Natural Language Process on Amazon reviews

James Liao 490851 Zaiyu Huang 490908 Freya Xu 489711 Eunice Wang 487717 Robin Li 499985

Executive summary

Amazon, the world's largest internet online retailer, is regarded as a treasury for customer real-time feedback. Research also shows that the ranking algorithm used by Amazon takes product reviews and ratings into account. We choose the Amazon customer review dataset as our research base. Regarding this project, we focused on the data of reviews from the Patio, Lawn, and Garden category, and we used the methodologies of Topic Modeling and Sentiment Analysis. Based on the result of our analysis process, we generated three main conclusions: 1. There is a strong correlation among different products within the Patio, Lawn, and Garden category. 2. The Amazon Patio, Lawn, and Garden categories have significant seasonality. 3. Using both supervised machine learning and unsupervised machine learning algorithms have higher test accuracy and precision compared to built-in Lexicon methods.

Data description

Our team used the data published by UCSD Computer Science Professor Julian McAuley, who has built this lab for students who are interested in computer science, in 2018. This lab has a lot of students who are currently working for large tech companies, such as Google and Microsoft, and they have more than 10 sponsors who support them. This dataset contains product reviews and metadata from Amazon, from 1996 to 2014, and includes reviews such as ratings and helpfulness votes, links, and so on.

Our team chose to use the data of Patio, Lawn, and Garden from this database. In this particular data set, we have 13000+ rows, including the "ProductID", "Reviewer ID", "Reviewer Name", "Review Text", "Review Time", "Product Rating", "Review Helpful Rating", dating from June 21, 2021, to June 24, 2014. We separated the data into the following three categories.1: Plants: such as fertilizer, seeds, lawnmowers. 2: Animals & Accessories: such as mousetraps, bird feeders, parts for tools repair. 3: Tools and furniture, such as barbecue grill, storage box, garden umbrella, and other tools. Since these data are all related to the garden and lawn area, our group would like to use them and to make the results clear and easy to understand. So, based on this separation of data, we can further eliminate sample selection biases as much as possible.

Project objectives

The overall objective of our project is to excavate the public's demands on the products categorized as "Patio, Lawn and Garden" via multiple analyses on customer reviews posted on Amazon. By extracting keywords on customer reviews and ratings, we could find out the characteristics which customers value most during the purchase and follow-up use, and the unmet needs on existing products from customers' complaints. Our second goal is to provide an efficient sentiment model for the company to predict customers' purchasing experience, or the sentiment, in order to monitor the performance of the products.

Methodology:

The main methodologies used in this project are Topic Modeling and Sentiment Analysis.

Topic Modeling- Data-Preprocessing

The first methodology we applied to the data is the LatentDirichletAllocation function, which stands for Topic Modeling Analysis. There are two individual parts of our topic modeling process. However, before we move on to these two parts, we need to divide our data by subcategory. According to the category management***, we generated three main categories:

1. Plants: Including Fertilizers, seeds, flower pots, sprinklers, weeders, etc.

2. Animals & Accessories: Including Mousetraps, bird feeders, various spare parts, pet food, etc.

3. Tools & Furnitures: Barbecue racks, storage boxes, patio umbrellas, various tools, etc.

In the following two parts of our analysis, we modified our data based on these three subsets

Topic Modeling - Model building

The model for this part is quite straightforward. After we imported the required packages including LDA, we defined the bag-of-words vectorizer and vectorized the normalized data. Then we set the following parameters:

n_components=3, max_iter=100, doc_topic_prior = 0.5, topic_word_prior = 0.9

Finally, we applied the display_topics function to generate the output. Now we can move on to the data analysis process.

Topic Modeling- Results

Part 1: Good & Bad Comments

In this part, we defined two types of reviews, which are Good Comments and Bad Comments. The Good Comments consists of all the reviews with overall rating scores of 5 and 4, while the Bad Comments consists of all the reviews with overall rating scores of 3, 2, and 1. For each section, we implemented the LatentDirichletAllocation function and created three topics with top 15 words respectively.

The result displays on the table: Topic Modeling Part 1 [1]. As we can see, there are some words that appear frequently, like “use”, “good”, “like” and “well”. Those words are just some common choice of words and not our main focus. So, we would ignore those meaningless words in our analysis. Here are some features for each category:

Plants: In this category, there are many words that are related to animals, like “feeder”, “squirrel”, and “bird” even though it's a plant-related category.

