e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

SENTIMENT ANALYSIS ON AMAZON PRODUCT REVIEWS

Akanksha Halde*1, Aditi Uttekar*2, Amit Vishwakarma*3

*1,2,3Department Of Information Technology Vasantdada Patil Pratishthan's College Of Engineering And Visual Arts, Eastern Express Highway, Near Everard Nagar,
Sion-Chunabhatti, Mumbai-400022, India.

ABSTRACT

Sentimental Analysis is employed all told online product firms. These reviews were taken into consideration by other users during their seek for products. These reviews were taken into consideration by other users during their look for products. Hence the industry has found the foundation of delivering the right product searched by the user supported the reviews of the users using the concept of sentimental analysis. Sentimental Analysis is that the concept of knowledge analysis where the collections of reviews are taken into consideration, and people reviews are analyzed, processed and recommended to the user. The reviews given are longer and which contains some paragraphs of content. This paper showcases a comparative study between different machine learning models to perform sentiment analysis on the customer reviews of Amazon products within the Electronics category. the first models we'll scrutinize for our analysis are Support Vector Machines, Naive Bayes Classifier, Random Forest Classifier, BERT Model, Stochastic Gradient Descent (SVM Linear) and Linear SVC models.

Keywords: Natural Language Processing (NLP), Sentiment Analysis, Tokenization, Product Reviews.

I. INTRODUCTION

Social media plays important role in almost everybody's day to day life. It allows the people to convey what they think and feel about the products in E-commerce website. An opinion or review aims to work out the mood of the customer, it maybe either positive or negative towards the merchandise. With new products popping into the market daily customers depend highly on customer reviews of products on eCommerce sites to form decisions about their purchase. it's important for brands to seem into the type of reviews products get and the way these affect the way the merchandise performs within the market. These reviews take into consideration various aspects of the merchandise namely the prize, the brand, the features, the competitive products, etc. we wish to extract key performance indicators from the review data and extract useful reviews which will help the purchasers when shopping online.

II. LITERATURE REVIEW

[1] Sentimental Analysis of Product reviews:

Authors: Najma Sultan & Pintu Kumar & Monika Rani Patra & Sourabh Chandra and S.K. Safikul Alam(2019)

This research paper aims at running sentimental analysis on customer reviews and labeling text as either "Positive", "Negative" or "Neutral". The paper discusses a theoretical approach to sentimental analysis and analyses different algorithms for the same with their corresponding accuracies. It also gives a brief history of different other approaches to sentimental analysis techniques. The solution descried in the paper involves constructing an ML model through three major parts: Data Filtration, Training model and testing model. Data filtration involves pre-processing the text to remove unwanted items and using only relevant textual content for the model. In training, all the feature words (verbs, adverbs and adjectives) are extracted and then classified. Training a dataset to classification algorithms like Naïve Bayes classification algorithm, Linear Model algorithm, SVM algorithm and Decision tree is also done to compare accuracies. In the testing phase, the user review is mapped to the saved feature set. Feature extraction is done using the frequency of occurrence.

[2] Sentimental Analysis on Large Scale Amazon Product Reviews:

Authors: Tanjim UI Haque & Nudrat Nawal Saber & Faisal Muhammed Shah,Dept of Computer Science and Engineering,Ahsanullah University of Science & Technology,Dhaka,Banglore(2018)

The paper conducts analysis on Amazon Review dataset which makes it an ideal reference point. They have used Multinominal Naïve Bayesian and SVM as the main classifiers for analyzing reviews. Using active learning

(semi-supervised learning algorithm) helps avoid bottleneck situations of unlabeled data as the model chooses what data it learns from. We find this paper very insightful for the fact that they have given a comparison between how well different approaches(SVM, MNB, Stochastic Gradient Descent, Random Forest etc.) faired for the particular dataset by comparing the accuracy, precision, recall, and F1 score.

