uOttawa

Faculté de génie
Faculty of Engineering

DTI5125: Data Science Applications

Final Project

Group 8

Table of Contents

Table of Figures:

1. Problem Formulation and system overview:

Nowadays, there are many books in many different fields, making it difficult to locate a decent book that is appropriate for the reader. So, we use a chatbot to recommend a book based on collaborative filtering. First, the chatbot recommends a random book for the user. Secondly, the user responds with a book review. Then, the chatbot calls API using the URL provided by NGROK to access the API running in Google Colab to classify the user's review into "Positive", "Neutral", or "Negative". Lastly, using the book name, user review, history reviews of other users, we got the recommendation and return it back to the user as shown in Figure 1.

The classification part is responsible for classifying the reviews into three classes ("Positive", "Neutral", "Negative"). The clustering part is responsible for clustering similar books based their descriptions; it can be used with the chatbot as a future work. The recommender part proposes five different ways to recommend the books (only the last way used with the chatbot). Dialogflow is used to develop the chatbot.

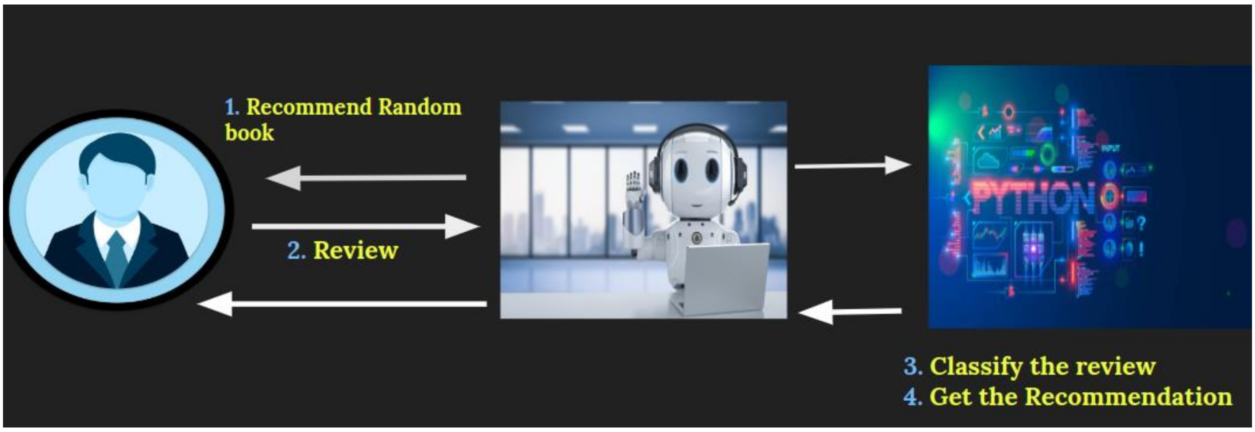


Figure 1 The overall sequence of the system

2. Dataset:

The used dataset is book reviews from Amazon [1]. Google books API is used to get the descriptions of the books that are needed by clustering.

3. Classification:

In this part, book reviews data from Amazon is used. The reviews of the books will be classified into three classes ("Positive", "Neutral", "Negative") based on user rating. The steps are as follows:

a) Data Preparation:

	Unnamed: 0	product/productId	product/title	product/price	review/userId	review/profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	0	1882931173	Its Only Art If Its Well Hung!	unknown	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	1	0826414346	Dr. Seuss: American Icon	unknown	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	2	0826414346	Dr. Seuss: American Icon	unknown	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	3	0826414346	Dr. Seuss: American Icon	unknown	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka "D...
4	4	0826414346	Dr. Seuss: American Icon	unknown	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...

Figure 2 The data set before preparation

1. One million reviews are taken from the dataset.
2. Removing extra columns. Seven columns out of eleven columns are removed.

	product/title	review/userId	review/score	review/text
0	Its Only Art If Its Well Hung!	AVCGYZL8FQQTD	4.0	This is only for Julie Strain fans. It's a col...
1	Dr. Seuss: American Icon	A30TK6U7DNS82R	5.0	I don't care much for Dr. Seuss but after read...
2	Dr. Seuss: American Icon	A3UH4UZ4RSVO82	5.0	If people become the books they read and if "t...
3	Dr. Seuss: American Icon	A2MVUWT453QH61	4.0	Theodore Seuss Geisel (1904-1991), aka "D...
4	Dr. Seuss: American Icon	A22X4XUPKF66MR	4.0	Philip Nel - Dr. Seuss: American IconThis is b...

Figure 3 The dataset after removing extra columns

3. Renaming columns.
4. Removing the users that don't have userID.
5. Removing duplicated rows.
6. Constructing the "class" column that have the class of the review ("Positive", "Neutral", "Negative").
7. Choosing balanced sample from the data (7000 row for each class).
8. Cleaning the reviews text by removing the digits, extra spaces, and special characters.

	title	userID	rating	review	class	review_words
0	If Israel Lost the War	A3A88NM6H1PRT5	1.0	why would anyone be interested in a book with ...	Negative	[why, would, anyone, be, interested, in, a, bo...
1	Manifold: Time	A1AE7AMOBC88L8	1.0	i was really distracted by the constant use of...	Negative	[i, was, really, distracted, by, the, constant...
2	The Devil Wears Prada	A90NXCIR72YZ9	1.0	this is the worst waste of ink in a long time ...	Negative	[this, is, the, worst, waste, of, ink, in, a, ...
3	Heart of the Trail: The Stories of Eight Wagon...	A3PBQUBQNC5S7T	1.0	this book was a big disappointment for me pers...	Negative	[this, book, was, a, big, disappointment, for,...
4	Caesar, the Gallic War	A2HCF28X296QA2	2.0	this review concerns the kessinger publishing'...	Negative	[this, review, concerns, the, kessinger, publi...

Figure 4 The dataset after preparation

b) Feature Engineering:

We used two types of feature Engineering as follows:

1. TF-IDF.
2. LDA. To determine the optimal number of topics, coherence measure is used to determine the optimal value for topic numbers.

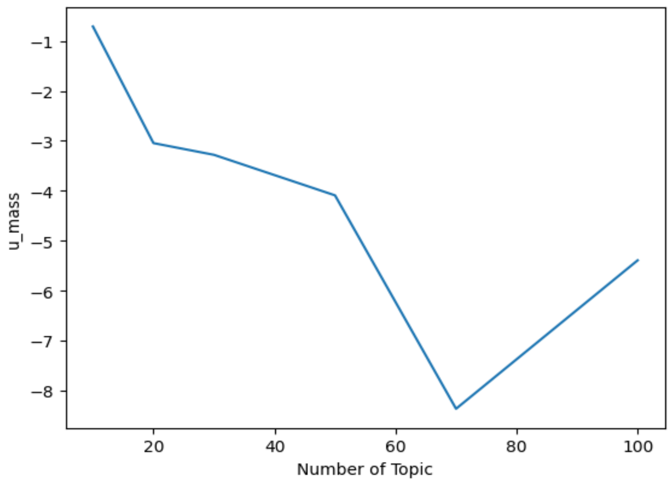


Figure 5 Coherence of LDA features using different number of topics

As the above figure shows the best value is 70 topics.

