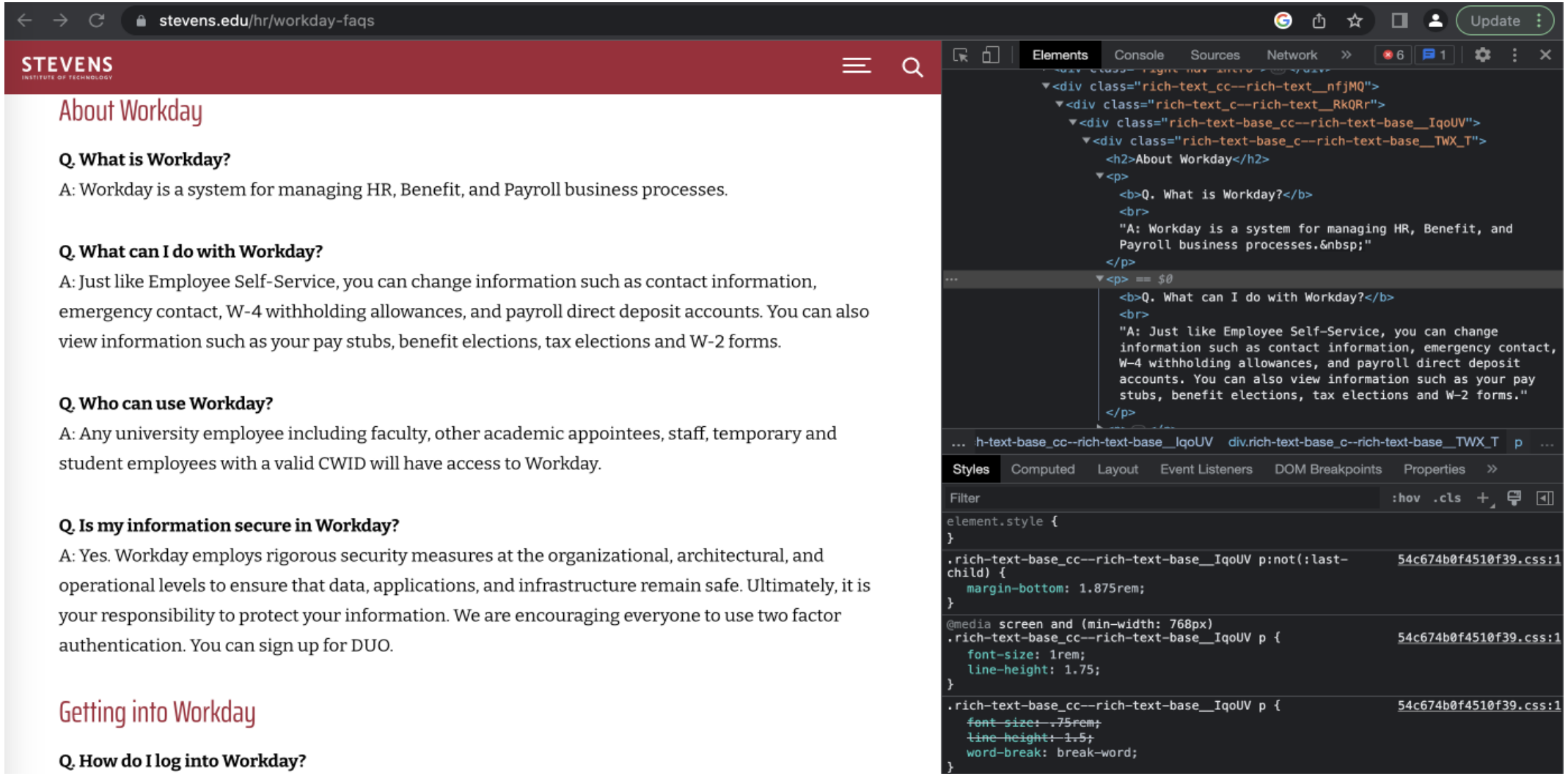


Source: https://www.hp.com/us-en/shop/tech-takes/what-is-a-chatbot

Data collection

- We scraped stevens website for FAQs from 24 different web pages of Stevens
- We also obtained a dataset from the Business School that contains questions, corresponding answers, and pre-assigned categories.
- The dataset has been preprocessed, including renaming columns and cleaning text data to remove unnecessary characters and stop words.

