Sentiment Analysis for Product Review using Machine Learning and Deep Learning Methods

Supervised By

Mrs. Manoara Begum

Senior Lecturer & Chairman (in-charge **)**
Department of CSE
Port city international university

Presented By

Md. Minhajur Rahman

Batch : 18th (C)

ID : CSE 01806748

CONTENTS

Introduction

- In today's digitalized world, e-commerce is growing popularity since it brings things closer to customers without requiring them to leave their homes.
- Sentiment analysis helps to analyze these opinioned data and extract some important insights which will help to other user to make decision.
- In this competitive business world, various industries especially e-commerce immensely use sentiment analysis to increase productivity and make better business decisions.

Motivation

- Sentiment analysis is currently at the leading position of scientific research. Its applications are beneficial in many cases.
- This thesis investigates the sentiment classification problem for online reviews, employing supervised approaches to determine the overall semantic of customer reviews by classifying them as positive, negative, or neutral.
- A model could be assist possible clients with settling on an informed decision on their purchase and organizations to improve their items or services .

Objective

- My research aims to do this by conducting sentiment analysis on customer reviews and classifying them as positive, negative, or neutral.
- I believe that my suggested approach would reduce consumer frustration while shopping online since they will be able to examine product reviews based on the ratio of previous customers positive, negative, and neutral comments.
- It can also be beneficial to the vendor because it allows him to spot product flaws and give better customer service.

State Of The Art

Title	Author	Date & Publisher	Findings
1. Sentiment Analysis For Product Review	Najma Sultana, Pintu Kumar, Monika Rani Patra, Sourabh Chandra and S.K. Safikul Alam	International Journal of Soft Computing, April 2019	**Algorithm :** Naive Bayes, Logistic Regression, Linear SVC ,Decision Tree **Accuracy :** 89.85%.(NB) **Dataset :** Stanford dataset for 50,000 movie reviews. **Limitation:** They use machine learning algorithm.The preprocess part have only Punctution remove and used one kinds of feature extraction approach.
2. Sentiment Analysis For Amazon.Com Reviews	Levent Guner, Emilie Coyne and Jim Smit	International Conference on Intelligent Computing and Control (I2C2), March 2019	**Algorithm :** Linear SVM, Multinomial NB,LSTM network **Accuracy :** 90%.(LSTM) **Dataset :** Kaggle dataset and Amazon review data. **Limitation:** They use only three algorithm and use one feature extraction.