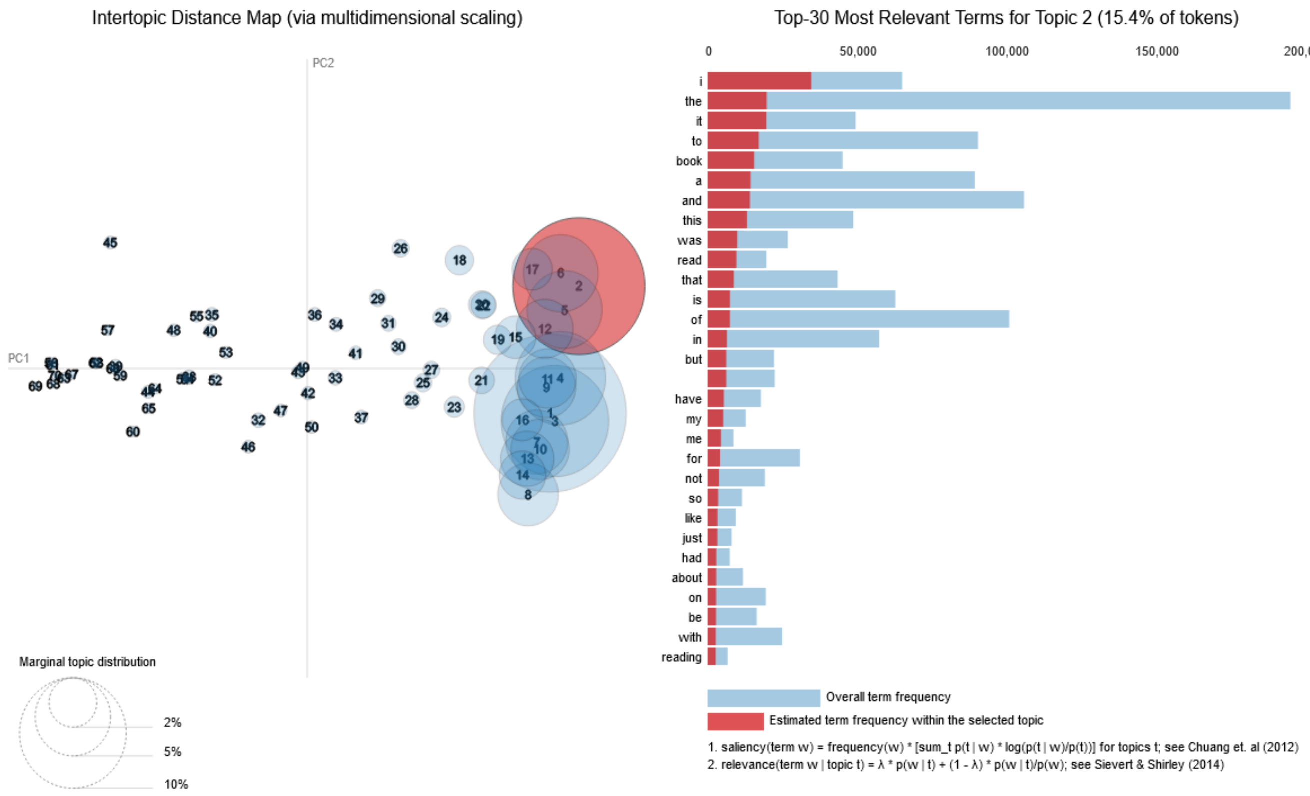


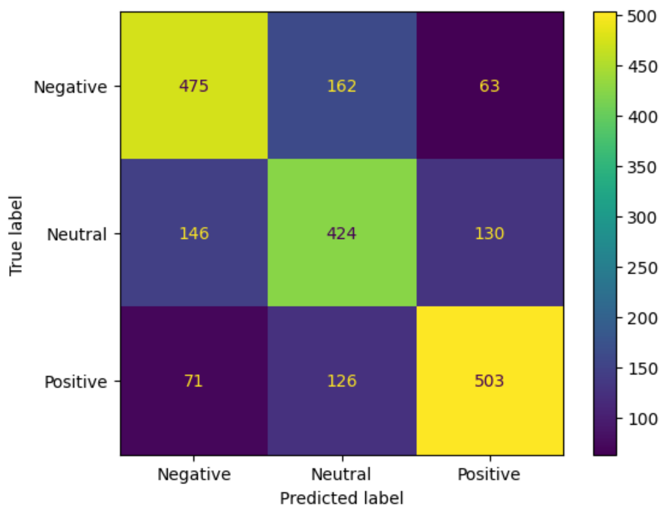
Figure 6 The frequency of top 30 words in topic 2

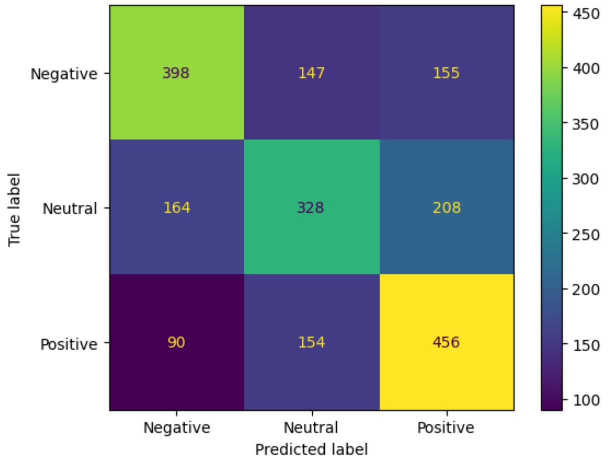
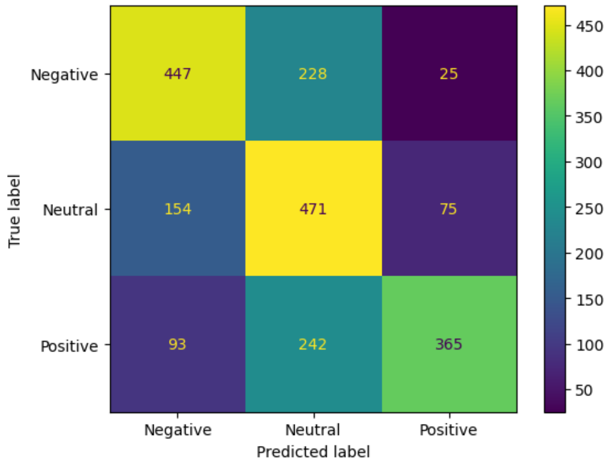
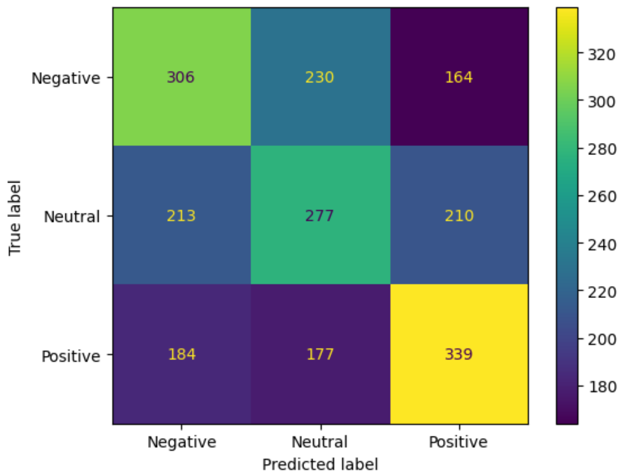
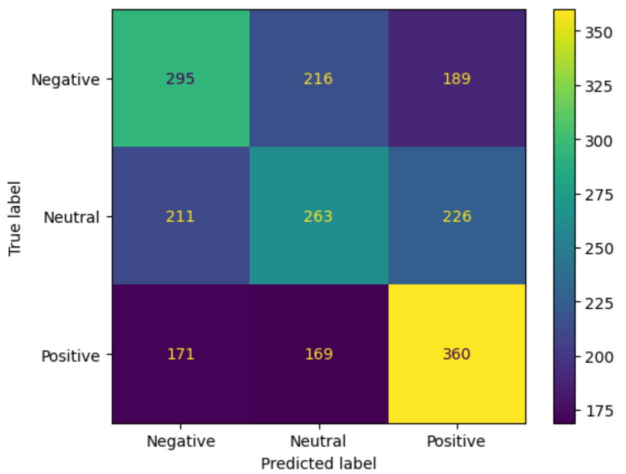
We visualized the results as shown in the above figure to show the top 30 words in each topic.