Data collection



Data collection

- Our dataset after collecting

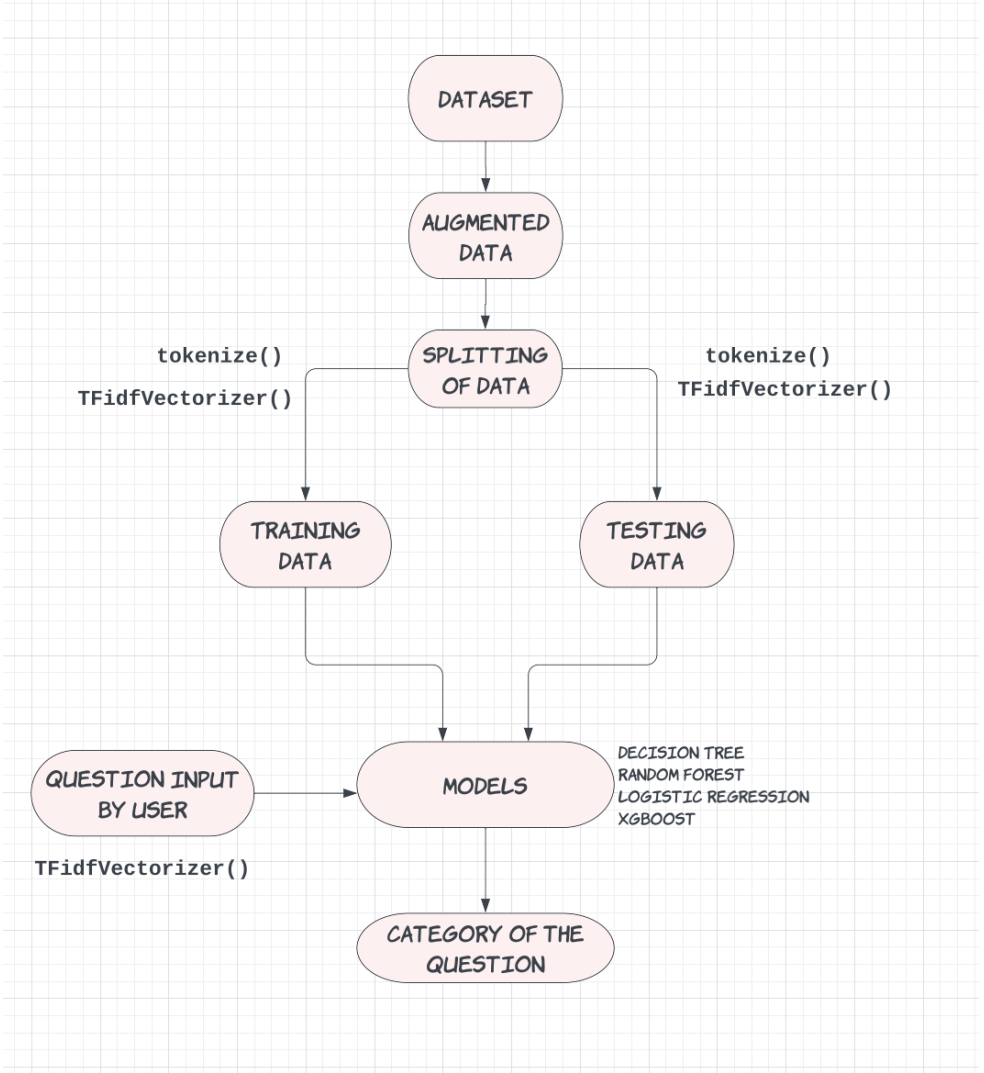
	Category	Questions	Answers
0	Housing and Dining	When will students find out about their housin...	Returning students (all students other than ne...
1	Housing and Dining	How long are the beds? What kind of sheets wil...	All the beds in the residence halls are xl twi...
2	Housing and Dining	Are pets allowed?	No. With the exception of documented service a...
3	Housing and Dining	Can I bunk my bed?	Yes, our beds are designed to be bunked. If yo...
4	Housing and Dining	Can furniture be removed from the room and rep...	No, the furniture provided may not be removed ...
...	...	...	...
328	Disability Services	If a student is accustomed to being accompanie...	Stevens, like most other universities, does no...
329	Disability Services	How are Accommodations determined?	Accommodations are developed for students on a...
330	Disability Services	If my accommodations for a course were sent to...	You should notify the ODS (disabilityservices@...
331	Disability Services	Since I am not sure where all my classes are l...	Yes, the ODS is happy to provide campus walkth...
332	Disability Services	Can I request additional time with a tutor if ...	Yes, students registered with the ODS can cont...

333 rows × 3 columns

Data pre-processing

- Using Data augmentation method to increase the size of dataset.
  - Replace letter
  - Delete letter
  - Add letter
- Creating some functions to extract token, and the cosine similarity between the paired question and answer by tokenize(), compute_tfidf(), and assess_similarity()

Model for predicting category:



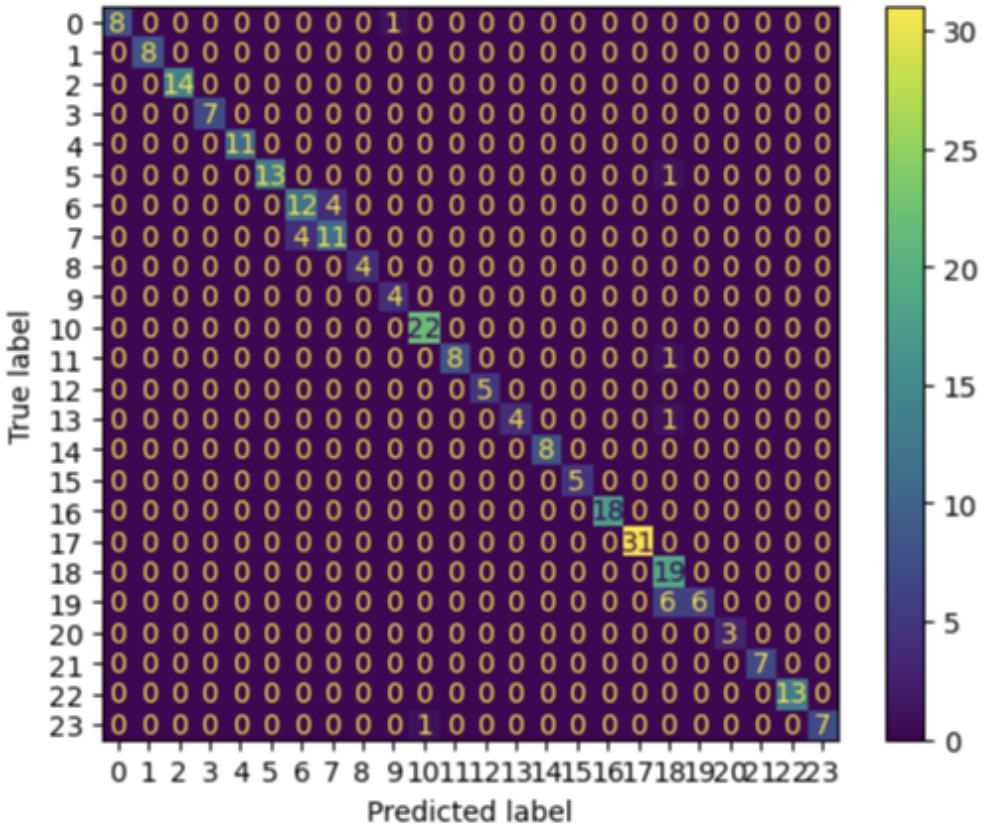
Model for predicting category:

- Create the dataset with 2 features: Category and Questions.
- Using this dataset training the model to predict category for question.
- Using cross_val_score() to find the best model
  among Decision Tree, Random Forest,Logistic Regression and XGBoost.

```
Average accuracy for Decision Tree model: 99.18%
Standard deviation of accuracy: 0.87%
---------------
Average accuracy for Random Forest model: 99.54%
Standard deviation of accuracy: 0.63%
---------------
Average accuracy for Logistic Regression model: 99.28%
Standard deviation of accuracy: 0.93%
---------------
Average accuracy for XGBoost model: 99.23%
Standard deviation of accuracy: 0.84%
```

	precision	recall	f1-score	support
0	1.00	0.89	0.94	9
1	1.00	1.00	1.00	8
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	7
4	1.00	1.00	1.00	11
5	1.00	0.93	0.96	14
6	0.75	0.75	0.75	16
7	0.73	0.73	0.73	15
8	1.00	1.00	1.00	4
9	0.80	1.00	0.89	4
10	0.96	1.00	0.98	22
11	1.00	0.89	0.94	9
12	1.00	1.00	1.00	5
13	1.00	0.80	0.89	5
14	1.00	1.00	1.00	8
15	1.00	1.00	1.00	5
16	1.00	1.00	1.00	18
17	1.00	1.00	1.00	31
18	0.68	1.00	0.81	19
19	1.00	0.50	0.67	12
20	1.00	1.00	1.00	3
21	1.00	1.00	1.00	7
22	1.00	1.00	1.00	13
23	1.00	0.88	0.93	8
accuracy			0.93	267
macro avg	0.95	0.93	0.94	267
weighted avg	0.94	0.93	0.93	267