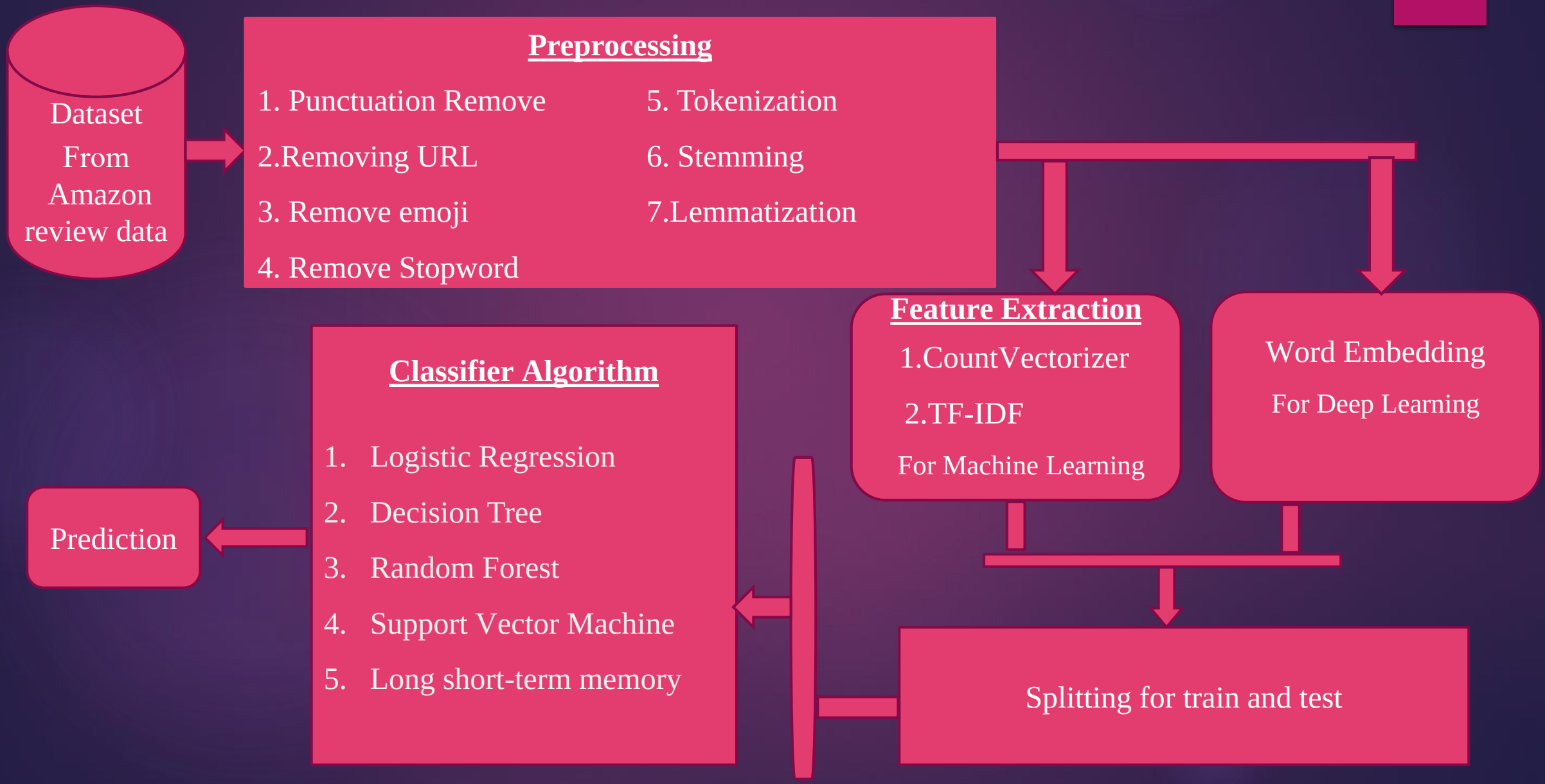
State Of The Art

Title	Author	Date & Publisher	Findings
3. Sentiment Analysis on Large Scale Amazon Product Reviews	Tanjim Ul Haque, Nudrat Nawal Saber and Faisal Muhammad Shah	International Conference on Innovative Research and Development (ICIRD), June 2018	**Algorithm :** Linear SVM, Multinomial NB, Stochastic Gradient Descent ,Random Forest, Logistic Regression, decision Tree **Accuracy :** 90%.(SVM) **Dataset :** Amazon review data**.** **Limitation:** They use a supervised learning method and used a mix of 2 kinds of feature extractor approach.
4.Sentiment Analysis on Product Reviews Using Machine Learning Techniques	Rajkumar S. Jagdale, Vishal S. Shirsat and Sachin N. Deshmukh	International Journal of Soft Computing, January 2019	**Algorithm :** Naive Bayes, SVM **Accuracy :** 98.17%(NB) **Dataset :** Amazon review data **Limitation:** They use only Two algorithm.

State Of The Art

Title	Author	Date & Publisher	Findings
5. Sentiment Analysis Of Customer Product Reviews Using Machine Learning	Zeenia Singla, Sukhchandan Randhawa and Sushma Jain	International Conference on Intelligent Computing and Control (I2C2), June 2017	**Algorithm :** Naive Bayes, SVM ,Decision Tree **Accuracy :** 81.77%. **Dataset :** Amazon review data. **Limitation:** Accuracy is low.

Methodology



Tools & Technologies

I. Python
II. Pandas
III. NumPy
IV. Matplotlib
V. Scikit-learn
VI. Google Colab

Dataset Collection

I collect the Amazon Review dataset from Julian McAuley,UCSD website. "Clothing, Shoes, and Jewelry" is the name of this dataset. This dataset contains 278 677 reviews. From this dataset, I took a total of 55000 reviews, which I used on my thesis. This dataset has been labeled with three classes, which are positive, negative, and neutral.

Data Preprocessing

Punctuation Remove

Removing URL

Remove emoji

Remove Stopword

Tokenization

Stemming

Lemmatization

Data Preprocessing

- Punctuation Remove : !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~

"love these glasses! perfect size"

↓

love these glasses perfect size

- Removing URL : https:\\.*|http:\\.*

Go to: https://example.com

↓

Go to

- Remove Stopwords :

I like this bag

↓

like , bag

- Tokenization : break sentences into parts

I love these glasses

↓

[I, love, these , glasses]

- Stemmng:

Operator, Going

↓

Operate, Go

- Lemmatization : Finding the same root

[Connection, Connected, Connecting]

↓

[Connect]

Feature Extraction

- Count Vectorizer : Vectorization is a process of converting the text data into a machine-readable form.

