Natural Language Processing: Classify Amazon reviews based on the customer's ratings.

By: JOE ANSON R. AQUINO

August 18, 2023

Springboard

Springboard Data Science

Introduction

Many times, ratings are represented by a numerical value () or stars (). However, the text feedback holds more value than the quantified ratings. Sometimes, the rating given may not accurately reflect the experience of the product. Given the text of the review of a product, we want to build a supervised, binary classifier model with the actual review text as the core predictor.

Approach

First, will gather data from Amazon Dataset contains the customer reviews for all listed *Electronics*, will do NLP Pre-Processing, Tokenization, Phrase Modeling, Vectorization before exploring the data before utilizing a machine-learning algorithm to predict if the review is negative or positive review. To justify this work, I conducted a comparative study and analysis using various classification algorithms, including Random Forest Classifier, xgBoost Classifier.

Amazon Dataset

- Contains customer reviews for all listed Electronics products from May 1996 up to July 2014.
- 1,689,188 reviews by 192,403 customers on 63,001 unique products

	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	0528881469	[0, 0]	5	We got this GPS for my husband who is an (OTR)...	06 2, 2013	AO94DHGC771SJ	amazdnu	Gotta have GPS!	1370131200
1	0528881469	[12, 15]	1	I'm a professional OTR truck driver, and I bou...	11 25, 2010	AMO214LNFCEI4	Amazon Customer	Very Disappointed	1290643200
2	0528881469	[43, 45]	3	Well, what can I say. I've had this unit in m...	09 9, 2010	A3N7T0DY83Y4IG	C. A. Freeman	1st impression	1283990400
3	0528881469	[9, 10]	2	Not going to write a long review, even thought...	11 24, 2010	A1H8PY3QHMQQA0	Dave M. Shaw "mack dave"	Great grafics, POOR GPS	1290556800

Unique ID of the product, *str*

Number of users that voted helpful; Total number of users that voted on the review, *list*

Reviewer's rating of product, *int*

Review text itself, *str*

Time when review was posted, *str*

Unique ID of the reviewer, *str*

Specified name of reviewer, *str*

Headline summary of review, *str*

Unix time when review was posted, *str*

Data Wrangling - NLP Pre-Processing

- The Final dataframe for the model will be drawn from the reviewText column.
- The overall column will serve as the ground truth labels

```
Let just start off by saying that I have tried the Kindle and although an OK hardware device, the interface was ter
rible.  It felt like the early days of the internet, with textual interfaces, underlined words, fiddling with butto
ns and keyboard to navigate, and all sorts of annoyances that got in the way of reading and enjoying the device. We
returned it within 3 days.Next we tried the original Nook Black&White;, and interface was much better, but the scre
en had poor contrast. (not the nice Pearl screen as the Kindle has)Enter Nook Touch.  WOW.  This thing has a top no
tch interface... they totally put their minds and heart into this one...  it is PERFECT.  Touchscreen makes a world
of a difference, and enables the slick interface that allows you to browse the store, personal library, and more.
On-screen Buttons and finger gestures make the magic happen, and this is one device you are sure to LOVE TO USE.  T
he screen is the new Pearl, so it looks just as good as the Kindle, with the added benefit of touchscreen!The size?
PERFECT!  Not too big, not too small.  Very thin.  Well built.  Feels durable.All I can say is this is finally the
ereader I've been waiting for...  excellent battery life, good Pearl screen, touchscreen, and perfect size.To boot,
the price is dead on!  $139 with Wifi.  If it gets stolen, I am buying another one immediately!
```

NLP Pre-Processing

- HTML Entities
- Lemmatization
- Accents
- Punctuations
- Lowercasing
- Stop Words
- Single Whitespaces

Tokenization

the document is broken down individually into words or tokens.

