Recommendation System and Sentimental Analysis for Grocery Application: gApp

Shraddha Chobhe[1], Jyoti Malhotra[2]
[1,2]*Department of Computer Science and Engineering, MIT-ADT University Pune*

Abstract—The **Internet is the most powerful thing in the world. Is it possible for any of us to envisage buying groceries online and preferring online shopping? All of this is feasible because of the internet. People prefer online grocery shopping to physical shopping, according to a new survey. This affected the business, and the use of new machine learning techniques has expanded. The sentiment analysis of reviews, as well as the product recommendation tool, has changed the way people shop online. Without reading the entire review, one may get a good idea of the product. So, in this research, we looked at sentiment analysis in the context of an online grocery review and recommendation system.**

***Index Terms*—Machine Learning, Sentiment Analysis, Recommendation System, Online Grocery**

I. INTRODUCTION

Online grocery apps and networks have most likely swept the world by storm, shrinking the earth's dimensions. People shop on a variety of grocery shopping services, including Groffers, Star Bazaar, Instacart, BigBasket, Amazon Pantry, and others. These rising rates of online purchasing reflect the interest and mood of those who engage in online shopping. This paves the door for a better knowledge of people who shop. Sentimental analysis, also known as opinion mining, is important for studying and comprehending the communication that occurs during transactions.

Sentiment analysis decrypt and predicts the feelings, emotions and opinions of customer who look at the product with text based information. It is difficult to manage a huge amount of data that includes the customer's opinion, which is difficult to foresee so that the client and the internet platform can profit.

Recommendation systems are used by the majority of e-commerce companies. These algorithms are used in these recommendation systems. The majority of e-commerce businesses use their algorithms for making product recommendations to users. Collaborative filtering and content-based filtering are two principles that are commonly utilized to construct recommendation systems. Both of these notions employ their methods for making recommendations to users.

Sentiment analysis and deep learning methods are used to collect complex features of natural language processing.The deep learning method is being used to extract information from the internet platform "in a big volume of data."

There are different techniques of Machine Learning[11], supervised, unsupervised and deep learning such as Linear Regression, Support Vector Machine, Random Forest, Naive Baiyes, Decision Tree, Apriori, K-mean, Convolutional Neural Network, Long-Short Term Memory, and Recurrent Neural Network are applied to predict the customer sentiment and also recommend products.

This research paper includes a section1: Introduction on the sentiment analysis and recommendation system in online grocery. In section 2: machine learning techniques such as sentimental analysis and recommendation system are addressed. Section 3 explains the the proposed solution. In section 4: discussion of the proposed solution methodology. The sections 5 and 6 indicate the result and conclusion.

II. MACHINE LEARNING TECHNIQUES

Various machine learning techniques are addressed in two categories: A. Sentiment Analysis and B. Recommendation System.

A. Sentiment Analysis:

A machine learning technology called sentiment analysis [18] examines texts for polarity, ranging from positive to negative. Machines automatically learn how to recognize sentiment without human input by training machine learning techniques with samples of emotions in text.

Computers can learn new talents using machine learning without having those skills explicitly built into them. Sentiment analysis models can be trained to recognize things like context, sarcasm, and improper word usage in addition to straightforward definitions.

Dual stage working of sentiment analysis is addressed in the figure 1.

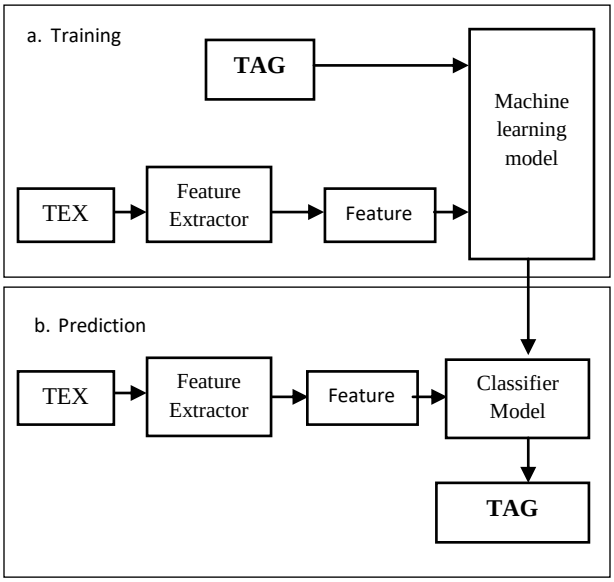

Fig 1. Sentiment analysis in Machine Learning

Following are the famous algorithms used for sentiment analysis:

