International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Utilizing NLP Sentiment Analysis Approach to Categorize Amazon Reviews against an Extended Testing Set

Arman Sarraf*

Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
Email: arman.hosseinsarraf@alumni.ucalgary.ca

Abstract

Sentiment analysis, also known as opinion mining, is a pivotal aspect of natural language processing (NLP). This method entails discerning the polarity of textual information and determining whether it conveys positive or negative sentiments. In one of the domains, e-commerce, sentiment analysis assumes paramount significance. It offers businesses a nuanced understanding of their brand and product sentiment as reflected in customer reviews, facilitating market comprehension and strategic decision-making. This study primarily focused on analyzing the Amazon food reviews dataset, augmenting the original dataset with newly generated data, and subsequently conducting data preprocessing tasks, encompassing text cleansing, removing stop words, lemmatization, and stemming. Subsequently, machine learning models were constructed, trained, and evaluated using NLP feature extraction techniques to address the sentiment analysis challenge and investigate the impact of increased data volume on model performance. Among the diverse methodologies employed for extracting features from textual data samples, this research integrated term frequency-inverse document frequency (TF-IDF), Word to Vector (W2V), and Bag of Words (BoW) techniques in the feature extraction phase. Furthermore, three distinct machine learning models, namely Logistic Regression, Decision Tree, and Random Forest, were designed, implemented, and assessed. The models' performance was scrutinized following hyperparameter optimization to determine the most effective approach. The outcomes revealed that the performance of the models was consistent, yielding accuracy rates ranging from 85% to 89% on the testing dataset. Nevertheless, the Logistic Regression model, employing BoW features, demonstrated superior performance compared to the other models. Following optimization of the logistic regression model, a remarkable accuracy of 89% was attained on the testing dataset by operating the BoW extracted features.

Keywords: Machine Learning; NLP; Sentiment Analysis.

Received: *1/26/2024*
Accepted: *3/26/2024*
Published: *4/5/2024*

---

* Corresponding author.

1. Introduction

The classification of sentiment in consumer reviews represents a burgeoning field within NLP on a global scale. Sentiment analysis, or opinion mining, is a computational technique to discern whether a written piece expresses positive or negative sentiment [1]. Our research offers a valuable resource for restaurants of varying sizes and market presence, enabling them to gain deeper insights into reviewers' sentiments regarding their products. Furthermore, it can be leveraged for diverse tasks beyond sentiment analysis, including the development of recommender systems [2]. The primary challenge in sentiment analysis classification revolves around categorizing sentiment polarity. This challenge entails assigning a specific sentiment label, such as positive or negative, to a given written text. Sentiment polarity categorization operates across three distinct levels based on the scope of the text: document-level, sentence-level, and entity and aspect-level categories [3]. At the document level, the focus is determining whether the document expresses a positive or negative sentiment. Sentence-level analysis involves classifying individual sentences within the text and assessing the sentiment conveyed in each sentence independently.

Conversely, entity and aspect-level analysis identify specific aspects or entities within the text that elicit positive or negative opinions from individuals. This study gathered an additional 2000 reviews from alternative food-ordering platforms to augment the original dataset [4]. Following data acquisition, samples were labeled positive or negative based on pertinent keywords in the dataset's text reviews. Discrepancies in labeling were resolved by majority voting from third parties, resulting in the selection of final labels for the records. Data generation procedures ensured alignment of key columns, including text, summary, and data type, among the original dataset. Text preprocessing, a critical step in NLP, was applied utilizing Python libraries such as NLTK [5], string, and regex to enhance data quality and improve machine learning algorithms' performance [6]. Initial preprocessing involved the removal of numbers and non-informative sentence markers, followed by word tokenization as per the problem statement. Subsequently, stop word removal using NLTK [5] eliminated commonly occurring but insignificant words, while punctuation removal was implemented using the string library. Lemmatization was then applied to standardize words to their base forms (lemmas), followed by stemming to reduce words to their root forms. Feature engineering was conducted using TF-IDF, W2V, and BoW to extract features from the raw data. Three classifiers - Logistic Regression, Decision Tree, and Random Forest - were trained using the extracted features.