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fc3e8ba0250>



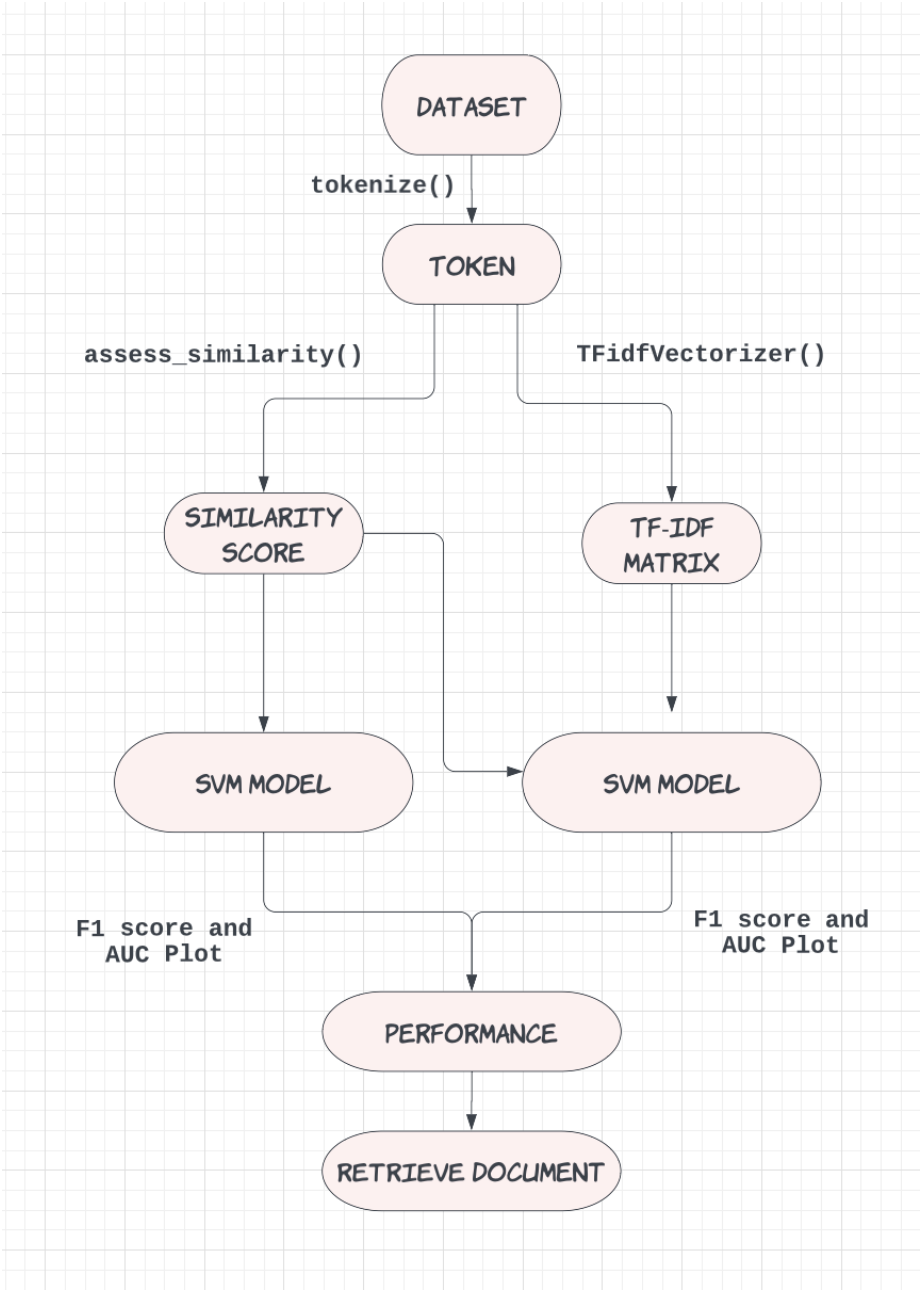
Let's test our Category Detection Model by asking different Questions

```
#The category for Below Question is 'Admissions', let's see what our model predicts ?
# Q.1

question = ["What happens if I miss the application deadline?"]
question_tfidf = tfidf.transform(tokenize(question, lemmatized=True, remove_stopword=Tru

# # Decode the predicted label for XGBoost model
predicted_label_xgb = model_xgb.predict(question_tfidf)
predicted_label_xgb = label_encoder.inverse_transform(predicted_label_xgb)

print("Logistic Regression prediction:", model_log.predict(question_tfidf))
print("Decision Tree prediction:", model_tree.predict(question_tfidf))
print("Random Forest prediction:", model_ran.predict(question_tfidf))
print("XGBoost prediction:", predicted_label_xgb)
```

```
Logistic Regression prediction: ['ADMISSIONS']
Decision Tree prediction: ['ADMISSIONS']
Random Forest prediction: ['ADMISSIONS']
XGBoost prediction: ['ADMISSIONS']
```

Inference- Looks good ! All models pick the same right category.

Model 1 - SVM

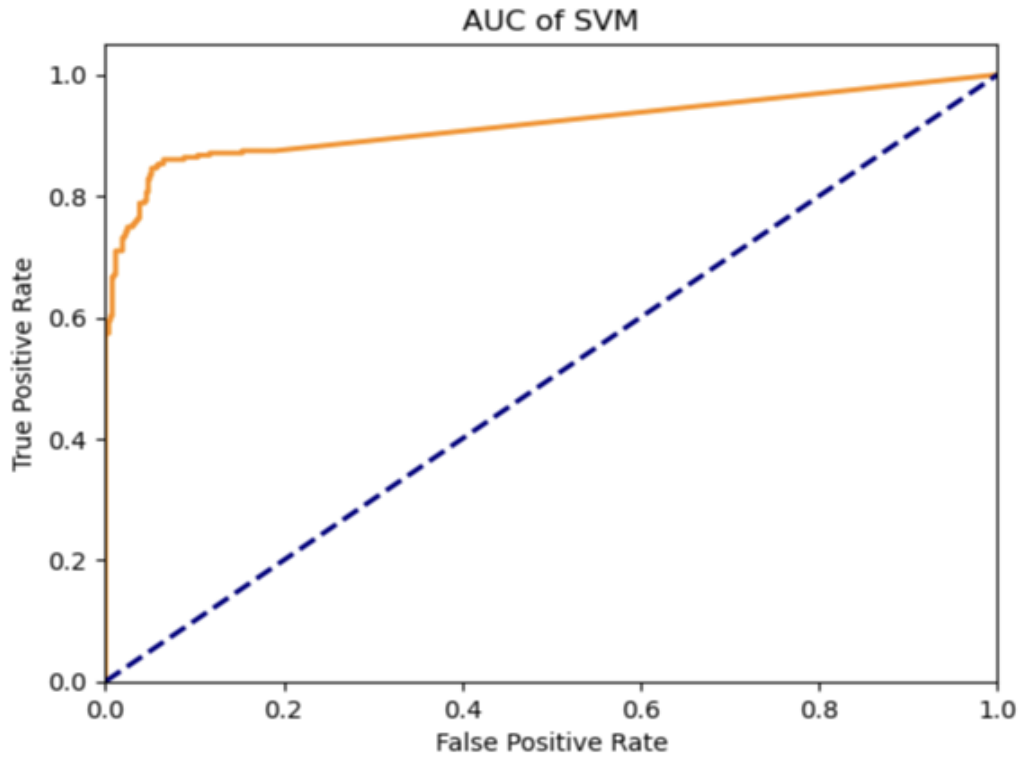

DATASET
tokenize()
TOKEN
assess_similarity()
TFidfVectorizer()
SIMILARITY SCORE
TF-IDF MATRIX
SVM MODEL
SVM MODEL
F1 score and AUC Plot
F1 score and AUC Plot
PERFORMANCE
RETRIEVE DOCUMENT