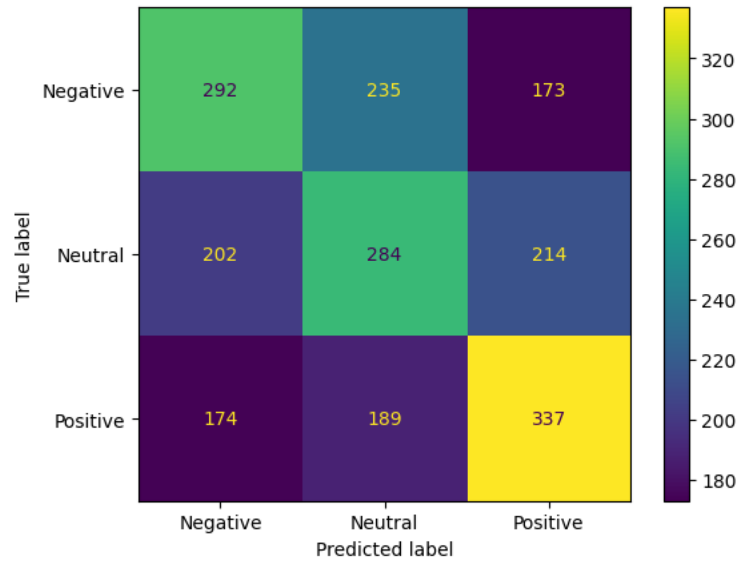
c) Models:

In this step we train multiple models (logistic regression, RandomForest Classifier, naive bayes) each model will be trained with each feature engineering techniques (TF-IDF and LDA).

d) Results:

Model	Confusion Matrix on validation data	Train Accuracy	Validation Accuracy
TF_IDF and logistic regression		84.94%	66.76%

<table>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>TF_IDF and RandomForest Classifier</td><td>
</td><td>74.60%</td><td>56.28%</td></tr>
<tr><td>TF_IDF and naive bayes</td><td>
</td><td>83.01%</td><td>61.09%</td></tr>
<tr><td>LDA and logistic regression</td><td>
</td><td>43.98%</td><td>43.90%</td></tr>
<tr><td>LDA and RandomForest Classifier</td><td>
</td><td>67.69%</td><td>43.71%</td></tr>
</table>

<table>
<tr><td>LDA and naive bayes</td><td>
</td><td>43.73%</td><td>43.47%</td></tr>
</table>

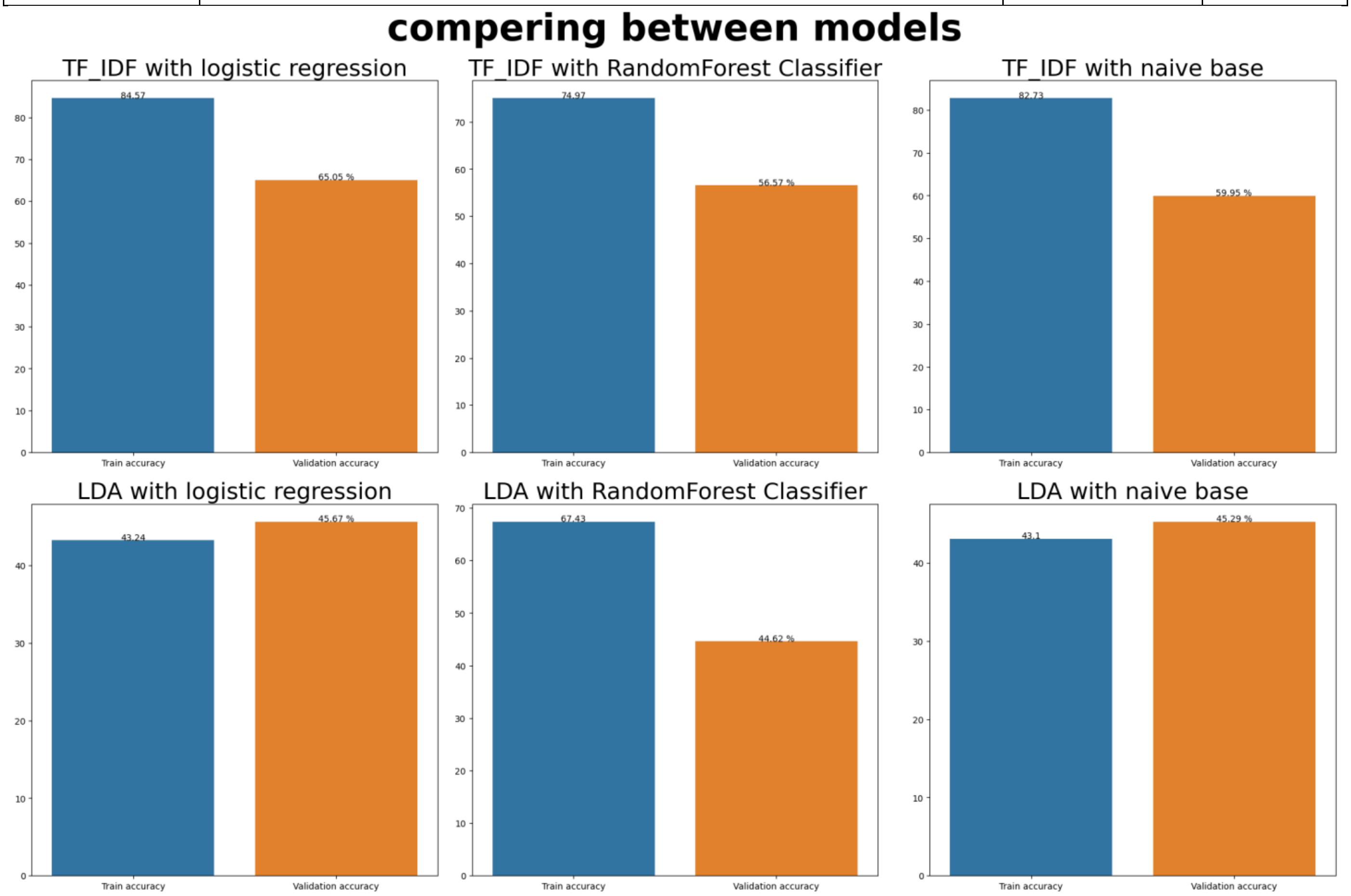


Figure 7 Training and Validation Accuracies of the models

From the above figure the champion model will be Logistic Regression using TF-IDF features (highest validation accuracy).

