FAANTUBE PRESENTING

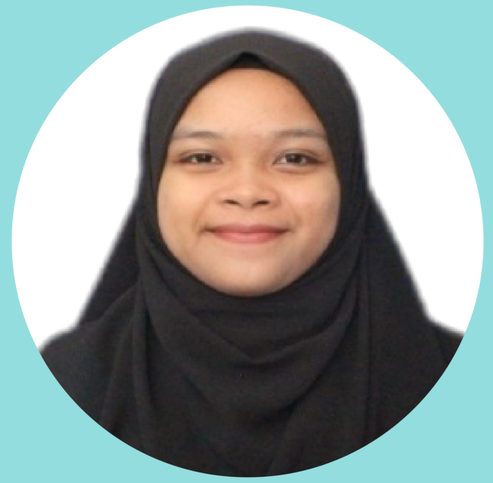
Movie Review Sentiment Analysis

Meet Our Team

FATIN NURUL AMALIN BINTI FAIZU
2021172885

MUHAMMAD NAIM BIN RIZALI
2021102747

SITI ASIAH BINTI BASARUDIN
2021113699

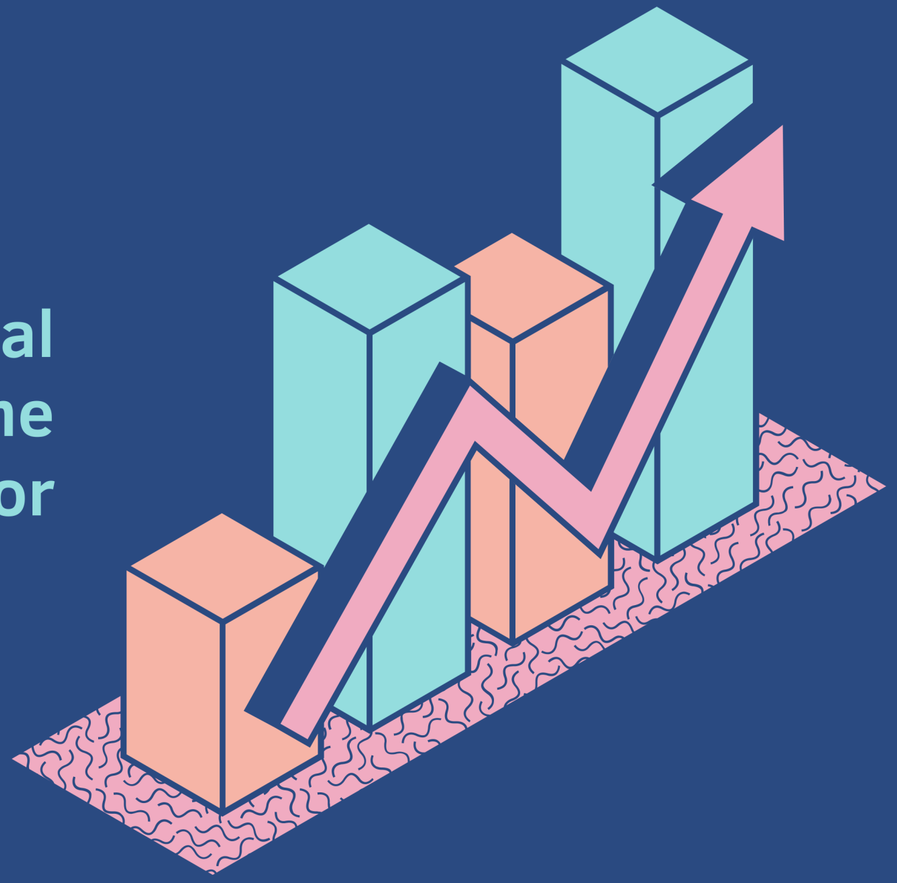


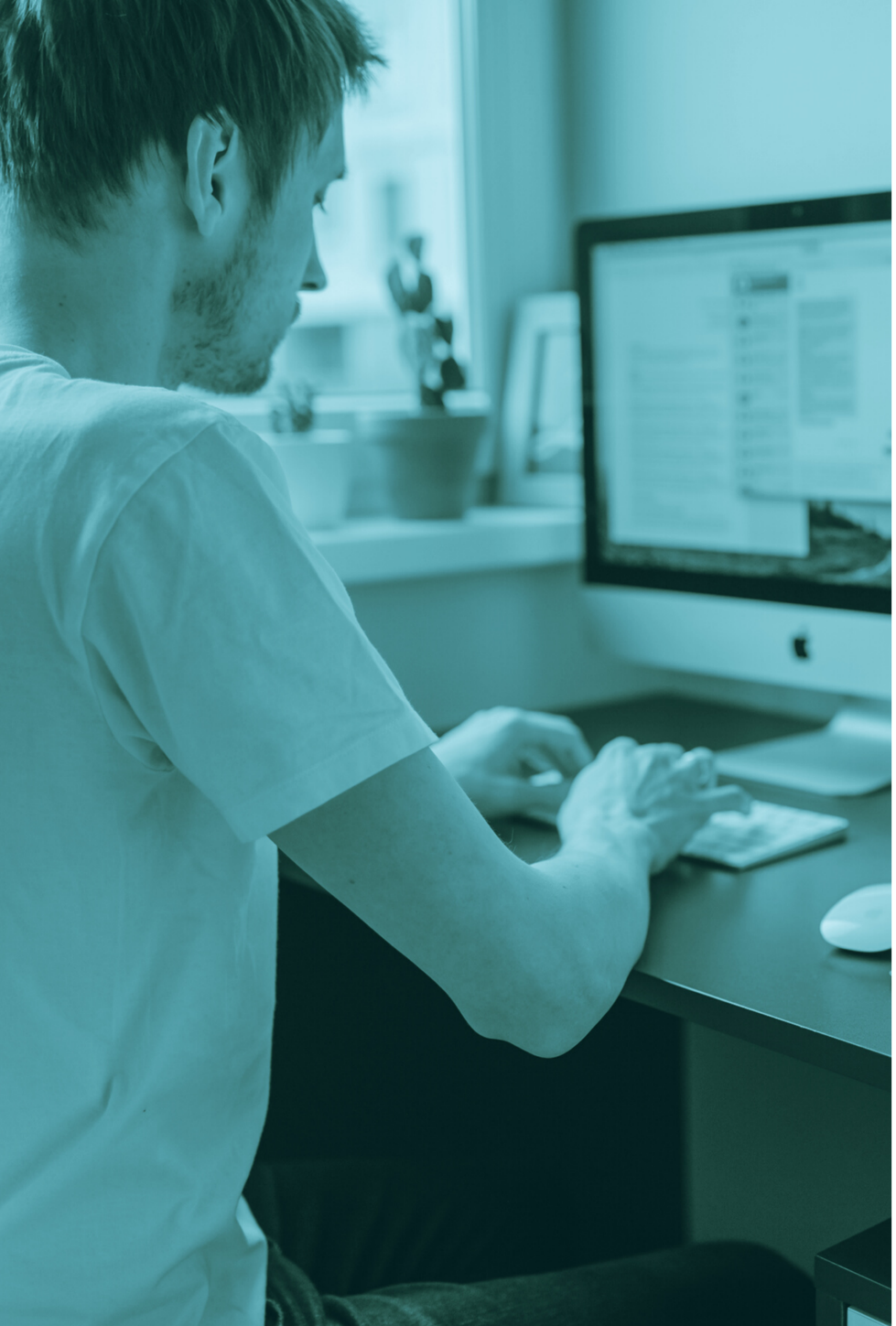
NURUL AMIRAH IZZAH BINTI SAMSU
2020490166

Table of Contents

Introduction

Movie Reviews are typically an individual giving their opinion of the movie. Some reviews include score (4 out of 5 stars) or recommendations (thumbs up).

