In-House Training

(**A Research Review**)

ON  
Applications of Artificial Intelligence

TITLED

"ROTTEN EGGS"



AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY

AMITY UNIVERSITY, SECTOR-125, NOIDA-201303

Submitted to: Submitted By:

Dr. Sushil Kumar Sureena Wadhwa

Enrollment No:A2305215027

**DECLARATION**

I certify that

1. The work contained in this report is original and has been done by me under the guidance of my supervisor.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. I have followed the guidelines provided by the Institute in preparing the report.
4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

(Sureena Wadhwa)

Signature of the Student

# **CERTIFICATE**

This is to certify that the Summer Internship Report entitled, **“Rotten Eggs”** submitted by **“Sureena Wadhwa“** to Amity University, Noida, Uttar Pradesh, India is a record of bonafide Project work carried out by him under my supervision and guidance and is worthy of consideration for the award of credits in partial fulfilment for the degree of Master of Technology in Computer Science and Engineering of the Institute.

(Dr. Sushil Kumar )

Mentor

# **ACKNOWLEDGEMENTS**

I would like to express my deepest sense of gratitude to my supervisor, Dr. Sushil Kumar under whose constant guidance the research review could be completed. He spent his valuable time encouraging me to write a paper undergoing thorough survey of the literature for collecting data requisite in writing a successfully reviewed paper.

I would be failing in my duty if I didn’t express my sincere thanks to our Department of Computer Science and same to JOINT HEAD ASET, HOD CSE, Prof. (Dr.) Abhay Bansal for providing me an opportunity to pursue my summer internship from Amity University, Noida, Uttar Pradesh under the guidance of distinguished faculties with valuable insights.

Lastly, I owe more than thanks to my family members who have been supportive throughout my life without which pursuing post-graduation studies wouldn’t have been seamless.

WEEKLY PROGRESS REPORT (WPR)

For the week commencing: 23rd May 2017 to 29th May 2017

WPR: 1

Enrollment number: A2305215027

Program: B.Tech (CSE)

Student Name: Sureena Wadhwa

Faculty Guide’s Name: Dr. Sushil Kumar

Co - Guide’s Name: N/A

Project Title:

TARGETS SET FOR THE WEEK

* To start a daily of diary work done and progress
* To find relevant sources of information to learn basics of natural language processing and python.

PROGRESS/ACHIEVEMENTS FOR THE WEEK

* Started the daily diary of work done.
* Relevant sources found and started to learn basics of natural language processing using python

FUTURE WORK PLANS

* Learn about more classifiers and python libraries for the project.

WEEKLY PROGRESS REPORT (WPR)

For the week commencing: 30th May 2017 to 5th June 2017

WPR: 2

Enrollment number: A2305215027

Program: B.Tech (CSE)

Student Name: Sureena Wadhwa

Faculty Guide’s Name: Dr. Sushil Kumar

Co - Guide’s Name: N/A

Project Title: Rotten Eggs

TARGETS SET FOR THE WEEK

* Learn about classifying data, find data sets of movie reviews.

PROGRESS/ACHIEVEMENTS FOR THE WEEK

* Learnt how to read a .csv file and process it

FUTURE WORK PLANS

* To further classify the data more
* To implement the layout plan and make a basic frame of the project

WEEKLY PROGRESS REPORT (WPR)

For the week commencing: 6th June 2017 to 13th June 2017

WPR: 3

Enrollment number: A2305215027

Program: B. Tech (CSE)

Student Name: Sureena Wadhwa

Faculty Guide’s Name: Dr.Sushil Kumar

Co - Guide’s Name:

Project Title: Rotten Eggs

TARGETS SET FOR THE WEEK

* To make a layout plan and basic frame of the project

PROGRESS/ACHIEVEMENTS FOR THE WEEK

* Basic frame completed
* Layout plan implemented.

FUTURE WORK PLANS

* Implement the functional part of the project

WEEKLY PROGRESS REPORT (WPR)

For the week commencing: 14th June 2017 to 21st June 2017

WPR: 4

Enrollment number: A2305215027

Program: B. Tech (CSE)

Student Name: Sureena Wadhwa

Faculty Guide’s Name: Dr. Sushil Kumar

Co - Guide’s Name: N/A

Project Title: Rotten Eggs

TARGETS SET FOR THE WEEK

* Implement the functionality module of the app.
* Test the app in as many different scenarios as possible.

PROGRESS/ACHIEVEMENTS FOR THE WEEK

* Functionality module partially complete.
* Tested only in one scenario.

FUTURE WORK PLANS

* Finish the functionality module of the app.
* Clean the app UI for final submission.
* Make the Report for Summer In-house practical training.

# ABSTRACT

Sentiment analysis is a crucial and latest research area. This paper associates prototype- based machine learning and supervised learning together. The algorithm was tested on movie reviews and it provides with the sentiment found within comments and reviews that the users input or give. The result is a hybrid of various classifiers and clustering techniques used in the program to get the correct sentiments of the users and review a movie.