e) Error Analysis:

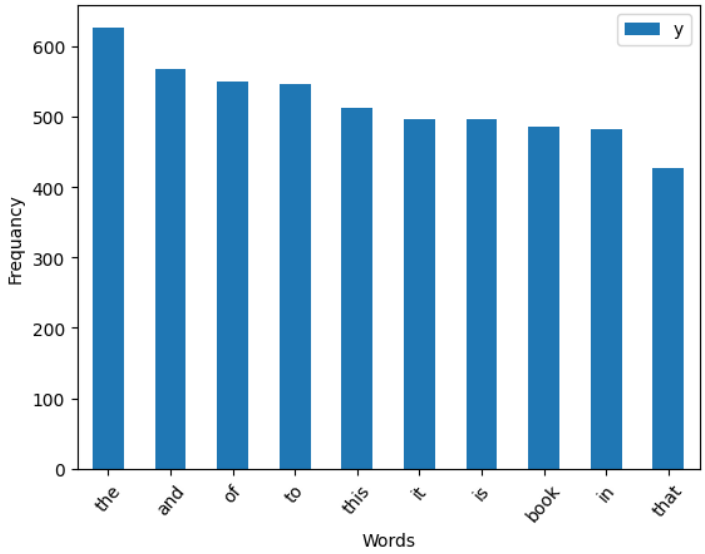
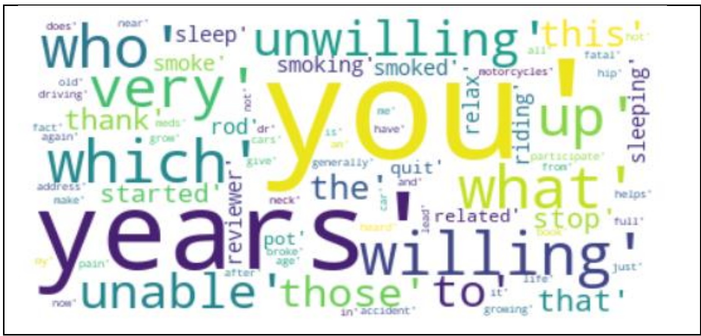
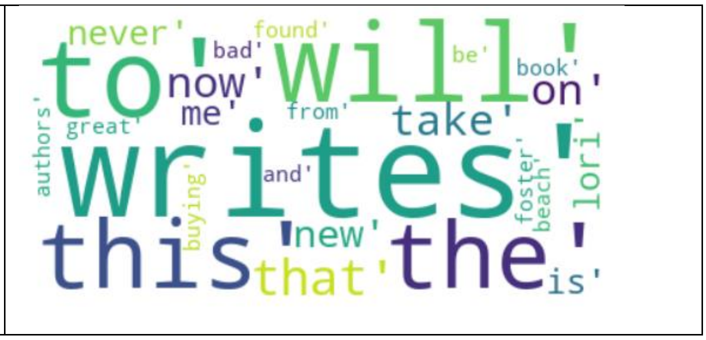


Figure 8 The top 10 words in misclassified reviews

From the above figure, it is concluded that the most frequent words in the misclassified reviews are stop words. So, it seems that removing these words from the beginning may increase the accuracy of the model.





The above figures show the word clouds of two misclassified reviews ("this" is common between them).

4. Clustering:

In this part, book description data that is got using Google Book API is used. The books are clustered based on these descriptions. The steps are as follows:

a) Data Preparation:

1. The Descriptions of ten thousand books are acquired.

	Unnamed: 0	Unnamed: 0.1	Title	description	authors	image	previewLink	publisher	publishedDate	infoLink	categories	ratingsCount
0	0	0	Its Only Art If Its Well Hung!	0 ...	0 [Julie...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=DykPAAAA...	0 NaN\n1 ...	0 1996\n1 2001\n2 2006-08...	0 http://books.google.com /books?id=DykPAAAA...	0 [Comics & Graphic Novels]\n1 [Comics &...	0 NaN\n1 NaN\n2 NaN\n3 3.0\n4 ...
1	1	1	Dr. Seuss: American Icon	0 Philip Nel takes a fascinating look into ...	0 [Philip Nel]\n1 ...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=IjvHQsCn...	0 A&C Black\n1 Bloom...	0 2005-01-01\n1 2004-01-01\n2 2...	0 http://books.google.com /books?id=IjvHQsCn...	0 [Biography & Autobiography]\n1 ...	0 NaN\n1 NaN\n2 NaN\n3 6.0\n4 ...
2	2	2	Wonderful Worship in Smaller Churches	0 This resource includes twelve principles ...	0 [David R. Ray]\n1 ...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=2tsDAAAA...	0 NaN\n1 T...	0 2000\n1 2009-06-01\n2 2010-07...	0 http://books.google.com /books?id=2tsDAAAA...	0 [Religion]\n1 [Religion]\n2 [Religi...	NaN
3	3	3	Whispers of the Wicked Saints	0 Julia Thomas finds her life spinning out ...	0 [Veronica Haddon]\n1 [Em...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=aRSIgJlq...	0 iUniver...	0 2005-02\n1 2019-04-02\n2 2020-04...	0 http://books.google.com /books?id=aRSIgJlq...	0 [Fiction]\n1 [Young Adult ...	0 NaN\n1 14.0\n2 3.0\n3 NaN\n4 ...
4	4	4	Nation Dance: Religion, Identity and Cultural ...	0 Addresses interplay of diverse spiritual,...	0 [Patrick...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=MZUMnoy5...	0 Indiana University Press\n1 Col...	0 2001\n1 2016-04-19\n2 2013-04...	0 http://books.google.com /books?id=MZUMnoy5...	0 [Social Science]\n1 [Philosophy]...	0 1.0\n1 NaN\n2 NaN\n3 NaN\n4 N...

