We declare that we have completed this assignment completely and entirely on our own, without any consultation with others. We have read the UAB Academic Honor Code and understand that any breach of the Honor Code may result in severe penalties.

We also declare that the following percentage distribution ***faithfully*** represents individual group members' contributions to the completion of the assignment.

Name	**Overall Contribution (%)**	**Major Work items completed by me**	**Signature or initials**	**Date**
Saugat Adhikari	20%	I trained Machine Learning Models and compared the performance. Contributed to Report Writing and making presentation slides.	SA	20th April 2022
Krishna Gahatraj	20%	I explored the dataset and did data visualization. Contributed to Report Writing and making presentation slides.	KG	20th April 2022
Shusma Kafle	20%	I did data visualization and tried model ensembling. Contributed to report writing and making presentation slides.	SK	20th April 2022
Jing Wu	20%	I worked on data preprocessing and oversampling the dataset. Contributed to report writing and making	JW	20th April 2022

		presentation slides.		
Lyuheng Yuan	20%	I trained Deep Learning Model and compared the performance. Contributed to Report Writing and making presentation slides.	LY	20th April 2022

SENTIMENT ANALYSIS OF ONLINE CUSTOMER REVIEWS

Saugat Adhikari, Krishna Gahatraj, Shusma Kafle, Jing Wu, Lyuheng Yuan

Department of Computer Science, The University of Alabama at Birmingham

ABSTRACT

Online customer reviews increasingly play an important role in making business decisions. Sentiment analysis is the most effective way to retrieve insights from reviews quickly and accurately. We downloaded a dataset of 34,660 consumer reviews for Amazon products from Kaggle, and preprocessed the dataset by adding target labels, dropping inconsistent interviews, and discarding trivial values. From the initial visualization of the dataset, we noticed that it was imbalanced, then SMOTE technique was applied in the subsequent classifications. Several different machine learning models and a deep learning model were utilized for training and inference to compare the performance based on different metrics.

Key words: online reviews, sentiment analysis.

1. INTRODUCTION AND BACKGROUND

From the Consumer Reviews of 2018 B2B Sales & Marketing Report, 90% of consumers say that their buying decisions are influenced by online reviews [1]. With more consumers than ever turning to online reviews to get fellow customers' feedback to help decide who to give their business to, it is essential for business to understand their online reviews and what their customers are saying about their products in order to better evaluate, bolster, and maintain their businesses' online presence.

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative, or neutral. Positive online reviews can have a significant impact on a business's revenue and bottom line; Negative and fake online reviews can deter potential customers and employees from engaging with a business or product; and Online reviews are often interpreted as an accurate reflection of the quality of a customer's experience [2].

The goal of our project is to gain insights from the review dataset through sentiment analysis by applying different machine learning models and a deep learning model to train and infer from the dataset and further compare the performance based on different metrics.

2. DATA COLLECTION AND PREPARATION

We downloaded the Amazon review dataset from Kaggle. It is a list of 34,660 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more from Datafiniti's Product Database updated between September 2017 and October 2018, including basic product information, rating, review text, and more for each product. Table 1 shows the features and their corresponding explanations of the dataset.

There are a total of 21 features in the original dataset. We selected only 4 relevant features as shown in Table 1.

Since the dataset only had ratings on the scale of 0-5, we added a new column(sentiment) as our target label based on the ratings. We assigned sentiment 1(positive) for ratings greater than 3, 0(negative) for ratings less than 3 and -1(neutral) for ratings equal to 3. We further preprocessed the dataset by dropping the reviews whose ratings and recommendation were not consistent i.e., if the rating is high but the user does not recommend the product or if the rating is low but the user recommends the product, then we discard such review from our dataset since they add noise.

We further preprocessed the review texts by discarding words with single letter and non-alphanumeric characters since they do not have much contribution on text sentiment analysis.

Table 1. **Selected features of the Amazon review dataset**

Feature	Explanation
reviews.doRecommend	A true/false for whether or not the reviewer recommends the product.
reviews.rating	A 1 to 5 start value for the review.
reviews.text	The full (or available) text of the review.
reviews.title	The review's title.

3. VISUALIZATION

Once we had the dataset ready, we perform data exploration as well visualizations to see the nature of the data set. We can see from the pie chart that the rating of the 5 dominates the entire dataset. Also, the most occurring data set also has most of the dataset with the positive words. Imbalance data set will have a poor performance on the data set that is the minority. In real life we care about the performance of those datasets. We use the smote (synthetic minority oversampling technique) to over come that issue.