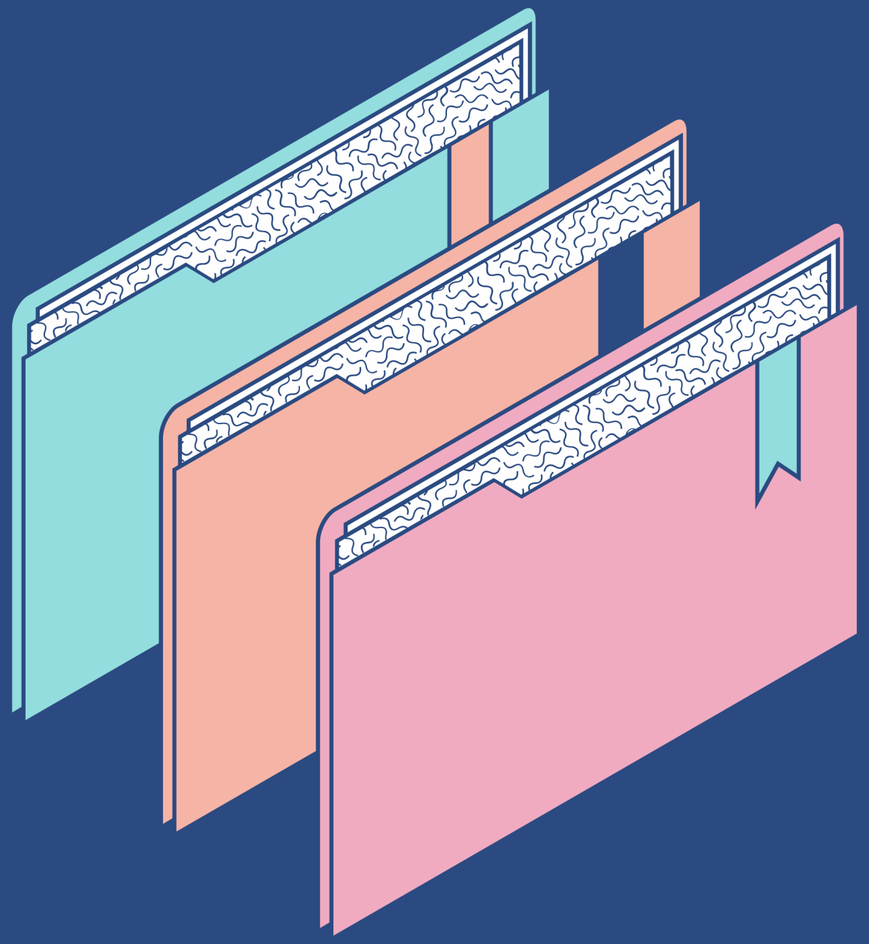




Summary of Study

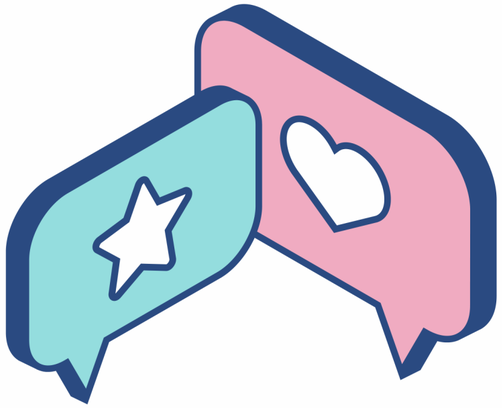
Iron Man 3 is a 2013 American superhero film based on the Marvel Comics character Iron Man. In this film, Tony Stark wrestles with the ramifications The Avengers (2012) events during a national terrorism campaign on the United States.

Iron Man 3 premiered in Paris on 2013, and released in the United States on May, 2013, as the first film in Phase Two of the MCU. It received positive reviews from critics, with praise for its action sequences, though there was criticism for its potrayal of the Mandarin.



Problem Statements

- Lack of proper evaluation of users in a movie review.
- The difficulty to decide whether the user review is genuine or skeptical.



Objectives

- To extract people opinion from a large review of Ironman 3 on IMDB.
- To classify the review into sentiment classes such as positive and negative.
- To create a model that predict the sentiment and gain insight on the movie review.

BACKGROUND OF STUDY

LITERATURE REVIEW

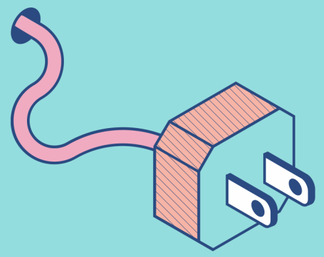
Writing a movie review is a great way to express an individual opinion of a movie. The purpose of most movie reviews is to help readers determine if they want to watch, rent, or buy the movie. The review should give enough details about the movie so that the reader can make an informed decision, without giving anyway any essentials such as the plot or any surprises.

Reviews analyze the effectiveness of the plot, theme, acting, direction, special effects, musical effects, cinematography, and all other elements that created the movie. There are qualities and guidelines that a critique of a movie should possess. Avoid the use of generalized opinions such as "it was a great movie" or "the acting was horrible," but rather give specific reasons and the whys.

COMPARISON OF MACHINE LEARNING BASED ON LITERATURE REVIEW

METHODS	ACCURACY OF MACHINE LEARNING		
	RAHMAN & HOSSEN (2019)	YASEN & TEDMORI (2019)	VENKATA & OBAIDAT (2020)
Support Vector Machine (SVM)	87.33	87.45	-
Naive Bayes (NB)	88.50	81.83	97.00
Decison Tree (DT)	80.17	91.28	65.00
Maximum Entropy (MaxEnt)	60.67	-	-

The three models that been chosen for this experiment is Support Vector Machine, Naive Bayes and Decision Tree.