Figure 9 The data set before preparation

2. Removing extra columns. two columns out of twelve columns are removed.

	Title	description	authors	image	previewLink	publisher	publishedDate	infoLink	categories	ratingsCount
0	Its Only Art If Its Well Hung!	0 ...	0 [Julie...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=DykPAAAA...	0 NaN\n1 ...	0 1996\n1 2001\n2 2006-08...	0 http://books.google.com /books?id=DykPAAAA...	0 [Comics & Graphic Novels]\n1 [Comics &...	0 NaN\n1 NaN\n2 NaN\n3 3.0\n4 ...
1	Dr. Seuss: American Icon	0 Philip Nel takes a fascinating look into ...	0 [Philip Nel]\n1 ...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=IjvHQsCn...	0 A&C Black\n1 Bloom...	0 2005-01-01\n1 2004-01-01\n2 2...	0 http://books.google.com /books?id=IjvHQsCn...	0 [Biography & Autobiography]\n1 ...	0 NaN\n1 NaN\n2 NaN\n3 6.0\n4 ...
2	Wonderful Worship in Smaller Churches	0 This resource includes twelve principles ...	0 [David R. Ray]\n1 ...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=2tsDAAAA...	0 NaN\n1 T...	0 2000\n1 2009-06-01\n2 2010-07...	0 http://books.google.com /books?id=2tsDAAAA...	0 [Religion]\n1 [Religion]\n2 [Religi...	NaN
3	Whispers of the Wicked Saints	0 Julia Thomas finds her life spinning out ...	0 [Veronica Haddon]\n1 [Em...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=aRSIgJlq...	0 iUniver...	0 2005-02\n1 2019-04-02\n2 2020-04...	0 http://books.google.com /books?id=aRSIgJlq...	0 [Fiction]\n1 [Young Adult ...	0 NaN\n1 14.0\n2 3.0\n3 NaN\n4 ...
4	Nation Dance: Religion, Identity and Cultural ...	0 Addresses interplay of diverse spiritual,...	0 [Patrick...	0 http://books.google.com /books/content?id=...	0 http://books.google.com /books?id=MZUMnoy5...	0 Indiana University Press\n1 Col...	0 2001\n1 2016-04-19\n2 2013-04...	0 http://books.google.com /books?id=MZUMnoy5...	0 [Social Science]\n1 [Philosophy]...	0 1.0\n1 NaN\n2 NaN\n3 NaN\n4 N...

Figure 10 The dataset after removing extra columns

3. Removing nulls.
4. One thousand books are chosen instead of ten thousand to reduce running time.
5. Cleaning the description text by removing the digits, extra spaces, and special characters.
6. The description column only will be used to cluster the books.

b) Feature Engineering:

We used TF-IDF feature engineering technique to convert descriptions text into numbers.

c) Models:

We used K-means model. WCSS and Silhouette will be used to determine which k would be better.

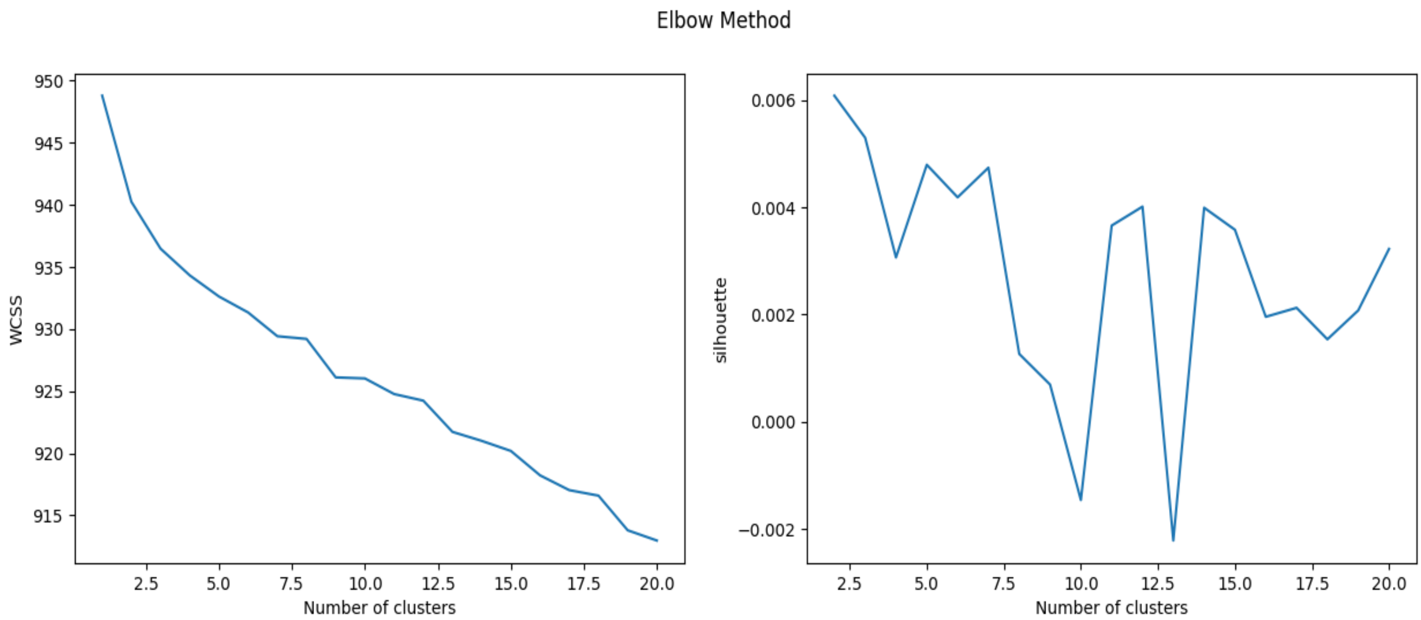


Figure 11 Elbow method with K-means

From the above figure, the chosen **k will be 7**

5. Recommender:

In this part, book reviews data from Amazon is used. We have implemented five different recommendation methods:

1. Popularity Based (Top In the whole collection).
2. Recommendation using Average Weighted Rating.
3. User-Item Collaborative Filtering Recommendation.
4. Recommendation using surprise library (specific to a user).
5. Recommendation using surprise library (not specific to a user).

But we used only the last one to connect it with our chatbot. The steps of recommendations are as follows:

a) Data Preparation:

1. One million reviews are taken from the dataset.
2. Removing extra columns. Seven columns out of eleven columns are removed.
3. Renaming columns.
4. Removing the users that don't have userID.
5. Removing duplicated rows.
6. Constructing the "class" column that have the class of the review ("Positive", "Neutral", "Negative").
7. Choosing balanced sample from the data (all three classes will have the same number of reviews as the minority class).
8. Choosing the most frequent 15 books.
9. Cleaning the reviews text by removing the digits, extra spaces, and special characters.

b) Building the recommender:

1. Popularity Based (Top In the whole collection):

This is an easy method that relies on counting the number of ratings for each book and recommending the top 15 books with the highest ratings.