['im', 'big', 'fan', 'brainwavz', 's1', 'actually', 'headphone', 'yet', 'disappoint', 'product', 's1', 'main', 'set', 'active', 'use', 'e', 'g', 'workouts', 'run', 'etc', 'since', 'flat', 'cable', 'durable', 'resistant', 'tangle', 's5', 'keep', 'good', 'feature', 's1', 'add', 'sound', 'quality', 'rich', 'well', 'define', 'thats', 'say', 's1', 'sound', 'poor', 'quite', 'good', 'fact', 's5', 'well', 'high', 'well', 'define', 'midrange', 'punch', 'bass', 'come', 'clearly', 'without', 'move', 'harsh', 'territory', 'volume', 'push', 's1s', 'overall', 'sound', 'quality', 'please', 'build', 'quality', 'seem', 'solid', 'solid', 's1', 'good', 'love', 'flat', 'cable', 'know', 'thats', 'something', 'appreciate', 'everyone', 'work', 'wonderfully', 'although', 'brainwavz', 'headset', 'come', 'excellent', 'hard', 'shell', 'case', 'usually', 'tote', 'earbuds', 'wrap', 'around', 'mp3', 'player', 'pocket', 'easy', 'carry', 'stressful', 'cable', 'lead', 'tangle', 'round', 'wire', 'flat', 'wire', 'especially', 'thick', 'jacket', 'survive', 'abuse', 'zero', 'problem', 'earbuds', 'sleekly', 'shape', 'style', 's1', 'comfort', 'line', 'customary', 'brainwavz', 'style', 'say', 'outstanding', 'come', 'wide', 'range', 'tip', 'fit', 'pretty', 'much', 'ear', 'plus', 'comply', 'foam', 'tip', 'favorite', 'fit', 'properly', 'end', 'zero', 'ear', 'irritation', 'plus', 'excellent', 'sound', 'isolation', 'bass', 'response', 'ear', 'design', 'much', 'like', 's1', 'never', 'use', 'design', 'prior', 's1', 'take', 'little', 'time', 'get', 'accustom', 'become', 'second', 'nature', 'quickly', 'design', 'lot', 'stable', 'exercise', 'conventional', 'ear', 'design', 'expensive', 's1', 'hear', 'difference', 'price', 'still', 'look', 'keep', 'cost', 'bit', 's1', 'excellent', 'performer', 'well', 'great', 'sound', 'great', 'comfort', 'wonderful', 'cable', 'design', 'come', 'solidly', 'make', 'case', 'lot', 'eartips', 'highly', 'recommend', 'sample', 'provide', 'review']

Phase Modeling

Bi-grams

['hdmi_dvi', 'lens_without', 'time_forget', 'like_return', '2_00', 'fast_run', 'make_convenient', 'point_think', 'matter_fact', 'although_make', 'actually_see', 'sure_problem', 'course_good', 'get_catch', 'take_find', 'include_product', 'problem_design', 'work_everything', 'standard_camera', '1080p_120hz', 'make_give', 'set_ipad', 'control_cable', 'nikon_brand', 'really_beat', 'game_also', 'tiny_size', 'tiny_camera', 'use_default', 'color_come', 'get_12', 'plug_network', 'piece_technology', 'light_fit', 'button_click', '4kb_qd', 'wheel_click', 'wish_purchase', 'hold_device', 'ipod_phone', 'might_break', 'work_need', 'big_small', 'tell_would', 'lot_high', 'noise_ratio', 'less_200', 'star_seem', 'design_camera', 'camera_function’]

Tri-grams

['play_blu_ray', 'samsung_galaxy_s4', 'old_macbook_pro', 'quality_top_notch', 'b_w_filter', 'one_living_room', 'mac_os_x', 'far_exceed_expectation', 'nexus_7_2013', 'cell_phone_use', 'customer_service_great', '5d_mark_iii', 'cell_phone_camera', 'macbook_pro_work', 'first_blu_ray', 'case_nexus_7', 'double_sided_tape', 'price_highly_recommended', 'almost_non_existent', '2_4ghz_5ghz', 'macbook_pro_13', 'customer_service_rep', 'samsung_840_pro', 'blu_ray_disk', 'use_third_party', 'n_uuml_vi', 'home_theater_pc', 'complete_waste_money', 'small_form_factor', 'use_home_theater', 'fast_forward_rewind', 'wi_fi_connection', 'amazon_return_policy', 'new_kindle_fire', '192_168_1', 'aps_c_sensor', 'ear_bud_come', 'mp3_player_work', 'mp3_player_use', 'use_macbook_pro', 'run_os_x', 'canon_5d_mark', 'blu_ray_movie', 'western_digital_passport', 'dd_wrt_firmware', 'inch_macbook_pro', 'heart_rate_monitor', 'great_mp3_player', 'kindle_fire_hd', 'samsung_galaxy_tab']