Model 1 - SVM

- Using only the cosine similarity to train the model.

	precision	recall	f1-score	support
0	0.89	0.96	0.93	521
1	0.91	0.79	0.85	279
accuracy			0.90	800
macro avg	0.90	0.87	0.89	800
weighted avg	0.90	0.90	0.90	800

AUC: 91.67% , PRC: 92.25%



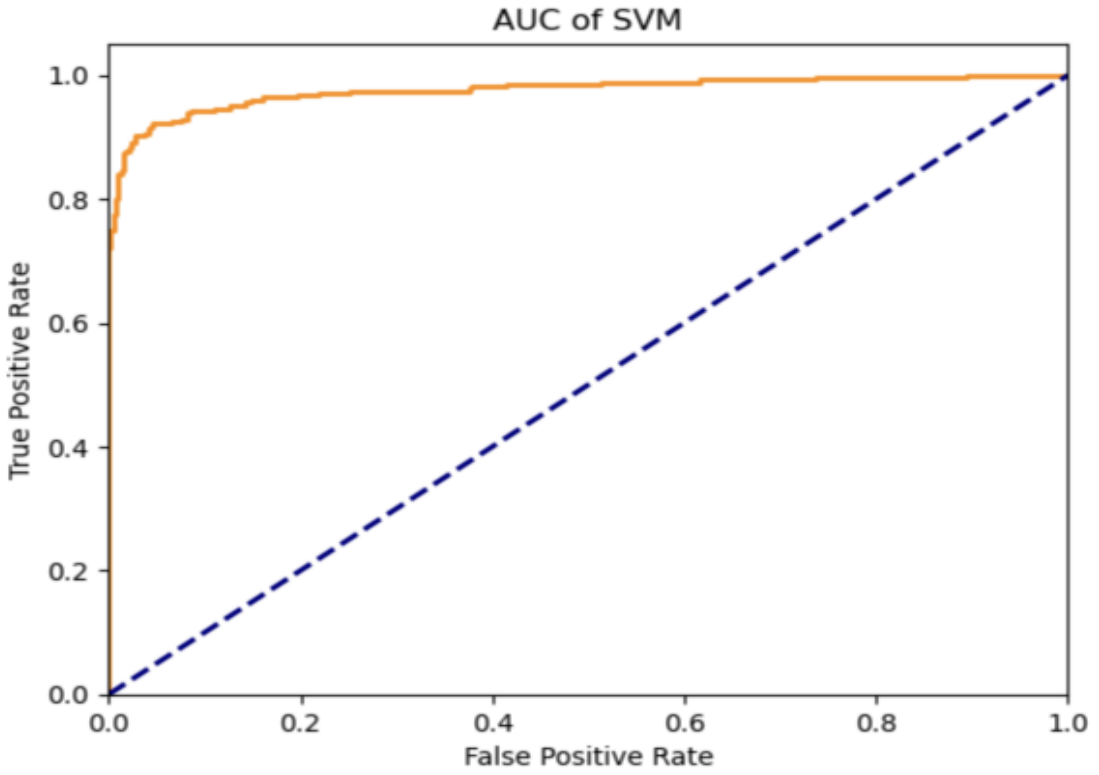
Model 1 - SVM

- Combine Tf-idf matrix and the cosine similarity to train the model.

```
tfidf_vector_1,clf_1=svm_model(X_train,Y_train,X_test,Y_test,\
                               remove_stopword=True,bigram = False)
```

	precision	recall	f1-score	support
0	0.93	0.98	0.96	521
1	0.97	0.87	0.92	279
accuracy			0.94	800
macro avg	0.95	0.93	0.94	800
weighted avg	0.95	0.94	0.94	800

AUC: 97.61% , PRC: 97.10%



Model 1 - SVM

- Create a new data to check the accuracy when retrieve document for questions.

	Category	Question Collect	Question Chat GPT	Answers
0	CPT	What is the difference between CPT and OPT?	What are the differences between CPT and OPT?	In short, CPT is employment that is directly r...
1	Seeking Help Off-Campus	Is it really private and secure?	How secure and confidential is the service?	Yes, we are HIPAA and FERPA compliant which me...
2	Seeking Help Off-Campus	Can I continue working with my counselor once ...	Is it possible to continue working with my cou...	Yes, you will have the option to continue work...
3	CPT	My internship ended early. What documents shou...	How should I report an early end date for my i...	If your CPT employment is ending or ended prio...
4	Innovation Expo	Is registration required to attend?	Is registration mandatory for attending the Expo?	Registration is NOT required, but setting up I...