# TABLE OF CONTENTS

[CERTIFICATE 3](#_Toc488621575)

[ACKNOWLEDGEMENTS 4](#_Toc488621576)

[ABSTRACT 9](#_Toc488621577)

[Introduction 12](#_Toc488621579)

[ARTIFICIAL INTELLIGENCE AND SENTIMENT ANALYSIS 13](#_Toc488621580)

[WHAT IS MACHINE LEARNING? 15](#_Toc488621581)

[Rotten Eggs 17](#_Toc488621582)

[The idea behind it: 17](#_Toc488621583)

[Languages used : 17](#_Toc488621584)

[Classifiers used and the concept behind them : 18](#_Toc488621585)

[1.NAÏVE BAYES CLASSIFIER : 19](#_Toc488621586)

[2.SUPPORT VECTOR MACHINE : 20](#_Toc488621587)

[3.BAG OF WORDS : 20](#_Toc488621588)

[4.TOKENIZATION : 21](#_Toc488621589)

[About the app 22](#_Toc488621590)

[RESULTS AND DISCUSSIONS 25](#_Toc488621591)

[FUTURE ASPECT OF THE PROJECT 26](#_Toc488621592)

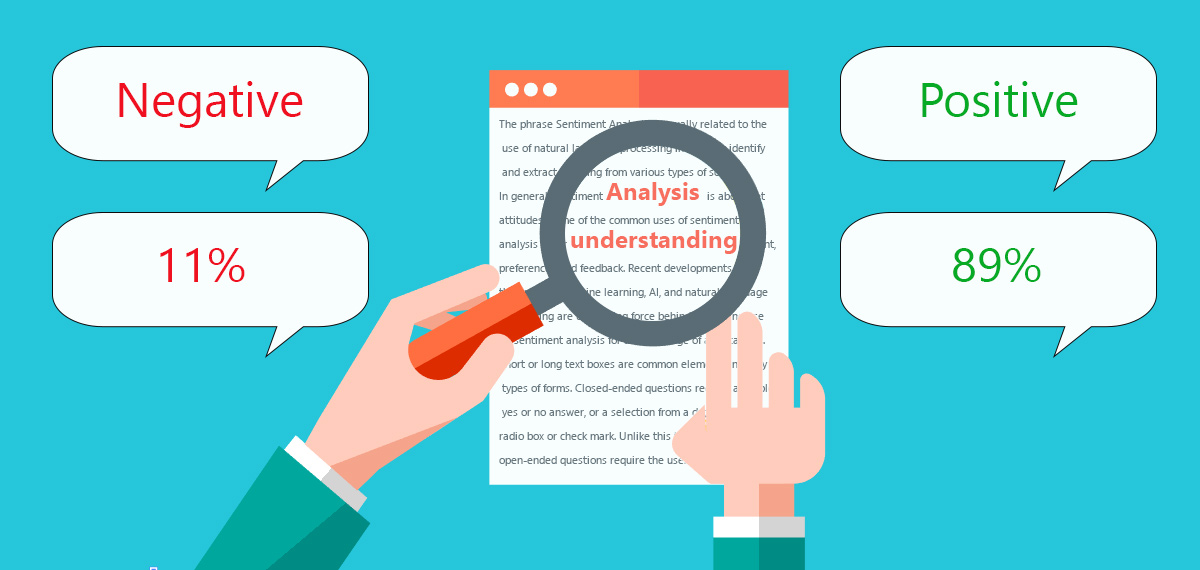
[REFERENCES 27](#_Toc488621593)

# Introduction

The sentiment found inside remarks, criticism or scrutinizes give helpful markers to various purposes. These estimations can be classified into two types: positive and negative; or into a n-point scale, e.g., great, great, tasteful, terrible, awful. In this regard, an opinion investigation undertaking can be deciphered as an arrangement errand where every class speaks to an estimation. Slant investigation gives organizations a way to evaluate the degree of item acknowledgment and to decide methodologies to enhance item quality. It additionally encourages strategy producers or legislators to break down open notions regarding approaches, open administrations or political issues.

This paper exhibits the observational consequences of a near report that assesses the adequacy of various classifiers, and demonstrates that the utilization of numerous classifiers in a half breed way can enhance the viability of conclusion examination. The technique is that in the event that one classifier neglects to characterize a report, the classifier will pass the archive onto the following classifier, until the point when the record is ordered or no other classifier exists.

# ARTIFICIAL INTELLIGENCE AND SENTIMENT ANALYSIS



Artificial intelligence (AI) is an area of computer science that focusses the creation of intelligent machines that work and react like humans. Some of the activities computers with artificial intelligence are designed for include:

* Speech recognition
* Learning
* Planning
* Problem solving

The advancement of artificial intelligence has opened various possibilities about how we use computers to help us detect, analyze and categorize online opinions/comments/reviews about products. The combination of big data and intelligent machines has made space for new tools in the market that help in running a business in an efficient way. Now that we have such valuable tools, we can have a better understanding about the users and the products available in the market and can have certain answers about questions like how they like the product, why they like the product or is it good for business? With Artificial Intelligence, a very basic algorithm can find out what people think using opinion mining.

AI and Machine Learning have emerged as a central aspect of analytics which is applied to multiple domains. AI and Machine Learning, Pattern classifiers and natural language processing (NLP) underpin Sentiment Analysis (SA); SA is a technology that makes rapid assessment of the sentiments expressed in news releases as well as other media sources such as Twitter and blogs. It wouldn't be wrong to say that Sentiment Analysis is a field of Artificial Intelligence where psychology meets technology.

There is a big difference in the way humans communicate with each other and the way humans communicate with a computer. We tend to use longer sentences while talking to a human whereas in a computer we input short and precise sentences which are to the point. We make a number of sentences which mean the same thing and it's not easy to comprehend what we want to say.

Now, talking to a computer isn't like how you talk to your friend. One has to follow a set of rules, follow syntax and structure, should have the desired software to make the computer understand. As artificial intelligence is digging deep into our devices, it has become very easy to talk to the computer in the language we are familiar with. Programmers can code, but a normal consumer apparently cannot. Consumers have to be able to speak their 'natural' language.