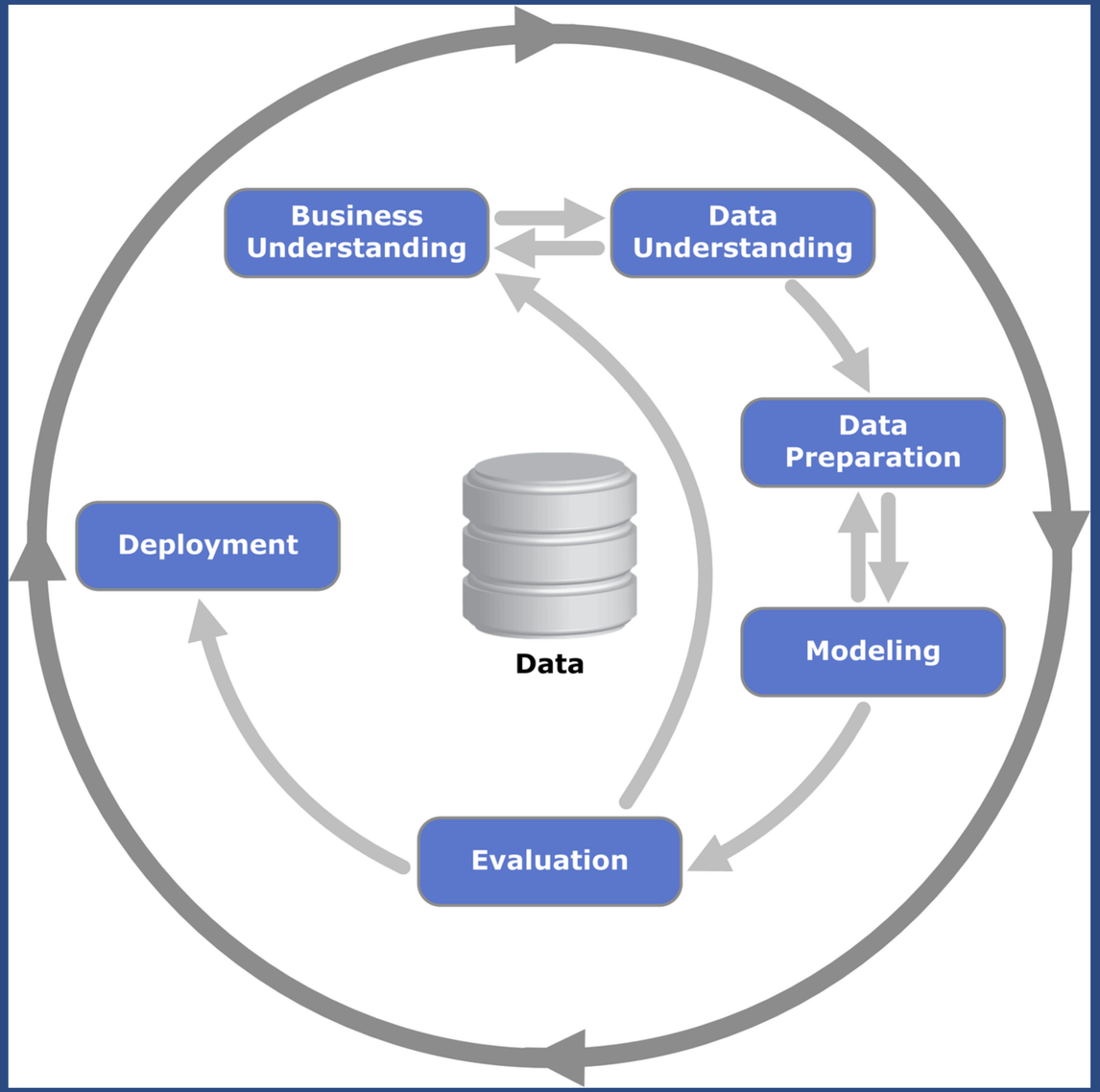


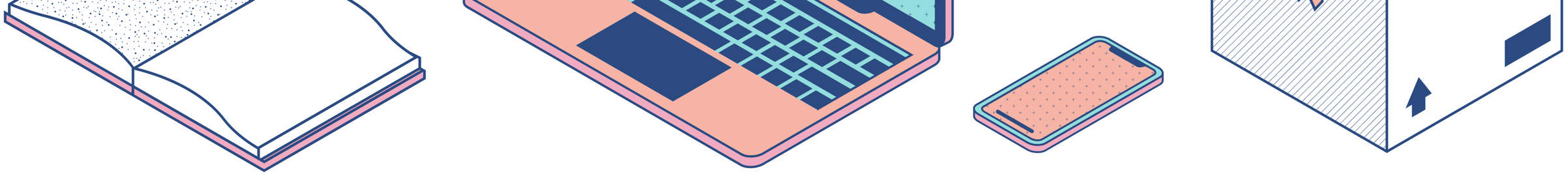
SENTIMENT ANALYSIS

Sentiment analysis is the process of text mining that collects natural language to be processed and analyzed. Based on Hamborg and Donnay (2021), this data can be used to understand general bias on a topic or different views on said topics. Methods of sentiment analysis usually use web-scraping on websites to gather natural language as data that will be turned into meaningful information (Pozzi et al., 2017).

In this project, web scraping will be done on a movie to collect reviews on this film. Mainly, focusing on the reviews on Iron Man 3. The data will be gathered and then later be analyzed through RapidMiner. From this analysis, we hope to better understand the review into sentiment classes such as positive and negative.

METHODOLOGY





BUSINESS UNDERSTANDING

What do we want to do?

- We want to see the reviews of Iron Man 3 from people who watched that movie.

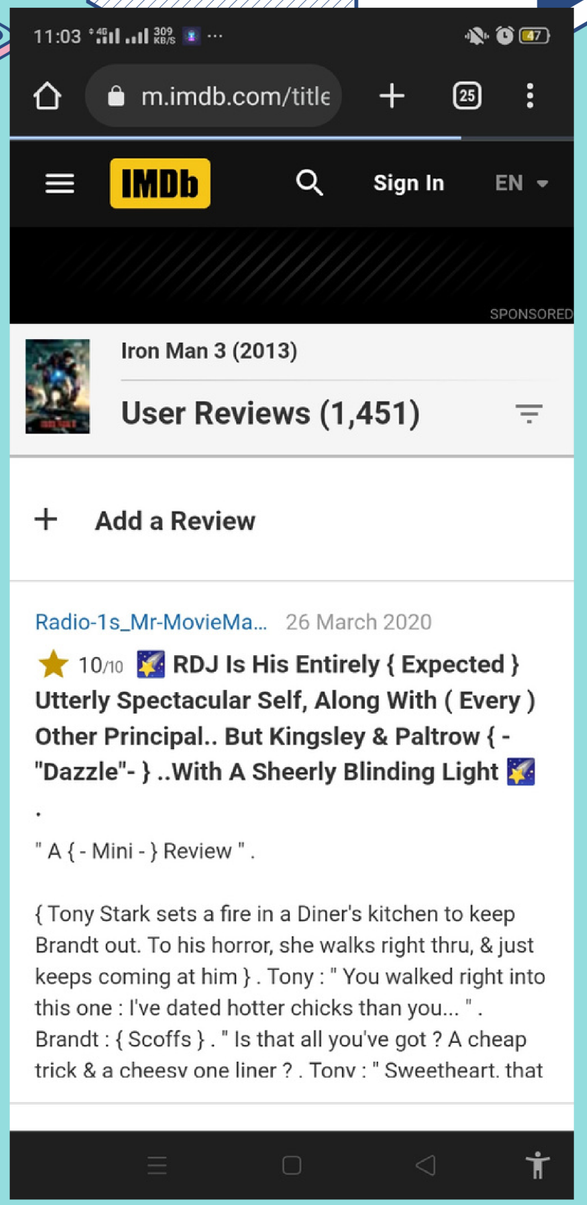
How would we do it?

- By collecting data from movie review to make a sentiment analysis. With the collected data, it will be categorized into positive and negative sentiments.