Count-based Feature Engineering

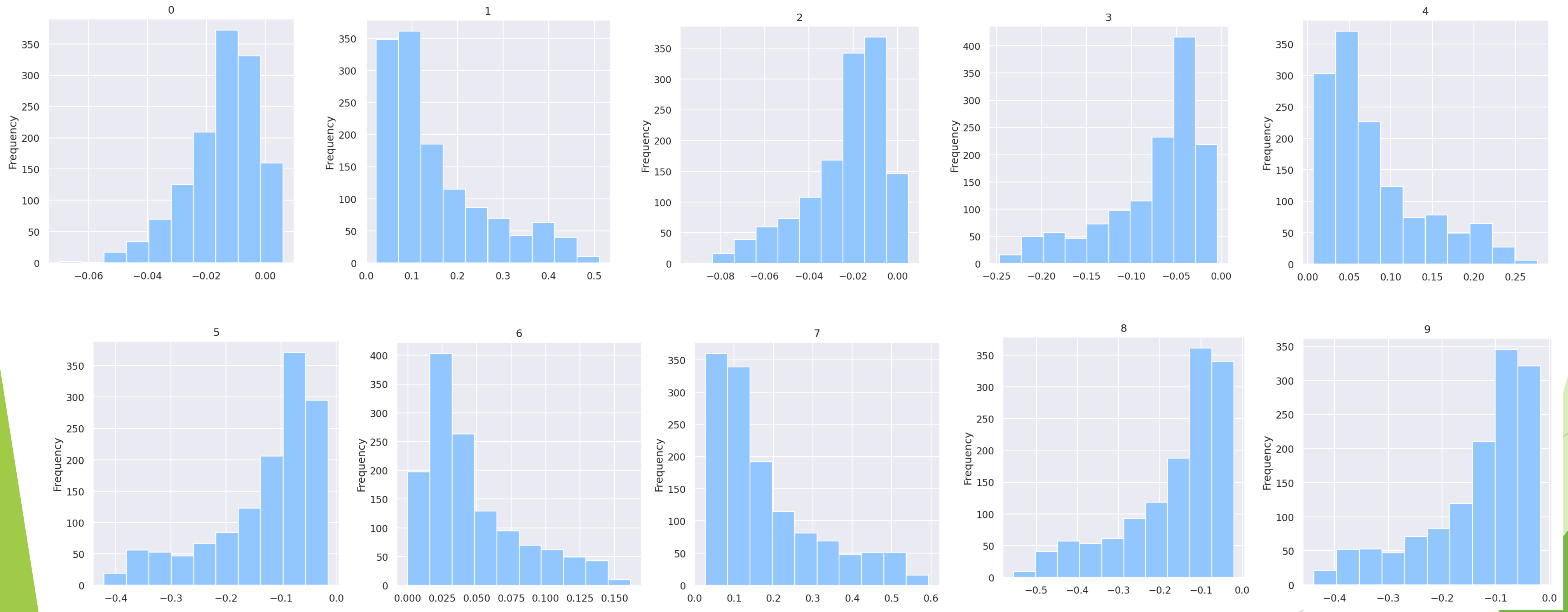
Bag of Words

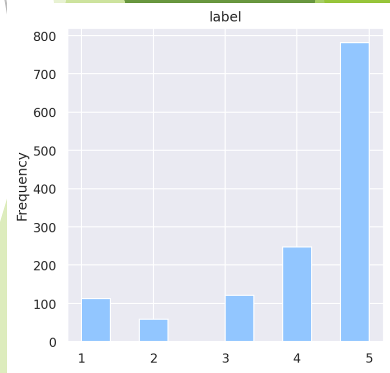
Word: address, Frequency: 1 Word: around, Frequency: 1 Word: arrive, Frequency: 1 Word: back, Frequency: 1 Word: bad, Frequency: 1 Word: big, Frequency: 2 Word: come, Frequency: 1 Word: contact, Frequency: 1 Word: could, Frequency: 1 Word: day, Frequency: 1 Word: earlier, Frequency: 1 Word: ease, Frequency: 2 Word: ect, Frequency: 1 Word: email, Frequency: 2 Word: exception, Frequency: 1 Word: exchange, Frequency: 1 Word: expect,