Model 1 - SVM

- Calculate the accuracy when retrieve document following the top-1, top-3, and top-5 answers.

```
score = test_retrive(final_dataset["Question Chat GPT"], top =1,\
             remove_stopword=False,bigram = False,change_para=True
print("Accuracy for the top 1 is ", sum(score)/len(score)*100)
```

```
Accuracy for the top 1 is  41.0
```

```
score = test_retrive(final_dataset["Question Chat GPT"], top =1,\
             remove_stopword=True,bigram = False,change_para=False)
print("Accuracy for the top 1 is ", sum(score)/len(score)*100)
```

```
Accuracy for the top 1 is  48.0
```

```
score = test_retrive(final_dataset["Question Chat GPT"], top =3,\
            remove_stopword=False,bigram = False,change_para=True)
print("Accuracy for the top 3 is ", sum(score)/len(score)*100)
```

```
Accuracy for the top 3 is  63.0
```

```
score = test_retrive(final_dataset["Question Chat GPT"], top =3,\
            remove_stopword=True,bigram = False,change_para=False)
print("Accuracy for the top 3 is ", sum(score)/len(score)*100)
```

```
Accuracy for the top 3 is  67.0
```

```
score = test_retrive(final_dataset["Question Chat GPT"], top =5,\
            remove_stopword=False,bigram = False,change_para=True)
print("Accuracy for the top 5 is ", sum(score)/len(score)*100)
```

```
Accuracy for the top 5 is  75.0
```

```
score = test_retrive(final_dataset["Question Chat GPT"], top =5,\
            remove_stopword=True,bigram = False,change_para=False)
print("Accuracy for the top 5 is ", sum(score)/len(score)*100)
```

```
Accuracy for the top 5 is  75.0
```

=> The accuracy will be better if we remove stop-words.

Model 1 - SVM

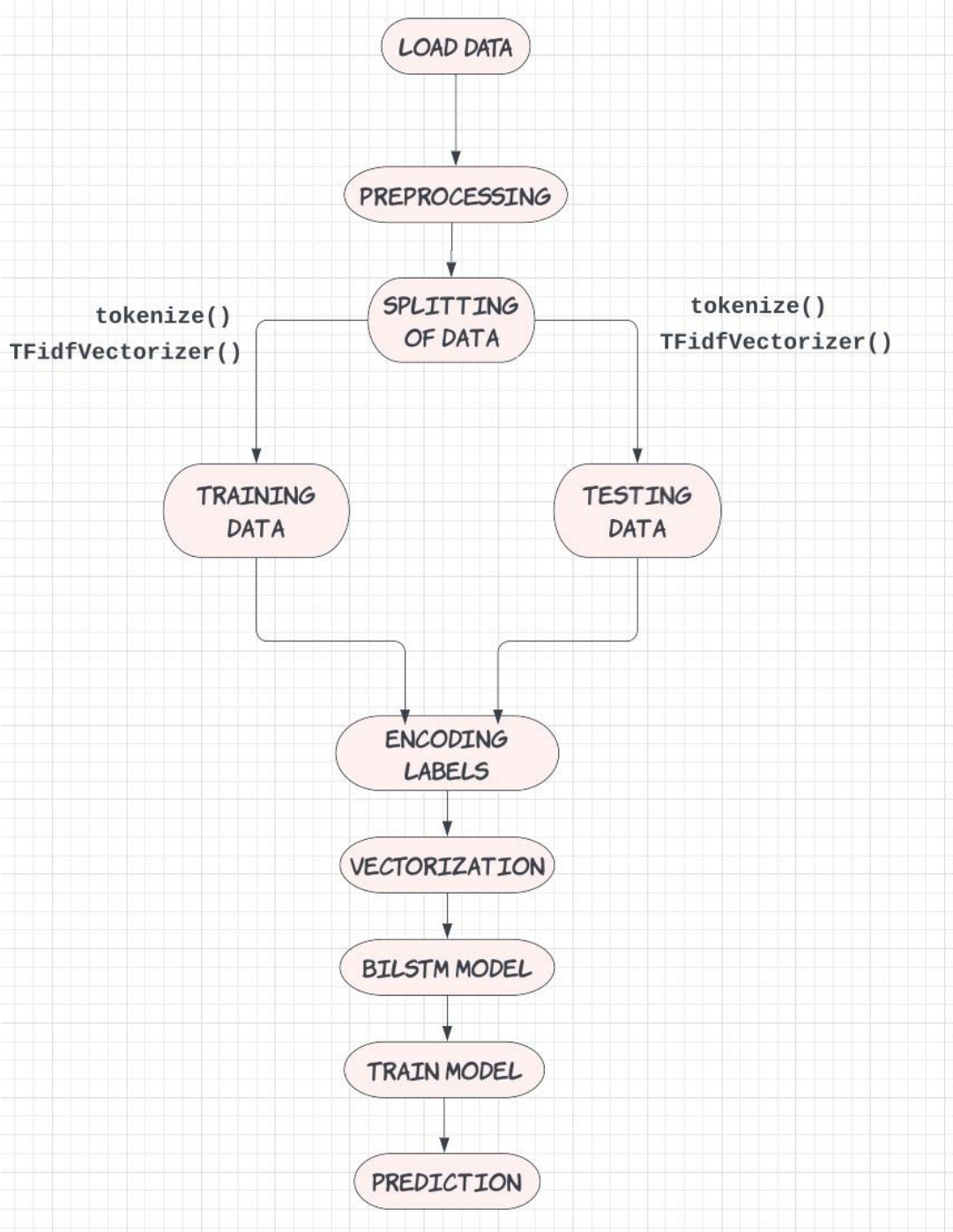
- Visualizing about retrieve the top-3 document for a query.