[3] Amazon Reviews Sentimental Analysis: A Reinforcement Learning Approach:

Authors: Roshan Pramod Samineedi Joseph (2020)

The researcher wants to classify Amazon Feedback into binary class and certain into multi-school utilizing reinforcement learning and pre-trained BERT model. The researcher utilizes the product-based data collection from amazon.com. The study aims to gauge and predict the sentiment behind the analysis using algorithms like BERT, LSM and to reinforce learning, classifying it as positive, negative and neutral. LSTM could be a specific type of RNN to avoid long-term dependency problem. It processes data passing on information because it propagates forward. The paper trees to come back to a conclusion on why this model is an apt approach for Sentimental Analysis

[4] Machine Learning-Based Sentiment Analysis for Twitter Accounts. Mathematical and Computational Applications:

Authors:Hasan, Moin, Karim, & Shamshirband (2018).

E-commerce platform also provides customers with the chance to post reviews and ratings about products. This contains a significant influence on future customers who want to shop for the identical product. Therefore, it enables companies to look at opinion mining in sentiment analysis of reviews and ratings to observe product sales and market price. At the sentencing stage, sentiment analysis was conducted, and a posh vocabulary was used with predefined positive and negative expressions to assist find the polarity of sentiment.

III. DATASET

A. About the dataset

For this project, we decided to travel with the 2018 updated version of the Amazon review dataset released in 2014. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

B. Review Data

- The total number of reviews is 233.1 million (142.8 million in 2014).
- Current data includes reviews within the range May 1996 - Oct 2018.

IV. PREVIOUS WORK AND CHALLENGES FACED

Previous iterations of our work contain testing multiple methods of data cleaning and preparation so as to get the cleanest possible text data. additionally to the fundamental pre-processing and data cleaning we had done to the text data, lemmatizing, tokenization and removal of unclean data(tags, punctuation, stopwords etc.) were performed. This alone ended up increasing the accuracy by over 18 points.EDA remains the identical for the bulk.

The main changes between previous and current iterations of our work are changes to the hyperparameters of all models tested to enhance accuracy even by a touch, inclusion of cross validation testing for all models and increase in dataset size.

Major problems that we saw were only during the data collection part. Combated by a 20gb dataset with an 11gb Metadata dataset that failed to run on any of our local machines led us to explore alternatives where we are able to do some basic cleaning and fragment the dataset into a more workable size. Furthermore, the dataset was originally in JSON format which didn't help with the memory limitation problem we faced. We solved this problem by running a SageMaker notebook with 256gb ram to parse, merge, clean and fragment the dataset. aside from this there have been no major roadblocks. Models and hyperparameter tuning went without too many hitches.

Fragmenting the data proposes a replacement problem of the sample not representing the population. we elect to use an easy random sample of the population for our analysis.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

In order to get verity overall sentiment towards a product, we'd like to assume that each one customer that buy the product leave a rating/review. Practically this can be not true. the bulk of consumers that buy a product don't review or rate the product. the idea would be that the bulk of reviews tend to air extremes of either side, but it seems that follow up emails and marketing for a product net multiple reviews within the range of average (3 stars) to excellent (5 stars). On the opposite hand, the length of a text review tends to be larger when the person is writing a chunk criticizing the product. this can be clearly seen in our EDA, and therefore the dataset itself tends to carry some bias towards positive reviews.

V. DATA- PREPROCESSING

A. DATA CLEANING

- The first step is to look into duplicates, missing value, and inconsistent data.
- We predict the importance of columns and drop/keep them accordingly, mainly retaining only text data.
- The number of rows containing NANs were extremely insignificant to the total number rows we had, so we dropped them.
- Renaming the columns, and reassigning values for better readability.

B. Preprocessing Text

The text is that the most unstructured kind of all the available data, various sorts of noise are present in it and also the data isn't readily usable with none pre-processing. the whole process of cleaning and standardization of text, making it noise-free and prepared for analysis is understood as text preprocessing and may be a mandatory to induce any coherent results.

The main columns we focus our analysis on is that the overall column(i.e. rating of the product) and therefore the review Text column(i.e textual description and products review)

The rating column comes in five categories(range 1-5), essentially 1 to five stars, which are transformed to three unique categorical values starting from 0 to 2, where 0 is for negative reviews, 1 for neutral reviews and 2 for positive review.