Performance comparisons with the original paper revealed that Logistic Regression achieved consistent performance with TF-IDF and BoW features while achieving the highest accuracy with W2V. Decision Tree surpassed all three features from the original paper, while Random Forest's performance with TF-IDF and BoW decreased compared to the original dataset results. Hyperparameter tuning was performed to optimize all models, resulting in similar performance across the board [7]. However, the test accuracy of the Decision Tree classifier marginally decreased, while the accuracy of Logistic Regression using W2V slightly increased [8]. Logistic Regression with BoW demonstrated the best performance among all classifiers, achieving an accuracy of 89%.

Nevertheless, it misclassified some reviews, prompting an investigation into the reasons for misclassification [9]. Analysis of 200 randomly selected misclassified reviews revealed that 118 reviews contained an equal number of

positive and negative keywords, making them ambiguous [10]. Additionally, 35 records exhibited more positive keywords despite naturally conveying negative sentiments, while 47 reviews conveyed positive meaning despite containing more negative keywords.

2. Data Generation

The dataset utilized in the primary study comprises 560,000 review entries sourced from the Amazon food platform captured from 1999 to 2012. Each entry encompasses various fields: summary, textual review, order timestamp, and review timestamp. These reviews were collected from around 250,000 users and about 74,000 products. Reviews include product and user information, ratings, and a plain text review (Table 1). To expand this dataset by 2000 additional entries, a decision was made to manually procure records bearing comparable information, particularly regarding summary and textual content. Diverse online platforms such as DoorDash, Skip the Dishes, and Uber Eats reviews were employed to collect and synthesize new sample entries shown in Table 2 [11]. A total of 2000 newly generated food reviews underwent a thorough examination against the entirety of the dataset to eliminate any instances of duplication [12]. Subsequently, the entire dataset was labeled based on the content of both the summary and text of the reviews.

It should be noted that keywords pivotal to the reviews carrying significant semantic value were considered during the labeling process. For instance, phrases such as "will not buy again," "not recommended," and "waste of money" were categorized as indicative of negative reviews. In contrast, expressions like "good products," "recommend it," and "buy again" were construed as positive indicators. Consequently, reviews with conflicting labels were juxtaposed against the extracted keywords, and a decision was reached to assign labels based on the predominant meaningful words present in the text [13].

Table 1: Summary Statistics of the Original Data

Features	Definition	Data Distribution
Id	Id	~560,000
ProductId	Unique identifier for the product	~74,000 unique values
UserId	Unqiue identifier for the user	~256,000 unique values
Score	Rating between 1 and 5	1(52k),2(29k),3(42k),4(80k),5(363k)
Summary	Summary of the review	~295,000 unique values
Text	Text of the review	~393,000 unique values

Table 2: Summary Statistics of the Generated Data

Features	Definition	Data Distribution
Id	Id	2000
Summary	Summary of the review	2000
Text	Text of the review	2000

3. Data Preprocessing

Several crucial steps were undertaken to clean the data during the preprocessing phase. Initially, the regex package was utilized to eliminate numbers and punctuation marks from the text, as they do not contribute significantly to the text's overall meaning [14]. Subsequently, the texts and paragraphs within each review were tokenized into individual words using the Word-Tokenization package from Spark. Following tokenization, the word tokens were cross-checked with lists of common stop words and punctuation errors, and superfluous words such as 'at,' 'the,' and 'a - known as stop words - were removed from the text and converted to lowercase [15]. After this step, two essential techniques, namely lemmatization and stemming, were employed across the entire dataset to ensure the extraction of the most accurate and clean text possible, thereby facilitating the extraction of better features from the data. Table 3 provides an overview of the steps involved in data preprocessing [16].

