Sentimental Analysis on Amazon Fine Food Reviews Using Machine Learning

MSc Research Project
Data Analytics

Venkata Ramya Bandaru
Student ID: X22151699

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Venkata Ramya Bandaru
Student ID:	X22151699
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Teerath Kumar Menghwar
Submission Due Date:	31/01/2024
Project Title:	Sentimental Analysis on Amazon Fine Food Reviews Using Machine Learning
Word Count:	5537
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	B.V.Ramya
Date:	30th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentimental Analysis on Amazon Fine Food Reviews Using Machine Learning

Venkata Ramya Bandaru
X22151699

Abstract

This research project looks at sentiment analysis of amazon fine food reviews using machine learning techniques. Sorting customer reviews based on whether they are good, or negative is the main goal; this is an important process for the e-commerce industry. The paper investigates many feature extraction and data preparation techniques using logistic regression, such as Word2Vec, TF-IDF, Bag of Words, and a hybrid strategy modified with L1 regularization for better model performance. Data loading, preprocessing, model training, feature engineering, and hyperparameter tweaking are all included in the implementation step. This all-encompassing strategy guarantees the model's efficacy in deciphering the intricate nature of human language in the reviews. Metrics including accuracy, precision, recall, F1-score, and the confusion matrix are used to assess the model's performance. To further guarantee the robustness of the model, regularization strategies and a multicollinearity perturbation test are carried out. The outcomes demonstrate the usefulness of several feature extraction methods as well as the influence of regularization on the model. The research offers a comparative analysis that clarifies the advantages and disadvantages of each technique. In summary, the study not only shows how logistic regression can be successfully applied to sentiment analysis, but it also opens up new avenues for future research in this area. It makes a substantial contribution to sentiment analysis in e-commerce and provides a starting point for more research and advancement in this field.

1 Introduction

Deep learning has been successful in many field such as image Kumar, Brennan and Bendechache (2022); Kumar, Mileo, Brennan and Bendechache (2023); Kumar et al. (n.d.); Roy et al. (2022); Ranjbarzadeh et al. (2023); Aleem et al. (2022); Kumar, Park, Ali, Uddin and Bae (2021); Turab et al. (2022); Singh, Ranjbarzadeh, Raj, Kumar and Roy (2023); Kumar, Park, Ali, Uddin, Ko and Bae (2021); Singh, Raj, Kumar, Verma and Roy (2023); Chandio et al. (2022); Khan et al. (2022); Roy et al. (2023), audio Kumar, Turab, Mileo, Bendechache and Saber (2023); Chandio et al. (2021); Park et al. (2020); Kumar et al. (2020) and many more Kumar, Turab, Raj, Mileo, Brennan and Bendechache (2023). Customer reviews play vital role in today age. Customer reviews have been at the vanguard of this revolutionary shift in consumer behavior since the beginning of internet purchasing. These reviews may be found in abundance in the Amazon Fine Food Reviews dataset, which also gives insights into the opinions of customers. This study untangles the intricate web of consumer feedback in the e-commerce industry by using machine learning algorithms to categorize these evaluations into positive or negative attitudes.

1.1 Establishing the Scene

Customer evaluations are becoming a useful resource for businesses and consumers in the digital age. They serve as a link between the quality of the product and the customer experience, affecting decisions to buy and building brand recognition. These reviews' sentiments are a crucial source of information on consumer satisfaction and product acceptability.

1.2 Historical Narrative

Over the past few decades, sentiment analysis has seen a significant evolution, especially in the e-commerce industry. Online retailers used to assess client feedback using crude methods, frequently restricted to simple keyword-based analysis or basic rating systems. However, the emergence of sites like Amazon and the explosion of user-generated material that followed sparked important developments in this field. To determine the tone of customer evaluations, early attempts in sentiment analysis relied on simple natural language processing (NLP) approaches, mainly concentrating on keyword frequency. With machine learning being included in sentiment analysis, the year 2000 saw a significant change. A deeper, more complex analysis of textual data was made possible by the advent of more advanced techniques like Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) during this era. The next ten years saw a significant expansion of sentiment analysis's application in e-commerce. When combined with machine learning regularization approaches, advanced NLP models such as Word2Vec provided context-aware and more accurate sentiment predictions. These advancements greatly improved the capacity to derive valuable insights from substantial amounts of client feedback. The field of sentiment analysis is currently at a fascinating turning point in its development, with more and more studies concentrating on improving these sophisticated computer methods. In this sense, the Amazon Fine Food Reviews dataset offers a special chance, providing a rich corpus for using and assessing different sentiment analysis techniques. With various featurization approaches and regularization methods in conjunction with logistic regression, this work seeks to add to this developing story by providing both theoretical and useful insights into the field of e-commerce. Sentiment analysis is a branch of natural language processing that started with early attempts to identify and classify human emotions in written communication. Sentiment analysis has advanced and become increasingly sophisticated because of the growth of internet platforms and user-generated content. It is now essential to comprehend customer behaviour. The trajectory of sentiment analysis, from simple polarity detection to sophisticated deep learning approaches, mirrors the development of AI and data science in the interpretation of human language.

1.3 Importance and Research Issue

With consumers depending more and more on evaluations from previous customers to make judgments about what to buy, sentiment analysis has become more important in the e-commerce space, especially on sites like Amazon. This is particularly true in the fine dining industry, where client feedback provides incredibly specific information on the subtleties of their experiences and preferences. Precisely examining these evaluations is essential for comprehending consumer behaviour as well as helping companies mould their product offerings and raise client retention. The difficulty, as usual with classic sentiment analysis approaches, is in efficiently processing and comprehending the nuances of real language in these assessments.

To tackle this problem, the study poses a crucial research question: **How can sentiment analysis accuracy in the e-commerce industry, specifically for Amazon's fine food product reviews, be optimized for logistic regression when applied with a range of featurization methods (Bag of Words, TF-IDF, Word2Vec, and TF-IDF-Word2Vec) and L1 regularization?**

The study aims to investigate and assess these cutting-edge techniques, with a particular emphasis on how well they extract and interpret consumer sentiment from textual data. To help e-commerce companies better understand consumer feedback and guide their product development, marketing, and customer engagement initiatives, the research hopes to provide insightful analysis and useful tools.