The reviewText column storing textual data goes through multiple preprocessing stages before we run it through the model.

1) Removing Punctuation

One important task in text normalization involves removing unnecessary and special characters including punctuation. the most reason for doing so is because although punctuation can provide useful insights into the sentiment of a review, even state of the art language processing models like BERT and GPT3 are unable to select up on this. The punctuation tends to throw the models off rather than helping them predict the right class, which is why removing the punctuation is a common practice

2) Removing Stopwords

Stopwords are usually words that find yourself occurring the foremost if you aggregated any corpus of text supported singular tokens and checked their frequencies. Articles (a, the, an), Pronouns (you, them, they, me), Prepositions then on are stopwords. they need little or no significance. they're usually aloof from text during processing so on retain words having maximum significance and context.

3) Lemmatization and Stemming

The process of lemmatization is reducing the inflectional varieties of each word into a typical base or root. Stemming usually refers to a crude process that chops off the ends of words within the hope of achieving this goal correctly most of the time, and sometimes includes the removal of derivational units (the obtained element is known because the stem).

Alternatively, lemmatization consists in doing things properly with the use of a vocabulary and morphological analysis of words, to return the underside or dictionary kind of a word, which is understood as the lemma. Although Lemmatization seems to be the route to travel, the difference that we obtained within the results were insignificant.

4) Tokenization

Tokenization is that the process of tokenizing or splitting a string, text into an inventory of tokens. Tokens are subunits of segments of a text document. they can be words, sentences, phrases etc.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

In our analysis we've used the method of Ngrams for tokenization. An n-gram could also be a contiguous sequence of n items from a given sample of text or speech. the items are often phonemes, syllables, letters, words or base pairs in line with the appliance.

5) Representing Textual Data in Numerical form

We used two approaches for converting textual data to numerical form so we can feed it to our models. TfidfVectorizer and CountVectorizer both are methods for converting text data into vectors as the models can process only numerical data.

In Count Vectorizer (Bag of words) we only count the number of times a word appears within the document which results in biasing in favour of most frequent words. This ends up in ignoring rare words which could have helped us in processing our data more efficiently.

For Term Frequency-Inverse Document Frequency the product of Term frequency and inverse document frequency is used. Term frequency is how frequently a term has appeared in a document. as an instance a term appears f times in a document with d words.

$$\text{Term Frequency} = f/d. \quad (1)$$

IDF is inverse document frequency. If a corpus contains N documents and therefore the term of our interest appears only in D documents then IDF is:

$$\text{IDF} = \log(N/D). \quad (2)$$

TF-IDF is a product of Term Frequency and Inverse Document Frequency. TF-IDF shows the rarity of a word within the corpus. If a word is rare then probably it's a signature word for a specific sentiment/information.

VI. EXPLORATORY DATA ANALYSIS

A. Key Relations and Insights from the dataset

1) People are more likely to review a product once they were highly impressed with it.
2) Another observation made is that a customer is more likely to jot down a review once they find a fault as compared to after they think the product is average.
3) Commonly used words and phrases permanently, bad and neutral reviews were analysed to induce clarity on what users use to express their sentiment.

B. Visualizations

1. Percentage of reviews in each category

Figure 1: Comparison of data distribution in various categories of review sentiment

Referring back to the introduction, the dataset contains more positive reviews as compared to average or bad reviews.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