DATA UNDERSTANDING

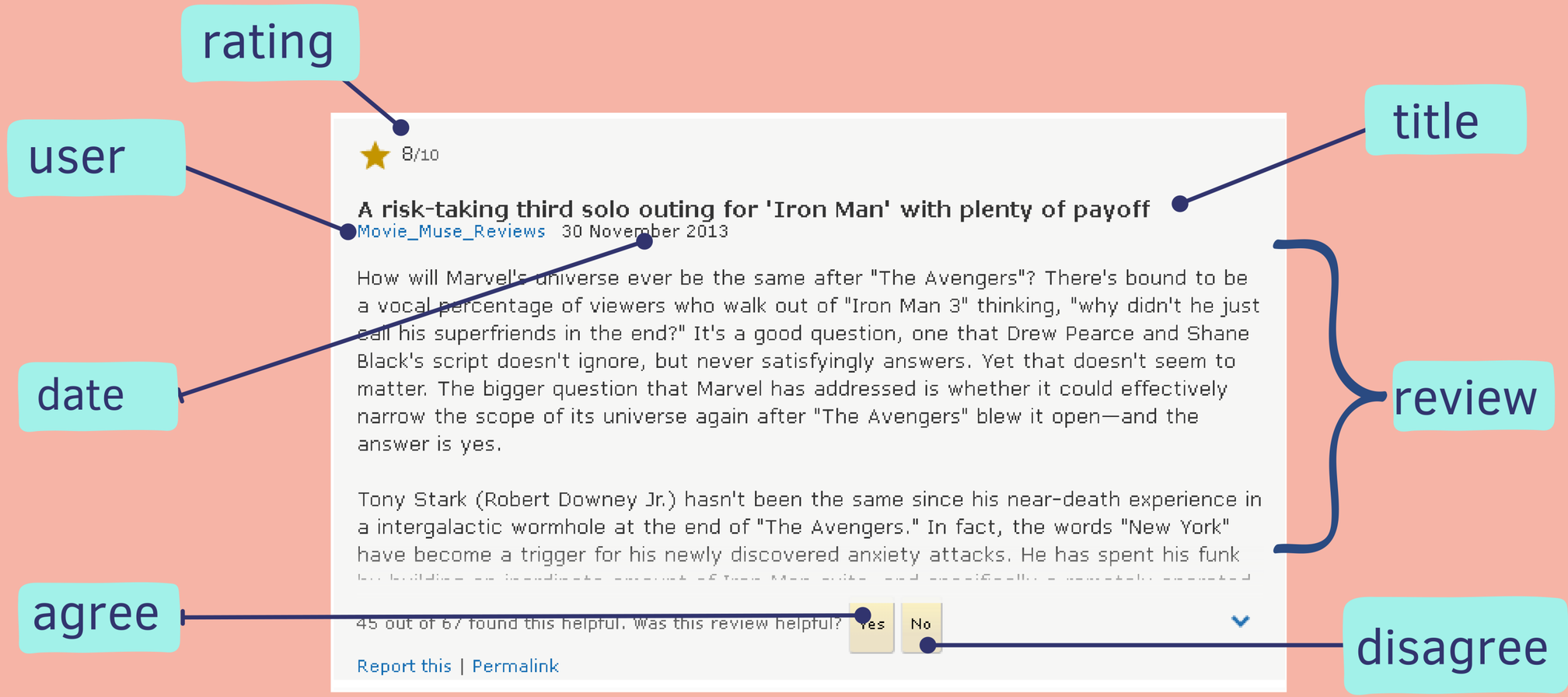
- The dataset was scrapped from IMDB.
- The dataset have attributes such as:

ATTRIBUTES NAME	DESCRIPTION	DATA TYPE
user	The username of the reviewer.	text
date	The date the review posted.	date
title	The title of the post.	text
review	The review of the movie by user.	text
rating	The rating over 10.	numerical
agree	The number of user agree with the review.	numerical
disagree	The number of user disagree with the review.	numerical



DATA PREPARATION

- Scrap data from movie review which have been chosen (IRONMAN 3).

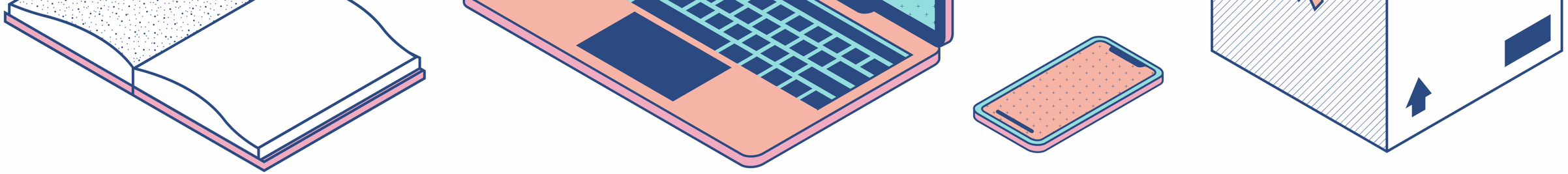


- As we found the location of attributes, we scrap the data using Python.

```
url = (
    "https://www.imdb.com/title/tt1300854/reviews/_ajax?ref_=undefined&paginationKey={}"
)
key = "g4w6ddbmqyzdo6ic4oxwjnzyr7q4yarz2modr67eb3e74vt5pjqo4wazcjsbx337lj4ydzfvjjqu3jpglisb5jlp5sjokngo57sb4vrluqzn53y"
data = {"user":[],"date":[],"title": [], "review": [],"rating":[], "agree":[],"disagree":[]}
#
while True:
    response = requests.get(url.format(key))
    soup = BeautifulSoup(response.content, "html.parser")
    # Find the pagination key
    pagination_key = soup.find("div", class_="load-more-data")
    if not pagination_key:
        break

    # Update the `key` variable in-order to scrape more reviews
    key = pagination_key["data-key"]
    for title, review, date, user ,rating, helpfulness in zip(
        soup.find_all(class_="title"), soup.find_all(class_="text show-more__control"),soup.find_all(class_="review-date"), soup.find_all(class_="display-name-link"),
        soup.find_all('span',class_="rating-other-user-rating"),soup.find_all('div',class_="actions text-muted")

    ):

        therating = rating.get_text(strip=True)
        rating = therating.rsplit('/',1)

        text = helpfulness.get_text(strip=True)
        newtext = re.sub(",","",text)
        arr = newtext.split()

        print(arr)
        agree = arr[0]
        disagree = int(arr[3]) - int(arr[0])

        data["user"].append(user.get_text())
        data["date"].append(date.get_text())
        data["rating"].append(rating[0])

        data["agree"].append(agree)
        data["disagree"].append(disagree)

        data["title"].append(title.get_text(strip=True))
        data["review"].append(review.get_text())

df = pd.DataFrame(data)
```

Beautiful Soup targetting certain attributes

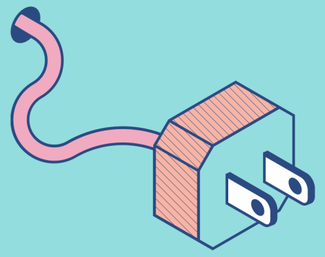
Compile unstructured data (text line) to structured data (csv)