[442]:
```
# try with retrive top-3 answers for question below
# If I have questions about Orientation, whom should I contact?
text = input("Hi, how can I help you?\n")
print("\n- The retrived doccuments:\n")
retrive_doc(text)
```

```
Hi, how can I help you?
 If I have questions about Orientation, whom should I contact?

- The retrived doccuments:

1. For all questions related to Orientation, please email student_life@stevens.edu

2. Each student will have access to our mobile app through Guidebook to view the Orientation Schedule.

3. Each student will receive information regarding new student arrival for Orientation in the summer. Residential Stu
dents: For residential students participating in Pre-Orientation, those students will move in on Thursday, August 24.
Specific move-in times will be communicated closer to the date. For residential students not participating in Pre-Ori
entation, those students will move in on Sunday, August 27 when Orientation begins. Specific move-in times will be co
mmunicated closer to the date.Commuter Students: For commuter students participating in Pre-Orientation, those studen
ts will arrive when their program begins on Thursday, August 24. Please note that there are not on-campus accommodati
ons for commuter students during Pre-Orientation. It is expected that students will travel to/from campus each day du
ring Pre-Orientation. For commuter students not participating, those students will arrive on campus on Sunday, August
27 when orientation begins. Specific arrival information for commuter students will be communicated closer to the dat
e.
```

Model 2 - BiLSTM Question-Answering Model

Run time – 2 min.



Model 2 - BiLSTM Model

- Problem Description: Our goal is to develop a fast question-answering system for a Chatbot that can accurately classify and provide answers to various questions based on a user's input

	CATEGORY	Question	Answer	Label
0	ADMISSIONS	What happens I if I miss the application daedl...	Your application to the BI&A program may be ap...	1
1	ADMISSIONS	How do I apply for the MS degree?	Apply online at the Graduate Admissions web si...	1
2	ADMISSIONS	Can I take BI&A form without a GRE or GMAT score?	Yes. No more than three graduate cerrtificate ...	1
3	ADMISSIONS	Where can i get help on admissions questions?	Contact Graduate Admissions at https://www.ste...	1
4	ADMISSIONS	Can I take BI&A courses without a GRE or GMAT ...	Yes. No more than three graduate cerrtificate ...	1
...	...	...	...	...
2431	STEVENS	What is s. smith stevens like?	Login to https://my.stevens.edu. with your Ste...	0
2432	STEVENS	When comprise Stevens established?	Login to https://my.stevens.edu. with your Ste...	0
2433	STEVENS	When When was Stevens established?	Send an email with a job posting to: To the Pr...	0
2434	STEVENS	When When was Stevens established?	Living costs are hard to estimate. Many stude...	0
2435	STEVENS	XX	Sorry, I don't have an answer to that please r...	1

Model 2 - BiLSTM Model

In [48]:
```
# Q.3
test_question_3 = "How good is placement?"
predicted_answer_3 = predict_answer_model1(test_question_3)

print(f"Question: {test_question_3}")
if isinstance(predicted_answer_3, str):
    print(f"Answer: {predicted_answer_3}")
else:
    print("Closest Answers:")
    for answer in predicted_answer_3:
        print(f"- {answer}")
```

```
Question: How good is placement?
Answer: Stevens Institute of Technology is the only school in the country to be ranked in the Top 20 for bo
Career Placement" (6th in the nation in Colleges that Pay You Back, 2016 edition) and "Best Career Services
in the nation in Best 380 Colleges, 2016 edition) by The Princeton Review.
```

Inference- For above Q.3, output on minor modification in pre-trained question looks good.

```python
# Q.5
test_question_5 = "Web Mining"
predicted_answer_5 = predict_answer_model1(test_question_5)

print(f"Question: {test_question_5}")
if isinstance(predicted_answer_5, str):
    print(f"Answer: {predicted_answer_5}")
else:
    print("Closest Answers:")
    for answer in predicted_answer_5:
        print(f"- {answer}")
```

```
Question: Web Mining
Closest Answers:
- All of the courses in the curriculum are relevant to data science. In particular there are two courses in data mi
ning and Machine Learning: MIS 637 Knowledge Discovery in Databases and BIA 656 Statistical Learning & Analytics.
- The 18 companies represented on the programs' Industry Advisory Board support the program by providing advice and
directions for our curriculum, and internships and full-time jobs for our students. Hundreds of other companies  ac
tively recruit our students and many provide visiting speakers. For more information see: https://www.stevens.edu/s
chool-business/masters-programs/business-intelligence-analytics/board-advisors
- The Hanlon Lab offers students and faculty access to real-time data feeds from leading providers of financial inf
ormation, such as Bloomberg, RealTick, Thomson Reuters and Gain Capital. Software providers and partners include De
cide-FS, Tripwire, Redseal and Fortify. Several BI&A Advisory Board companies provide opportunities for students to
work on their prorietary data sets.
```