Animals & Accessories: In this category, there are more animal words and also some words related to plants.

Tools & Furniture: In this category, there are still many words related to the previous two categories.

Here are our recommendations and ideas towards the results:

1. For the manufacturer, improve customer experience by simplifying the installation process and using processes. For example, some plant feeders could be installed on garden hoses to spray the chemicals easily.

2. Weber grill appeared many times in good reviews, this could be regarded as a leading brand of the industry.

3. Easy to mount is also the key demand among the customers.

Part 2 Brief: Keywords by Quarter

In this part, we divided the data into four parts by quarter, which are Q1, Q2, Q3 and Q4. We'd like to know if the season would affect the focus of customers.

The result displays on the table: Topic Modeling Part 2 [2]

According to the result, we could see that the focus of the reviews differed by different quarters. For example, in the Quarter 4 of Plant category, there are some keywords of “smoke”, “charcoal” and “axe”, those are the words related to making a fire, which makes sense because Quarter 4 is winter and there is demand for warming. In the Quarter 2 and 3, the frequency of animal words increased, which also makes sense because during the summer the animals' activities increased a lot.

Sentiment Analysis- Data Pre-processing

Our data has an imbalance problem due to the discrepancy in the tendency of leaving the comments. Results show that people tend to leave more 5-star ratings than 1-star ratings, with respect numbers of 6276 and 448.[3] Hence, it is necessary to modify the sample data. Here, we simply implement the undersampling method, randomly selecting the positive ratings to make each rating balance. Afterward, we transform our ratings into 3 classes: negative, neutral, and positive. The final sample data includes 1000 in each class.[4] For the training and testing dataset, we shuffle the data and separate it to 90% for training and 10% for testing due to the limited number of data.

Sentiment Analysis- Model Building

Sentiment analysis is a prominent method for analyzing customers' experiences and their market demand. The filtered text can serve as the variables for modeling merchandises' rating, or level of purchasing experiences, which is positive, neutral, or negative. In this project, we aim at using three different approaches combined with three indicators for efficiency comparison: Testing Accuracy, precision rate, and recall rate.

1. **VADER Lexicon + multi-class stochastic gradient descent classifier**

In this model, we first use the VADER lexicon to filter out the points of each word, then calculate the total score of the text. Because VADER Lexicon can only identify binary classes, that is whether the text is positive or negative but not neutral. Hence, we use a multi-class stochastic gradient descent classifier to classify multiple categories.

2. Feature extraction + multi-class stochastic gradient descent classifier

VADER Lexicon has a chance to be biased due to the original score of each word. The meaning and intensity may change from time to time. Hence, we will use feature engineering to filter out the words and transform them into the normalized array.[5] Then we will split the array into training data and testing data and fit it into machine learning algorithms- using the multi-class stochastic gradient descent method.

3. Text encoding + Recurrent Neural Network

The recurrent neural network has been proved useful for text analysis and sentiment prediction. Different from the traditional neural network, the recurrent neural network has more inherence characteristics in the learning process. The values learned from the hidden layers become the input of the same hidden layer in the next run.[6] By doing so, the model can gradually train and predict the entire text sequence instead of considering a single word only at once.

For the input data, we use the word encoding method [7] instead of bags of words. The reason is that if we use the normalized bag of words, the dimension of input data will be too large for neural networks. We further select 20 words as the longest sequence for encoding. After building up the layers of the neural network, we use grid search to optimize the best solution for the hyperparameters: Batch_sizes and epochs. Also, we add the dropout method to prevent the overfitting problem.

The neural networks contain one embedding layer, two simple hidden layers with a dropout function, and one output layer. The output layer is three-dimensional constructed by the activation function softmax, due to our dependent variables dummies (overall rating: positive, neutral, negative). The result will select the highest possibility in each dummy as that category. After optimizing by grid search, The Bath_sizes is 400, and the epochs are 45 [8]

Sentiment Analysis- Results

For our first model, we have a simple Vader Lexicon + multi-class stochastic gradient descent classifier for the prediction. The table shows that the model is not good at classifying the neutral categories, the testing data has no neutral prediction. The testing accuracy is only 0.42. The precision rate is 0.33, while the recall rate is 0.8.[9] It is in no doubt a bad method for solving multi-class text mining problems.