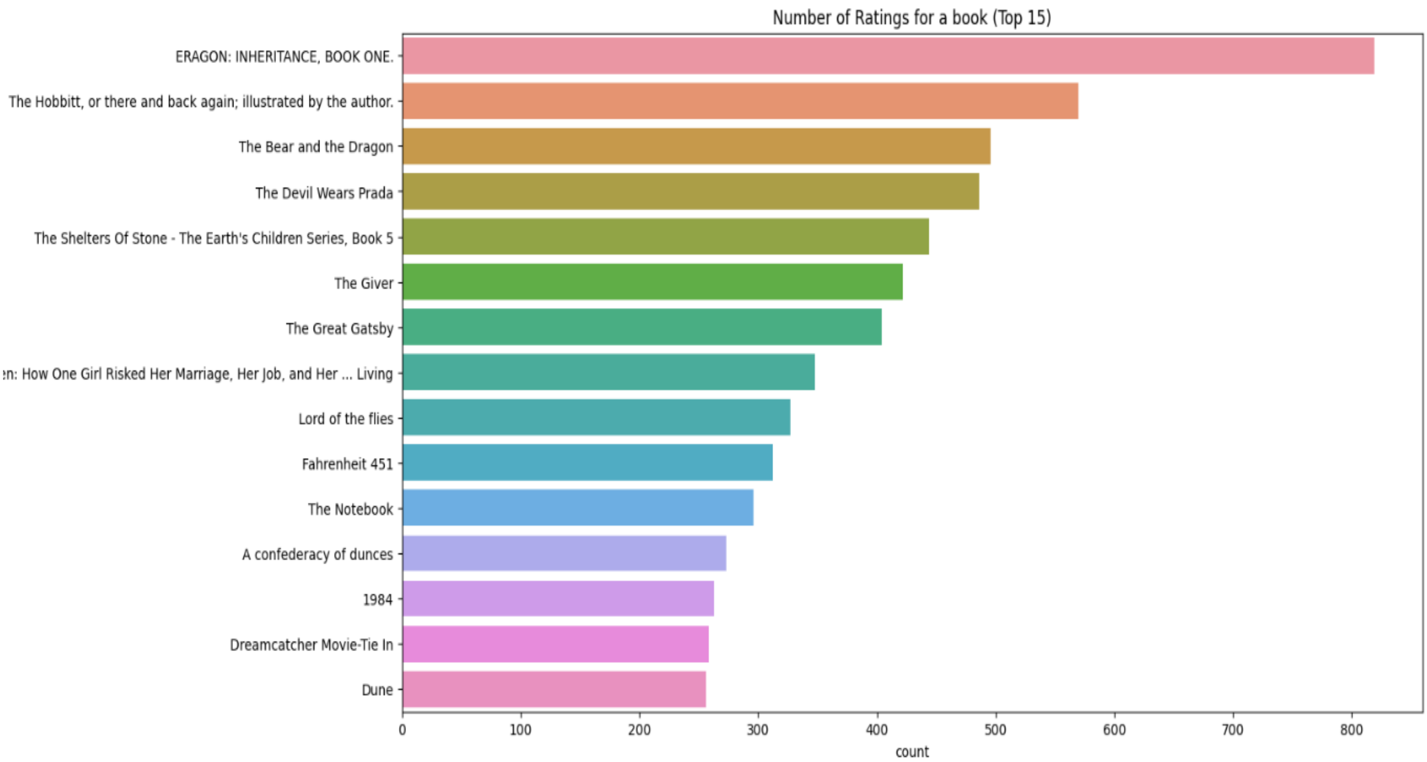


Figure 12 Number of ratings for each book

Recommendation Results are as follows.

title	rating
ERAGON: INHERITANCE, BOOK ONE.	819
The Hobbitt, or there and back again; illustrated by the author.	570
The Bear and the Dragon	496
The Devil Wears Prada	486
The Shelters Of Stone - The Earth's Children Series, Book 5	444
The Giver	422
The Great Gatsby	404
Julie and Julia: 365 Days, 524 Recipes, 1 Tiny Apartment Kitchen: How One Girl Risked Her Marriage, Her Job, and Her ... Living	348
Lord of the flies	327
Fahrenheit 451	312

Figure 13 The recommended books

2. Recommendation using Average Weighted Rating:

In this method, for all the books we calculated a weighted score using the below formula:

$$score = t/(t+m) * a + m/(m+t) * c$$

where:

t represents the total number of ratings received by the book,
m represents the minimum number of total ratings considered to be included,
a represents the average rating of the book and,
c represents the mean rating of all the books.

using Average Weighted Rating and recommended the books with the highest score.

Recommended Books using Average Weighted Rating method:

	title	Total_Ratings	Book_Title	Average_Rating	score
1	570	570	The Hobbitt, or there and back again; illustra...	3	2.675552
0	819	819	ERAGON: INHERITANCE, BOOK ONE.	2	2.132509

Figure 14 Recommended books using Average Weighted Rating

3. User-Item Collaborative Filtering Recommendation:

Using this method, we could recommend by considering user ratings and finding cosine similarities in ratings by several users.

Example 1:

```
Input Book:
ERAGON: INHERITANCE, BOOK ONE.

RECOMMENDATIONS:

Dreamcatcher Movie-Tie In
The Giver
The Notebook
The Hobbitt, or there and back again; illustrated by the author.
The Shelters Of Stone - The Earth's Children Series, Book 5
The Devil Wears Prada
The Bear and the Dragon
Lord of the flies
Julie and Julia: 365 Days, 524 Recipes, 1 Tiny Apartment Kitchen: How One Girl Risked Her Marriage, Her Job, and Her ... Living
The Great Gatsby
```

Figure 15 The recommended books given "ERAGON: INHERITANCE, BOOK ONE." book as an input

Example 2:

```
Input Book:
The Devil Wears Prada

RECOMMENDATIONS:

Julie and Julia: 365 Days, 524 Recipes, 1 Tiny Apartment Kitchen: How One Girl Risked Her Marriage, Her Job, and Her ... Living
ERAGON: INHERITANCE, BOOK ONE.
The Hobbitt, or there and back again; illustrated by the author.
Dune
1984
A confederacy of dunces
The Notebook
Lord of the flies
The Shelters Of Stone - The Earth's Children Series, Book 5
The Great Gatsby
```

Figure 16 The recommended books given "The Devil Wears Prada" book as an input

4. Recommendation using surprise library (specific to a user):

We used the surprise library twice in our code, the first time to implement a recommender regarding a specific user.

Example:

If we have a user ID

```
'A3CU0U5RY99NXJ'
```

Figure 17 Example for user ID

The recommendation:

	0	1	2	3	4
A3CU0U5RY99NXJ	(The Hobbitt, or there and back again; illustr...	(1984, 3.812257028693454)	(The Giver, 3.666742984891319)	(The Great Gatsby, 3.426340944463574)	(Fahrenheit 451, 3.383444463144665)

Figure 18 The five recommended books for this specific user

5. Recommendation using surprise library (not specific to a user):

In this method, we also used the surprise library to implement a general recommender not specific to a user.

In the beginning, we performed some visualization for the average ratings.

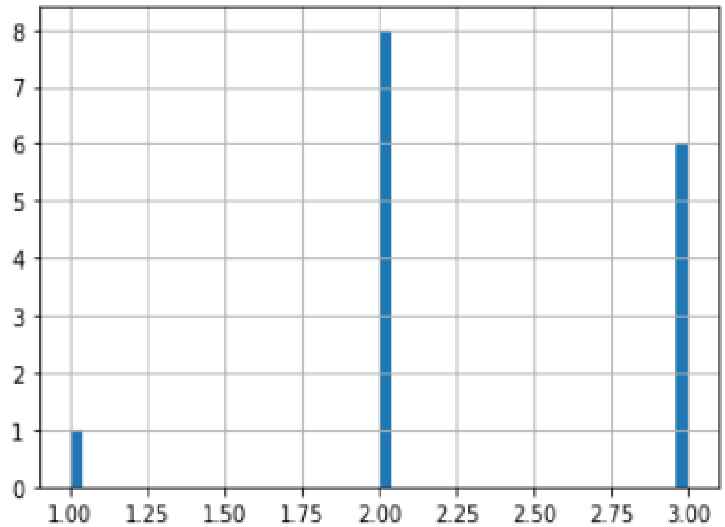


Figure 19 The number of books vs the average rating

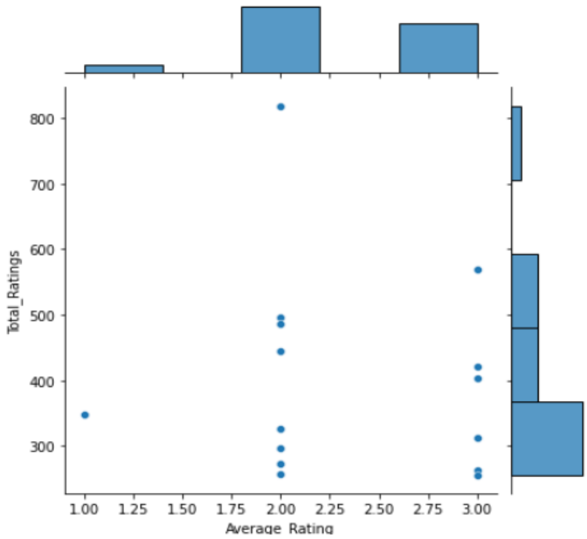


Figure 20 Scatterplot shows the total number of ratings and the average rating for each book

