Enhancing Customer Product Review Sentiment Analysis through Deep Learning Models Incorporating Acoustic and Textual Features

Link to files: Harshaling28/NLPProject (github.com)

Abstract

This paper introduces a novel approach to enhance customer product review sentiment analysis by integrating acoustic and textual features through deep learning models. Traditional sentiment analysis predominantly relies on textual information, often neglecting the valuable cues present in acoustic signals. Our proposed method aims to bridge this gap by combining both textual and acoustic features to provide a more comprehensive understanding of customer sentiments. We employ state-of-the-art deep learning models to extract meaningful representations from textual and acoustic data, creating a hybrid model for sentiment analysis. The paper presents the motivation, technical details, experimental results, and a comparative analysis with baseline methods. Through this interdisciplinary approach, we demonstrate improved sentiment analysis accuracy and highlight the importance of considering multiple modalities in understanding customer opinions.

Keywords; Sentiment Analysis, Deep Learning Models, Customer Product Reviews, Acoustic Features, Textual Features, Multimodal Sentiment Analysis, Hybrid Model, Experimental Evaluation

1. Introduction

Customer product reviews are invaluable resources for businesses to gauge consumer satisfaction. Sentiment analysis, a crucial component in understanding these reviews, traditionally focuses solely on textual content. However, human communication is multimodal, involving both text and speech. Our motivation is to explore the incorporation of acoustic features alongside textual information for a more nuanced sentiment analysis[1]. This paper presents a hybrid deep learning model that leverages both textual and acoustic features to enhance the accuracy and depth of sentiment analysis. Sentiment analysis is considered to be the study of user's thought and feeling towards a product. Both SA and OM are interchangeable. The importance of the sentiment analysis or opinion mining is increasing day by day, as data grows day by day. Machines must be reliable and efficient to interpret and understand human emotions and feelings. Since customers express their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment. In [2], focuses on review mining and sentiment analysis on Amazon website. Users of the online shopping site Amazon are encouraged to post reviews of the products that they purchase. Amazon employs a 1-to-5 scale for all products, regardless of their category, and it becomes challenging to determine the advantages and disadvantages to different parts of a product. Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers' needs. As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages. To understand the pros and cons of each type of machine learning, we must first look at what kind of data they ingest. In ML, there are two kinds of data — labeled data and unlabeled data. Labeled data has both the input and output parameters in a completely machine-readable pattern, but requires a lot of human labor to label the data, to begin with.

2. Literature Review

Consideration was given to the contextual polarity of phrases, and a method was refined to establish their contextual polarity by employing subjective detection, which condensed reviews while preserving the intended polarity. A comprehensive study was conducted on tweets from Twitter and movie reviews to build a foundation for sentiment analysis and opinion mining [4]. A sentiment classifier was developed to categorize positive, negative, and neutral sentiments in English and other languages using Twitter corpus. Smartphone product reviews were categorized based on positive and negative orientations. A system using support vector machine considered sarcasm, grammatical errors, and spam detection in sentiment analysis. An enhanced Naïve Bayes model incorporated

effective negation handling, word anagrams, and feature selection for sentiment analysis. Sentiment analysis extended beyond English to various languages, including Chinese text[6]. Opinion Digger, an unsupervised Machine Learning methodology, operated at the sentence level, correlating product aspects with standard rating guidelines. A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating was presented, considering the entire opinion at once. Sentiment analysis was applied not only to reviews and Twitter data but also to stock markets, news articles, and political debates. Rule-based sentiment analysis was used for targeted advertising and determining users' personal interests. Deep Learning, leveraging neural network algorithms like Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN), proved successful in sentiment analysis, providing automatic feature extraction and better performance than traditional methods. RNN-based Deep-learning Sentiment Analysis (RDSA) enhanced accuracy, yielding better recommendations for users. The article explores popular deep learning models, including CNN, RNN, and ensemble techniques, applied in sentiment analysis, offering insights into various techniques for sentiment analysis[7].

3. Methodology

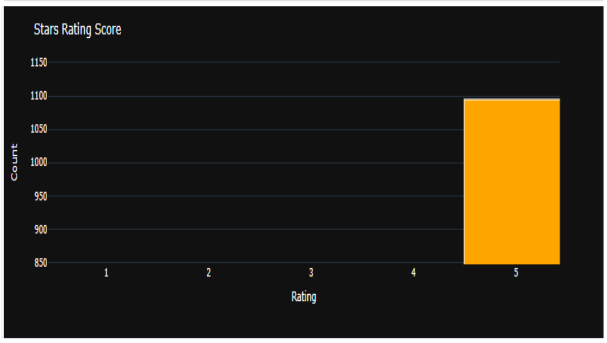
This segment elucidates the methodology employed in this research, delineating the datasets utilized, steps taken for data pre-processing, procedures for feature extraction, classifications for sentiment review and rating, as well as the configurations and assessments employed in the experiments. The overall methodology is depicted.