1.4 Objectives and Methods

The main goal of this study is to improve sentiment analysis for e-commerce reviews of gourmet culinary products, with a special emphasis on Amazon's large review database. The goal is to combine several text feature extraction techniques with logistic regression, a reliable and popular machine learning method. These techniques include Word2Vec, Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BOW), and a hybrid strategy that combines Word2Vec and TF-IDF. The study also investigates the use of regularization methods, particularly L1 regularization, to improve the logistic regression model's generalizability and prediction accuracy.

The methodological approach entails a thorough examination of the Kaggle-sourced Amazon Fine Food Reviews dataset. This dataset is a perfect case study for this research since it offers an extensive collection of user-generated evaluations. The preprocessing of the text data will be the primary focus of the investigation, after which various featurization approaches will be applied to convert the textual data into a format appropriate for logistic regression analysis. Standard measures including accuracy, precision, recall, and F1-score will then be used to assess the logistic regression model's performance under each featurization approach and with the addition of L1 regularization. Through this procedure, the study hopes to evaluate how well each method performs in precisely identifying the sentiment of the evaluations and investigate the wider ramifications.

1.5 Recognizing Own Limitations

Interpretability and Complexity of the Model: Although logistic regression is a reliable method for binary classification, it may not be as good at capturing intricate correlations in the data as some more complicated models. This could restrict sentiment analysis's depth, particularly when it comes to comprehending the complex context of text evaluations.

Featurization Constraints: Word2Vec, TF-IDF-Word2Vec, Bow, TF-IDF, and Word2Vec text featurization methods are directly related to the efficacy of logistic regression. Every technique has its bounds. For example, Word2Vec-based approaches may overlook the syntactical subtleties of the reviews, while BOW and TF-IDF may oversimplify the text, missing context, and semantic connotations.

Impact of Regularization: To improve the model's performance, the study applies L1 regularization. This may result in the omission of potentially significant features, particularly in a big and varied dataset, even while it can aid in feature selection and prevent overfitting. This could have an impact on the model's generalizability to other kinds of reviews.

Dependency on Data Preprocessing: The effectiveness of logistic regression is significantly influenced by the calibre of data pretreatment. The accuracy of the sentiment analysis in this study may be impacted by procedures like eliminating duplicate rows, HTML elements, and punctuation, and utilizing Porter Stemmer for stemming, which may introduce biases or exclude important data.

Restricted Scope of Evaluation Metrics: The accuracy, precision, recall, and F1-score are the main metrics used in the study to assess the model's performance. Even though these are conventional measures, it's possible that they don't completely convey how well the model handles unbalanced data or the nuances of various sentiment expressions in the reviews.

1.6 An Overview of the Assumptions

Several fundamental presumptions guide the sentiment analysis using logistic regression that is carried out on the Amazon Fine Food Reviews dataset. First, it is presumed that the text reviews sufficiently capture the attitude stated by the reviewers after they have been preprocessed and feature-rich utilizing techniques such as BOW, TF-IDF, Word2Vec, and TF-IDF-Word2Vec. This involves the implicit presumption that the selected featurization methods accurately capture the subtleties of context and semantic meanings in natural language. Another key premise is that there is a linear relationship between the characteristics and the log chances of the review sentiments, which makes logistic regression applicable for binary sentiment categorization into positive or negative categories.

Furthermore, it is anticipated that the preprocessing procedures which include stemming, HTML tag removal, and duplication elimination will improve the model's performance without appreciably sacrificing pertinent data. These presumptions provide the basis for the study's analytical methodology and are essential for applying the model and interpreting its results.

1.7 Report Structure

Section 2 of the report discusses the current works done in the field. Section 3 Methodology, details the description of data preprocessing steps, feature engineering techniques, and model training processes. Section 4 is a Comparative analysis of the performance of each model, discussing strengths and weaknesses. The section 5 section presents the results in terms of classification metrics for each model. Whereas, Interpretation of the findings, and exploration of their implications for the fine food sector in e-commerce, including consumer behaviour and product strategy is discussed in section 6. Finally, the Conclusion and Future Work are discussed in the last section.

2 Related Work

2.1 Enhanced POS tagging on Twitter

According to (Sun et al.; 2012), 2012 paper ,"Twitter Part-of-Speech Tagging Using Pre-classification Hidden Markov Model", pre-classification hidden Markov models (HMMs) provide a novel method for part-of-speech (POS) tagging on Twitter. Through a pre-classification HMM combined with the outcomes of polarity classification, the suggested approach added subjectivity information to the POS tagging task. The testing sentence's degree of subjectivity was found to improve the POS tagging performance when it was employed as a combination factor to select a suitable value from the interval.

2.2 Unsupervised Machine Learning using SentiWordNet

The authors present an unsupervised opinion mining method that use SentiWordNet to find emotion words and phrases in text reviews in their 2015 paper "Unsupervised Opinion Mining From Text Reviews Using SentiWordNet" ((Soni and Patel; 2014)). Sentiment words and phrases are identified by first segmenting the text into sentences and then analyzing their SentiWordNet scores. The method then use a co-occurrence matrix to pinpoint sentiment word and phrase clusters that are most likely to correspond to features of the product. Ultimately, the methodology leverages the clusters to extract product reviews pertaining to every facet of the product.

2.3 Sentimental Analysis in E-Commerce

In the 2015 publication "Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms," ((Mm et al.; 2020). offer a thorough examination of the algorithms and approaches used in sentiment analysis in e-commerce. The two main methods for sentiment analysis that are thoroughly examined in this research are lexicon-based and machine learning-based techniques. Lexicon-based techniques use dictionaries and predetermined sentiment lexicons to rate the sentiment of words or phrases. Although this method is simple and effective, it might not be sophisticated enough to deal with intricate or changing linguistic patterns.