For our second model, we have a Feature extraction + multi-class stochastic gradient descent classifier. The feature-engineering model performs better than the Vader Lexicon model. The testing accuracy rises to 0.55, with a precision rate of 0.60 and a recall rate of 0.64.[10]

For our last model, we have Text encoding + Recurrent Neural Network[11] for sentiment prediction. The neural network model performs slightly better than the second model with a testing accuracy of 0.69, with a precision rate of 0.72 and a recall rate of 0.61.[8]

Comparing the three models, we can easily find that the Vader lexicon model has the worst outcome. It even doesn't predict a single output in the class- neutral. The low precision rate also implies that the model is not good at predicting positive class correctly. The second model performs better than the first model, resulting in higher accuracy and higher recall rate. It is probably because the machine learning approach is able to learn the pattern of the words. Finally, the recurrent neural network model has both the highest accuracy and precision rate. stating that the neural network model has stronger learning adaptability and capability to deal with the enormous unstructured datasets.

Conclusion:

Based on the result of our analysis, we generated the following conclusions:

1. There is a significant correlation among the different subcategories within the Patio, Lawn, and Garden category. The consumers of one specific product have a strong potential to be the consumers of other products under this category. So, expanding the business variety would be a good strategy to improve the revenue for the companies of this industry.

2. The Amazon Patio, Lawn, and Garden category has high seasonality. For each quarter, the corresponding keywords reflected the main demand. The operators within this category could utilize this kind of seasonality to optimize their schedule and arrange the supply chain to meet the demand at different time points.

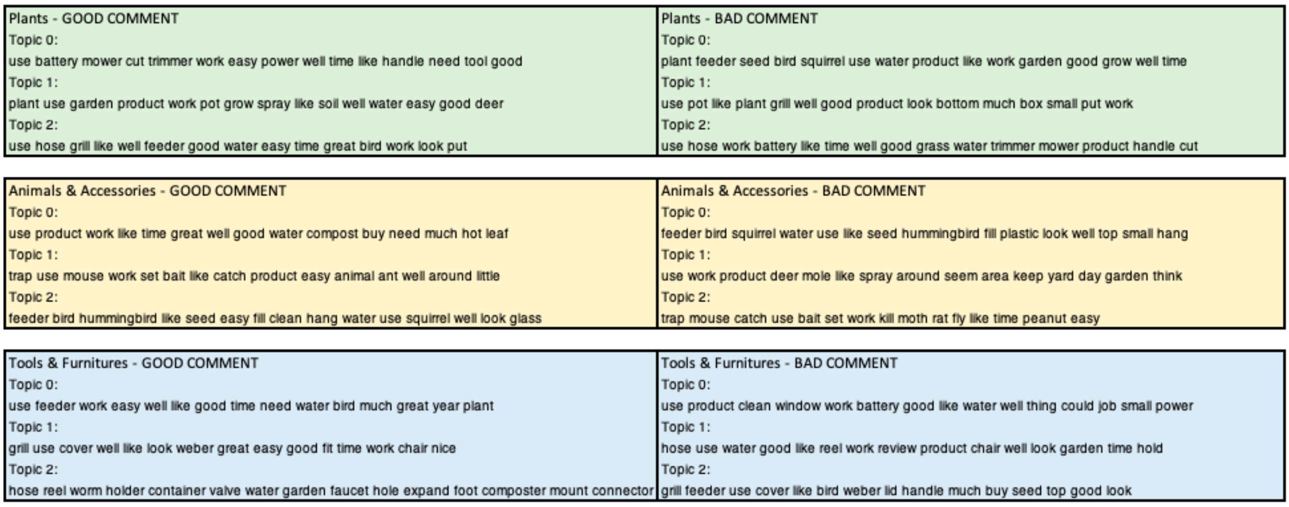
3. For the sentiment analysis, the result shows that using both supervised machine learning and unsupervised machine learning algorithms has higher test accuracy and precision compared to built-in Lexicon methods. Our models also show that the company can use recurrent neural networks as the base model to predict and monitor users' experiences.

By combining our topic model and sentiment analysis model, the company can figure out what customers demand at a different stage and plan out the marketing strategies in advance. For those comments that are left on social media such as Twitter, Facebook, Instagram, our model also provides the company to simulate the customers' satisfaction by training the reviews.