Natural language processing is only half the battle though. Human communication isn’t just words and their explicit meanings. Human communication is nuanced and complex. You can tell based on the way a friend asks you a question whether they’re bored, angry, or curious. You can tell based on word choice and punctuation whether a customer is getting exasperated, even in a completely text-based chat. You can read a review for a book and understand whether the reviewer liked or disliked it even if they never directly say so.

# WHAT IS MACHINE LEARNING?

The fundamental distinction with machine learning is that quite recently like measurable models, the objective is to comprehend the structure of the information – fit hypothetical conveyances to the information that are surely known. Along these lines, with factual models there is a hypothesis behind the model that is numerically demonstrated, however this requires information meets certain solid suppositions as well. Machine learning has created in view of the capacity to utilize PCs to test the information for structure, regardless of the possibility that we don't have a hypothesis of what that structure resembles. The test for a machine learning model is an approval mistake on new information, not a hypothetical test that demonstrates an invalid theory. Since machine adapting regularly utilizes an iterative way to deal with gain from information, the learning can be effortlessly computerized. Passes are run through the information until the point that a continuous pattern is found.

Machine learning calculations can make sense of how to perform critical undertakings by summing up from illustrations. This is often possible and financially savvy where manual writing computer programs is most certainly not. As more information winds up noticeably accessible, more aggressive issues can be handled. Therefore, machine learning is generally utilized as a part of software engineering in different fields. There are two major categories in machine learning:

* Unsupervised Learning :

In machine learning, unsupervised learning is a class of issues in which one looks to decide how the information is sorted out. The algorithm requires unlabeled data and is able to group new data.

* Supervised Learning :

A machine learning technique whereby a system uses a set of training examples to learn how to correctly perform a task. The algorithm requires labeled data to be trained and is able to classify new data on that basis.

Supervised machine learning is the scan for calculations that reason from remotely provided examples to deliver general speculations, which at that point make forecasts about future occasions. As such, the objective of regulated learning is to assemble a compact model of the conveyance of class marks as far as indicator highlights. The subsequent classifier is then used to relegate class names to the testing examples where the estimations of the indicator highlights are known, however the estimation of the class mark is obscure.

* What does a classifier do?

In the definition of machine learning, classification is viewed as an occasion of supervised learning, i.e. realizing where a preparation set of accurately recognized perceptions is accessible. A calculation that actualizes grouping, particularly in a solid execution, is known as a classifier.

# Rotten Eggs

## The idea behind it:

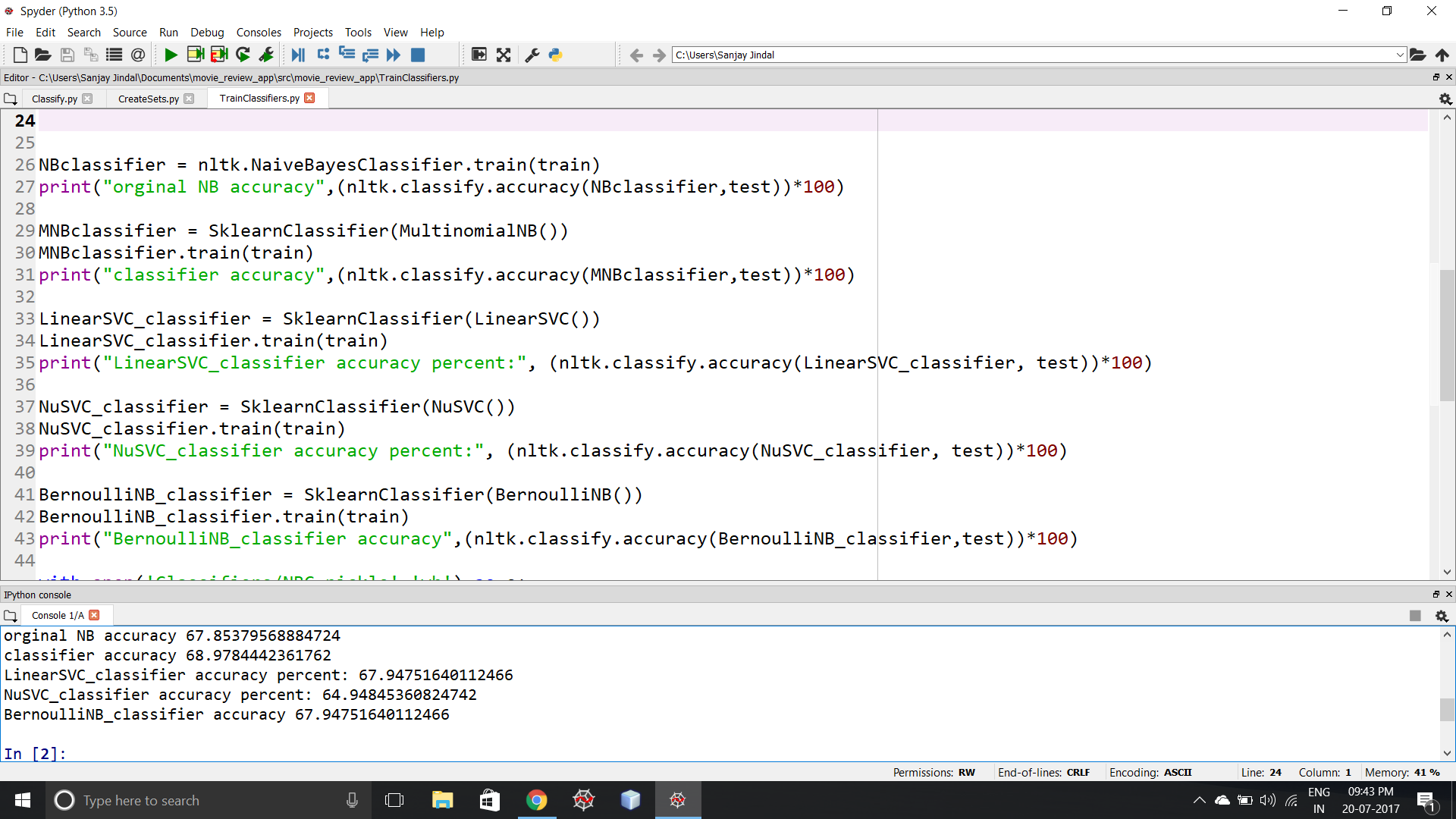
For the summer project, I wanted to explore in the field of Artificial Intelligence. So I decided to work with Natural language processing using python. While working on it, I decided to build a software which analyses sentiments and classifies them into positive or negative. Sentiment analysis is a very broad application of Artificial Intelligence which can be used for various purposes in businesses for analyzing different products and their perception in the market. Sentiment Analysis (SA) is a two-way technology which works for the user and the business organization both. It helps people, general consumers and the product-makers on both sides of the tunnel. As for the users, it lets them be a part of the organization by taking input of their reviews and categorizing their views. It's a sign by the organizations/companies that their customers matter and their opinion is the most important thing. As for the businesses, it helps them take in account real time reviews of customers about a product and further work on it.