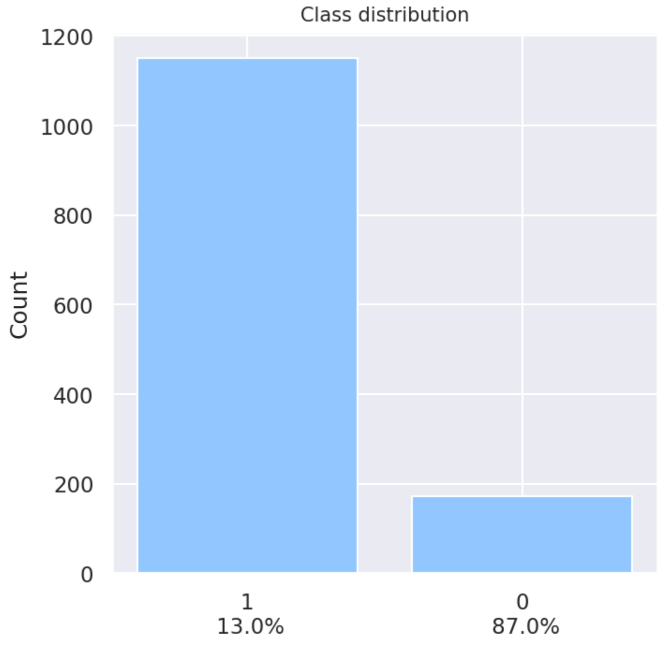
TF-IDF

Word: address, Weight: 0.113 Word: around, Weight: 0.060 Word: arrive, Weight: 0.093 Word: back, Weight: 0.051 Word: bad, Weight: 0.068 Word: big, Weight: 0.126 Word: come, Weight: 0.046 Word: contact, Weight: 0.103 Word: could, Weight: 0.054 Word: day, Weight: 0.061 Word: earlier, Weight: 0.141 Word: ease, Weight: 0.220 Word: ect, Weight: 0.181 Word: email, Weight: 0.213 Word: exception, Weight: 0.131 Word: exchange, Weight: 0.132 Word: expect, Weight: 0.067 Word: freeze, Weight: 0.259 Word: get, Weight: 0.028 Word: glitch, Weight: 0.141 Word: gps, Weight: 0.102 Word: great, Weight: 0.059 Word: however, Weight: 0.064

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
0	-0.042715	0.449554	-0.058469	-0.210948	0.221794	-0.406341	0.152388	0.559175	-0.487808	-0.415364	-0.309774	-0.355006	-0.060860	-0.071051	-0.072163	-0.282156	0.037355	-0.639570	-0.387473	-0.854952	0.251314	0.505311	-0.0
1	-0.058009	0.447886	-0.060280	-0.213946	0.239694	-0.400263	0.140340	0.551776	-0.484460	-0.416487	-0.322385	-0.357648	-0.065022	-0.071069	-0.066520	-0.268836	0.042529	-0.635724	-0.382373	-0.840551	0.241458	0.491586	-0.0
2	-0.042507	0.462787	-0.075884	-0.228007	0.244245	-0.383637	0.138473	0.547498	-0.502998	-0.400627	-0.338452	-0.356055	-0.046745	-0.082042	-0.062373	-0.293004	0.040697	-0.641002	-0.398294	-0.861069	0.231807	0.502349	-0.0
3	-0.051971	0.446820	-0.068522	-0.217371	0.231471	-0.385490	0.139727	0.540269	-0.464324	-0.403209	-0.304717	-0.333408	-0.049585	-0.087214	-0.057092	-0.276166	0.045084	-0.632086	-0.375502	-0.814678	0.233380	0.483056	-0.0
4	-0.046612	0.502452	-0.093345	-0.246627	0.276240	-0.374039	0.146505	0.557952	-0.554885	-0.401337	-0.378426	-0.362731	-0.063099	-0.092559	-0.087104	-0.311212	0.044388	-0.688007	-0.426816	-0.916759	0.238222	0.535650	-0.0







Class Distributions :