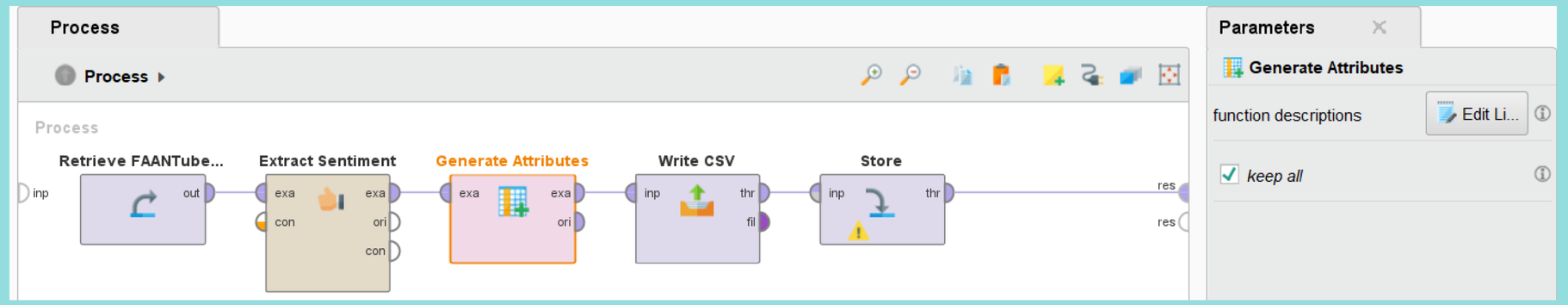
MODELING

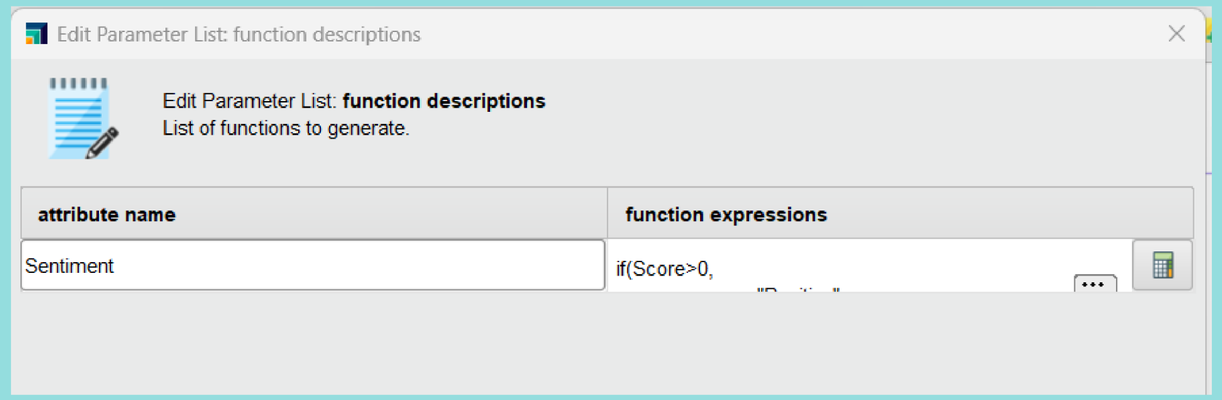
- In this project, regarding to sentiment analysis, we want to find out the highest model that gain the highest accuracy using RapidMiner.

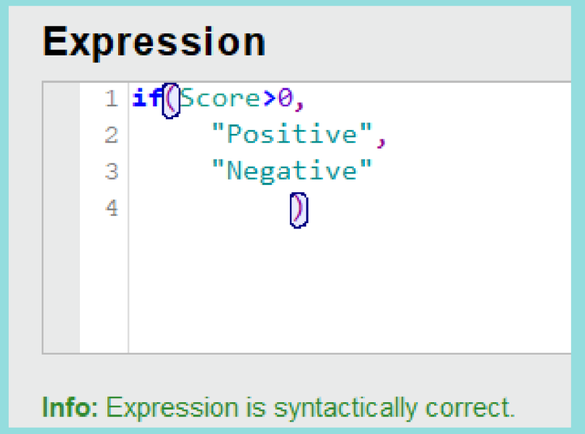
- Three modeling that we choose to test the accuracy are:
  - Decision Tree
  - Naive Bayes
  - Support Vector Machine



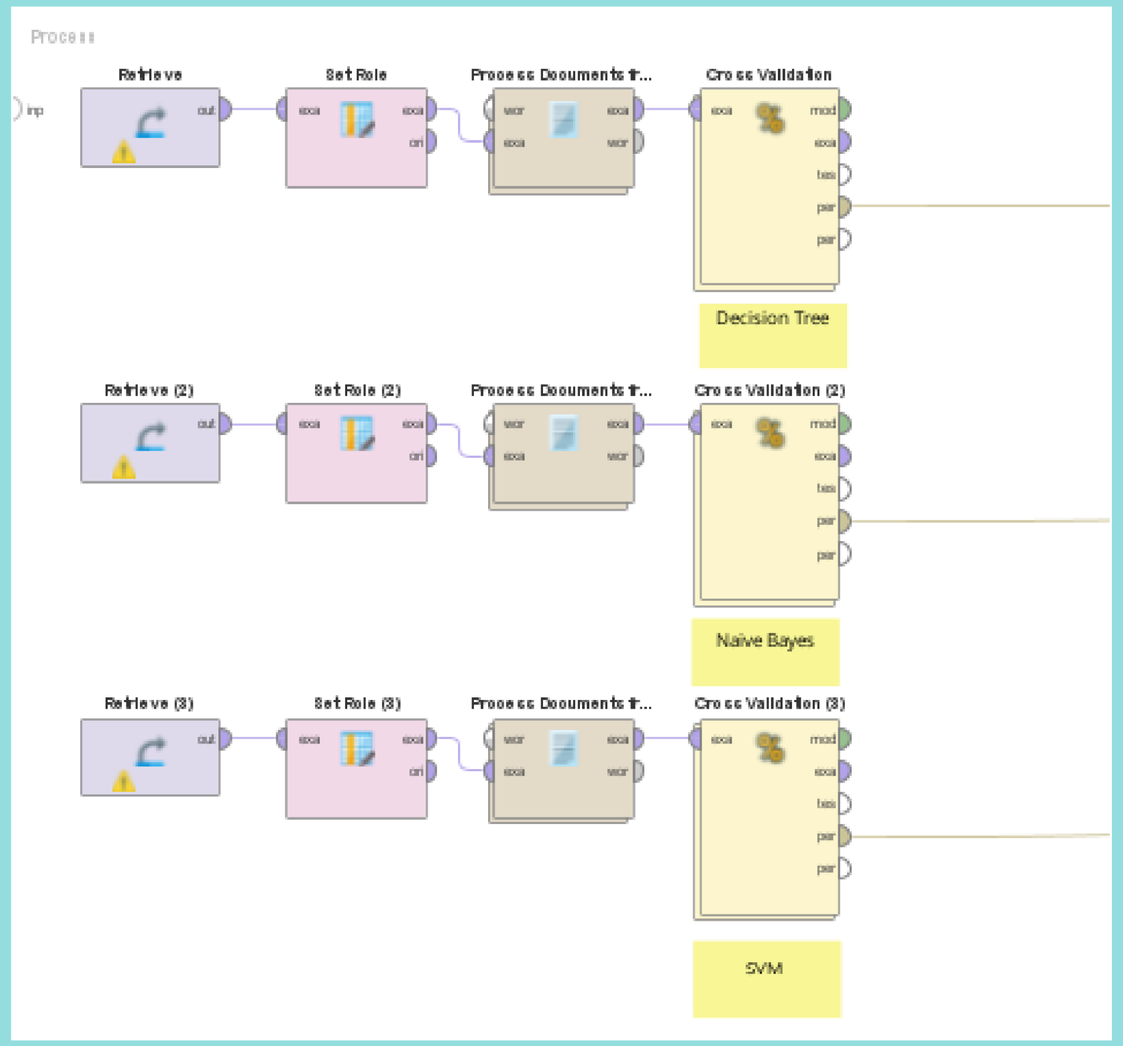
- Extract the sentiment analysis using Extract Sentiment operator.
- Create a new attribute named "Sentiment" using Generate Attributes operator.