Text Input : [‘ this pair is great ‘]

Vocabulary : {'this': 3, 'pair': 2, 'is': 1, 'great': 0}

Encoded Text is: [[1 1 1 1]]

- TF-IDF : It is used to tokenize the documents learn the vocabulary and inverse the document frequency weightings, and allow to encode new documents.

Text Input : ['this pair is great', 'i love these glasses']

Vocabulary : {'this': 6, 'pair': 4, 'is': 2, 'great': 1, 'love': 3, 'these': 5, 'glasses': 0}

Encoded Text is: [1.40546511 1.40546511 1.40546511 1.40546511 1.40546511 1.40546511 1.40546511]

Classification Algorithm

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- Long short-term memory (LSTM)

Experimental Output

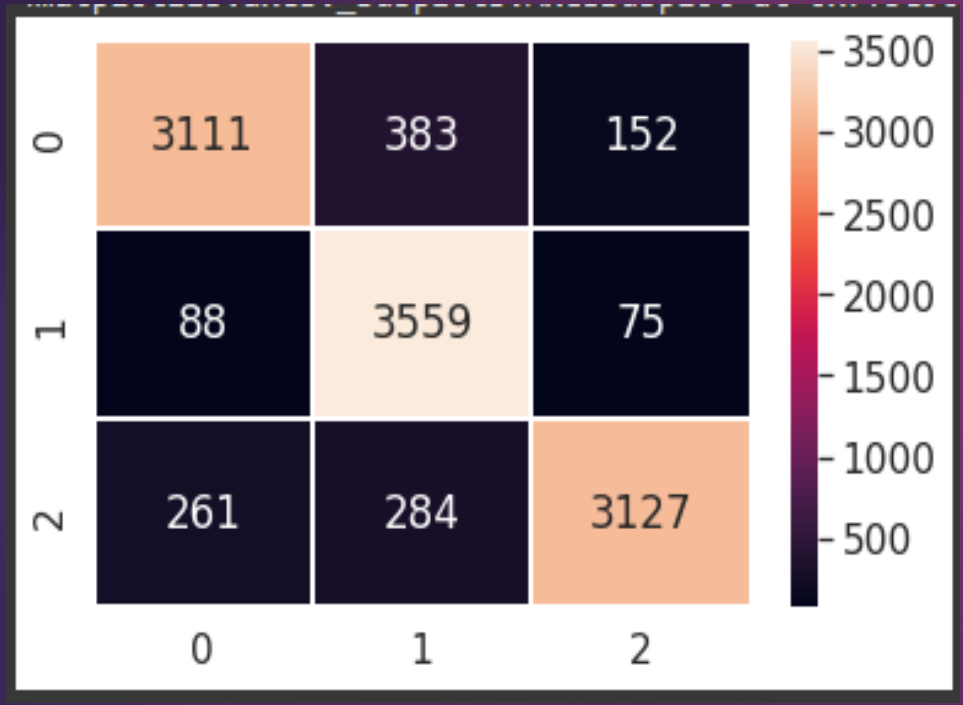

```
Input Text: Bad qualilty..... material qualilty is very poor
Output : Negative
```

```
Input Text :  ['Great quality']
Output :Positive
```

```
Input Text :  ['I buy this product from bangladesh']
Output : Neutral
```