0 is Positive review 87%

1 is Negative review 13%

Classification Report

Classification Report for TRAINING Data

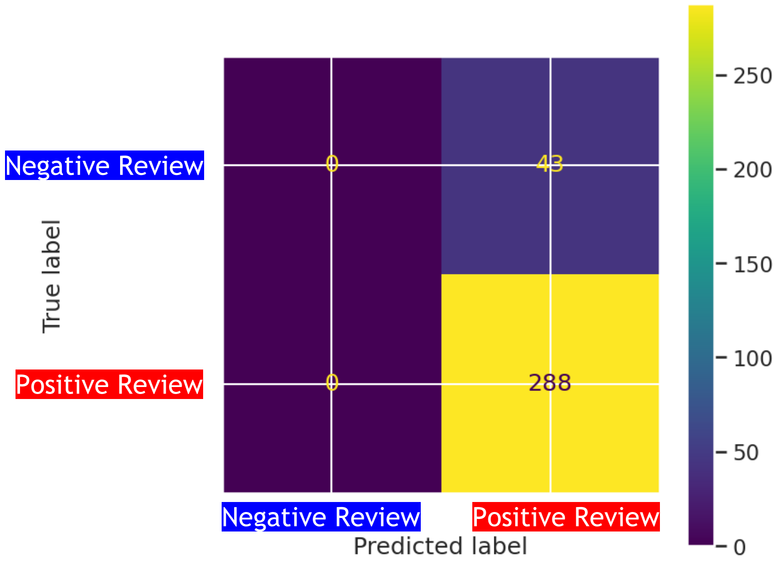
	Precision	Recall	F1-Score	Support
Negative	0.00	0.00	0.00	132
Positive	0.88	1.00	0.93	924
Accuracy			0.88	1056
Macro avg	0.44	0.50	0.47	1056
Weighted avg	0.77	0.88	0.82	1056

Classification Report for TEST Data

	Precision	Recall	F1-Score	Support
Negative	0.00	0.00	0.00	39
Positive	0.85	1.00	0.92	265
Accuracy			0.85	265
Macro avg	0.43	0.50	0.46	265
Weighted avg	0.73	0.85	0.79	265

Classification Prediction Report

Confusion Matrix



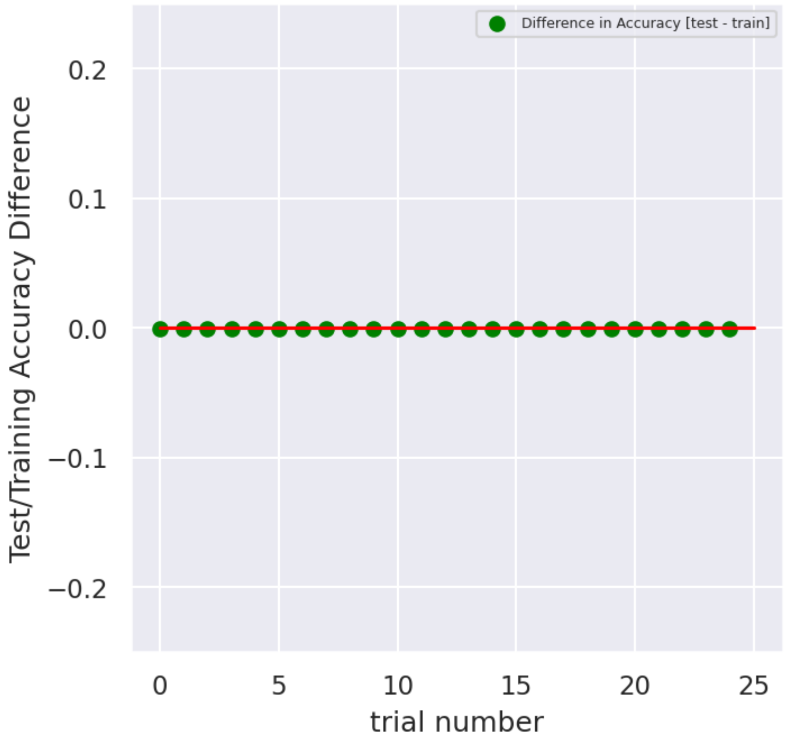
Classification Report for TRAINING Data

	Precision	Recall	F1-Score	Support
Negative	0.00	0.00	0.00	128
Positive	0.87	1.00	0.93	862
Accuracy			0.87	990
Macro avg	0.44	0.50	0.47	990
Weighted avg	0.76	0.87	0.81	990