Also, we plotted the most frequently occuring words as shown in Figure 2 and see that the words like great, love, like, works, best are dominant which plays a vital role in identifying the sentiment. Also, the words with negative sentiment are rare in our dataset which further supports the fact that the dataset is imbalanced and we need to balance it before training the models.

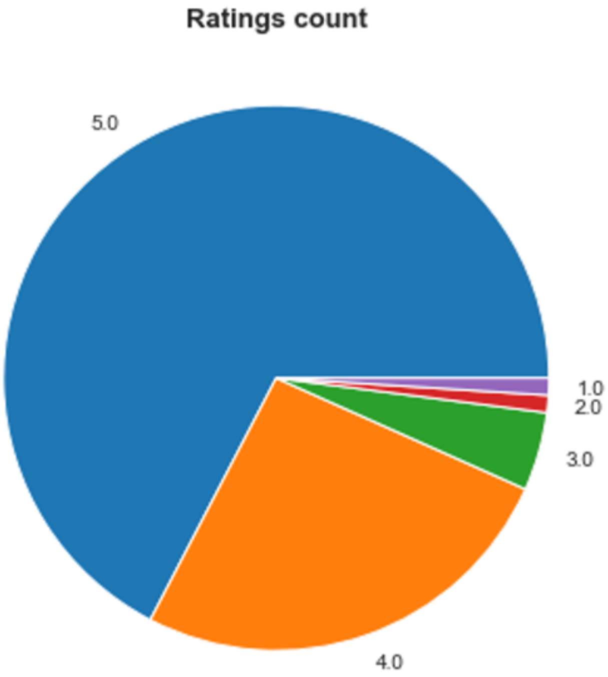


Fig. 1.

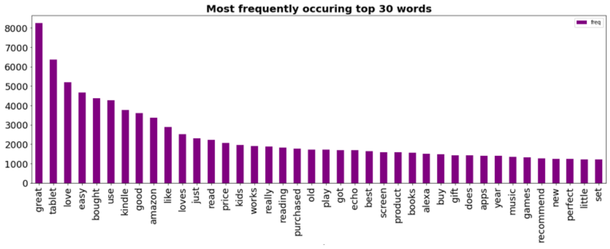


Fig. 2.

4. MODELS TRAINING AND RESULTS

We use different machine learning models and a deep-learning model for training and inference to compare and contrast the performance based on different metrics. We used machine learning models like Naive-Bayes Classifier, Logistic Regression Classifier, Stochastic Gradient Descent Classifier, Decision Tree Classifier, and k-Nearest Neighbor Classifier[3].

First we used the original dataset[4] of Amazon Customer Reviews to see the performance of different machine learning models using k-fold cross validation technique where k = 10. Since, the dataset was highly imbalanced with majority of the reviews dominated by class 'positive', the f1-score for all the models are quite low as shown in Table 1.

Table 2: Metrics on Original Dataset

Model	F1-Score	Accuracy
Naive-Bayes	0.3295	0.9333
Logistic Regression	0.4815	0.9372
Stochastic Gradient Descent	0.3429	0.9338
Decision Tree	0.4303	0.8983
k-Nearest Neighbor	0.3284	0.9327

Even though the accuracy is high, we cannot rely on this because of class imbalance problem. To overcome this problem we used SMOTE[5] oversampling to oversample the minority class and make the training dataset more balanced. The test results on f1-score on oversampled data is shown in Table 2.

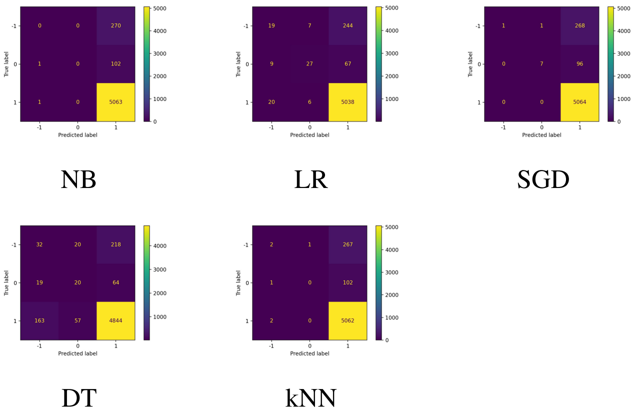
Table 3: Metrics on Oversampled Dataset

Model	F1-Score	Accuracy
Naive-Bayes	0.83	0.78
Logistic Regression	0.86	0.81
Stochastic Gradient Descent	0.85	0.80
Decision Tree	0.87	0.85
k-Nearest Neighbor	0.21	0.14

After trying individual models, we also experimented on model ensemble using three different techniques namely Voting Classifier, Bagging, and Boosting(XGBoost Classifier)[6]. Here also we checked the metrics on both original dataset and oversampled dataset. The results are shown on Table 3 and Table 4 respectively.