## Languages used :

1. Python 3.5
2. Java 8.0

# Classifiers used and the concept behind them :

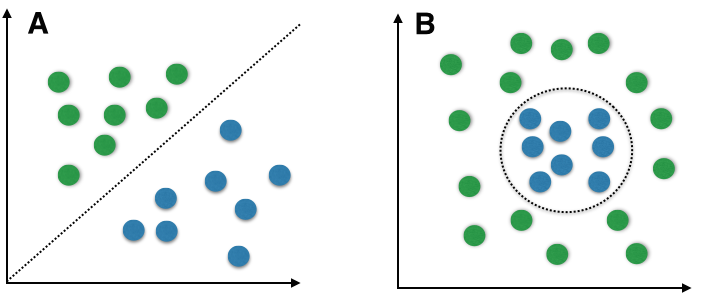
I have used 5 classifiers, from the family of Naïve Bayes and Support Vector Machine classifiers. The review is fed to all 5 classifiers. A mode is taken of all the resulting predictions and then classified.



1. NAÏVE BAYES CLASSIFIER :

It is a family of classifies based on Bayes theorem are known for creating easy yet efficient models in the area of classifying and predicting. Naïve Bayes classifiers are straight classifiers that are known for being fundamental yet astoundingly successful. The probabilistic model of straightforward Bayes classifiers relies upon Bayes' speculation, and the graphic word unsophisticated starts from the assumption that the segments in a dataset are ordinarily free. Essentially, the opportunity supposition is frequently ignored, yet blameless Bayes classifiers still tend to perform incredibly well under this nonsensical doubt. Especially for little illustration sizes, guileless Bayes classifiers can beat the all the more viable choices. Being for the most part intense, easy to execute, brisk, and correct, naïve Bayes classifiers are used as a piece of an extensive variety of fields. A couple of delineations consolidate the finish of infirmities and settling on decisions about treatment shapes, the gathering of RNA plans in requested examinations, and spam filtering in email clients. Regardless, strong encroachment of the flexibility suppositions and non-straight course of action issues can provoke to a great degree poor shows of honest Bayes classifiers. We have to recall that the sort of data and the sort issue to be enlightened direct which arrange demonstrate we have to pick. Eventually, it is continually recommended to consider different game plan models on the particular dataset and consider the desire displays and furthermore computational capability.

*posterior probability = (conditional probability \* prior probability)/evidence*



### SUPPORT VECTOR MACHINE :

A Support Vector Machine (SVM) is a directed machine learning estimation that can be used for both course of action and backslide purposes. SVMs are more frequently used as a piece of request issues and in that limit, this is the thing that we will focus on. SVMs rely upon finding a hyperplane that best segments a dataset into two classes, as showed up in the photo underneath.



### BAG OF WORDS :

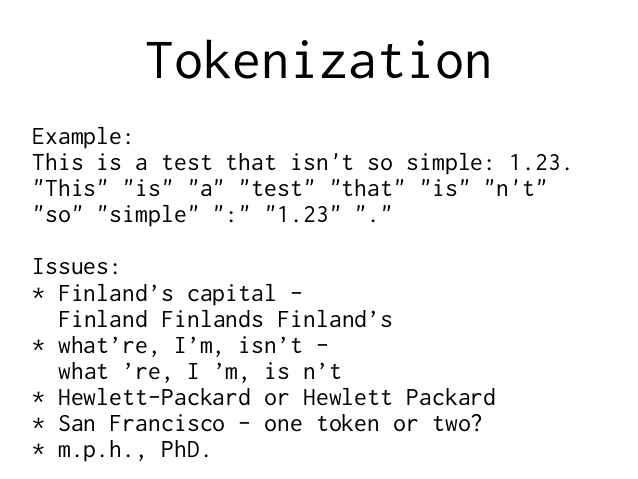
A standout amongst the most essential sub-errands in design characterization are include extraction and choice; the three primary criteria of good elements are recorded beneath:

* *Salient*. The features are important and meaningful with respect to the problem domain.
* *Invariant*. Invariance is often described in context of image classification: The features are insusceptible to distortion, scaling, orientation, etc.
* *Discriminatory.* The selected features bear enough information to distinguish well between patterns when used to train the classifier.