Table 3: Summary statistics of the data preprocessing

Data Preprocessing Step	Example Input	Output
Sentence Mark Removal	Unbelievable!! 100% Great product for that price.	Unbelievable 100 Great product for that price
Number Removal	Unbelievable 100 Great product for that price	Unbelievable Great product for that price
Tokenization	Unbelievable Great product for that price	‘Unbelievable’ ‘Great’ ‘product’ ‘for’ ‘that’ ‘price’
Stop Word Removal	‘Unbelievable’ ‘Great’ ‘product’ ‘for’ ‘that’ ‘price’	‘Unbelievable’ ‘Great’ ‘product’ ‘price’
Punctuation Removal	‘Unbelievable’ ‘Great’ ‘product’ ‘price’	‘unbelievable’ ‘great’ ‘product’ ‘price’
Lemmatization	‘unbelievable’ ‘great’ ‘product’ ‘price’	‘unbelieve’ ‘great ‘produce’ ‘price’
Stemming	‘unbelieve’ ‘great ‘produce’ ‘price’	‘unbeliev’ ‘grate‘‘produc’ ‘pric’

4. Feature Extraction

Three distinct techniques, Term Frequency-Inverse Document Frequency (TF-IDF), W2V, and BoW, were employed on the preprocessed text data to extract features for model training. Consistent with the original paper, a set number of features, namely 1000, were utilized for all feature extraction processes [17]. We recorded train accuracy, test accuracy, precision, recall, and F1-score to assess and compare the performance of the models after extracting features followed by the training phase.

5. Results

After comparing the results of our best models with the performance metrics outlined in the original paper, several observations were made, focusing on F1-score and test accuracy. Logistic regression models utilizing TF-IDF and BoW features achieved identical performances, with 88% and 90% test accuracies. Notably, the Word2Vec features surpassed the original paper's reported performance, achieving 88% accuracy on our test dataset compared to the previous 60%. Conversely, the original paper's decision tree model surpassed all three feature

types. While TF-IDF and BoW models exhibited a modest improvement of 3% in accuracy, the W2V feature model demonstrated a significant enhancement, achieving a 15% improvement. However, overfitting was observed with TF-IDF and BoW features for random forest models, resulting in a 3% decrease in performance compared to the original paper's results. Nevertheless, random forest models still outperformed the W2V feature model, as the former achieved 70% and 85% accuracy on the test dataset in the previous study. Table 4 presents the performance of models trained on the original data, while Table 5 illustrates the performance of models with default parameters trained on the extended data.

Table 4: Original Data Model Performance for Various Evaluation Metrics

Model	TF-IDF	W2V	BoW
Logistic Regression	Test accuracy: 0.88	Test accuracy: 0.57	Test accuracy: 0.90
	Train accuracy: 1	Train accuracy: 0.85	Train accuracy: 1
Decision Tree	Test accuracy: 0.83	Test accuracy: 0.88	Test accuracy: 0.83
	Train accuracy: 0.92	Train accuracy: 0.73	Train accuracy: 0.85
Random Forest	Test accuracy: 0.88	Test accuracy: 0.88	Test accuracy: 0.90
	Train accuracy:0.95	Train accuracy: 0.90	Train accuracy: 0.94

Table 5: Extended Data Model Performance for Various Evaluation Metrics

Model	TF-IDF	W2V	BoW
Logistic Regression	Test accuracy: 0.88	Test accuracy: 0.88	Test accuracy: 0.89
	F1-Score: 0.87	F1-Score: 0.87	F1-Score: 0.88
	Precision: 0.86	Precision: 0.86	Precision: 0.88
	Recall: 0.88	Recall: 0.88	Recall: 0.89
	Train accuracy: 0.89	Train accuracy: 0.88	Train accuracy: 0.90
Decision Tree	Test accuracy: 0.85	Test accuracy: 0.86	Test accuracy: 0.86
	F1-Score: 0.83	F1-Score: 0.85	F1-Score: 0.84
	Precision: 0.82	Precision: 0.83	Precision: 0.83
	Recall: 0.85	Recall: 0.86	Recall: 0.86
	Train accuracy: 0.86	Train accuracy: 0.86	Train accuracy: 0.86
Random Forest	Test accuracy: 0.85	Test accuracy: 0.85	Test accuracy: 0.85
	F1-Score: 0.78	F1-Score: 0.85	F1-Score: 0.86
	Precision: 0.72	Precision: 0.85	Precision: 0.87
	Recall: 0.85	Recall: 0.85	Recall: 0.85
	Train accuracy: 0.85	Train accuracy: 0.85	Train accuracy: 0.85