Authors	Models	Limitations
(Sun et al.; 2012)	Naive Bayes Classifier	limited evaluation datasets, overemphasis on subjectivity information, neglected domain-specific bias, and static modeling
(Soni and Patel; 2014)	SentiWordNet, Sentence Segmentation and Clustering	It is based only on bag-of-words representation of text and other forms of sentiment modification are not handled
(Mm et al.; 2020)	Naïve bayes, SVM, Random Rorest, RNN, CNN	Overemphasis on Naive Bayes, Lack of Comparative Analysis of Lexicon-based and Machine Learning Approaches, costly to compute
(Farhan et al.; 2023)	KNN with an accuracy of 81.8%	small datatset, single algorithm used, doamin speciifc naunces are not identiifed
(Sharma et al.; 2023)	SVM, Naive Bayes accuracy of 82.1 and 83.7%	single datatset, no broader applications are used
(Muhammad et al.; 2021)	word2vec with LSTM achieved accuracy of 85.96%	single evaluation metric is used
(Zhao and Sun; 2022)	BERT model achieved accuracy of 89.34%	couldnt interpret deep learning models, limited use in real time application
(Parthasarathy et al.; 2019)	hybrid(Word2Vec and CNN) model achieved accuracy of 81.60%	difficult to debug and improve performance
(Joshi et al.; 2023)	Naive Bayes model achieved accuracy of 84.93%	Domain-Specific Bias, smaller datasets
(Kumar, Sahoo, Mahapatro, Awasthi and Sahoo; 2022)	BERT, LSTM, CNN model achieved accuracy of 90.8 89.3 and 87.2%	DL models require large data, interpretability issues

2.4 Sentimental Analysis of Practo App Reviews

A sentiment analysis technique based on k-nearest neighbors (KNN) and word2vec is presented in the paper ”Sentiment Analysis of Practo App Reviews using KNN and Word2Vec” by ((Farhan et al.; 2023). It achieves an excellent accuracy of 81.8%, outperforming current lexicon-based approaches. However, its generalization is limited by the use of a single algorithm (KNN) for assessment and a small dataset. More research should be done to increase the dataset’s usefulness and investigate a greater variety of algorithms, particularly those that are customized to account for peculiarities unique to a particular domain, such as medical terminology.

2.5 Machine Learning based Sentimental Analysis of Movie Reviews

The sentiment analysis of movie reviews from IMDb is examined in the paper ”Sentimental Analysis of Movie Reviews Using Machine Learning” by ((Sharma et al.; 2023). The analysis involves examining a dataset of movie reviews to determine sentiment polarity, extract important opinion phrases, and classify reviews according to their sentiment. Models obtain 82.1% and 83.7% accuracy for SVM and Naive Bayes classifiers, respectively. The study highlights the significance of comprehending audience preferences for movie production decisions by identifying the most prevalent sentiment patterns and important opinion expressions in both positive and negative reviews.

2.6 Harnessing Sentimental Analysis of Hotel Reviews using LSTM

The paper ”Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews” by ((Muhammad et al.; 2021) investigates the sentiment analysis of hotel reviews in the Indonesian language. In this the proposed method combines word embedding using Word2vec with recurrent neural networks (RNNs) using Long Short-Term Memory (LSTM) units. This outperforms traditional lexicon-based methods and displays the robustness of combining Word2vec and LSTM for sentiment analysis in Indonesian. this research shows the effectiveness of combining Word2vec and LSTM for capturing semantic and syntactic information in the Indonesian language, leading to improved sentiment classification with an accuracy of 85.96

2.7 BERT driven Sentimental Analysis

The paper “Amazon Fine Food Reviews with BERT Model” by ((Zhao and Sun; 2022) investigates the sentiment analysis of food reviews from Amazon’s online marketplace. It utilizes the BERT (Bidirectional Encoder Representations from Transformers) a neural network model to calculate sentiment polarity. It illustrates how BERT can effectively capture contextual linkages and rich semantic information in food evaluations, improving the accuracy of sentiment classification. The results have potential applications in the food sector, including better product creation, identifying consumer preferences, and creating focused advertising campaigns.

2.8 Convolutional Neural Network(CNN) Approach for Medical Reviews

Sentimental Analysis Using Convolution Neutral Network through Word to Vector Embedding for Patients Dataset" by ((Parthasarathy et al.; 2019) takes patient reviews from medical websites. This research suggests a technique to classify reviews into good, negative, and neutral categories by combining convolutional neural networks (CNN) with word embedding using Word2vec. The proposed method can be used to various tasks in health care industry.

2.9 Sentimental Analysis using Supervised Machine Learning

The use of different machine learning (ML) algorithms for sentiment analysis is examined in the study "Sentiment Analysis using Machine Learning Algorithms" by ((Joshi et al.; 2023). It investigates the suitability of several machine learning methods for categorizing twitter user text into positive, negative, and neutral sentiment categories, such as Naive Bayes, Support Vector Machines (SVMs), and Support Vector Regression (SVR). Particularly noteworthy for its ease of use and efficiency in text categorization problems is Naive Bayes. It draws attention to how well various machine learning methods handle ambiguous language and extract sentiment from textual input. Opinion mining, consumer feedback analysis, and social media analytics can all benefit from the findings.

2.10 Sentimental Analysis using Deep Learning

An important addition to the use of deep learning (DL) for sentiment analysis in e-commerce applications is made by the publication "Sentimental Analysis of Amazon Customers using Deep Learning Techniques" by (Kumar, Sahoo, Mahapatro, Awasthi and Sahoo; 2022). It investigates how to categorize reviews into positive, negative, and neutral sentiment categories using a variety of deep learning models, such as Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs). It highlights the advantages of deep learning (DL) models over more conventional machine learning (ML) techniques and illustrates how DL may be used to better analyze customer sentiment and inform business decisions.

3 Methodology

In Figure 1, the methodology followed in this research is explained. The six steps steps involved in this research are data collection, data preprocessing, featurization techniques, model training, model evaluation and model validation.