- Naïve Bayes [6]: A probabilistic algorithm named as Naïve Bayes by classifying depicts given words or phrase as positive or negative i.e. sentiment of the word or phrase.
- Linear regression [12]: Linear regression showcases the relation of input X as words and phrases and the output Y as polarity.
- SVM [6]: SVM is more advanced as compare to linear regression.It simply uses input/output prediction after classifying the text i.e. reviews.

B. Recommendation System:

Recommendation engines: are a type of machine learning that is used to rank or rate products and users. A recommender system, in a broad sense, is a system that predicts how a user would rate a certain item. After then, the predictions will be ranked and returned to the user.

Recommender systems are frequently referred to as "black boxes," and the models developed by these major corporations are difficult to comprehend. The generated results are frequently recommendations for the user for things that they need/want but are unaware of until it is recommended to them.

Types of recommendation systems:

- A recommendation system based on similar content is known as a *content-based recommendation system*. If a user is viewing a movie, the system will look for additional films with comparable content or genres. When checking for comparable content, a variety of basic attributes are employed to compute similarity.
- *Collaborative-based recommendation system:* It is regarded as one of the most intelligent recommender systems that are based on the similarity between different users and also things that are commonly utilized as e-commerce websites and also online movie websites.

III. PROPOSED SOLUTION: gApp

Figure 2 depicts the proposed system's gApp a grocery Application work flow:

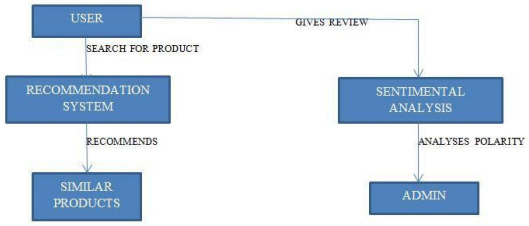

Fig. 2: gApp work flow

A recommendation system and sentimental analysis modules are designed for the application.

The recommendation module employs a content-based recommendation system, whereas the *sentimental analysis module* employs an LSTM model.

The dataset was used to train a cosine similarity model for a recommendation system in this work. The sentiment analysis model was then trained using a different dataset and an LSTM model. The sentimental analysis model will assess the polarity of the text and output a positive, negative, or neutral tag to the admin dashboard using the product name or category name as input. Following the completion of product reviews, the sentimental analysis model will

assess the text's polarity and return a positive, negative, or neutral tag to the admin dashboard.

- *Dataset*:

Two datasets are used in this research. One is the product dataset used for the recommendation system and the other dataset is used for sentimental analysis. Where in, the LSTM model is generated. Product dataset contains the details of product id, product name, sub-categories of the product, size, final price, image URL, comb, and available. The dataset used for sentimental analysis consists of reviews.

- *Data preprocessing*:

For applying the algorithm for the recommendation model, we have combined two columns and generated a new column that is a comb column. Then using the count vectorizer, the newly formed column is converted to a count matrix that is to numbers. Here the pre-preprocessing of recommendations comes to an end.

Natural Language Tool Kit is required for Sentiment Analysis pre-processing (NLTK)[1]. Popular Python software for working with human language data is called NLTK. A typical Python library for linguistics and Natural Language Programming is provided here.This library helps to import stop words Stop words are the popular words used in any language. 'A, an, the, if, or' are examples of such terms. They're employed in Text Mining and Natural Language Processing (NLP) to filter out terms that aren't very useful. The column of Comments is then tokenized with the Tfidf Vectorizer.