2. Common Phrases used in review according to categories.

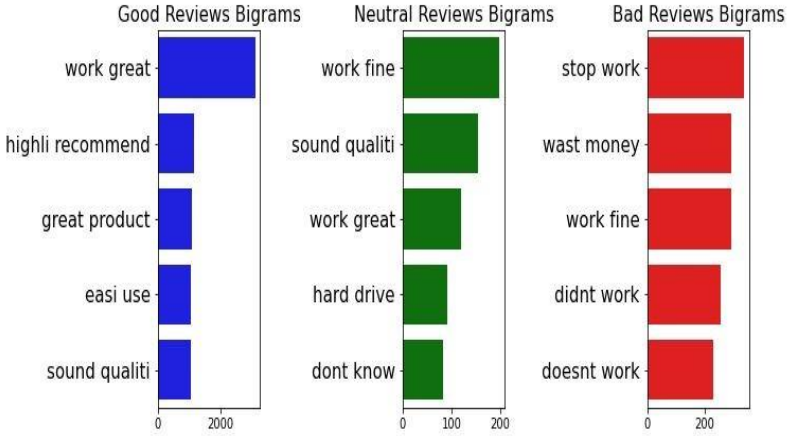

Figure 2: Different Phrases used in reviews and what sentiment they project.

This depicts the frequency of certain words and phrases occurring in each category of reviews. It is observable that there are product specific phrases like sound quality occurring frequently as the dataset is for electronics, which is dominated by headphones. In negative reviews people tend to point out damaged products, products that fail to work and product that don't match its price value.

3. Word Clouds

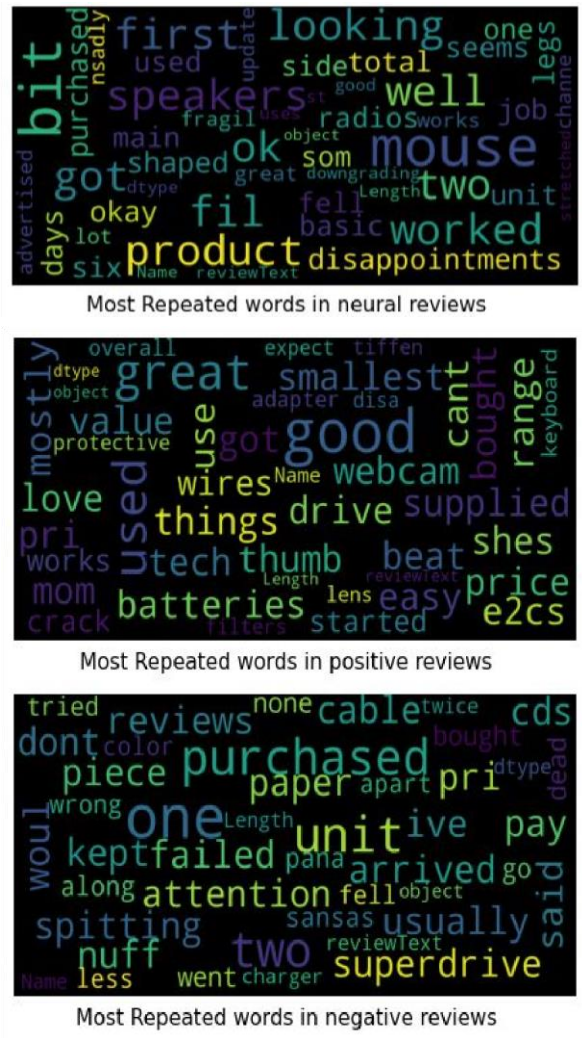

Figure 3: Comparative study of different word clouds generated.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

4. Sentiment Distribution

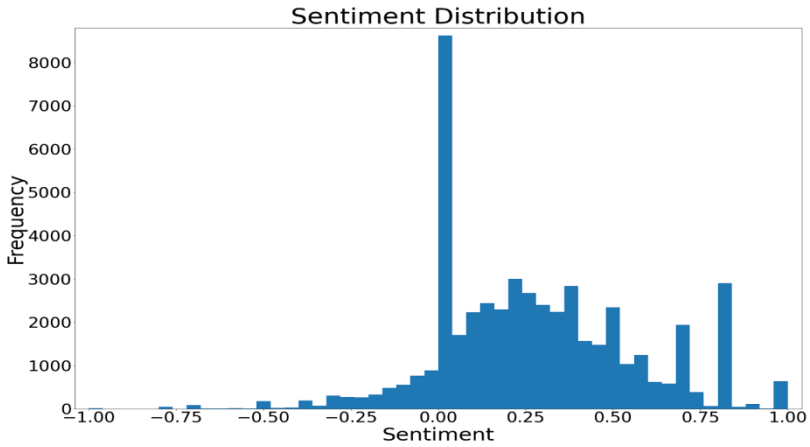

Figure 4: Variation of word frequency with respect to varying sentiment.