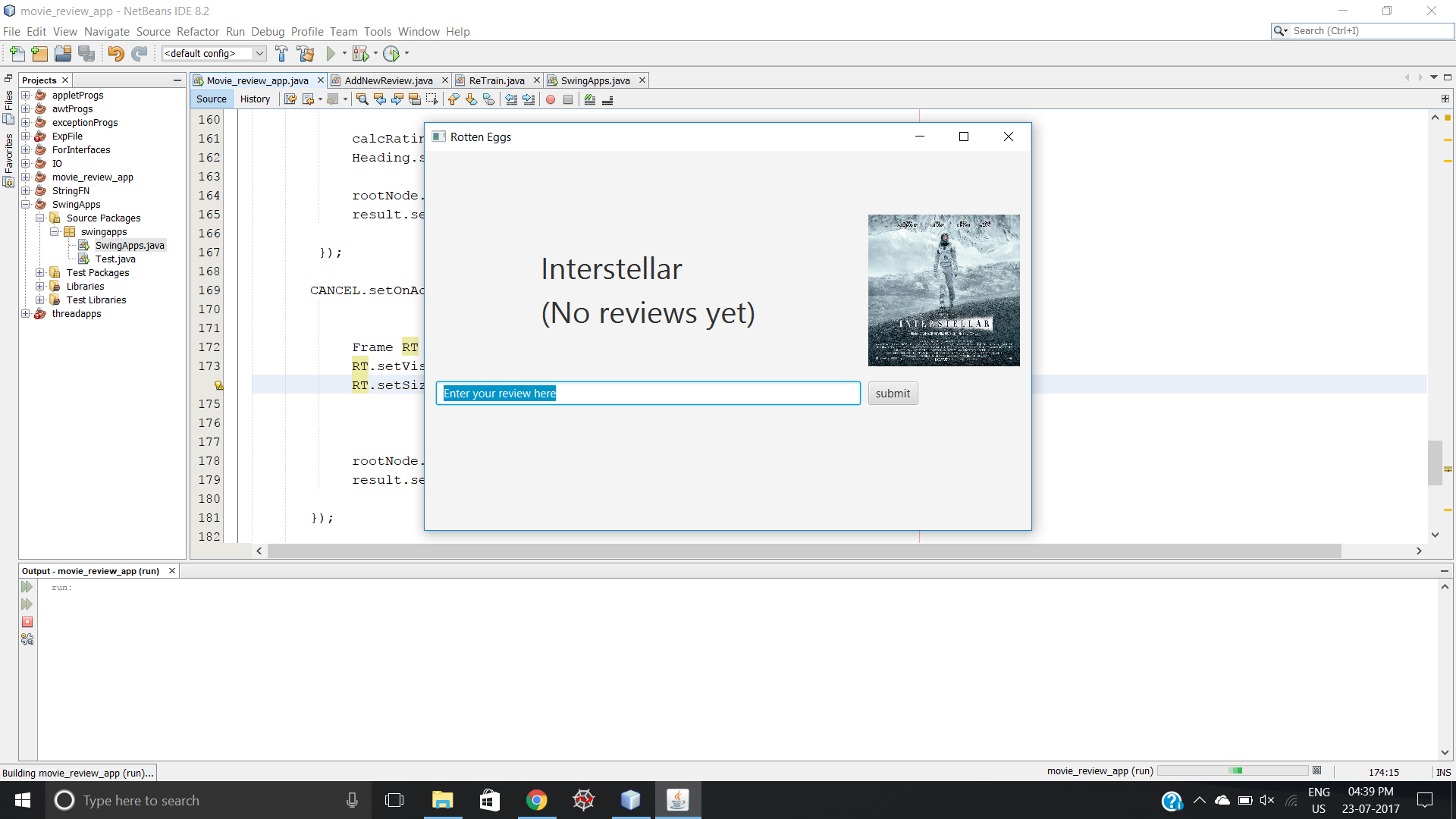
Before fitting the model and utilizing machine learning calculations for preparing, we have to consider how to best speak to a content record as a component vector. A normally utilized model in Natural Language Processing is the supposed sack of words display. The thought behind this model truly is as straightforward as it sounds. To begin with comes the formation of the vocabulary — the accumulation of every single diverse word that happen in the preparation set and each word is related with a tally of how it happens. This vocabulary can be comprehended as an arrangement of non-excess things where the request doesn't make a difference.