We also used this method to be connected with our chatbot so, we have implemented a function that takes an initial book name and user review classification results for this initial book and returns back the recommendation.

```
b = list(df.title)[0]
recommend(b,"Positive")
```

Figure 21 Passing the book name and "Positive" to get the book recommendation

```
'Julie and Julia: 365 Days, 524 Recipes, 1 Tiny Apartment Kitchen: How One Girl Risked Her Marriage, Her Job, and Her ... Living'
```

Figure 22 The recommended book

```
b = list(df.title)[0]
recommend(b,"Neutral")
```

Figure 23 Passing the book name and "Neutral" to get the book recommendation

```
'The Hobbitt, or there and back again; illustrated by the author.'
```

Figure 24 The recommended book

If you noticed here that we used the same book twice:
The first time with positive feedback from the user
The second time with neutral feedback.
In the two cases, we got different recommendations because while implementing this function we set the return to be the most correlated book if the feedback was positive, and on the other hand we return the least correlated book if the feedback was negative or even neutral.

6. Chatbot design:

a) Intents design:

Four Intents are developed (“GetName”, “recommender”, “recommender followup”, and “thanking”) besides the two default intents as shown in Figure 25. The Contexts are used to force a specific sequence for the chat. The Webhook call is enabled for “recommender” and “recommender followup” intents. “recommender” intent is responsible for recommending a random book and “recommender followup” is responsible for recommending a book based on the feedback of the user about the previous recommended book.

Figure 25 The developed intents

The training phrases for the developed intents ("GetName", "recommender", "recommender followup", and "thanking") are shown in the following table.

GetName Intent	thanking Intent
omar my name is omar The given name by the user will be stored in name parameter.	iam happy, thanks! thank you thanks
recommender Intent	**recommender followup Intent**
i like reading, do you know a good book to read? do u know a good book to read recommend me a book i want to buy a book, what do u recommend? can you recommend me a good book? i want to read a book i want you to recommend me a book	it seems good i didn't like it no, i want another book i liked it it's very nice

The input and output contexts of the intents are shown in the following table.

Default Welcome Intent	GetName Intent
Default Welcome Intent Contexts Add input context 1 awaiting_name Add output context	GetName Contexts awaiting_name Add input context 1 awaiting_question Add output context
recommender Intent	**recommender followup Intent**
recommender Contexts awaiting_question Add input context 1 awaiting_feedback Add output context	recommender followup Contexts awaiting_feedback Add input context Add output context

The responses of the intents are shown in the following table.

Default Welcome Intent	GetName Intent
Text Response 1 Hello! What's your name? 2 Enter a text response variant	Text Response 1 How can I help you, $name? 2 Enter a text response variant The chatbot will use the name parameter to call the user by his name.
recommender Intent	**recommender followup Intent**
Text Response 1 check 2 Enter a text response variant This response will be generated only if there is a problem in the Webhook call.	Text Response 1 Server Offline 2 Enter a text response variant This response will be generated only if there is a problem in the Webhook call.
thanking Intent	
Text Response 1 You are welcome. 2 Enter a text response variant	

The URL provided by NGROK will be put in the fulfilment as shown in the following figure.

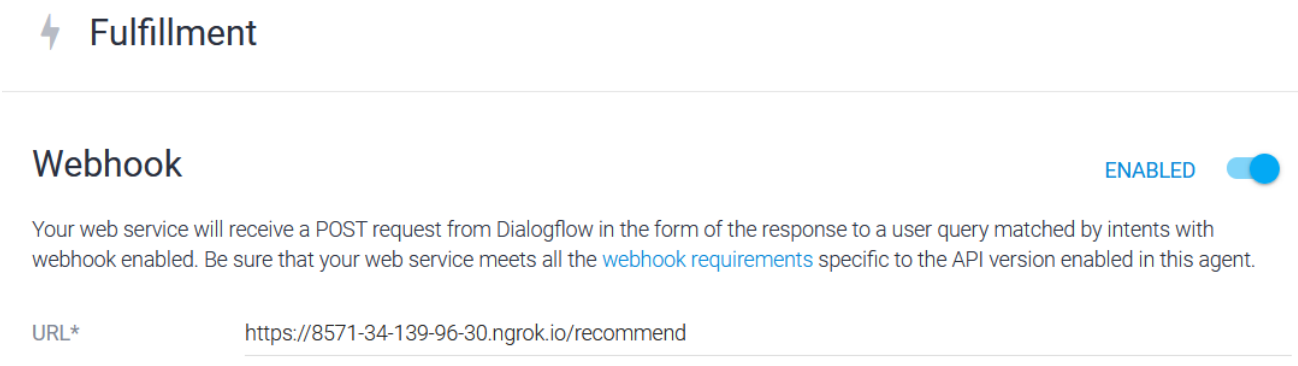
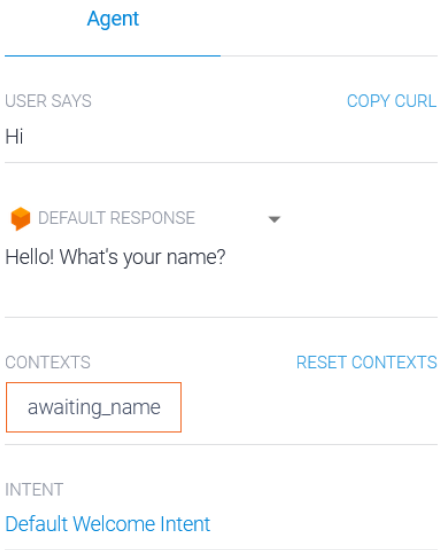
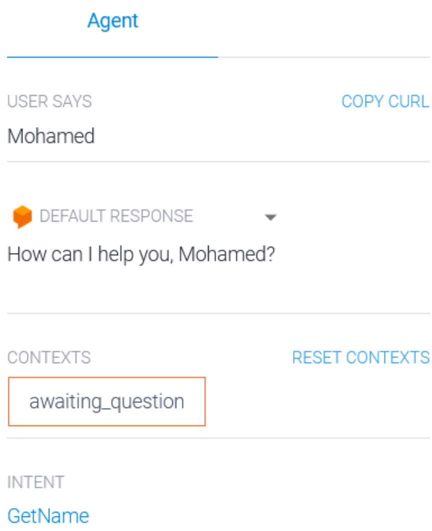
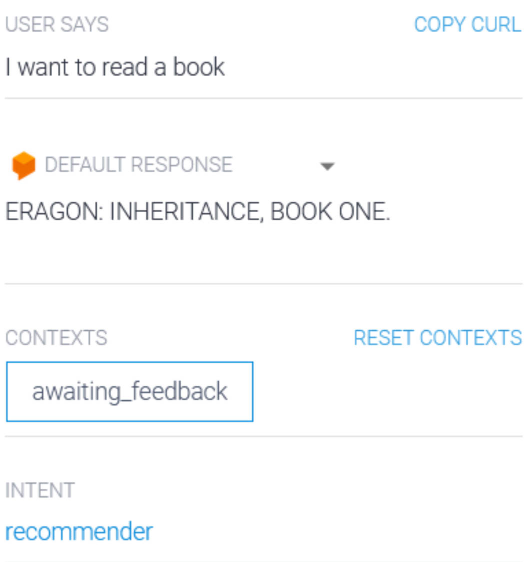
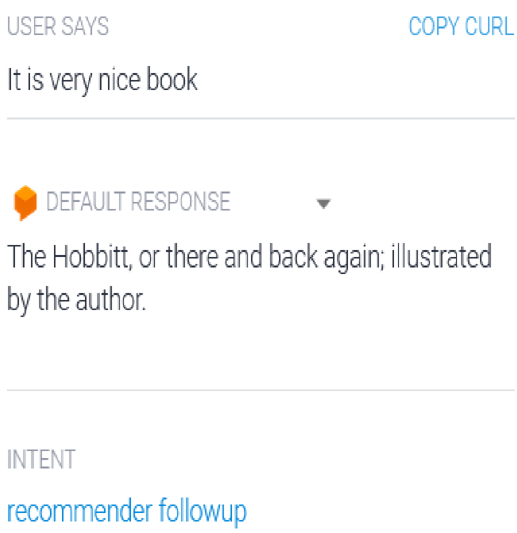
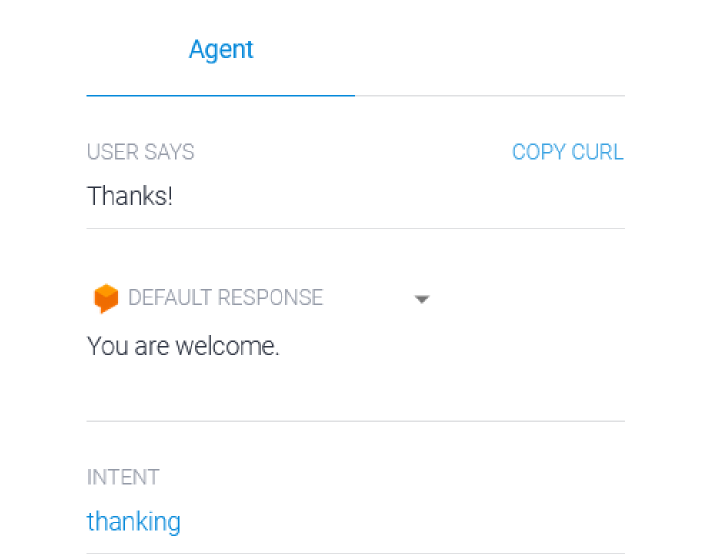