In data collection stage the data is collected from Kaggle and stored into sqllite database. Using jupyter notebook as a platform, the data is extracted to a dataframe from database. Later, pre-processing techniques like eliminating duplicate rows, HTML tag removal, stemming, lemmatization and categorization are performed.

In Featurization stage, Bag of words(BOW), Term Frequency Inverse Document Frequency(TF-IDF) are performed on textual data to convert to numerical data. Techniques like Word2Vec, Avg-Word2Vec are performed to extract semantic meaning of the

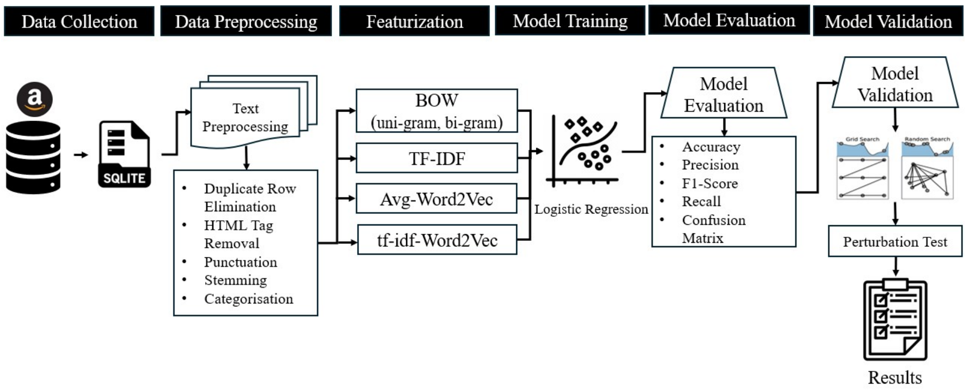


Figure 1: Research Methodology

data. Hybrid methods like tf-idf-Word2Vec is also performed to improve accuracy of the model.

In model training phase, Logistic regression is used for its fast prediction and training over large datasets. In the next step, model is trained on the data and evaluated using metrics such as accuracy, precision, recall, f1-score.

In the final step, the model is evaluated by adding noisy data which is known as performing perturbation test. In this stage GridSearchCV, RandomizedSearchCV are performed by using L1,L2 regularization techniques. Performing this creates more robust and generalised machine learning model. Lastly, the results are cross verified on different featurization techniques.

3.1 Logistic regression application

This project uses the predictive analytic technique of logistic regression to categorize Amazon food reviews as either favourable or negative. The process begins with choosing several feature sets for the model, including Average Word2Vec, Term Frequency-Inverse Document Frequency (TF-IDF), and Bag of Words (BoW) containing unigrams and bigrams.
justification: The selection of logistic regression stems from its effectiveness in solving binary classification issues and its potential to yield outcome probabilities, which are valuable in comprehending the degree of confidence associated with predictions.

3.2 Cross-Validation Techniques

Grid Search CV: By experimenting with every possible combination of preset settings, hyperparameters were fine-tuned using this extensive search technique. It guarantees that the model has good generalization and optimization.

As a quicker substitute for Grid Search, Randomized Search CV selects a predetermined number of parameter settings from predefined distributions. With a great deal less computing, this approach can get results that are comparable.

3.3 Evaluation Metrics for Performance

Several metrics, including accuracy, F1-score, precision, and recall, were used to assess the model's performance. These measures offer a comprehensive assessment of the model's effectiveness, considering both false positives and false negatives.

Confusion Matrix: Using Seaborn for visualization, confusion matrices were shown to reveal the proportion of accurate and inaccurate predictions, so providing more insight into the kinds of errors the model is producing.

3.4 Regularization and Sparsity

Observation on Sparsity: The study investigates how the degree of regularization affects the degree of sparsity in the model's coefficients. Regularization facilitates feature selection, simplifies the model, and guards against overfitting.

Regularization Techniques: Special attention has been given to the L1 regularizer (Lasso regression), which can decrease the number of features by setting the coefficients of less significant predictors to zero.

3.5 Perturbation Test for Multicollinearity

To determine if there was multicollinearity among the features, a perturbation test was run. It was investigated whether there were substantial changes in the coefficients, suggesting multicollinearity, by adding noise to the data and reevaluating the model.

4 Design Specification

4.1 Data Summary Coverage of the dataset

568,454 reviews of 74,258 goods from 256,059 consumers between October 1999 and October 2012 are included in the dataset. To verify robustness, a sample of data from users who submitted more than 50 reviews was also added.

Features: The ID, product ID, user ID, profile name, helpfulness numerator and denominator, score, time, summary, and the entire review text were among the ten elements that were mentioned. Every property has a function, either as a target for categorization or for feature extraction.

4.2 Imports from Libraries

Essential Python libraries for database operations (sqlite3), data management (numpy, pandas), and visualization (matplotlib, seaborn). gensim for natural language processing tasks and sklearn.metrics for model evaluation.

4.3 Metrics of Performance

The precision, recall, accuracy, confusion, precision score, `f1_score`, and recall scores are among the metrics used to evaluate the predictive performance of the logistic regression model.

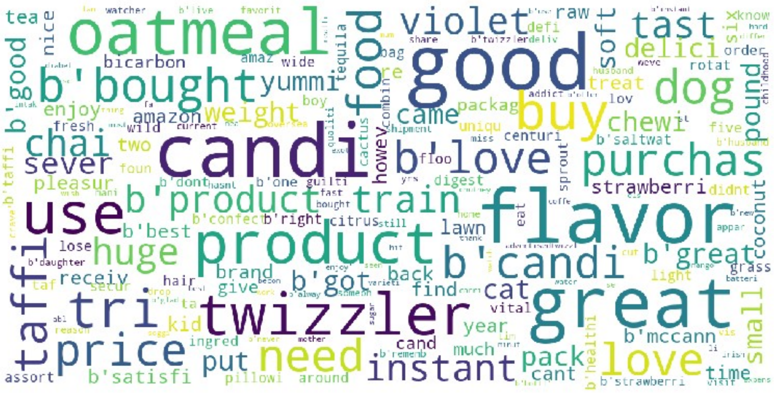


Figure 2: Word Cloud Plot on Positive Reviews

4.4 Utilities and configuration

Suppressing alerts that aren't essential.For in-notebook visualization, specify `%matplotlib` inline and filterwarnings("ignore"). pickle to help with effective workflow management by loading and storing model items.