Experimental result

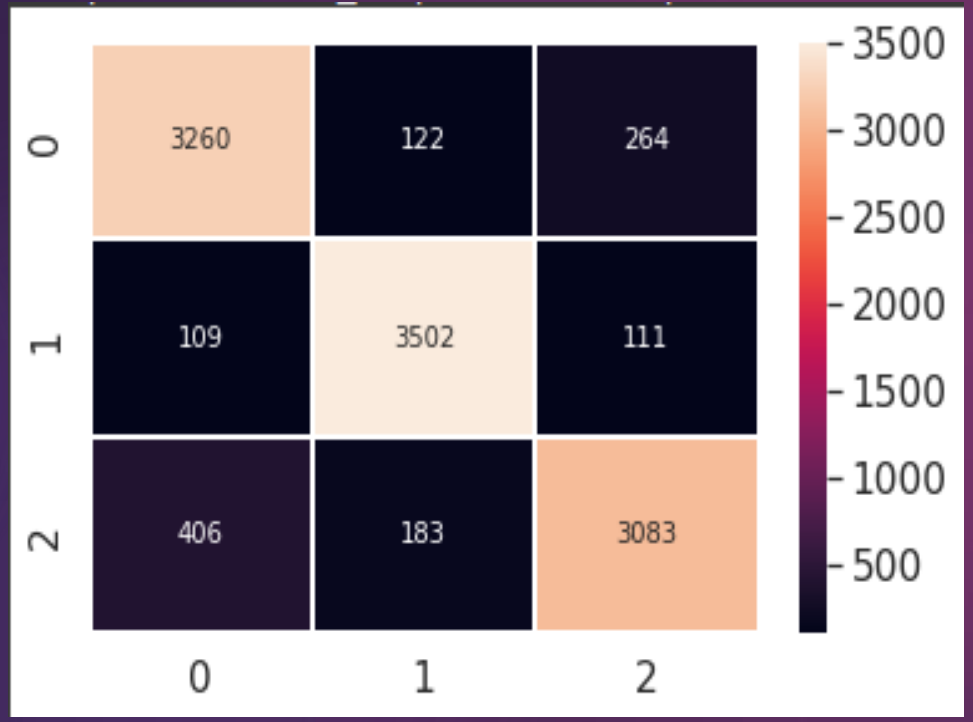
- Confusion Matrix of Logistic Regression



	precision	recall	f1-score	support
0	0.90	0.85	0.88	3646
1	0.84	0.96	0.90	3722
2	0.93	0.85	0.89	3672
accuracy			0.89	11040
macro avg	0.89	0.89	0.89	11040
weighted avg	0.89	0.89	0.89	11040

Experimental result

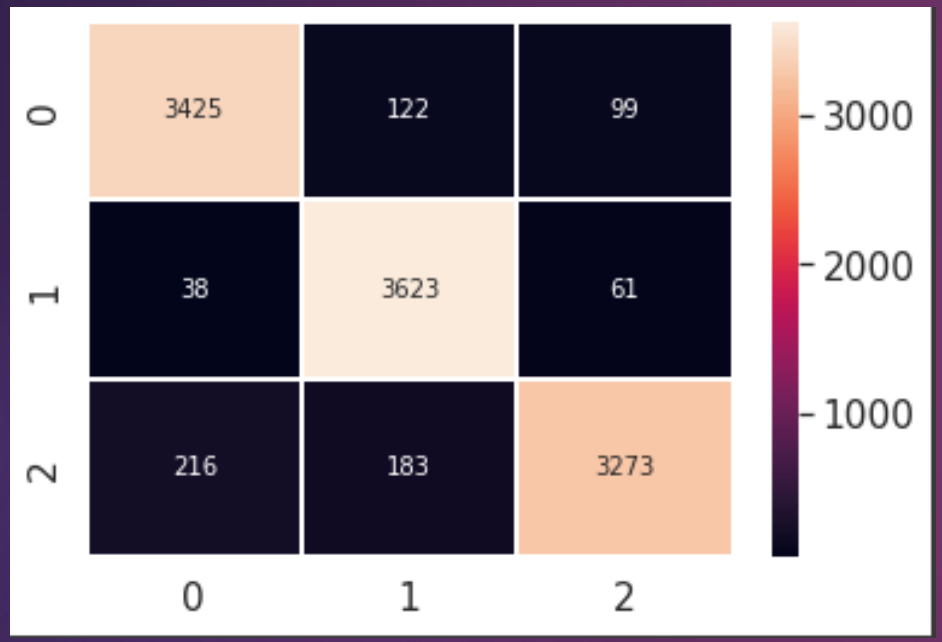
- Confusion Matrix of Decision Tree



	precision	recall	f1-score	support
0	0.86	0.89	0.88	3646
1	0.92	0.94	0.93	3722
2	0.89	0.84	0.86	3672
accuracy			0.89	11040
macro avg	0.89	0.89	0.89	11040
weighted avg	0.89	0.89	0.89	11040