Once again this shows that people tend to leave ratings when they are satisfied with the product.

5. Correlation between review length and sentiment

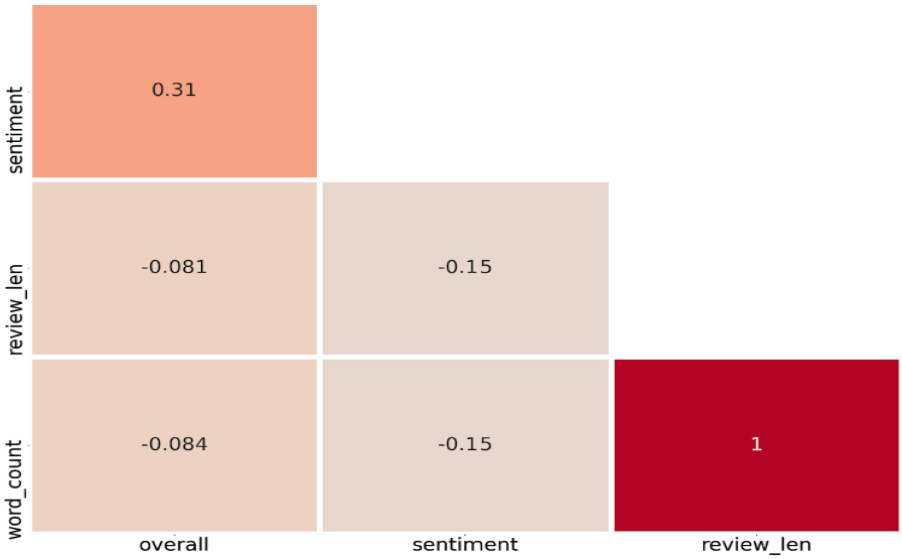

Figure 5: Correlation heatmap to show varying correlation strength and direction between various data attributes.

Correlation between review length and rating (bad products have longer reviews)

VII. EXPERIEMNTAL RESULTS

In all models implemented below we've got implemented selection of k highest scored features of the data to undergo our model. Further on we've tried to tune parameters of previously executed models and tested it using both CountVectorizer and TF-IDF Tokenizer.

A. Naive Bayes

Bayes' Theorem provides the simplest way that we are able to calculate the probability of a chunk of data belonging to a target class using past knowledge. Bayes' Theorem is stated as:

P(class|data) = (P(data|class) * P(class)) / P(data)

Where P(class|data) is that the probability of class given the provided data. Naive Bayes is a classification algorithm for binary (two-class) and multiclass classification problems. Probabilities for every class are simplified i.e. they're considered independent of every other.

We worked with three variants of Naive Bayes algorithm - Multinomial Naive Bayes, Complement Naive Bayes and Bernoulli Naive Bayes using (1,1) ngrams and TF-IDF and CountVectorizer.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

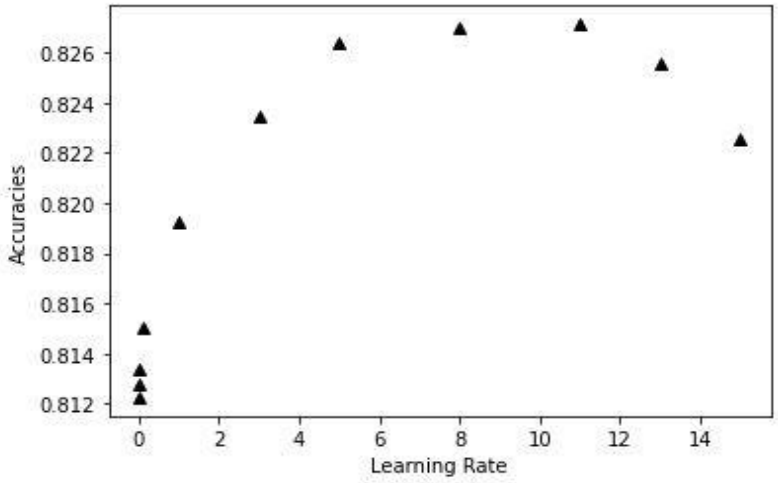

Figure 6: Graph plotted to best estimate hyperparameters when using CountVectorizer tokenization (Learning rate vs accuracy).