5.1. Tuning Phase

During this stage, hyperparameter tuning was conducted for each model to determine if the highest performance

could be achieved with optimal parameters. Table 6 enumerates the parameters utilized during the tuning phase for each model. Following parameter tuning, the results closely resembled those obtained with the default parameters for random forest and logistic regression models. However, the test accuracy of the decision tree model exhibited a decrease. Table 7 provides an overview of the models' performance after hyperparameter tuning, while Table 8 outlines the best parameters determined through hyperparameter tuning.

Table 6: Selected Parameters for Each Classifier

Classifiers	Parameters
Logistic Regression	Regression Parameters: [0, 0.02, 0.08]
	Max Iteration: [10, 20]
	ElasticNet Parameters: [0.2, 0.6, 0.8]
Decision Tree	Max Depth: [2, 10, 20, 30]
	Max Bin: [10, 20, 40, 80]
Random Forest	Number of Trees: [5, 15, 20]

Table 7: Models' Performance After Hyperparameter Tuning

Models	TF-IDF	W2V	BOW
Logistic Regression	Test accuracy: 0.88	Test accuracy: 0.88	Test accuracy: 0.89
	F1-Score: 0.86	F1-Score: 0.87	F1-Score: 0.88
	Precision: 0.86	Precision: 0.87	Precision: 0.88
	Recall: 0.88	Recall: 0.88	Recall: 0.89
	Train accuracy: 0.89	Train accuracy: 0.88	Train accuracy: 0.90
Decision Tree	Test accuracy: 0.84	Test accuracy: 0.84	Test accuracy: 0.85
	F1-Score: 0.83	F1-Score: 0.83	F1-Score: 0.84
	Precision: 0.82	Precision: 083	Precision: 0.83
	Recall: 0.84	Recall: 0.84	Recall: 0.85
	Train accuracy: 0.94	Train accuracy: 0.93	Train accuracy: 0.93
Random Forest	Test accuracy: 0.85	Test accuracy: 0.85	Test accuracy: 0.85
	F1-Score: 0.78	F1-Score: 0.85	F1-Score: 0.86
	Precision: 0.72	Precision: 0.84	Precision: 0.87
	Recall: 0.85	Recall: 0.85	Recall: 0.85
	Train accuracy: 0.85	Train accuracy: 0.85	Train accuracy: 0.85

Table 8: Models' Parameters After Hyperparameter Tuning

Models	TF-IDF	W2V	BOW
Logistic Regression	Regression Parameters: 0 Max Iteration: 10 ElasticNet Parameters: 0.2	Regression Parameters: 0 Max Iteration: 20 ElasticNet Parameters: 0.2	Regression Parameters: 0 Max Iteration: 10 ElasticNet Parameters: 0.2
Decision Tree	Max Depth: 30 Max Bin: 10	Max Depth: 10 Max Bin: 10	Max Depth: 30 Max Bin: 10
Random Forest	Number of Trees: 20	Number of Trees: 20	Number of Trees: 20

Considering the preceding sections and tables, the logistic regression model utilizing the BoW features exhibited superior performance compared to other models, achieving an accuracy of 89% on the test dataset. To elucidate the distribution of misclassified labels and discern the areas where the model above faltered in label prediction, a subset of 200 randomly selected misclassified data records about the BoW feature was analyzed. Three primary rationales were identified to elucidate the model's failure: First, the instances where the reviews contained an equal frequency of positive and negative keywords led to ambiguity in the model's class assignment. Secondly, cases where the count of positive keywords exceeded that of negative ones, contrary to the actual negative sentiment expressed in the review, contributed to misclassification. Thirdly, Instances where the sentences exhibited a naturally positive tone but incorporated a higher prevalence of negative keywords also contributed to misclassification.