The insights gained from this project can be applied by various organizations in the industry, including manufacturers, distributors, and retailers, etc. By improving the service on product making, selling, transportation, after-sale service, and so on, companies could strengthen their customer management, which means attracting new customers and consolidating old customers, in order to increase the revenue and profit.

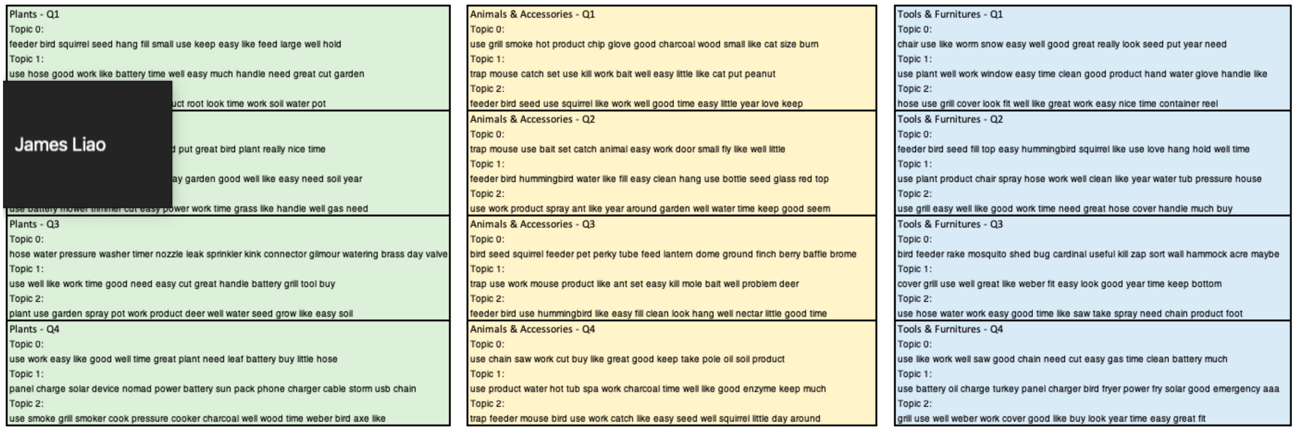
Appendices: Tables and Plots

[1] Topic Modeling Part 1 - Result



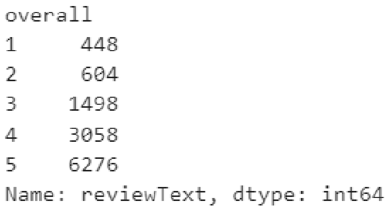
Plants - GOOD COMMENT Topic 0: use battery mower cut trimmer work easy power well time like handle need tool good Topic 1: plant use garden product work pot grow spray like soil well water easy good deer Topic 2: use hose grill like well feeder good water easy time great bird work look put	Plants - BAD COMMENT Topic 0: plant feeder seed bird squirrel use water product like work garden good grow well time Topic 1: use pot like plant grill well good product look bottom much box small put work Topic 2: use hose work battery like time well good grass water trimmer mower product handle cut
Animals & Accessories - GOOD COMMENT Topic 0: use product work like time great well good water compost buy need much hot leaf Topic 1: trap use mouse work set bait like catch product easy animal ant well around little Topic 2: feeder bird hummingbird like seed easy fill clean hang water use squirrel well look glass	Animals & Accessories - BAD COMMENT Topic 0: feeder bird squirrel water use like seed hummingbird fill plastic look well top small hang Topic 1: use work product deer mole like spray around seem area keep yard day garden think Topic 2: trap mouse catch use bait set work kill moth rat fly like time peanut easy
Tools & Furnitures - GOOD COMMENT Topic 0: use feeder work easy well like good time need water bird much great year plant Topic 1: grill use cover well like look weber great easy good fit time work chair nice Topic 2: hose reel worm holder container valve water garden faucet hole expand foot composter mount connector	Tools & Furnitures - BAD COMMENT Topic 0: use product clean window work battery good like water well thing could job small power Topic 1: hose use water good like reel work review product chair well look garden time hold Topic 2: grill feeder use cover like bird weber lid handle much buy seed top good look