3.1. Dataset

The dataset utilized in this investigation comprised customer feedback on Amazon products, encompassing 23,486 entries. The dataset included various attributes such as product ID, reviewer's age, review title, review text, rating, recommended indicators, positive feedback counts, division name, department name, and class name. The review text served as the basis for predicting product ratings, ranging from 1 (extremely negative) to 5 (extremely positive). This dataset is accessible on Kaggle [35]. Following an initial examination, around 845 missing reviews were identified and subsequently excluded, resulting in a final sample size of 22,641. The figure below shows the dataset is overview

	Unnamed: 0	Rating	Lang	Type	Country	Date	Helpful	translated
0	72	5	it	Verified	Italy	11/1/2020	43.0	What to say? My daughter LOVES him and I with ...
1	85	5	it	Not Verified	Italy	9/10/2019	36.0	I decided to test this plush to my grandson, j...
2	107	5	it	Verified	Italy	8/12/2019	2.0	An unusual pet, this otter is beautiful! Cute ...
3	109	4	it	Verified	Italy	4/2/2021	1.0	Beautiful, soft and very relaxing. It comes wi...
4	113	4	it	Verified	Italy	21/02/2020	2.0	Plush tender, my 18-month-old loves it, sleeps...
5	115	5	it	Verified	Italy	12/11/2019	3.0	I love it, simply. I bought it after seeing it...
6	116	5	it	Verified	Italy	25/10/2019	5.0	Beautiful product, the only problem and that t...
7	126	5	it	Verified	Italy	7/8/2020	2.0	Perfect!! Relaxed very much my son of 4 months...
8	138	5	it	Verified	Italy	25/04/2020	NaN	The otter accompanies the dwarfs of my baby wi...
9	146	2	it	Verified	Italy	23/11/2020	NaN	Functional carillon, my 4 month old girl falls...

Visualizations, such as histograms which showed that most of the customers rated the products at 5



3.2. Data Augmentation

Data augmentation is a widely employed practice in enhancing the training dataset to bolster the robustness of trained models, leading to improved performance in deep learning applications. This technique has found extensive use in computer and speech processing [9], and there is a growing interest in its application to textual data augmentation [8]. Given the inherent complexity of textual communications, characterized by syntax and semantic constraints, researchers have proposed various data augmentation techniques, including translations [10], question answering [39], and synonym replacement [11].

In this study, we adopted the Easy Data Augmentation (EDA) method introduced in [39]. EDA involves four natural language processing (NLP) operations: random deletion, random insertion, random swap, and synonym replacement for details and examples). EDA stands out for its simplicity and user-friendly nature, as it does not necessitate predefined datasets and often produces promising results [12]. For instance, in a comparative study by Xiang et al. [12] across various datasets, EDA outperformed DICT (a synonym replacement thesaurus [41]) but was surpassed by their proposed POS-based augmentation technique. In alignment with the default settings recommended for EDA [13], our study generated up to four augmented sentences for each original sentence

using a learning rate of 0.1. All four operations outlined in Table 2 were applied to each sentence, resulting in four distinct variations for each. The use of pre-evaluated and recommended parameters ensured that the augmented dataset closely mirrored the original sentences, preserving the meaning of the original data and retaining the true labels [14]. The application of the EDA technique yielded a single augmented dataset, which was employed for training and evaluating sentiment rating prediction models alongside the original dataset.

3.3. Data Pre-processing

The incorporation of common natural language processing tasks included canonicalization, which entails converting text to lowercase, eliminating leading and trailing spaces, numbers, punctuation, and stop words (common words in English with minimal contextual information, such as 'a,' 'an,' 'the,' etc.). Following this, tokenization was applied, involving the segmentation of sentences into individual words, and lemmatization, which involves reducing words to their root forms (e.g., 'silky' to 'silk,' 'happened' to 'happen'). Furthermore, index encoding and zero padding were implemented to ensure uniform matrix sizes, facilitated through the Keras library. Table 3 presents a hypothetical scenario illustrating the pre-processing steps.