Classification Report for TEST Data

	Precision	Recall	F1-Score	Support
Negative	0.00	0.00	0.00	43
Positive	0.87	1.00	0.93	288
Accuracy			0.87	331
Macro avg	0.44	0.50	0.47	331
Weighted avg	0.76	0.87	0.81	331

The difference in Accuracy [Test – Train]



Logistic Regression Classifier

Logistic Regression Classifier Accuracy : 0.8701				
	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	0
1	1.00	0.87	0.93	331
Accuracy			0.87	331
Macro avg	0.50	0.44	0.47	331
Weighted avg	1.00	0.87	0.93	331

Logistic Regression with GridSearchCV Accuracy : 0.8701				
	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	0
1	1.00	0.87	0.93	331
Accuracy			0.87	331
Macro avg	0.50	0.44	0.47	331
Weighted avg	1.00	0.87	0.93	331

Light GBM Classifier

LGBM Classifier Accuracy : 0.861				
	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	3
1	0.99	0.87	0.93	328
Accuracy			0.86	331
Macro avg	0.49	0.43	0.46	331
Weighted avg	0.98	0.86	0.92	331

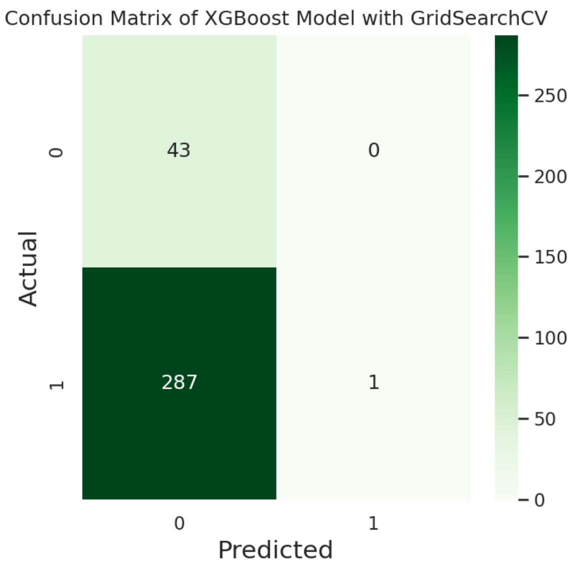
LGBM Classifier with GridSearchCV Accuracy : 0.864				
	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	2
1	0.99	0.87	0.93	328
Accuracy			0.86	331
Macro avg	0.50	0.43	0.46	331
Weighted avg	0.99	0.86	0.92	331