[2] Topic Modeling Part 2 - Result



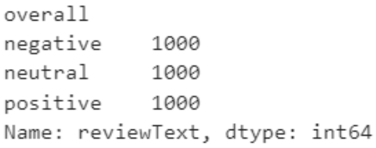
Plants - Q1 Topic 0: feeder bird squirrel seed hang fill small use keep easy like feed large well hold Topic 1: use hose good work like battery time well easy much handle need great cut garden [illegible] [illegible] uct root look time work soil water pot	Animals & Accessories - Q1 Topic 0: use grill smoke hot product chip glove good charcoal wood small like cat size burn Topic 1: trap mouse catch set use kill work bait well easy little like cat put peanut Topic 2: feeder bird seed use squirrel like work well good time easy little year love keep	Tools & Furnitures - Q1 Topic 0: chair use like worm snow easy well good great really look seed put year need Topic 1: use plant well work window easy time clean good product hand water glove handle like Topic 2: hose use grill cover look fit well like great work easy nice time container reel
[illegible] [illegible] [illegible] d put great bird plant really nice time [illegible] [illegible] ay garden good well like easy need soil year [illegible] use battery mower trimmer cut easy power work time grass like handle well gas need	Animals & Accessories - Q2 Topic 0: trap mouse use bait set catch animal easy work door small fly like well little Topic 1: feeder bird hummingbird water like fill easy clean hang use bottle seed glass red top Topic 2: use work product spray ant like year around garden well water time keep good seem	Tools & Furnitures - Q2 Topic 0: feeder bird seed fill top easy hummingbird squirrel like use love hang hold well time Topic 1: use plant product chair spray hose work well clean like year water tub pressure house Topic 2: use grill easy well like good work time need great hose cover handle much buy
Plants - Q3 Topic 0: hose water pressure washer timer nozzle leak sprinkler kink connector gilmour watering brass day valve Topic 1: use well like work time good need easy cut great handle battery grill tool buy Topic 2: plant use garden spray pot work product deer well water seed grow like easy soil	Animals & Accessories - Q3 Topic 0: bird seed squirrel feeder pet perky tube feed lantern dome ground finch berry baffle brome Topic 1: trap use work mouse product like ant set easy kill mole bait well problem deer Topic 2: feeder bird use hummingbird like easy fill clean look hang well nectar little good time	Tools & Furnitures - Q3 Topic 0: bird feeder rake mosquito shed bug cardinal useful kill zap sort wall hammock acre maybe Topic 1: cover grill use well great like weber fit easy look good year time keep bottom Topic 2: use hose water work easy good time like saw take spray need chain product foot
Plants - Q4 Topic 0: use work easy like good well time great plant need leaf battery buy little hose Topic 1: panel charge solar device nomad power battery sun pack phone charger cable storm usb chain Topic 2: use smoke grill smoker cook pressure cooker charcoal well wood time weber bird axe like	Animals & Accessories - Q4 Topic 0: use chain saw work cut buy like great good keep take pole oil soil product Topic 1: use product water hot tub spa work charcoal time well like good enzyme keep much Topic 2: trap feeder mouse bird use work catch like easy seed well squirrel little day around	Tools & Furnitures - Q4 Topic 0: use like work well saw good chain need cut easy gas time clean battery much Topic 1: use battery oil charge turkey panel charger bird fryer power fry solar good emergency aaa Topic 2: grill use well weber work cover good like buy look year time easy great fit

[3] Rating Distribution



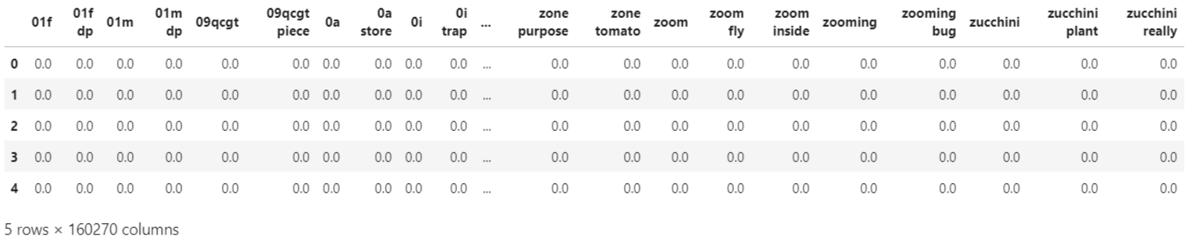
```
overall
1      448
2      604
3     1498
4     3058
5     6276
Name: reviewText, dtype: int64
```