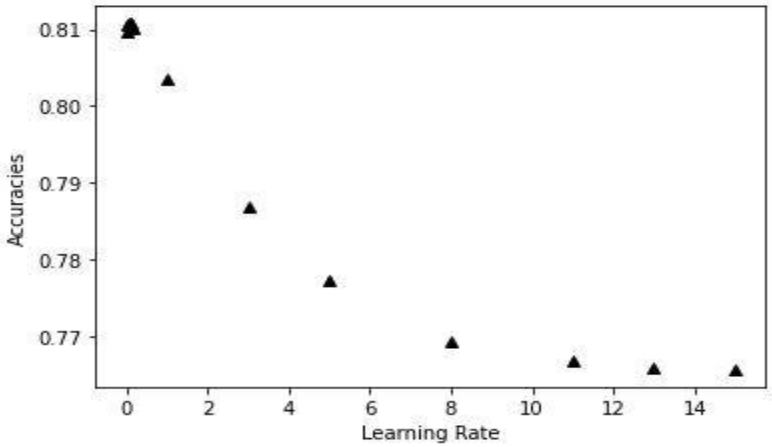

Figure 6: Graph plotted to best estimate hyperparameters when using TF-IDF tokenization (Learning rate vs accuracy).

Table 1. Naive Bayes Accuracy Results

MODEL TYPE	TUNING (k=2890)	
	CountVectorizer Learning Rate :11	TF-IDF Learning Rate:0.1
ultinomial	82.71%	81.07%
Complement	81.15%	78.05%
Bernoulli	75.45%	74.20%

Figure 8: Comparative study of different variants of Naive Bayes Algorithm

B. Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. Most accurate results obtained using the most effective parameters derived from GridSearch method

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

Table 2. Random Forest Accuracy Results

TUNING (k=1200)		
PARAMETERS	**CountVectorizer**	**TF-IDF**
n-estimators : 100 criteria : Gini index No pruning	82.98	83.33
n-estimators : 100 criteria : Entropy No pruning	82.17	N.A.

Figure 9: Comparative study of different variants of Random Forest Classifier

C. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions like (linear) Support Vector Machines and Logistic Regression. even though SGD has been around within the machine learning community for an extended time, it has received a substantial amount of attention within the near past in the context of large-scale learning.

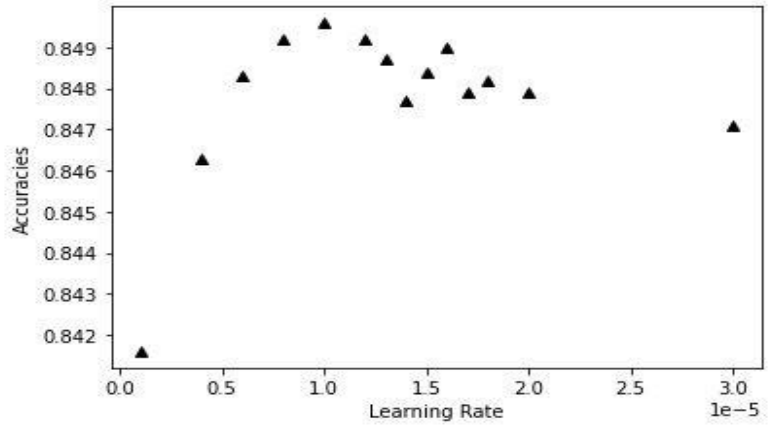

Figure 10: Graph plotted to best estimate hyperparameters when using Support vector machine model(Learning rate vs accuracy)