4.5 Context of Application

The code is ready for sentiment analysis, which uses logistic regression as the classification technique to categorize text reviews from Amazon as either positive or negative. In data science, this procedure is essential to comprehending consumer input.

5 Implementation

5.1 Loading of the data

Using Python's sqlite3 package, a SQLite database connection is made during the data loading stage. This enables SQL-type interactions with databases straight from Python. The `read_sql_query` method in pandas is used to run a SQL query that retrieves the dataset, which has probably previously been cleaned and preprocessed. To concentrate just on positive or negative feelings, this query selects all entries from the evaluations database, leaving out neutral evaluations with a score of 3. Tahen, using df.head(), the first few items of the dataset are shown, giving a glimpse of the structure of the data, which contains review texts, ratings, user and product IDs, and other metadata. To ensure that the data is appropriately obtained and prepared before moving on to analysis and modeling, this step is essential.

5.2 Performing EDA

The statistical summary of the numerical columns, which includes the mean, standard deviation, and range of values, is produced by the describe() function and aids in the understanding of the data distribution. The result indicates that the dataset has 364,171 data points.

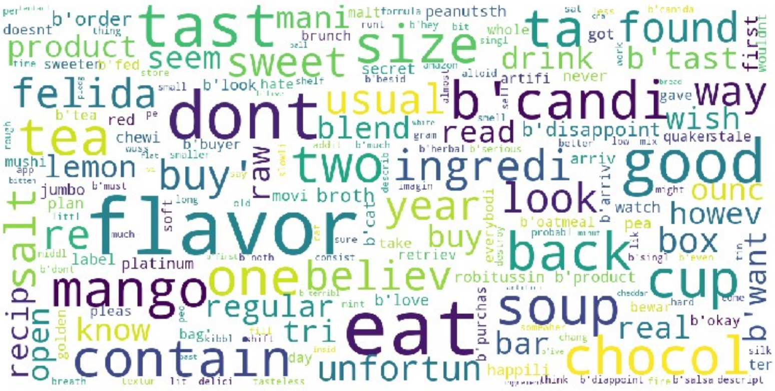


Figure 3: Word Cloud Plot on Negative Reviews

The structure of the dataset is next verified, and the total number of scores which ought to equal the number of data points is reported after that. This suggests that every data point has a corresponding score.

During the preprocessing stage of sentiment analysis, textual score representations are transformed into numerical binary classes using a function called polarity. 'Positive' scores are given a value of 0, while all other scores are given a value of 1. For machine learning models to do sentiment analysis which aims to infer a positive or negative sentiment from the review text, binary classification is necessary. Also, The word cloud plotted for both positive and negative reviews to understand the major words contributing to the categories as shown in Figure 2 and Figure 3.

The map() method is then used to apply this binary mapping to every score in the dataset, thus getting the dataset ready for additional analysis or model training. It is possible that the df.head() function is invoked once more to confirm that the mapping has been applied accurately; however, the extract does not show this portion of the output.

Then, to preserve chronological order, the dataset is sorted according to the dates of the reviews. This guarantees that any trends or changes over time are recorded and considered throughout the modeling process, which is essential for time series analysis. The dataset is then stored to disk, enabling the model to be trained on reliable data from many algorithm trials. This phase is crucial for evaluating the performance of several algorithms on the same dataset and ensuring repeatability. To preserve consistency in the assessment of the models, the stored data may then be reloaded from the disk to guarantee that the same sample is utilized in all algorithm testing.

5.3 Logistic Regression model

Bag of Words

Vectorization (BoW): First, CountVectorizer from the sklearn module is used to transform the set of text documents into a matrix of token counts. Word order and grammar are ignored in favor of tracking word frequency, which is used as a feature by the classifier.

Data Splitting: train test split is used to divide the dataset into training and test sets. Typically, 30% of the dataset is put aside for testing. For assessing the model's performance on hypothetical data, this distinction is essential.

Normalization: Preprocessing is used to standardize the training and test datasets.standardize after sklearn. By adjusting the feature vectors' size, normalization ensures that every feature contributes equally to the logistic regression algorithm's distance computations. When opposed to data normalization, which centers the data around zero, this can frequently result in higher accuracy.

Training and Test Size: The size of the datasets that will be utilized for modeling is confirmed by printing the training and test data's form following preprocessing. In order to ensure that the split was done successfully and that the datasets are prepared for training and assessment, this step is crucial.

Hyperparameter tuning: Prior to training, machine learning models frequently require the setting of hyperparameters. The performance of a model can be greatly affected by the selection of hyperparameters. In this instance, we are adjusting the logistic regression model's "C" (regularization strength) and "penalty" (kind of regularization) hyperparameters. By adjusting these hyperparameters, one may increase model accuracy by determining the best configuration for your dataset.

Temporal structure characterizes time-ordered data, such as financial time series or sequential sensor data, and this is achieved by **Time Series Cross-Validation** (Forward Chaining). Conventional random data splitting might result in erroneous model assessments and data leakage into training and testing sets. By using time series cross-validation, one can make sure that the data you test the model on arrives after the training data in time. This is essential for practical uses where the model needs to make predictions on future data points.

6 Evaluation

The main purpose of this section is to provide a comprehensive analysis of the results and main findings of the study as well as the implications of these finding both from academic and practitioner perspective are presented. Only the most relevant results that support the research question and objectives are presented. Provided an in-depth and rigorous analysis of the results.