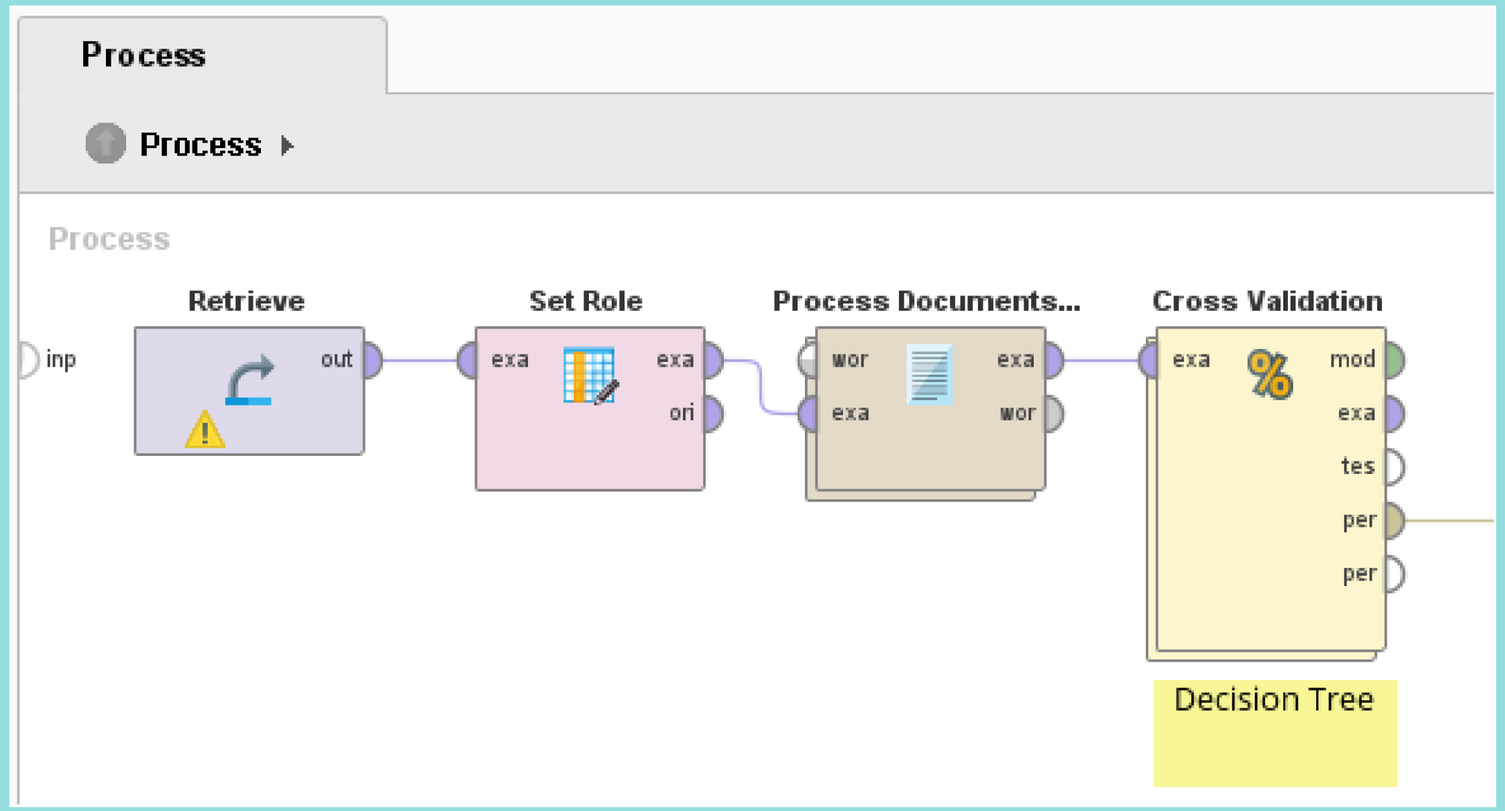


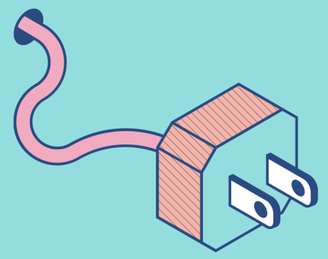
The three models have been combined to find out the comparison of their performance.



Each model has been through the steps below:

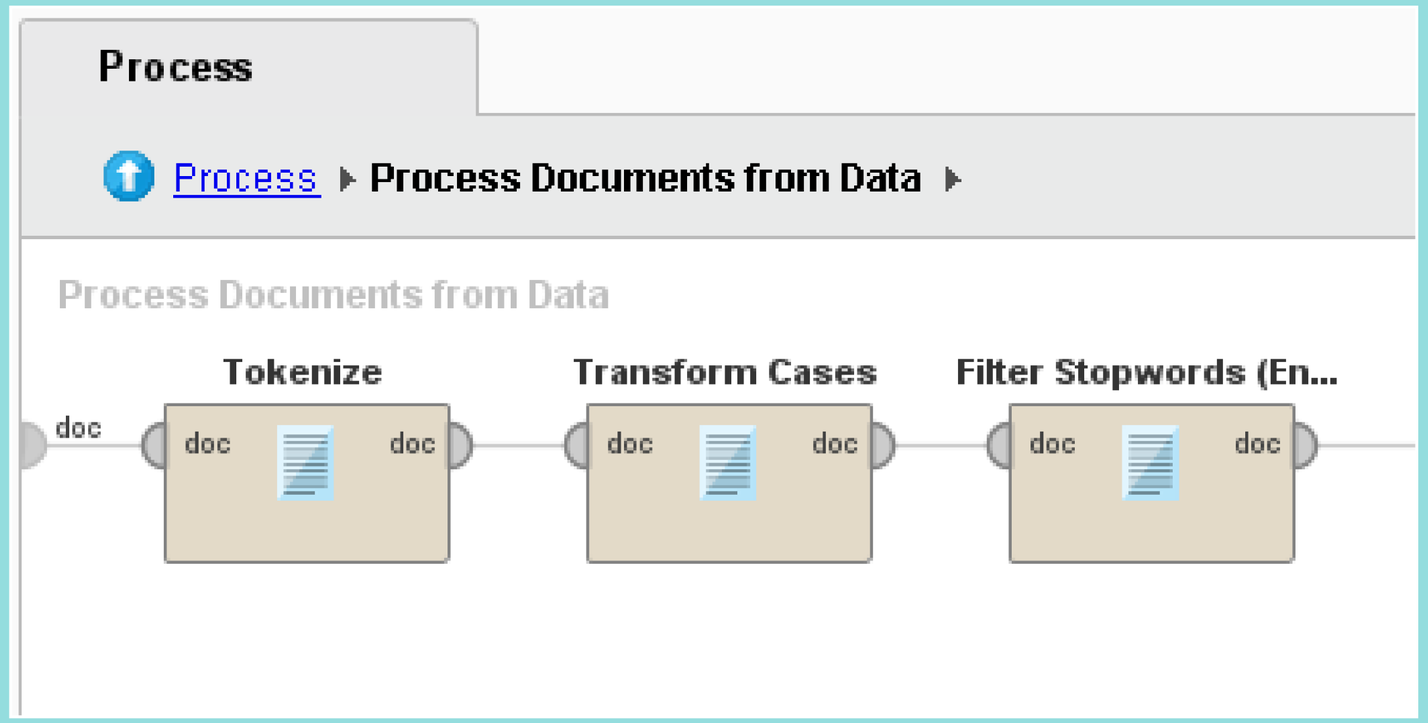
- Retrieve data from movie review dataset.
- The sentiment have been set as label in Set Role.
- Process Documents help remove unnecessary word.