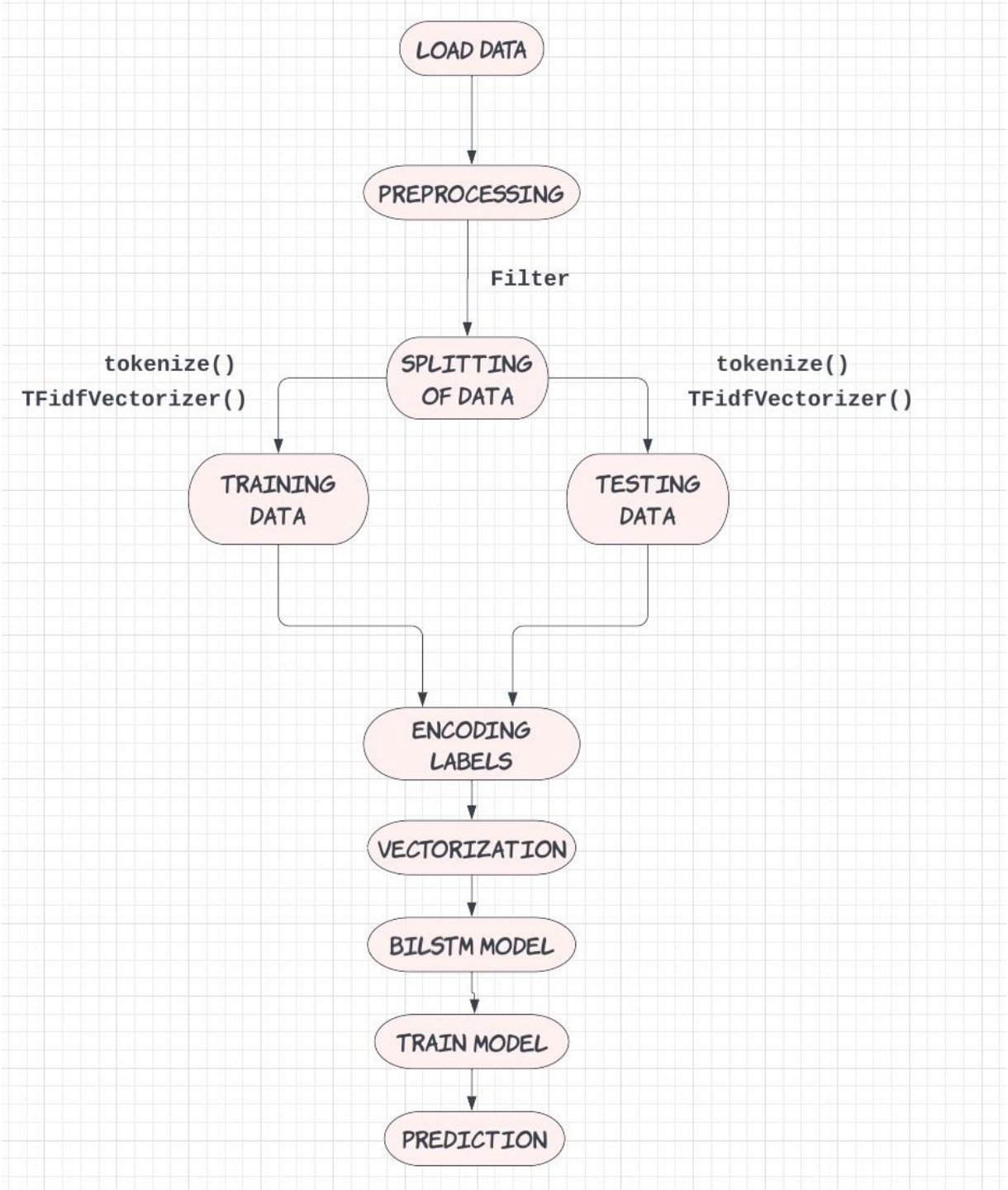
```python
# Q.5
test_question_5 = "Group D"
predicted_answer_5 = predict_answer_model1(test_question_5)

print(f"Question: {test_question_5}")
if isinstance(predicted_answer_5, str):
    print(f"Answer: {predicted_answer_5}")
else:
    print("Closest Answers:")
    for answer in predicted_answer_5:
        print(f"- {answer}")
```

```
Question: Group D
Answer: Sorry, I don't have an answer to that please reach us at phone:877.376.9534 email:online@stevens.edu
```

Inference- For above Q.5, when the predicted result doesn't pass the required threshold it is trained to give a generalised response.

Model 3 – BiLSTM+Category Identification Question-Answering Model

Run time – 5 sec.



Conclusion -

•Overall, the models' performance demonstrates the potential for automating question categorization and answering in a business school context.

•The SVM, Random Forest model performed well in category detection, while the SVM and BiLSTM model showed effectiveness in predicting answers based on question similarity.

• Further improvements can be made by considering keyword importance and handling cases where no suitable category or answer is found.

Thank you