Experimental result

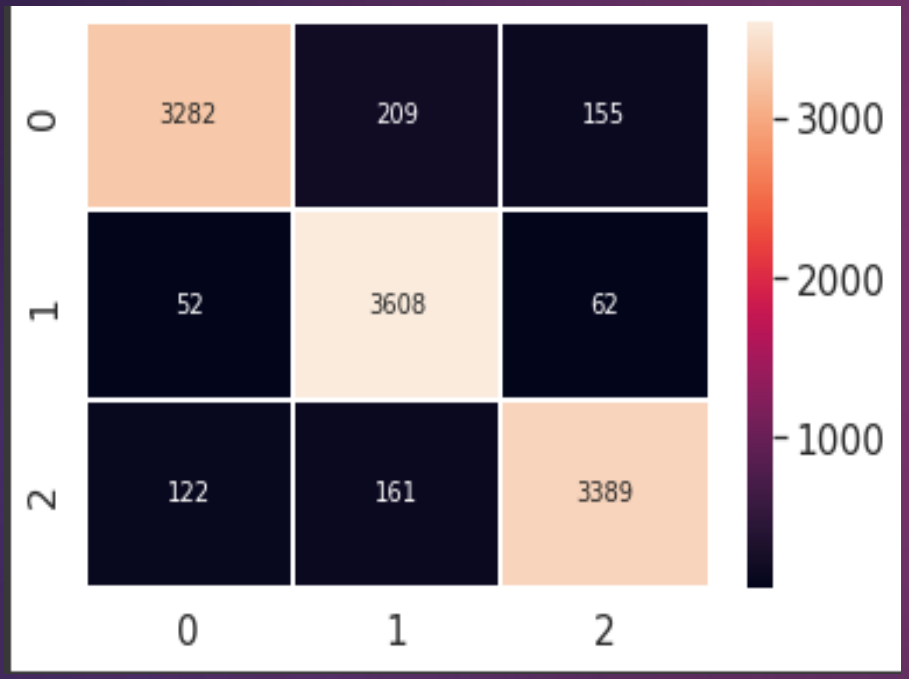
- Confusion Matrix of Random Forest



	precision	recall	f1-score	support
0	0.93	0.94	0.94	3646
1	0.92	0.97	0.95	3722
2	0.95	0.89	0.92	3672
accuracy			0.93	11040
macro avg	0.94	0.93	0.93	11040
weighted avg	0.94	0.93	0.93	11040

Experimental result

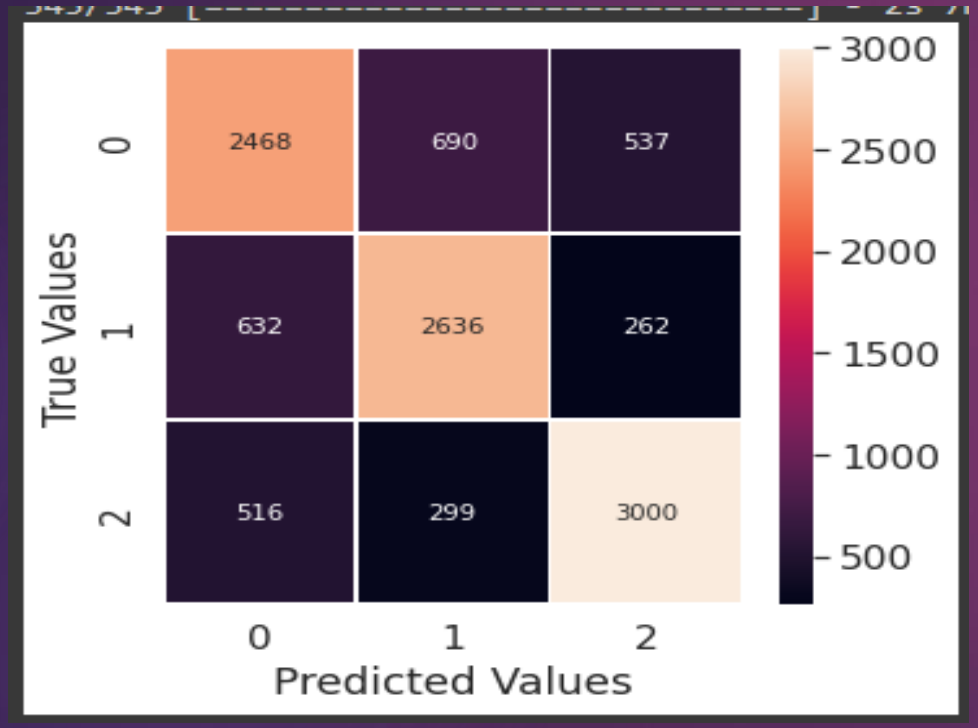
- Confusion Matrix of Support Vector Machine



	precision	recall	f1-score	support
0	0.95	0.90	0.92	3646
1	0.91	0.97	0.94	3722
2	0.94	0.92	0.93	3672
accuracy			0.93	11040
macro avg	0.93	0.93	0.93	11040
weighted avg	0.93	0.93	0.93	11040