[4] Undersampling result



```
overall
negative    1000
neutral     1000
positive    1000
Name: reviewText, dtype: int64
```

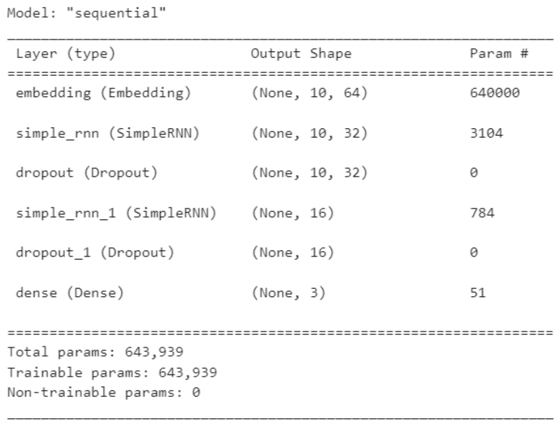
[5] Normalized Text Table (top5)



	01f	01f dp	01m	01m dp	09qcgt	09qcgt piece	0a	0a store	0i	0i trap	...	zone purpose	zone tomato	zoom	zoom fly	zoom inside	zooming	zooming bug	zucchini	zucchini plant	zucchini really
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

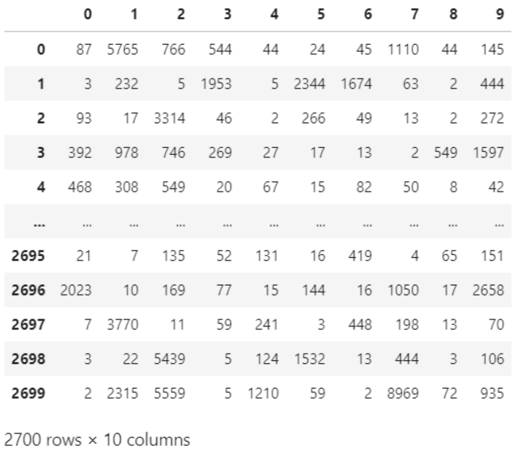
5 rows × 160270 columns

[6] Recurrent Neural Network Model



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 embedding (Embedding)       (None, 10, 64)            640000    
                                                                 
 simple_rnn (SimpleRNN)      (None, 10, 32)            3104      
                                                                 
 dropout (Dropout)           (None, 10, 32)            0         
                                                                 
 simple_rnn_1 (SimpleRNN)    (None, 16)                784       
                                                                 
 dropout_1 (Dropout)         (None, 16)                0         
                                                                 
 dense (Dense)               (None, 3)                 51        
                                                                 
=================================================================
Total params: 643,939
Trainable params: 643,939
Non-trainable params: 0
_________________________________________________________________
```

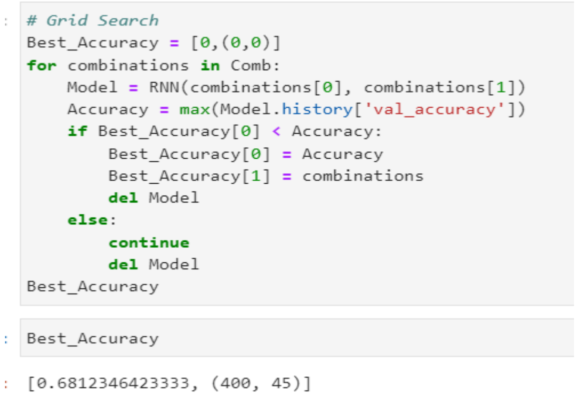
[7] Text Encoding Table (Training Data)



	0	1	2	3	4	5	6	7	8	9
0	87	5765	766	544	44	24	45	1110	44	145
1	3	232	5	1953	5	2344	1674	63	2	444
2	93	17	3314	46	2	266	49	13	2	272
3	392	978	746	269	27	17	13	2	549	1597
4	468	308	549	20	67	15	82	50	8	42
...	...	...	...	...	...	...	...	...	...	...
2695	21	7	135	52	131	16	419	4	65	151
2696	2023	10	169	77	15	144	16	1050	17	2658
2697	7	3770	11	59	241	3	448	198	13	70
2698	3	22	5439	5	124	1532	13	444	3	106
2699	2	2315	5559	5	1210	59	2	8969	72	935

2700 rows × 10 columns

[8] Grid Search Code- Batches sizes and Epochs Optimization and Accuracy-Model3

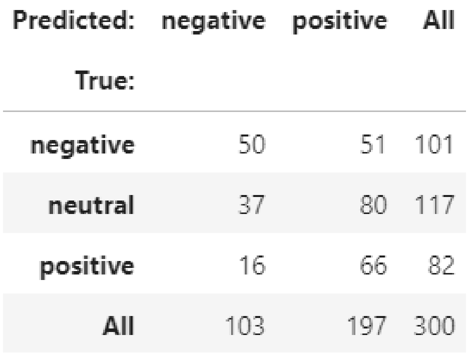