### TOKENIZATION :

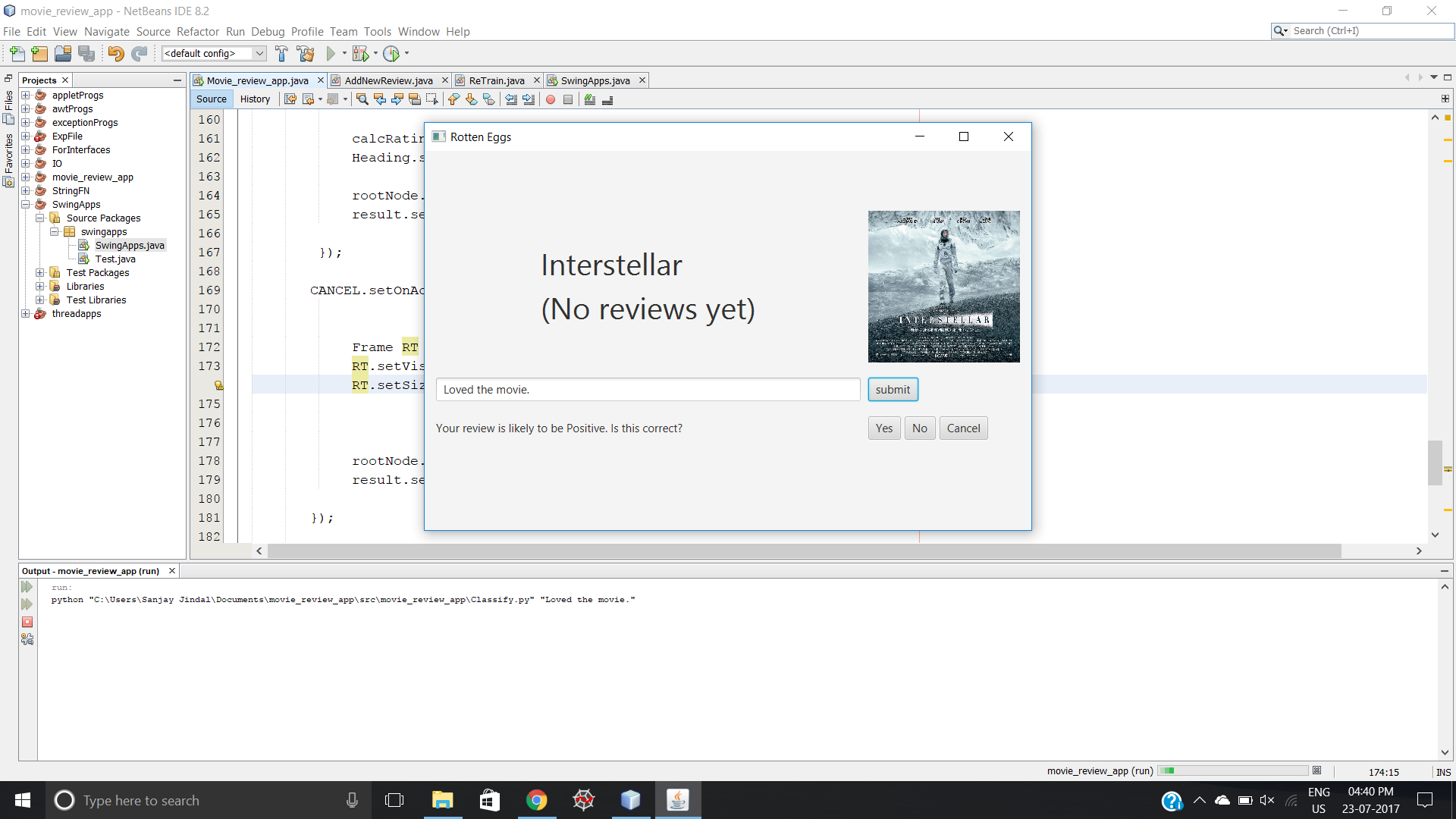
Tokenization portrays the general procedure of separating a content corpus into singular components that fill in as contribution for different regular dialect handling calculations. More often than not, tokenization is joined by other discretionary preparing steps, for example, the expulsion of stop words and accentuation characters, stemming or lemmatizing, and the development of n-grams. The following is a case of a basic however run of the mill tokenization step that parts a sentence into singular words, expels accentuation, and changes over all letters to lowercase.