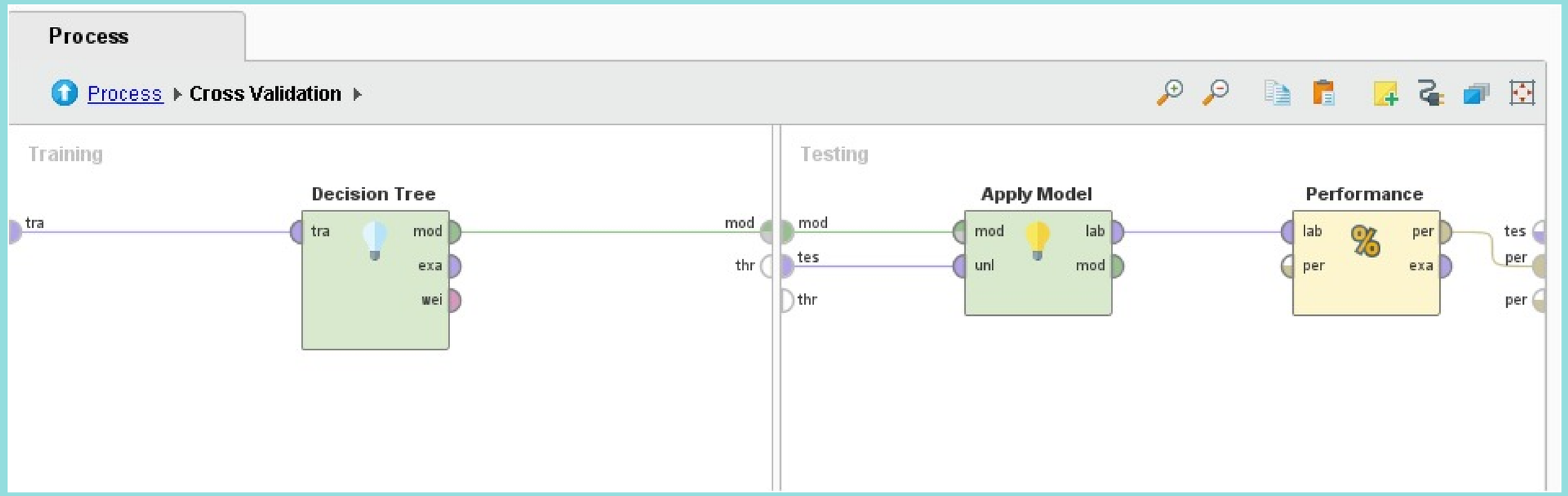
In Process Documents from Data, it went through 3 process:

- Tokenize -splits the text of a document into a sequence of tokens.
- Transform Cases-transforms all characters in a document to either lower case.
- Filter Stopword-filters English stopwords from a document by removing every token which equals a stopword from the built-in stopword list.



Cross Validation

- Cross Validation is used for testing and training data.
- In each Cross Validation, we include the model we want to use and apply the model to measure the performance.



The outputs on the :

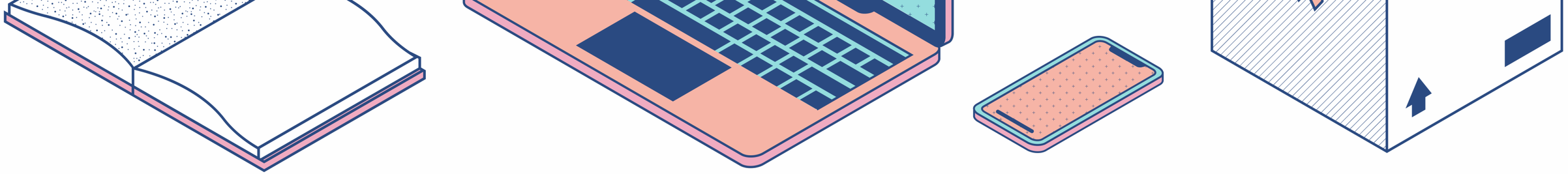
- The performance, including accuracy and AUC.

- The model applied to the document. The result is the prediction based on the vector that numerically represents the text. The vector is also included in the output.

EVALUATION

In this project, confusion matrix will be implemented to see the accuracy of the model. Below is the example of calculating accuracy using confusion matrix :

Accuracy = (TP + TN) / (TP + TN + FP + FN)



Decision Tree Confusion Matrix

Table View Plot View

accuracy: 76.26% +/- 2.11% (micro average: 76.26%)

	true Negative	true Positive	class precision
pred. Negative	24	36	40.00%
pred. Positive	285	1007	77.94%
class recall	7.77%	96.55%	

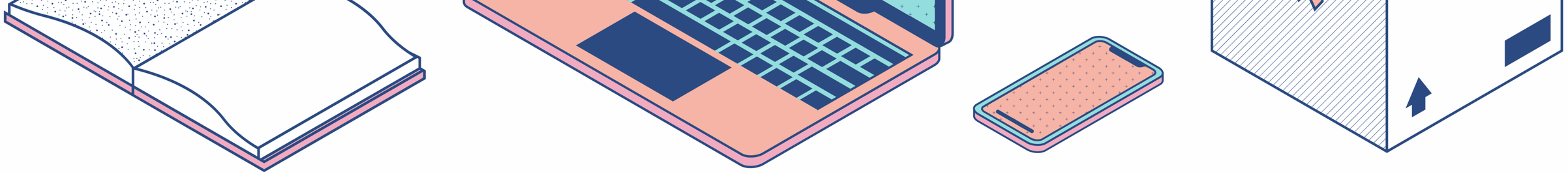
Accuracy: 76.26

TP:1007

TN: 24

FP: 285

FN: 36



Naive Bayes Confusion Matrix

Table View Plot View

accuracy: 71.97% +/- 2.18% (micro average: 71.97%)

	true Negative	true Positive	class precision
pred. Negative	61	131	31.77%
pred. Positive	248	912	78.62%
class recall	19.74%	87.44%	