IV. METHODOLOGY

Recommendation system [1]

We'll examine the operation of the cosine similarity algorithm to gain a better understanding of the methods employed in recommendation systems.

- Cosine similarity [3]:

Cosine similarity uses vectors as data objects in data sets, and it determines how similar they are by defining them in a product space. Higher similarity is associated with closer proximity, while lower similarity is associated with greater proximity.

The direction is determined by the angle between two vectors and is measured in 'θ'. This angle θ can be calculated by using equitation.

$$\text{Cos}(x, y) = x \,.\, y \,/\, ||x|| * ||y||$$

Sentimental Analysis:

- LSTM model[116]:

A more sophisticated RNN that allows for knowledge retention is the Long Short Term Memory Network (LSTMN). It can fix the RNN's vanishing gradient issue. Recurrent neural networks, or RNNs, are employed for permanent memory. Applications like stock forecasting, speech recognition, and natural language processing are only a few examples of regularly used ones.

- KNN[12]:

First, let's define the term "neighbors." The neighborhood of data samples is determined by their closeness/proximity. Depending on the problem at hand, there are numerous methods for calculating the proximity/distance between data points. Most familiar and popular is the straight-line distance (Euclidean Distance).

- Random Forest[9]:

Ensemble classification methods are learning algorithms that create a group of classifiers rather than a single classifier and then classify fresh data points by voting on their predictions. Bagging, Boosting, and Random Forest is the most often used ensemble classifiers (RF). Random forest is an ensemble learning-based supervised machine learning technique. The random forest algorithm combines several methods of the same sort.Both regression and classification problems can benefit from the random forest algorithm.

V. RESULTS

The Cosine Similarity algorithm will give accurate result to suggest comparable products to those entered by the user.The similarity score is matched with the input product similarity score from the dataset after vectorization, and the matches are shown on the screen in ascending order. That is, the highest cosine similarity value should be used first, followed by the lowest.We'll look at how recommendation works in this section.Consider the following image (refer to figure 3):

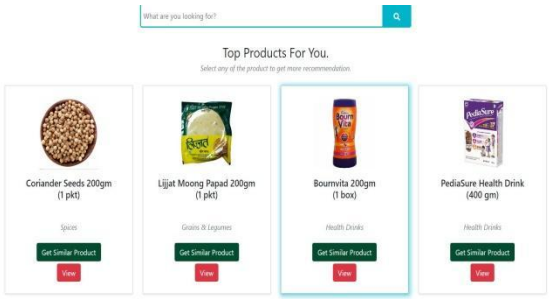

Figure 3: Before Similarity model's recommendation

We can observe that merchandise from many categories is available. When the user clicks on the "*get similar product"* button in the first product of spices category, the model displays all of the products in that category.

Let's have a look at the cosine similarity model's recommended output as mentioned in the figure 4.

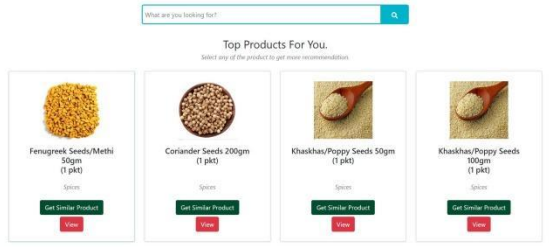

Figure 4: After Similarity model's recommendation

The full spice category's recommended products can be seen in the recommendation.

The gApp model used three algorithms to train the sentiment model. The following is the training model (refer to figure 5(a) for KNN, 5(b) for Random Forest and figure 5(c) for LSTM):

```
**********************************
     KNN Classification Report
**********************************
              precision    recall  f1-score   support

           0       0.52      0.36      0.43     40094
           1       0.51      0.40      0.45     39906

   micro avg       0.51      0.38      0.44     80000
   macro avg       0.51      0.38      0.44     80000
weighted avg       0.51      0.38      0.44     80000
 samples avg       0.38      0.38      0.38     80000
```