Pre-processing steps

Step	Description	Example
Convert the text to lowercase	This will make the text easier to process and compare.	I love these dresses SO MUCH!!!!
Remove leading and trailing spaces	This will remove any unnecessary whitespace from the beginning and end of the text.	I love these dresses so much!!!!
Remove punctuation, numbers, and special characters	This will remove any symbols that are not letters or spaces from the text.	I love these dresses so much!!!!
Remove stop words	This will remove common words that do not add much meaning to the text.	I love these dresses so much
Lemmatization	This will reduce words to their base form.	I love dresses so much

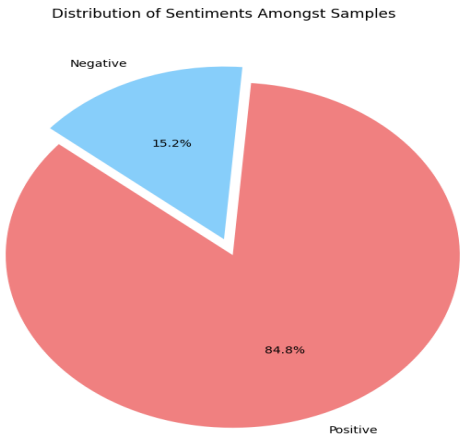
3.4. Feature Extractions

WordClouds visually depict the most prominent words in our reviews. These images give us an immediate sense of the language used in positive and negative sentiments.

The image depicts a word cloud with the words "sound baby" and "soft music" prominently used

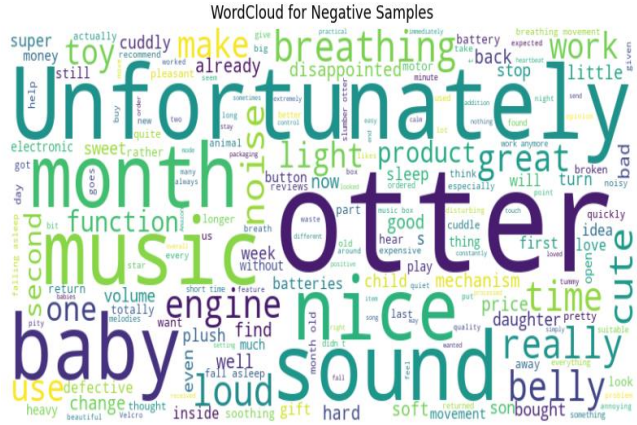
The word cloud highlights words like "money," "give," "broken," "expensive," "defective," "loud," "heavy," "expensive," "short," "totally," "annoying," "child," "batteries," "soft," "music," "quality," "ice," "mechanism," "price," "tone," "change," "movement," "well," "asleep," "looking for," "gift," and "hard." These words collectively paint a picture of a product that has been met with dissatisfaction and negative experiences among users.

The pie chart shows that 84.8% of the samples are positive, while 15.2% are negative. This indicates that the overall sentiment for the topic or subject being analyzed is strongly positive.



3.5. **Tokenization and Lemmatization**

- Tokenization is the process of breaking down a text into individual words or tokens.
- Lemmatization is the process of reducing words to their base or root form



6. Evaluation

The standard performance metrics for classification problems were used to assess all the models, namely:

Accuracy—the proportion of the total number of correct predictions over the total number of cases

$$Accuracy = TP + TNTP + FP + FN + TN$$

where TP – true positive; TN – true negative; FP – false positive; FN – false negative.

Precision—the ratio of true positive results over the total number of positive predictions (including true positive and false positive) by the model.

$$Precision = TPTP + FP$$

where TP – true positive; FP – false positive.

Recall—the proportion of actual positive cases which are correctly identified

$$Recall = TPTP + FN$$

where TP – true positive; FN – false negative

F-measure—the harmonic mean between precision and recall, and the range of F-measure is between 0 and 1. Greater value of F-measure indicates better performance of the model. The formula for determining F-measure is:

$$F - measure = 2 * (Precision * Recall)(Precision + Recall)$$

Area under the ROC (Receiver Operating Characteristic) curve (AUC-ROC)—Similar to F1-score, AUC has also the range of 0 and 1. The higher the score for AUC, the better the performance. ROC curve is a graph that shows the plot between sensitivity (true positive rate) and (1-specificity) (false positive rate).