Experimental result

- Confusion Matrix of Long Short-Term Memory



	precision	recall	f1-score	support
0	0.68	0.67	0.67	3616
1	0.76	0.71	0.74	3625
2	0.80	0.77	0.78	3799
micro avg	0.75	0.72	0.73	11040
macro avg	0.74	0.72	0.73	11040
weighted avg	0.75	0.72	0.73	11040
samples avg	0.72	0.72	0.72	11040

Experimental result

Performance comparison of different classification algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression (LR)	89%	89%	89%	89%
Decision Tree (DT)	89.14%	89%	89%	89%
Random Forest (RF)	93.55%	94%	93%	93%
Support Vector Machine (SVM)	93.13%	93%	93%	93%
Long Short-Term Memory (LSTM)	73.41%	74%	72%	73%

Experimental result

Performance comparison of the target paper dataset and my dataset

Algorithm	My Dataset	Target Paper Dataset
Logistic Regression (LR)	89%	89.51%
Decision Tree (DT)	89.14%	71.91%
Random Forest (RF)	93.55%	85%
Support Vector Machine (SVM)	93.13%	89.88%
Long Short-Term Memory (LSTM)	73.41%	87.97%

Conclusion

- My research aims to do this by conducting sentiment analysis on customer reviews and classifying them as positive, negative, or neutral.
- Out of the five classifiers, i.e. RF, DT, SVM, LR and LSTM, predictive accuracy of RF is found to be the best. The highest value of accuracy achieved was 93.55% for RF among the five models.
- I believe that my suggested approach would reduce consumer frustration while shopping online since they will be able to examine product reviews based on the ratio of previous customers' positive, negative, and neutral comments. It can also be beneficial to the vendor because it allows him to spot product flaws and give better customer service.

Future work

This Work can be extended in the following manner in future

- Try to use more feature extraction.
- To use more Deep Learning Models.
- Improve deep learning model accuracy
- Adding more data.

References

- Najma Sultana, Pintu Kumar, Monika Rani Patra, Sourabh Chandra And S.K. Safikul Alam, Sentiment Analysis For Product Review, researchgate.net, ICTACT Journal On Soft Computing, April 2019.
- Levent Guner,Emilie Coyne,Jim Smit,Sentiment analysis for Amazon.com reviews,researchgate.net, Big Data in Media Technology ,March 2019.
- Tanjim Ul Haque, Nudrat Nawal Saber, Faisal Muhammad Shah, Sentiment Analysis on Large Scale Amazon Product Reviews, IEEE, International Conference on Innovative Research and Development (ICIRD),June 2018.
- Rajkumar S. Jagdale, Vishal S. Shirsat and Sachin N. Deshmukh,Sentiment Analysis on Product Reviews Using Machine Learning Techniques,researchgate.net,January 2019.
- Zeenia Singla, Sukhchandan Randhawa, Sushma Jain, Sentiment Analysis Of Customer Product Reviews Using Machine Learning, IEEE,June 2017.
- Palak Baid, Apoorva Gupta and Neelam Chaplot on a research paper named ‘Sentiment Analysis of Movie Reviews using Machine Learning Techniques’, researchgate.net, International Journal of Computer Applications (0975 – 8887), December 2017.

References

- LI YANG ,YING LI , JIN WANG AND R. SIMON SHERRATT ‘Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning’ ,IEEE, January 27, 2020.
- Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat and Priyanka Badhani ’ Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python’, researchgate.net, International Journal of Computer Applications, May 2017.
- Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain,’ STATISTICAL AND SENTIMENT ANALYSIS OF CONSUMER PRODUCT REVIEWS’,IEEE, December 2017.
- Mst. Tuhin Akter , Manoara Begum and Rashed Mustafa ‘ Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors’IEEE, International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), April 2021.
- Dataset name “ Clothing, Shoes, and Jewelry ”. Available : http://jmcauley.ucsd.edu/data/amazon/index_2014.html.

Thank You