Fig 5(a): Accuracy report of KNN algorithm

```
**********************************
Random Forest Classification Report
**********************************
              precision    recall  f1-score   support

           0       0.56      0.55      0.56     40094
           1       0.56      0.51      0.53     39906

   micro avg       0.56      0.53      0.55     80000
   macro avg       0.56      0.53      0.55     80000
weighted avg       0.56      0.53      0.55     80000
 samples avg       0.53      0.53      0.53     80000
```

Fig. 5(b): Accuracy Report of Random Forest Algorithm

```
Epoch 00010: val_loss did not improve from 0.22640
125/125 [==============================] - 244s 2s/step -
<keras.callbacks.History at 0x7f17c52fa4d0>
loss: 0.1530 - acc: 0.9435 - val_loss: 0.2313 - val_acc: 0.9125

y_pred = model.evaluate(X_test,y_test,batch_size=2048)

40/40 [==============================] - 24s 592ms/step -
loss: 1.0392 - acc: 0.6251
```

Fig. 5(c): Accuracy Report of LSTM algorithm

The following (refer to figure 6) which depicts ROC curve that scores nearly to the value 1.

Table 1 summarizes the comparison between the models namely KNN, Random Forest and LSTM in terms of the performance metrics (Accuracy, Precision, and Recall) and is demonstrated in the figure 7.

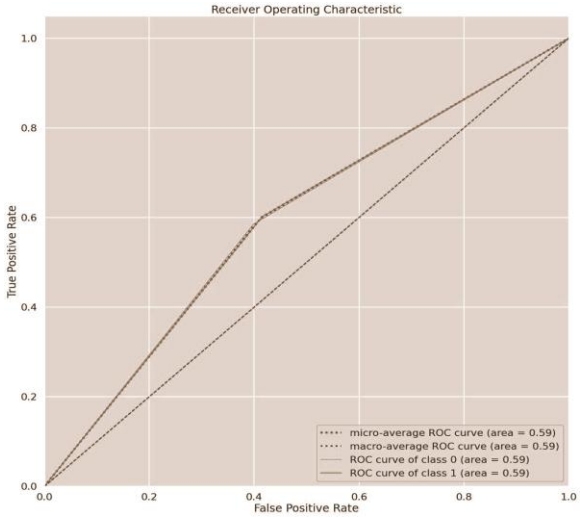

Figure 6: ROC

Table 1: Comparison

Algorithm	Accuracy	Precision	Recall
KNN	0.51	0.51	0.40
Random Forest	0.56	0.56	0.51
LSTM	0.91	0.82	0.81

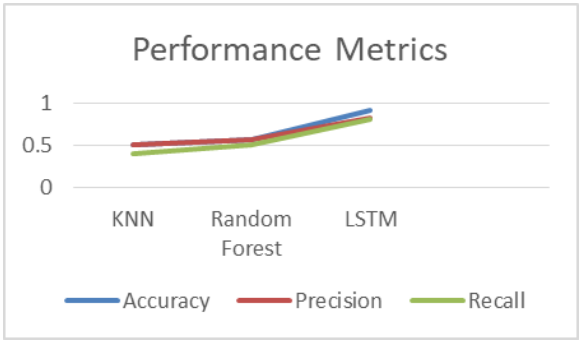

Figure 7: Performance metrics

According to the findings, it is observed that the suggested LSTM model outperforms KNN and Random Forest.

Following that, the study shows the outcomes of the LSTM algorithm when employed for sentimental analysis in figure 8.

It denotes a mix of positive, negative, and neutral polarity in various assessments for example say the product reviews could be – (Bad smell, Too strong smell, I liked it very much, Nice fragrance, Long lasts, etc). The sentiment analysis on reviews is not accessible to users in the proposed solution and can only be seen by the admin to determine whether goods have a business and what needs to be changed to increase profit and user outreach.

Figure 8: Reviews for Sentiment Analysis

VI. CONCLUSION

There are two primary portions to this work. One is focused on a system for recommending products, and the other is focused on sentiment analysis. The paper thoroughly analyses both systems and comes to some significant findings.The Cosine Similarity algorithm was employed in the product recommendation system to suggest items based on various categories that are pertinent to the product the user entered.Even after multiple tests, Cosine Similarity has shown reasonable findings and has been quite accurate in recommending products by product name and category name.