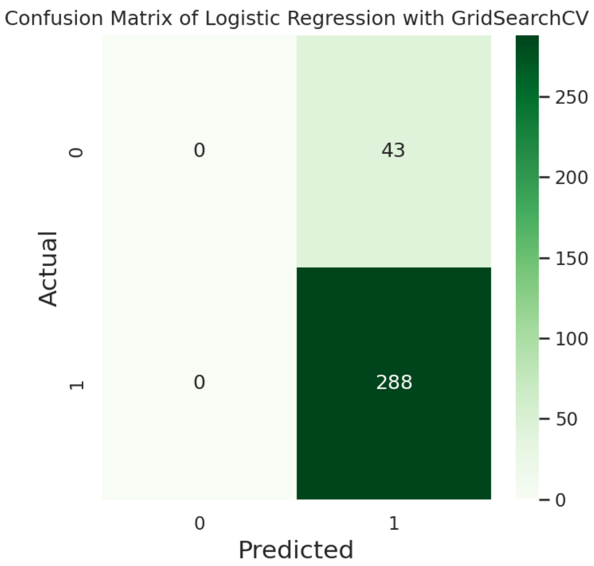
XGBoost Classifier

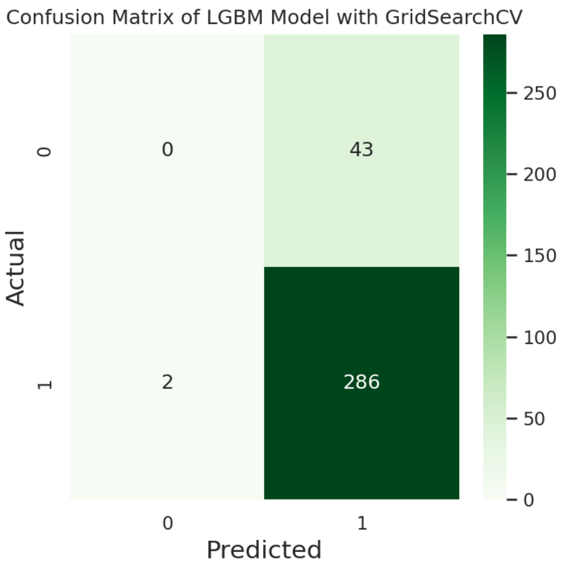
XGBoost Classifier Accuracy : 0.861				
	Precision	Recall	F1-Score	Support
0	0.05	0.29	0.08	7
1	0.98	0.87	0.92	324
Accuracy			0.86	331
Macro avg	0.51	0.58	0.50	331
Weighted avg	0.96	0.86	0.91	331

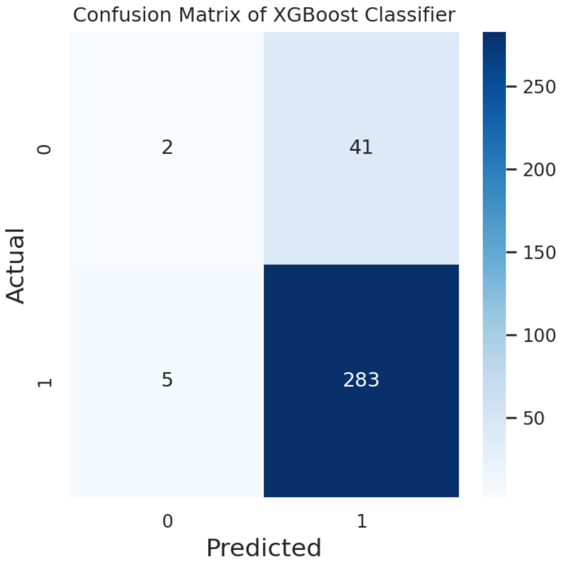
XGBoost Classifier Accuracy with GridSearchCV Accuracy				
	Precision	Recall	F1-Score	Support
0	1.00	0.13	0.23	330
1	0.00	1.00	0.01	1
Accuracy			0.13	331
Macro avg	0.50	0.57	0.12	331
Weighted avg	1.00	0.13	0.23	331

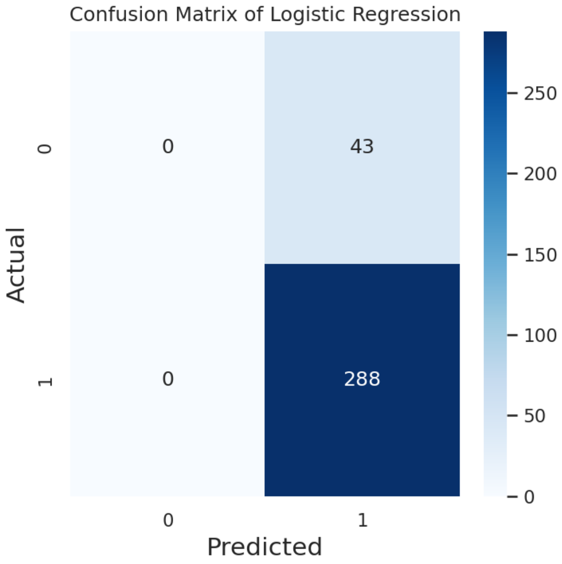
index	Train_Accuracy	Test_Accuracy	Train_Precision	Test_Precision	Train_Recall	Test_Recall	Train_F1-score	Test_F1-score
Logistic_Regression	0.870707	0.870090	0.870707	0.870090	1.0	1.0	0.810528	**0.809648**
LGBM	0.870707	0.870091	0.870707	0.870091	1.0	1.0	0.810529	**0.80**
XGBoost	0.870707	0.870091	0.870707	0.870091	1.0	1.0	0.810529	**0.809648**