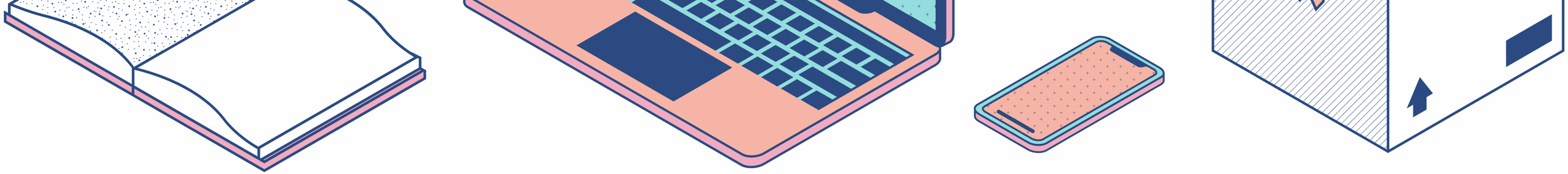
Accuracy: 71.97

TP:912

TN:61

FP:248

FN:131



Support Vector Machine Confusion Matrix

Table View Plot View

accuracy: 77.37% +/- 0.40% (micro average: 77.37%)

	true Negative	true Positive	class precision
pred. Negative	4	1	80.00%
pred. Positive	305	1042	77.36%
class recall	1.29%	99.90%	

Accuracy: 77.37

TP:1042

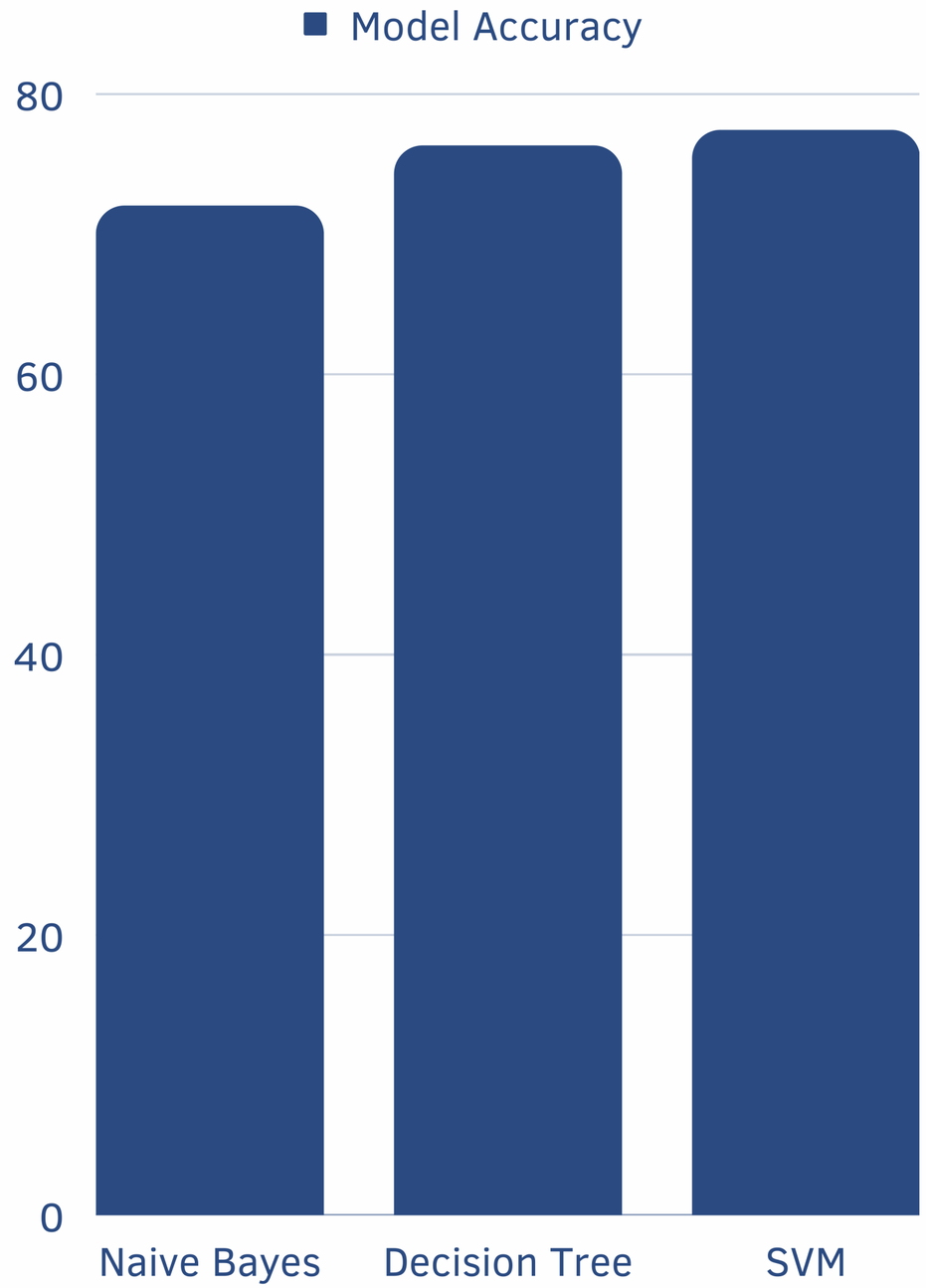
TN:4

FP:305

FN:1

Comparison of Model Accuracy

From the bar graph, we can conclude that SVM (77.37) have the highest accuracy compared to Decision Tree and Naive Bayes.



DEPLOYMENT

- Model is ready to be used.
- Model is implemented to the system.

CONCLUSION

- The research goal of this work is to address Sentiment Analysis by constructing an approach that can classify movie review sentiment and then compare the results in an inclusive study of three well-known classifiers. To evaluate the proposed model, IMDB reviews real dataset was utilized. Tokenization was applied on the dataset to transfer strings into word vectors, transform cases and filter stopwords are used to process the document, then the words gain ratio was applied to the dataset as an attribute selection algorithm. Then, the data was split into training and testing datasets using the Cross Validation with 70 training and 30% respectively. To evaluate the model, accuracy was used.
- It can be concluded that the Support Vector Machine ave the highest accuracy compared to Decision Tree and Naive Bayes.

References

Hamborg, F., & Donnay, K. (2021). NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. ACL Anthology. Retrieved January 19, 2023, from https://aclanthology.org/2021.eacl-main.142/

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Challenges of sentiment analysis in social networks. Sentiment Analysis in Social Networks, 1–11. https://doi.org/10.1016/b978-0-12-804412-4.00001-2

Rahman, A., & Hossen, M. S. (2019). Sentiment Analysis on Movie Review Data Using Machine Learning Approach. 2019 International Conference on Bangla Speech and Language Processing (ICBSLP). doi:10.1109/icbslp47725.2019.201470

References

Venkata Krishna, P., & Obaidat, M. S. (Eds.). (2020). Emerging Research in Data Engineering Systems and Computer Communications. Advances in Intelligent Systems and Computing. doi:10.1007/978-981-15-0135-7

Yasen, Mais; Tedmori, Sara (2019). [IEEE 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) - Amman, Jordan (2019.4.9-2019.4.11)] 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) - Movies Reviews Sentiment Analysis and Classification. , (), 860–865. doi:10.1109/JEEIT.2019.8717422