Figure 26 The fulfillment

b) Evaluation (overall system):

 The triggered intent, output context, and the chatbot response are as intended.	 The triggered intent, output context, and the chatbot response are as intended.
 The triggered intent and output context are as intended. The chatbot managed to recommend a random book for the user. One drawback of the chatbot, that it does not ask the user about his feedback about the recommended book.	 The triggered intent is as intended. The chatbot managed to recommend a book for the user based on the feedback of the user about the previous book.



The triggered intent and the chatbot response are as intended.

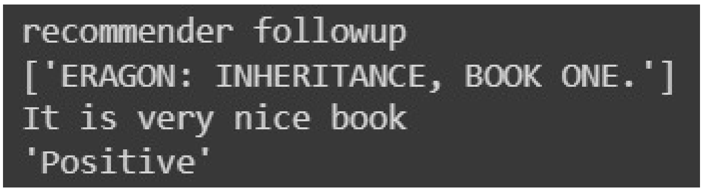
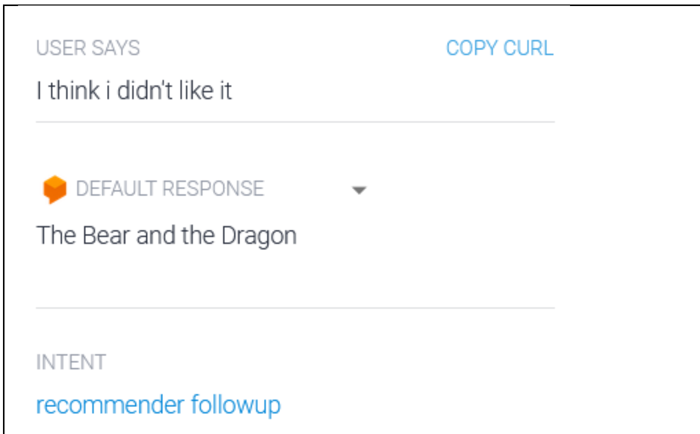
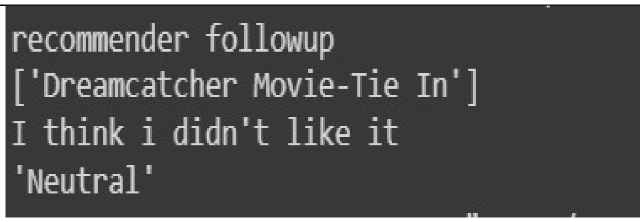


Figure 27 The book name and the user feedback that the "recommender followup" intent took and the classifier prediction of the feedback

From the above figure, it is concluded that the classifier model managed to classify the feedback from the user correctly.

c) Error Analysis (overall system):





In the previous example, the classifier did not manage to predict the class of the feedback correctly (predicted "Neutral" for obvious "Negative" class). So, the classifier needs to be improved as its accuracy is not high enough.

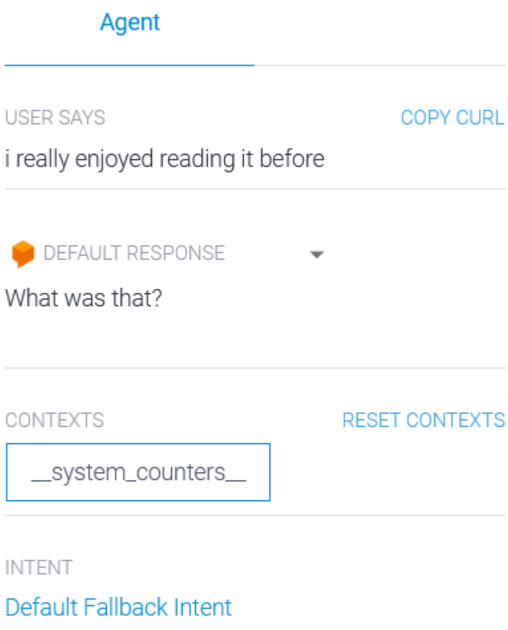


Figure 28 The response of the chatbot for the user feedback

From the previous figure, it is concluded that the "recommender followup" intent had not triggered, and the Default Fallback Intent is triggered instead. So, the training phrases of the "recommender followup" intent need to be improved.

References:

[1] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.