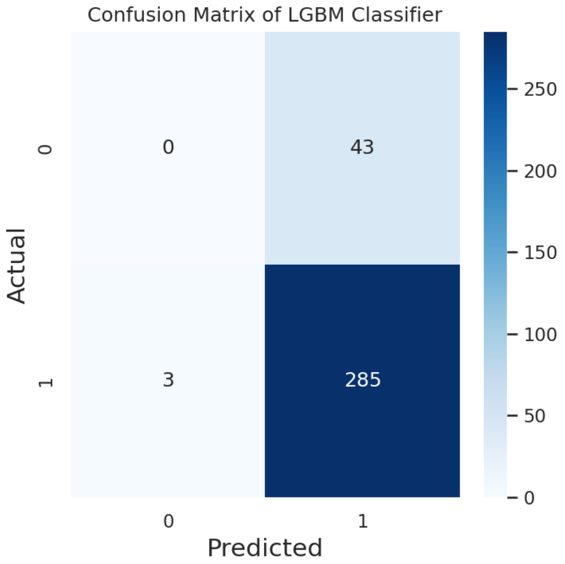

Confusion Matrix of XGBoost Model with GridSearchCV
43
0
287
1
Actual
Predicted
0
1
250
200
150
100
50


Confusion Matrix of Logistic Regression with GridSearchCV
0
43
0
288
Actual
Predicted
0
1
250
200
150
100
50


Confusion Matrix of LGBM Model with GridSearchCV
0
43
2
286
Actual
Predicted
0
1
250
200
150
100
50

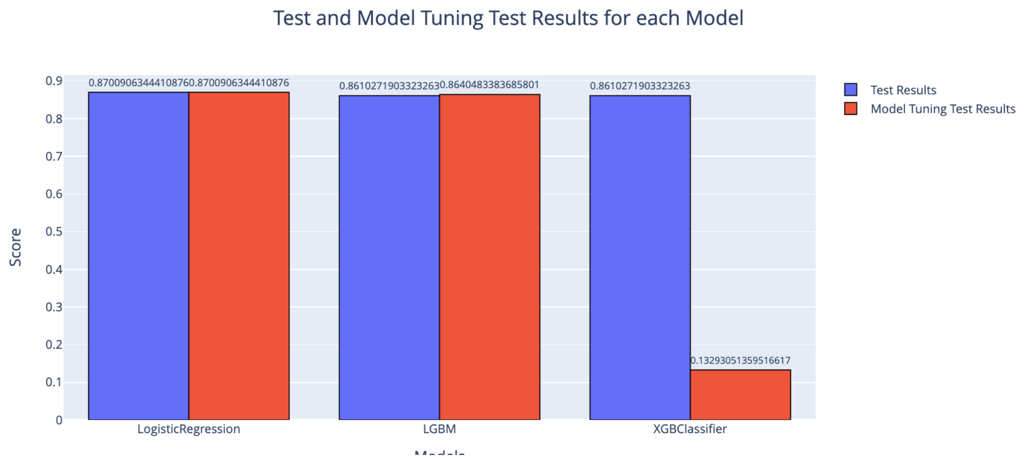

Confusion Matrix of XGBoost Classifier
2
41
5
283
Actual
Predicted
0
1
250
200
150
100
50


Confusion Matrix of Logistic Regression
0
43
0
288
Actual
Predicted
0
1
250
200
150
100
50


Confusion Matrix of LGBM Classifier
0
43
3
285
Actual
Predicted
0
1
250
200
150
100
50

Results

Model	Default	With GridSearchCV
Logistic Regression	0.870091	0.870091
LGBM	0.861027	0.864048
XGBoost	0.861027	0.132931



Conclusion and Recommendation

- Various NLP techniques and concepts were explored in the study.
- Though word embedding was central to building the model, pre-processing steps were crucial.
- The model extracts and quantifies context; therefore, the essence of a review by its words is the final Dataframe.
- Using deep learning techniques to predict more accurate review.

- It is recommended to have **Additional data** from many sources could be taken so that the models would be able to predict more accurate reviews.
- Use **Logistic Regression and XGBoost**, which had the best performance, could be deployed in real-time to provide doctors with faster inference results. This could aid in the diagnosis of whether a person is suffering from heart disease or not.