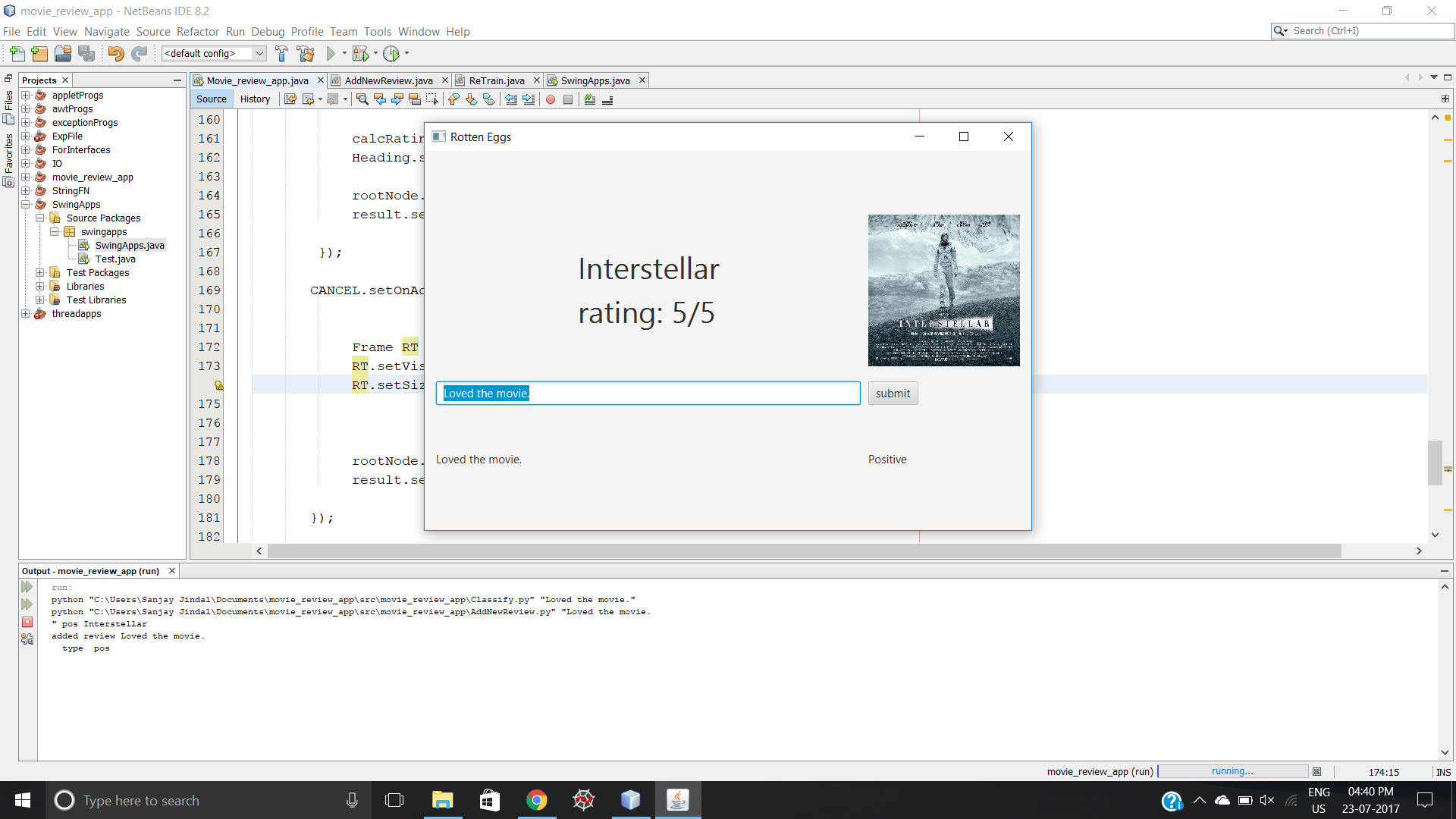
## About the app



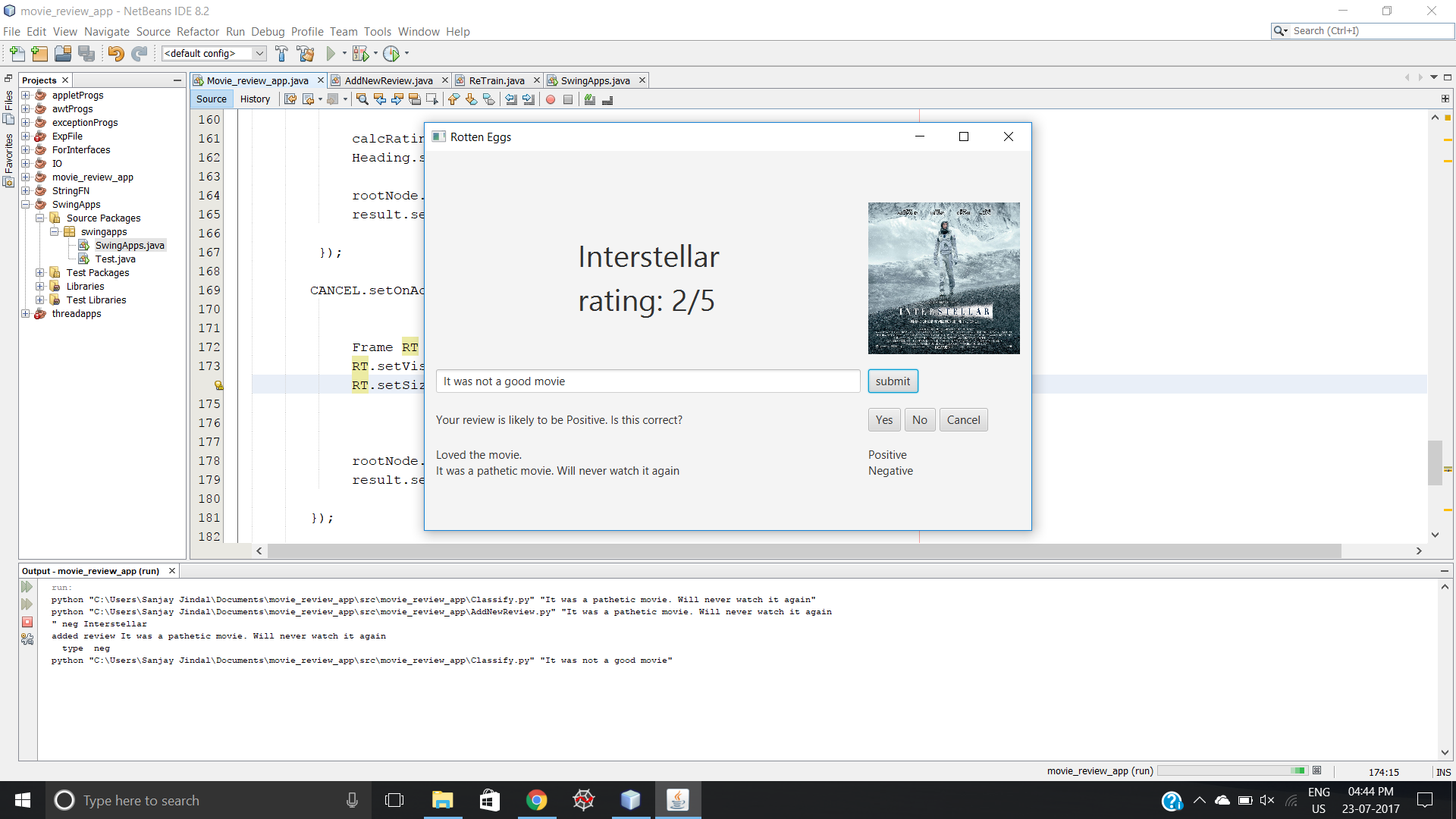
This is the interface of 'Rotten Eggs'. The front-end application is java based. The user can see the name of the movie, it's poster and can add reviews in the text box. All previous reviews of the movie can be viewed below the text box.