Optimal Hyperparameters: Using the given dataset as a guide, the algorithm determines the ideal hyperparameters for your logistic regression model. The ideal hyperparameters in your situation were "C": 10 and "penalty": "l2." To ensure that the logistic regression model has the best accuracy possible, these hyperparameters are essential. The Change in Misclassification Error in L1 and L2 Regularization is shown in Figure 4.

Increased Model Accuracy: One may make sure that your model is appropriate for the temporal nature of the information by utilizing Time Series Cross-Validation and the optimal hyperparameters. As a result, the model becomes more precise and trustworthy. The greatest accuracy of 92.16% was attained in this instance, demonstrating the model's ability to forecast time-ordered data.

In order to predict models performance, accuracy, precision, f1-score, recall plays an important role. To calculate this, true positives, true negatives, false positives, false negatives scores are needed. The score of bigrams model building using L2 regularizer with regularizer parameter C set to 100 is displayed in Figure 5.

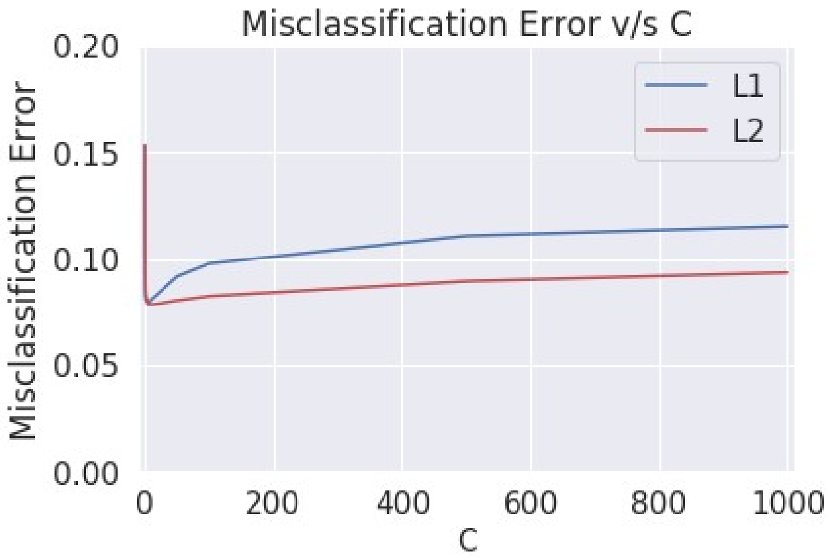


Figure 4: Change in Misclassification Error in L1 and L2 Regularization for bigram

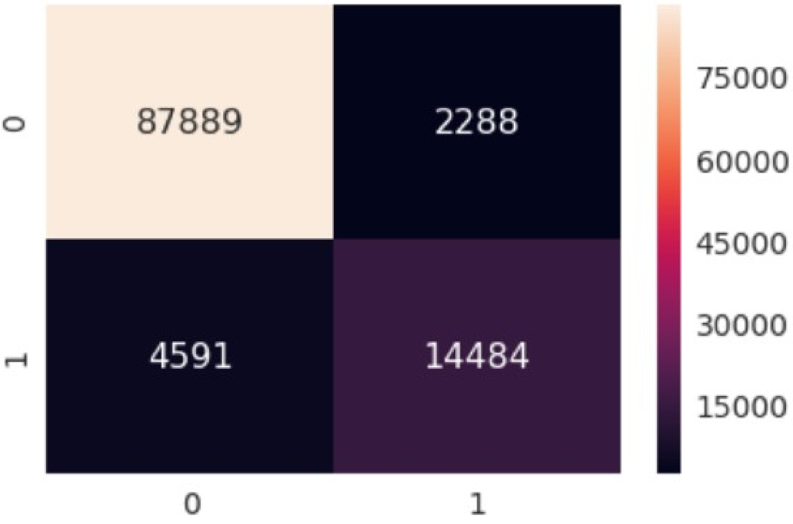
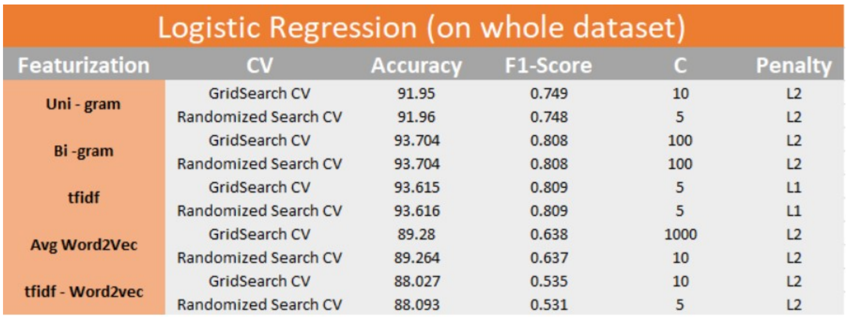


Figure 5: heatmap representation for bigrams



Logistic Regression (on whole dataset)

Featurization	CV	Accuracy	F1-Score	C	Penalty
Uni - gram	GridSearch CV	91.95	0.749	10	L2
	Randomized Search CV	91.96	0.748	5	L2
Bi -gram	GridSearch CV	93.704	0.808	100	L2
	Randomized Search CV	93.704	0.808	100	L2
tfidf	GridSearch CV	93.615	0.809	5	L1
	Randomized Search CV	93.616	0.809	5	L1
Avg Word2Vec	GridSearch CV	89.28	0.638	1000	L2
	Randomized Search CV	89.264	0.637	10	L2
tfidf - Word2vec	GridSearch CV	88.027	0.535	10	L2
	Randomized Search CV	88.093	0.531	5	L2

Figure 6: Comparison of the model performance

6.1 Relevence

Generalization: Time Series Cross-Validation and hyperparameter adjustment let your model generalize to new data. In time series data, where future data points may exhibit distinct patterns from prior data, this is particularly crucial.

Preventing Data Leakage: If you train and test on overlapping time periods, data leakage may happen. Time series cross-validation helps avoid this. It guarantees that the performance of the model is assessed realistically, as if it were generating predictions in an actual situation.