Table 3. Stochastic Gradient Descent Accuracy Results

LOSS FUNCTION	TUNING (k value=1800)		
	PARAMETERS	**CountVectorizer**	**TF-IDF**
Support Vector Machine	Learning Rate : 1e-05 Epochs: 1000	84.14%	84.96%
Logistic Regression	Learning Rate : 1e-05 Epochs: 1000	83.13%	84.74%

Figure 11: Comparative study of different variants of Stochastic Gradient Descent Algorithm

D. BERT (Bidirectional Encoder Representations from Transformers) Algorithm

BERT is a model designed by researchers at Google AI Language. It presents state-of-the-art leads to a good type of NLP tasks. BERT makes use of Transformer which learns contextual relations between words (or sub-words) in a text. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), with the assistance of Transformers the whole word is read at once. This characteristic allows the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

model to learn the context of a word based on all of its surroundings (left and right of the word), as opposed to only left context based learning of GPT3. This double-sided context is what led us to choose BERT over its competitors as there's no prediction involved, allowing the utmost context to be derived from each item within the dataset

Table 4. Bert Accuracy Results

TUNING **Hyperparameters Chosen:** Batch Size : 3 Steps: 0 Learning Rate: 1e-5 Adam Epsilon: 1e-8	
EPOCHS	**ACCURACY**
Epochs:1	88.89%
Epochs:2	89.05%
Epochs:3	89.25%

Figure 12: Comparative study of different variants of BERT Algorithm

Due to computational constraints, we didn't run further iterations but we believe the accuracy can be increased with more number of epochs.

VIII. RESULTS AND DISCUSSION

After getting the best set of hyperparameters for each of the models we drew a comparison on how they perform with respect to each other to see what model we can expand on in our further studies.

Table 5. Best Model Variants And Their Accuracies

MODEL	**ACCURACY**
Random Forest Classifier (TF-IDF)	83.33%
Naive Bayes Classifier - Multinomial (Bag Of Words)	82.71%
Stochastic Gradient Descent - Support Vector Machines	84.96%
Stochastic Gradient Descent - Logistic Regression	84.74%
BERT Text Classifier	89.25%

Figure 13: Comparative study of different ML Models for Text Classification

IX. CONCLUSION

After multiple model iterations and testing, we believe that the BERT classifier does the best job at estimating the sentiment of a review, with an accuracy of almost 90%. Even though the entire testing and analysis was done at a very basic level, we believe that this will be really useful in various fields of product and user relationship analysis. A good application of this would be in recommendation systems, where users can be clustered based on the similar reviews that they give on sites like amazon.

X. REFERENCES

[1] Najma Sultana &Pintu Kumar & Monika Rani Patra & Sourabh Chandra and S.K. Safikul Alam (2019): Sentimental Analysis of product reviews

[2] Haque, Tanjim & Saber, Nudrat & Shah, Faisal. (2018). Sentiment analysis on large scale Amazon product reviews.10.1109/ICIRD.2018.8376299.

[3] Dublin, Griffith & Joseph, Roshan. (2020). Amazon Reviews Sentiment Analysis: A Reinforcement Learning Approach. 10.13140/RG.2.2.31842.35523.

[4] Tim Althoff , Cristian Danescu-Niculescu-Mizil , Dan Jurafsky Stanford University, Max Planck Institute SWS : How to Ask for a Favor: A Case Study on the Success of Altruistic Requests

[5] Dey, Sanjay and Wasif, Sarhan and Tonmoy, Dhiman and Sultana, Subrina and Sarkar, Jayjeet and Dey, Monisha(February 2020) A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews

[6] Najma Sultana &Pintu Kumar & Monika Rani Patra & Sourabh Chandra and S.K. Safikul Alam (2019) : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[7] http://jmcauley.ucsd.edu/data/amazon/

[8] https://towardsdatascience.com/sentiment-analysis-introduction-to-naive-bayes-algorithm-96831d77ac91

[9] https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31

[10] https://towardsdatascience.com/understanding-randomforest-58381e0602d2