6. Discussion

In the contemporary era of globalization, sentiment analysis presents itself as a potent tool for addressing myriad real-world challenges [18]. Among these challenges, Customer Feedback Analysis stands out prominently. Traditionally, customer feedback is often distilled into numerical scores, with stakeholders relying on average scores to gauge product performance [19]. However, such an approach encounters complexities when the textual feedback accompanying these scores diverges from the numerical rating. For instance, a customer might assign a moderate score of '6' out of 10 while appending commentary such as 'The product is decent,' thus introducing a discordance between the numerical rating and the qualitative assessment [20]. The solution advanced within this study offers a remedy to this discordance, achieving an 89% accuracy in accurately classifying sentiments expressed within reviews [21].Furthermore, our proposed models offer a pathway to address another significant real-world challenge: discerning market trends. Market trends are pivotal in discerning the trajectory of business sectors and industries, enabling stakeholders to anticipate shifts in consumer preferences and identify businesses at risk of attrition. Leveraging the predictive capabilities of our models, we can extrapolate insights from customer reviews to ascertain prevailing market sentiments towards specific products. By aggregating and analyzing the sentiment polarity of reviews, we can discern emerging trends, thereby forecasting potential upticks or downturns in the popularity of products. This proactive approach empowers businesses to adapt strategies by evolving market dynamics, enhancing their competitive resilience.

7. Conclusion

The core objective of this project was to extend the findings of a previously selected journal paper titled "Machine Learning Model for Classifying L_Text Using NLP (Amazon and his colleagues.)." A dataset comprising 2000 new records was generated and independently labeled to achieve this. Furthermore, oversampling techniques were employed to address imbalances in the dataset to ensure an equitable distribution of target values. The text data underwent preprocessing using the Natural Language Toolkit (NLTK) library to facilitate model training. Feature engineering methodologies were implemented to extract relevant features from the dataset, including Term Frequency-Inverse Document Frequency (TF-IDF), W2V, and BoW (BoW). Subsequently, three classification algorithms - Random Forest, Logistic Regression, and Decision Tree - were chosen for model training and evaluation. Following model evaluation on the test dataset, hyperparameter tuning was performed to optimize the performance of each classifier by identifying the most effective parameter configurations. The results were then juxtaposed with those reported in the original paper for comparative analysis. The Logistic Regression classifier with BoW (BoW) features emerged as this project's most effective classification model, surpassing other classifiers in performance and achieving the highest accuracy. Notably, the Decision Tree classifiers in this project demonstrated superior performance across all extracted features compared to the results reported in the original paper.

References

[1] A. Sarraf and A. Abbaspour, “ChatGPT Application In Summarizing An Evolution Of Deep Learning Techniques In Imaging: A Qualitative Study,” *arXiv Prepr. arXiv2312.03723*, 2023.

[2] A. Sarraf, M. Azhdari, and S. Sarraf, “A Comprehensive Review of Deep Learning Architectures for Computer Vision Applications,” *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 77, no. 1, pp. 1–29, 2021.

[3] S. Sarraf, “Hair color classification in face recognition using machine learning algorithms,” *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 26, no. 3, pp. 317–334, 2016.

[4] C. Prabhavathi, N. Vishali, P. S. Reddy, and J. V Chandramouli, “Machine Learning Model for Classifying L _ Text Using Nlp (Amazon Product Reviews),” *Int. Res. J. Comput. Sci.*, vol. 6, no. 4, pp. 161–178, 2019.

[5] E. Loper and S. Bird, “Nltk: The natural language toolkit,” *arXiv Prepr. cs/0205028*, 2002.