Table 4: Metrics for Ensemble models on original Dataset

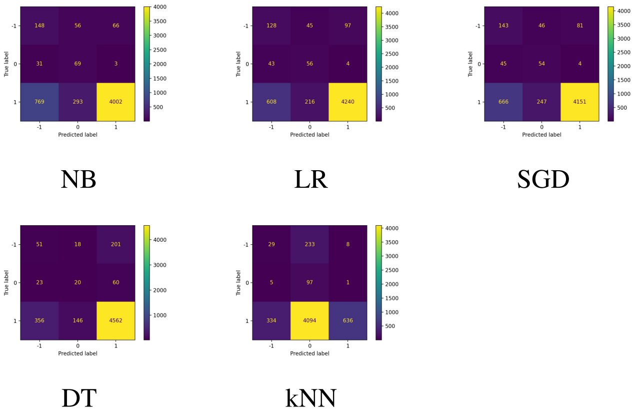
Model	F1-Score	Accuracy
Voting	0.3328	0.9335
Bagging Decision Tree	0.4071	0.9254
XG Boost	0.4238	0.9346



NB LR SGD

DT kNN

Fig. 3. Confusion Matrix on Original Dataset



NB LR SGD

DT kNN

Fig. 4. Confusion Matrix on Oversampled Dataset

Table 5: Metrics for Ensemble models on Oversampled Dataset

Model	F1-Score	Accuracy
Voting	0.85	0.80
Bagging Decision Tree	0.88	0.88
XG Boost	0.91	0.91

Also, in Figure 1 we show the confusion matrix for different models on original dataset as well as on oversampled dataset.

From the confusion matrix in Fig 1, we can see that there are many zeroes in true label vs predicted label for almost all the classifiers which further proves that we cannot rely on the accuracy and need to compare f1-score to get better idea about model performance.

In order to achieve more accurate classification performance on the review dataset, we consider using modern machine learning models, such as deep learning, which has been commonly used in various domains. The model we used here consists of 3 dense layers, with a ReLU activation function between every dense layer. The input dimension of the model is 1000, same as the raw feature dimension. The output dimension is 3, which corresponds to 3 categories, i.e., positive, neutral and negative. We apply cross entropy loss for classification task, Adam optimizer with learning rate of 0.01 and train 20 epochs.

The performance turns out to much better than most of the traditional machine learning models as shown in Table 6 with an accuracy of 93%. We also observe the similar improvements for precision, recall and F-score. Surprisingly, the model doesn't seem to be affected by overfitting, which suggests deep learning models can learn some implicit attributes that are extremely hard for classic models.

Table 6: Metrics for Deep Learning Model on Original Dataset

Model	F1-Score	Accuracy
Simple Neural Network	0.48	0.91

Table 7: Metrics for Deep Learning Model on Oversampled Dataset

Model	F1-Score	Accuracy
Simple Neural Network	0.91	0.93

5. DISCUSSIONS AND CONCLUSIONS

Based on the results we achieved we can conclude that the text classification problems work well only if we have the balanced dataset otherwise the prediction will be biased towards one particular class only. Also, F1-score is a good metrics to measure and compare the model performance rather than accuracy. We see that the Decision Tree Classifier performs the best among other traditional machine learning models for the text sentiment analysis problem. Also, simple deep learning models work well for these types of problem.

6. FUTURE WORK

As a future work, we can develop a recommendation system based on text sentiment analysis. We can develop a prediction model which can take any particular product as an input and as an output it predicts the recommendation for that product.

7. ACKNOWLEDGMENTS

We would like to Dr. Zhang for introducing us to different Data Mining techniques in the class and providing us with guidance about the project idea.

8. REFERENCES

[1] K. McCabe, “Consumer reviews - 2018 b2b sales marketing report,” available: https://learn.g2.com/consumer-reviews.

[2] Minc Law, “Importance of online reviews: 50+ key statistics,” available: https://www.minclaw.com/importance-online-reviews-statistics.

[3] M. Kamber J. Han and J. Pei, “Data mining: Concepts and techniques,” .

[4] “Consumer reviews of amazon products,” available: https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products.

[5] L. O. Hall W. P. Kegelmeyer N. V. Chawla, K. W. Bowyer, “Smote: Synthetic minority over-sampling technique,” available: https://arxiv.org/pdf/1106.1813.pdf.

[6] M. Kamber J. Han and J. Pei, “Data mining: Concepts andtechniques. 2012.,” .