Optimal Performance: One may maximize your model's performance by determining the ideal hyperparameters. In fields like banking, where precise forecasts are extremely valued, this might be crucial.

Sparsity with L1 Regularizer: The text describes how, when employing L1 regularization, the degree of sparsity in the model rises when the regularization parameter "lambda" (or, inversely, "C") is changed. It draws attention to how drastically the number of non-zero weights in the model changes as 'lambda' changes.

Using **Randomized Search Cross-Validation (CV)** to determine the optimal hyperparameters for a machine learning model is mentioned in the context of Mention. Using a random sample from a given parameter space, this approach effectively searches for the ideal hyperparameters.

Perturbation Test: According to the text, 42 features had weight shifts of more than 30% after a perturbation test was carried out. This shows that certain characteristics are multicollinear, which may have an impact on the interpretability of the model.

Significance of Features: The Multinomial Naive Bayes model's 'coef_' attribute is explored in particular in relation to the model's features' significance. It illustrates that in a binary classification issue, greater values of this characteristic suggest more significant features for the positive class.

N-grams: The term "bi-gram" implies that the feature engineering process considered n-gram features, particularly bigrams.

Other approaches: Word2Vec and TF-IDF are two other approaches that are frequently used for natural language processing applications, and they are briefly mentioned in the article.

Dataset: Although the specifics of the dataset are not given in this article, the analysis and approaches discussed are applied to a particular dataset.

6.2 Observations

In terms of accuracy and F1-score, bi-gram featurization in conjunction with logistic regression produced the greatest results as shown in the Figure 6. This implies that the model's predictive value is increased when bigrams—pairs of consecutive words are considered as they capture more context than unigrams, which are just single words.

Additionally, TF-IDF fared well, almost matching the bi-gram results, particularly in conjunction with L1 regularization that facilitates feature selection and model simplicity. When compared to other approaches, average Word2Vec and TF-IDF Word2Vec featurizations produced lower accuracy and F1-scores. This might be because weighting word vectors with TF-IDF scores or averaging them results in the loss of semantic meaning.

Given that Randomized Search and GridSearch CV perform almost identically, Randomized Search may be a more cost-effective option than GridSearch, particularly in situations when the dataset is huge.

Various text representations may need various regularization techniques to enhance performance, as seen by the variations in the choice of the hyperparameter C and the penalty type (L1 or L2).

6.3 Efficiency of Featurization

Bi-gram feature extraction produced the best F1-score and accuracy. This implies that bigrams (pairs of words) are important for capturing context in sentiment analysis. It enables the model to comprehend sentences with distinct meanings that could not be conveyed by a single word (uni-gram). "Not good" and "good," for example, have different sentiments, and bi-gram does a fantastic job of capturing these differences.

6.4 Regularization and Complexity of Models

L2 regularization was applied to various featurization's, especially those where the highest accuracies were noted. This suggests that L2 regularization which penalizes the square of the coefficient magnitude, maintains the complexity of the model without appreciably reducing its predictive power and aids in preventing overfitting.

TF-IDF featurization demonstrated the efficacy of L1 Regularization. This kind of regularization can result in sparser solutions by pushing some coefficients to zero, which essentially performs feature selection. It lessens the impact of less informative phrases while emphasizing the significance of terms (features) in sentiment analysis.

6.5 Adjusting Hyperparameters

Across featurization, there were notable differences in the values of the hyperparameter C, which is the inverse of regularization strength. A lower C denotes more robust regularization. The model's ability to balance bias and variation depends critically on the selection of C. For instance, TF-IDF Word2Vec and Avg Word2Vec required a greater value of C (less regularization), presumably because the averaging procedure intrinsically lowered the feature set for both approaches.

6.6 Assessing and Choosing Models

Randomized Search can be a computationally economical option for this dataset and the logistic regression model without sacrificing the model's performance, as evidenced by the fact that neither GridSearch CV nor Randomized Search CV significantly differed in performance.

6.7 Using Multicollinearity and Sparsity

The sparsity study revealed that the model got sparser, especially with L1 regularization, as the regularization strength rose (or C dropped). This is critical when dealing with high-dimensional data, such as text, because many characteristics could be superfluous. Features showed multicollinearity, according to the perturbation test. Although multicollinearity may be accommodated by logistic regression to a certain degree, its existence implies that some characteristics can be redundant. This is an important discovery for dimensionality reduction and figuring out which data points are responsible for the sentiment analysis.

6.8 Compliance with the Research question

Accurately categorizing evaluations as positive or negative was the goal of the study. The results demonstrate that logistic regression can accomplish this with high accuracy when proper feature extraction and regularization are applied. The best-performing models (TF-IDF and bi-gram) closely match the goal, indicating that the approach is appropriate for the given binary classification job.

7 Conclusion and Future Work

- Features are multi-collinear i.e. they are co-related
- Bigram Featurization performs best with accuracy of 93.704 and F1-Score of 0.808
- Sparsity increases as we increase lambda or decrease C when L1 Regularizer is used
- Algorithm like Logistic Regression performed best on this data

The findings show that logistic regression is a reliable model for text data sentiment analysis. To achieve high performance, selecting the right text feature, regularization strategy, and hyperparameter optimization are essential. The results validate the idea that effectively assessing the sentiment of reviews requires a grasp of the context (with bi-grams) and the significance of certain phrases (via TF-IDF). The employed technique effectively achieves the study's goal, offering a solid framework for precisely categorizing review feelings on e-commerce sites such as Amazon.

For future work, other departmental reviews of hospitality or business sector can be opted. More datasets can be used to make model more robust. Other machine learning and deep learning models can be applied to achieve better accuracy results.

Acknowledgement

My sincere thanks to the research supervisor Teerath Kumar Menghwar, all the faculty of Research In Computing (National College of Ireland) and friends who supported me to

get clear understanding of the project and its documentation in the best ways possible.

References

Aleem, S., Kumar, T., Little, S., Bendechache, M., Brennan, R. and McGuinness, K. (2022). Random data augmentation based enhancement: a generalized enhancement approach for medical datasets, *arXiv preprint arXiv:2210.00824* .