[6] S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. E. Anderson, M. Kabia, and A. D. N. Initiative, “OViTAD: Optimized vision transformer to predict various stages of Alzheimer’s disease using resting-state fMRI and structural MRI data,” *Brain Sci.*, vol. 13, no. 2, p. 260, 2023.

[7] S. Sarraf, “French Word Recognition Through a Quick Survey on Recurrent Neural Networks Using Long-Short Term Memory RNN-LSTM,” *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 39, no. 1, pp. 250–267, 2018.

[8] A. Sarraf, A. E. Jalali, and J. Ghaffari, "Recent Applications of Deep Learning Algorithms in Medical Image Analysis," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 72, no. 1, pp. 58–66, 2020.

[9] A. Sarraf, "Binary Image Classification Through an Optimal Topology for Convolutional Neural Networks," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 68, no. 1, pp. 181–192, 2020.

[10] S. Sarraf, "Binary Image Segmentation Using Classification Methods: Support Vector Machines, Artificial Neural Networks and K th Nearest Neighbours," *Int. J. Comput.*, vol. 24, no. 1, pp. 56–79, 2017.

[11] Y. Zhang, "Qualitative Analysis of DoorDash," in *2021 3rd International Conference On Economic Management And Cultural Industry (ICEMCI 2021)*, 2021, pp. 65–68.

[12] S. Sarraf, D. D. DeSouza, J. Anderson, G. Tofighi, and A. D. N. Initiativ, "DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI," *BioRxiv*, p. 70441, 2016.

[13] S. Sarraf and G. Tofighi, "Deep learning-based pipeline to recognize Alzheimer's disease using fMRI data," in *2016 Future Technologies Conference (FTC)*, 2016, pp. 816–820, doi: https://doi.org/10.1109/ftc.2016.7821697.

[14] S. Sarraf and M. Kabia, "Optimal Topology of Vision Transformer for Real-Time Video Action Recognition in an End-To-End Cloud Solution," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1320–1339, 2023.

[15] E. Moosavi-Zadeh, A. Rahimi, H. Rafiee, H. Saberipour, and R. Bahadoran, "Effects of fennel (Foeniculum vulgare) seed powder addition during early lactation on performance, milk fatty acid profile, and rumen fermentation parameters of Holstein cows," *Front. Anim. Sci.*, vol. 4, p. 1097071, 2023.

[16] S. H. Sarraf, M. Soltanieh, and H. Aghajani, "Repairing the cracks network of hard chromium electroplated layers using plasma nitriding technique," *Vacuum*, vol. 127, pp. 1–9, 2016.

[17] S. H. Sarraf, S. Rastegari, and M. Soltanieh, "Deposition of mono dispersed Co–CeO2 nanocomposite coatings by a sol-enhanced pulsed reverse electroplating: process parameters screening," *J. Mater. Res. Technol.*, vol. 23, pp. 3772–3789, 2023.

[18] S. Sarraf, D. D. Desouza, J. A. E. Anderson, and C. Saverino, "MCADNNet: Recognizing stages of cognitive impairment through efficient convolutional fMRI and MRI neural network topology models," *IEEE Access*, vol. 7, pp. 155584–155600, 2019, doi: https://doi.org/10.1109/access.2019.2949577.

[19] S. Sarraf, "Analysis and Detection of DDoS Attacks Using Machine Learning Techniques," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 66, no. 1, pp. 95–104, 2020.

[20] X. Yang, S. Sarraf, and N. Zhang, "Deep learning-based framework for Autism functional MRI image

classification," *J. Ark. Acad. Sci.*, vol. 72, no. 1, pp. 47–52, 2018.

[21] S. H. Sarraf, M. Soltanieh, and S. Rastegari, "Reactive air aluminizing of a nickel-based superalloy (IN738LC): Coating formation mechanism," *Surf. Coatings Technol.*, vol. 456, p. 129229, 2023.