In this research, sentiment analysis is also crucial. Its main goal is to categorize reviews as good, negative, or neutral. For the same, three algorithms were utilized. KNN, Random Forest, and LSTM are three of these algorithms. The primary goal of employing three algorithms is to determine which the best method for classifying reviews is. Because reviews vary greatly, it is critical to select the correct classification algorithm. Finally, the results reveal that the LSTM Algorithm outperforms the KNN and Random Forest algorithms in terms of accuracy.

ACKNOWLEDGMENT

It is my privilege to express my sincerest regards to my project guide, Dr. Jyoti Malhotra, for her valuable input, encouragement, and support throughout this research. I remain immensely obliged for her guidance and supervision which made this research happen.

REFERENCES

[1] N Pavitha, VithikaPungliya, AnkurRaut, Rosita Bhonsle, AtharvaPurohit, Aayushi Patel, R Shashidhar - Movie Recommendation and Sentiment Analysis Using Machine Learning || Global Transitions Proceedings (2022), DOI: https://DOI.org/10.1016/j.gltp.2022.03.012

[2] Pranavi Satheesan; Prasanna S. Haddela; JesuthasanAlosius - Product Recommendation System for Supermarket || 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), DOI: 10.1109/ICMLA51294.2020.00151

[3] Prof. Suja Panikar, Prayag Mane, Chetan Pakhale, Shubham Fulzele, Akshay Rathi- Online Grocery Recommendation system || IJSRD - International Journal for Scientific Research & Development| Vol. 4, Issue 04, 2016 | ISSN (online): 2321-0613

[4] Gautam Srivastava- A study on reviews of online grocery stores during COVID-19 pandemic using sentiment analysis || International Journal of Logistics Economics and Globalisation, DOI: February 4, 2022pp 205-222

[5] Shanshan Yi, XiofangLui- Machine learning-based customer sentiment analysis for recommending shoppers, shops based on customers' review || springer.com/article/10.1007/s40747-020-00155-2

[6] Vijay Kumar Soni, SmitaSelot. "A Comprehensive Study for the Hindi Language to Implement Supervised Text Classification Techniques" ,2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021
[7] Ebrahimi M, Yazdavar AH, Sheth A (2017) Challenges of sentiment analysis for dynamic events. IEEE IntellSyst 32(5):70–75
[8] Le, N.-B.-V.; Huh, J.-H. Applying Sentiment Product Reviews and Visualization for BI Systems in Vietnamese E-Commerce Website: Focusing on Vietnamese Context. Electronics 2021, 10, 2481
[9] Saberi, Bilal, and SaidahSaad. "Sentiment analysis or opinion mining: a review." *Int. J. Adv. Sci. Eng. Inf. Technol* 7.5 (2017): 1660-1666.
[10] Ahmad, SitiRohaidah, Azuraliza Abu Bakar, and MohdRidzwanYaakub. "A review of feature selection techniques in sentiment analysis." *Intelligent data analysis* 23.1 (2019): 159-189.
[11] TayybahaQuyyam*, Dr. Hamid Ghous, "Sentiment Analysis of Amazon Customer Product Reviews: A Review." *IJSRED 2021*. Volume 4 Issue 1.
[12] Peter Appiene, Stephen Afrifa, Emmanuel AkwaKyei, Peter Nimbe. "Understanding the Uses, Approaches and Applications of Sentiment Analysis", Research Square Platform LLC, 2022
[13] https://www.synthesio.com/blog/sentiment-analysis-reveals-insights-grocery-delivery-physical-stores/
[14] https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining
[15] https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
[16] https://www.analyticsvidhya.com/blog/2021/06/natural-language-processing-sentiment-analysis-using-lstm/
[17] https://towardsdatascience.com/sentiment-analysis-using-lstm-step-by-step-50d074f09948
[18] https://monkeylearn.com/blog/sentiment-analysis-machine-learning/