Chandio, A., Gui, G., Kumar, T., Ullah, I., Ranjbarzadeh, R., Roy, A. M., Hussain, A. and Shen, Y. (2022). Precise single-stage detector, *arXiv preprint arXiv:2210.04252* .

Chandio, A., Shen, Y., Bendechache, M., Inayat, I. and Kumar, T. (2021). Audd: audio urdu digits dataset for automatic audio urdu digit recognition, *Applied Sciences* **11**(19): 8842.

Farhan, M., Purbolaksono, M. D. and Astuti, W. (2023). Sentiment analysis of practo app reviews using knn and word2vec, *Building of Informatics Technology and Science (BITS)* **1**.

Joshi, V., Patel, S., Agarwal, R. and Arora, H. (2023). Sentiments analysis using machine learning algorithms, *International Conference on Electronics and Renewable Systems (ICEARS)* pp. 1425–1429.

Khan, W., Raj, K., Kumar, T., Roy, A. M. and Luo, B. (2022). Introducing urdu digits dataset with demonstration of an efficient and robust noisy decoder-based pseudo example generator, *Symmetry* **14**(10): 1976.

Kumar, S., Sahoo, R. R., Mahapatro, R., Awasthi, S. and Sahoo, S. (2022). Sentimental analysis of amazon customers using deep learning techniques, *nternational Conference on Machine Learning, Computer Systems and Security (MLCSS), Bhubaneswar, India* pp. 259–266.

Kumar, T., Brennan, R. and Bendechache, M. (2022). Stride random erasing augmentation, *CS & IT Conference Proceedings*, Vol. 12, CS & IT Conference Proceedings.

Kumar, T., Mileo, A., Brennan, R. and Bendechache, M. (2023). Rsmda: Random slices mixing data augmentation, *Applied Sciences* **13**(3): 1711.

Kumar, T., Park, J., Ali, M. S., Uddin, A. and Bae, S.-H. (2021). Class specific autoencoders enhance sample diversity, *Journal Of Broadcast Engineering* **26**(7): 844–854.

Kumar, T., Park, J., Ali, M. S., Uddin, A. S., Ko, J. H. and Bae, S.-H. (2021). Binary-classifiers-enabled filters for semi-supervised learning, *IEEE Access* **9**: 167663–167673.

Kumar, T., Park, J. and Bae, S.-H. (2020). Intra-class random erasing (icre) augmentation for audio classification, *Korean Society of Broadcasting and Media Engineering conference proceedings* pp. 246–249.

Kumar, T., Turab, M., Mileo, A., Bendechache, M. and Saber, T. (2023). Audrandaug: Random image augmentations for audio classification, *arXiv preprint arXiv:2309.04762* .

Kumar, T., Turab, M., Raj, K., Mileo, A., Brennan, R. and Bendechache, M. (2023). Advanced data augmentation approaches: A comprehensive survey and future directions, *arXiv preprint arXiv:2301.02830* .

Kumar, T., Turab, M., Talpur, S., Brennan, R. and Bendechache, M. (n.d.). Forged character detection datasets: Passports, *DRIVING LICENCES AND VISA STICKERS* .

Mm, M., Batcha, N. K. and and, M. R. (2020). Sentiment analysis in e-commerce: A review on the techniques and algorithms, *ResearchGate* **6**.

Muhammad, P. F., Kusumaningrum, R. and Wibowo, A. (2021). Sentiment analysis using word2vec and long short-term memory (lstm) for indonesian hotel reviews, *Procedia Computer Science* **179**: 728–735.

Park, J., Kumar, T. and Bae, S.-H. (2020). Search for optimal data augmentation policy for environmental sound classification with deep neural networks, *Journal of Broadcast Engineering* **25**(6): 854–860.

Parthasarathy, G., Preethi, D., Christo, M. S. and Soumya, T. R. (2019). Sentimental analysis using convolution neutral network through word to vector embedding for patients dataset, *Emerging Trends in Computing and Expert Technology* **106**: 1042–1051.

Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Tataei Sarshar, N., Tirkolaee, E. B., Ali, S. S., Kumar, T. and Bendechache, M. (2023). Me-ccnn: Multi-encoded images and a cascade convolutional neural network for breast tumor segmentation and recognition, *Artificial Intelligence Review* pp. 1–38.

Roy, A. M., Bhaduri, J., Kumar, T. and Raj, K. (2022). A computer vision-based object localization model for endangered wildlife detection, *Ecological Economics, Forthcoming* .

Roy, A. M., Bhaduri, J., Kumar, T. and Raj, K. (2023). Wildect-yolo: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection, *Ecological Informatics* **75**: 101919.

Sharma, H., Pangaonkar, S. A., Gunjan, R. and Rokade, P. (2023). Sentimental analysis of movie reviews using machine learning, *ITM Web of Conferences* **53**: 1–2.

Singh, A., Raj, K., Kumar, T., Verma, S. and Roy, A. M. (2023). Deep learning-based cost-effective and responsive robot for autism treatment, *Drones* **7**(2): 81.

Singh, A., Ranjbarzadeh, R., Raj, K., Kumar, T. and Roy, A. M. (2023). Understanding eeg signals for subject-wise definition of armoni activities, *arXiv preprint arXiv:2301.00948* .

Soni, V. and Patel, M. R. (2014). Unsupervised opinion mining from text reviews using sentiwordnet, *International Journal of Computer Trends and Technology* **11**: 234–238.

Sun, A. S., Liu, H., Lin, H. and Abraham, A. (2012). Twitter part-of-speech tagging using pre-classification hidden markov model, *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* pp. 1118–1123.

Turab, M., Kumar, T., Bendechache, M. and Saber, T. (2022). Investigating multi-feature selection and ensembling for audio classification, *arXiv preprint arXiv:2206.07511* .

Zhao, X. and Sun, Y. (2022). Amazon fine food reviews with bert model, *Procedia Computer Science* **208**: 401–406.