```
# Grid Search
Best_Accuracy = [0,(0,0)]
for combinations in Comb:
    Model = RNN(combinations[0], combinations[1])
    Accuracy = max(Model.history['val_accuracy'])
    if Best_Accuracy[0] < Accuracy:
        Best_Accuracy[0] = Accuracy
        Best_Accuracy[1] = combinations
        del Model
    else:
        continue
        del Model
Best_Accuracy
```

```
Best_Accuracy
```

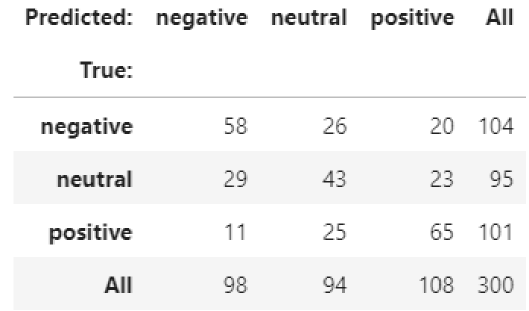
```
[0.6812346423333, (400, 45)]
```

[9] Testing Accuracy- Model1



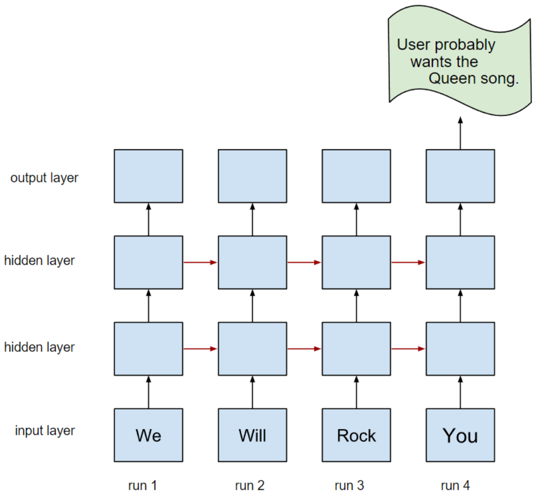
Predicted:	negative	positive	All
True:			
negative	50	51	101
neutral	37	80	117
positive	16	66	82
All	103	197	300

[10] Testing Accuracy- Model2



Predicted:	negative	neutral	positive	All
True:				
negative	58	26	20	104
neutral	29	43	23	95
positive	11	25	65	101
All	98	94	108	300

[11] Recurrent Network Concept



(Machine Learning Glossary | Google Developers)

References:

https://www.jaggaer.com/blog/category-management-strategy-golden-rules/

https://www.mytotalretail.com/article/applying-principles-of-category-management-to-e-commerce/

https://blog.tophatter.com/sellerblog/seasonality-how-to-increase-ecommerce-sales

https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf

https://www.statista.com/topics/846/amazon/

https://www.inc.com/craig-bloem/84-percent-of-people-trust-online-reviews-as-much-.html

https://www.journals.uchicago.edu/doi/abs/10.1086/665825

https://www.supplychaindive.com/news/success-category-management-procurement-ISM/552409/

https://www.aaai.org/Papers/Workshops/2007/WS-07-08/WS07-08-011.pdf

https://www.integromat.com/en/blog/6-effective-tips-to-manage-your-seasonal-inventory-over-the-holidays

https://medium.com/analytics-vidhya/sentiment-analysis-on-amazon-reviews-using-tf-idf-approach-c5ab4c36e7a1

https://medium.com/@am.benatmane/keras-hyperparameter-tuning-using-sklearn-pipelines-grid-search-with-cross-validation-ccfc74b0ce9f

https://towardsdatascience.com/recurrent-neural-networks-and-natural-language-processing-73af640c2aa1