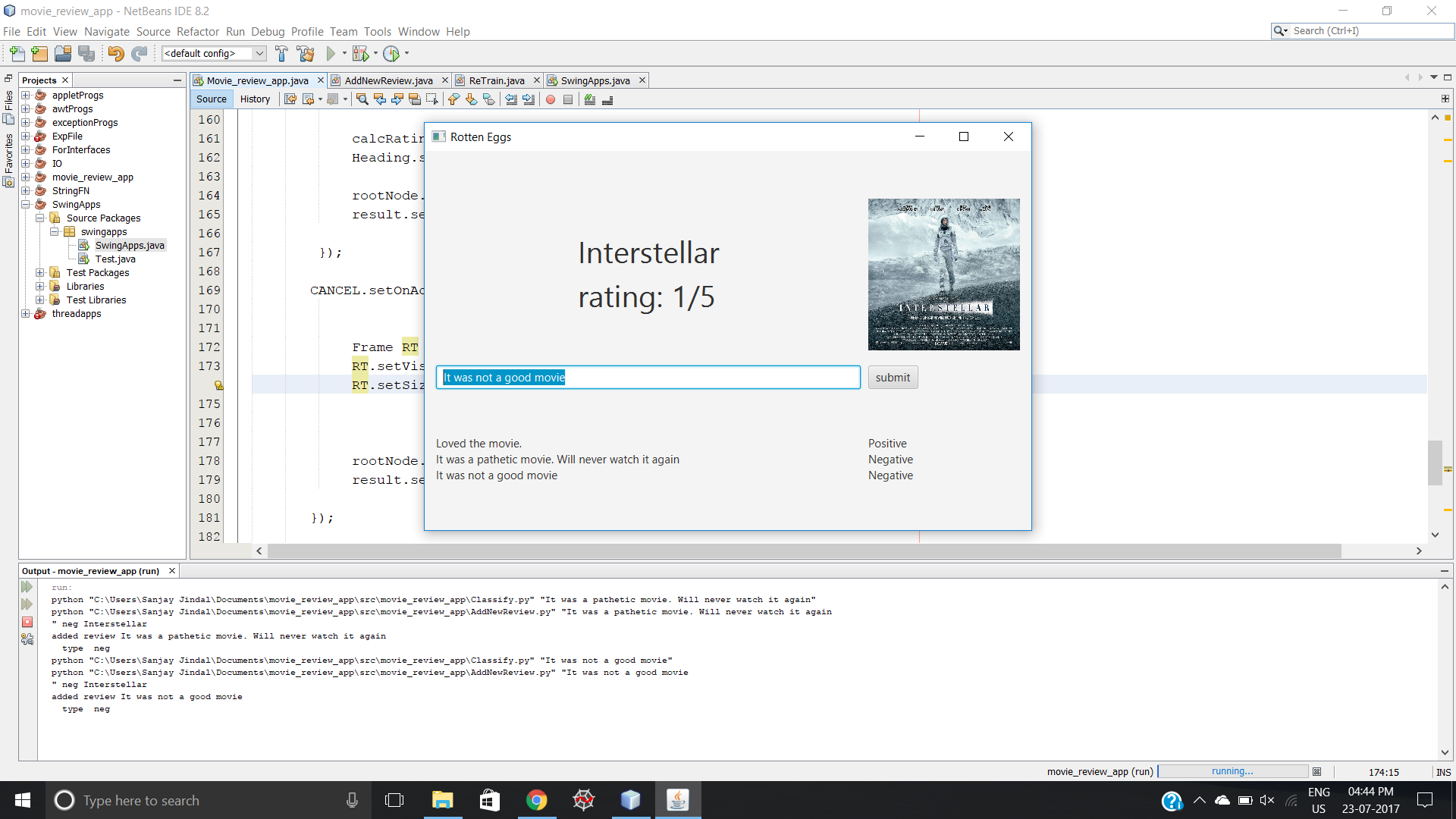
I added a review and clicked on submit. The classifier, which is based python working at the back-end classified the review as a positive one. The application gives an option to the user to correct the prediction if it's wrong.



Since, I added a positive review and the classifier predicted it correctly , I clicked on 'Yes' and the review was added in the reviews section. The classification is shown adjacent to the review. The app also calculates rating based on all reviews.



To test, I also added a negative review which was predicted correctly and was added to the review section which became the second review. Now, I added a new review and clicked on submit but this time it predicted it wrong. So, I have provided with a feature to correct this prediction manually by pressing on 'No'. The review will be added to the review section with the corrected classification as shown below.



At this point, the classifier is retrained taking input of the new reviews/data that was entered. Hence, the application becomes more intelligent every time it's being used.

# RESULTS AND DISCUSSIONS

In the end, the result was an application which uses Artificial Intelligence to make reviewing movies easier and effortless, It helps in giving real-time reviews and is helpful for the movie industry as wells. This software makes calculating rating and categorizing reviews automated perhaps reducing labor. This application collects movie reviews by the users (classified) which makes classifiers smarter and more powerful with time.

The final goal of the project was to explore in the field of Artificial Intelligence using python.

# FUTURE ASPECT OF THE PROJECT

Artificial Intelligence has a limitless scope in the future as it is digging deep into our devices. In the future, we can build neural networks to get even better results. Neural networks require large amounts of data which would be collected by this app over time. Sentiment analysis can be used for various fields, not just for movie reviews. It is a broad application of AI. Further, the user interface can be improved for better user experience and more features can be added.

# REFERENCES

<https://pythonprogramming.net/>

<https://www.youtube.com/watch?v=FLZvOKSCkxY&t=1s>

<http://www.cs.cornell.edu/people/pabo/movie-review-data/>

<https://www.youtube.com/watch?v=rISOsUaTrO4&t=5s>