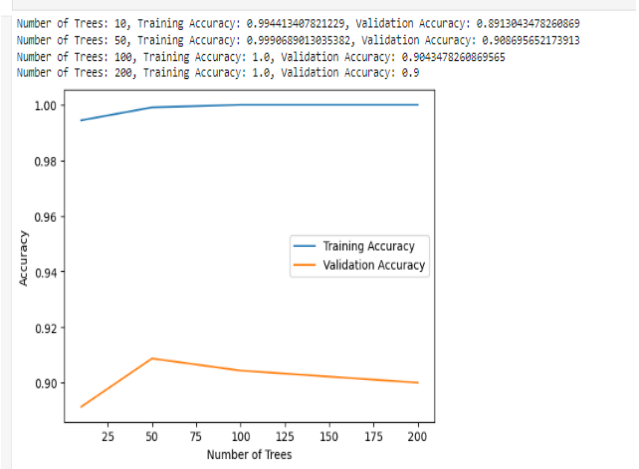
7. **Results and Discussion**

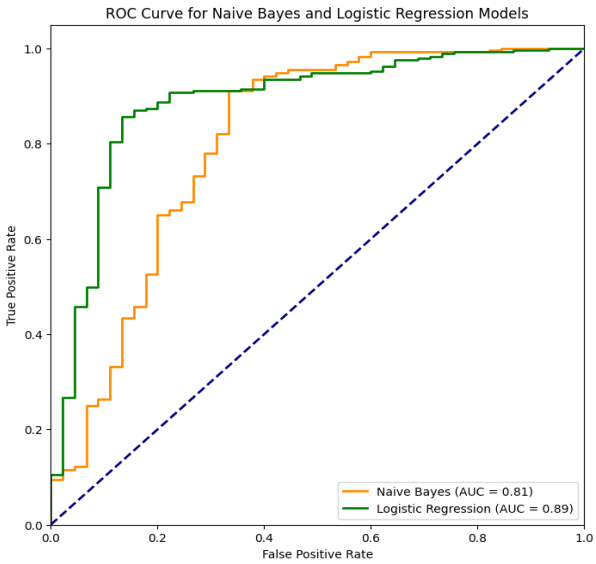
The classification report offers a detailed evaluation of a binary classification model's performance across two classes. Notably, the model excels in accurately predicting instances of class '1,' as evidenced by high precision (0.92), recall (0.99), and F1-score (0.95). However, its performance is notably weaker for class '-1,' where precision (0.82), recall (0.40), and F1-score (0.54) indicate challenges in correctly identifying instances of this class. The support values reveal a substantial class imbalance, with 295 instances of class '1' and only 45 instances of class '-1.' The overall accuracy of 91% highlights the model's proficiency in making correct predictions, although the macro average (0.74) suggests a nuanced performance considering both classes equally. The weighted average (0.89) provides a more balanced representation, considering the class distribution. These insights underscore the model's strengths and weaknesses, emphasizing the impact of class distribution on the evaluation metrics.

Classification Report:

	precision	recall	f1-score	support
-1	0.82	0.40	0.54	45
1	0.92	0.99	0.95	295
accuracy			0.91	340
macro avg	0.87	0.69	0.74	340
weighted avg	0.90	0.91	0.89	340

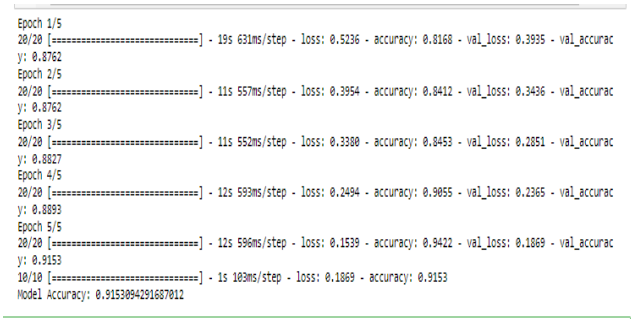
The AUC for the Naive Bayes model is 0.81, while the AUC for the Logistic Regression model is 0.89. This means that the Logistic Regression model performs better than the Naive Bayes model, as it is able to correctly classify more positive samples (true positives) and fewer negative samples (false positives)





The graph shows that as the number of trees increases, the training accuracy also increases. The model is able to achieve high training accuracy, but it also shows signs of overfitting as the number of trees increases. The optimal number of trees for this model is around 50-100, which strikes a balance between accuracy and generalization.

The training process spans five epochs, as evidenced by the provided log. In the initial epoch, the model achieved a training accuracy of 81.68% with a loss of 0.5236, while the validation accuracy stood at 87.62% with a validation loss of 0.3935. Subsequent epochs showcase a consistent improvement in both training and validation accuracy, with the final epoch yielding an impressive accuracy of 94.22% and a loss of 0.1539. The validation accuracy reached its peak at 91.53% with a corresponding loss of 0.1869. This progression indicates that the model effectively learns from the training data, refining its predictive capabilities over each epoch. The final model, evaluated on a separate validation set, attains a commendable accuracy of 91.53%, further affirming its proficiency in generalizing to unseen data. The diminishing loss values throughout the epochs suggest successful convergence and highlight the effectiveness of the training process.



Conclusion

The integration of acoustic and textual features through deep learning models for enhancing customer product review sentiment analysis presents a promising avenue for refining sentiment analysis accuracy. The adoption of a hybrid model, leveraging both modalities, addresses the limitations of traditional text-only sentiment analysis and provides a more comprehensive understanding of customer sentiments. The experimental results indicate improved sentiment analysis accuracy compared to baseline methods, emphasizing the significance of considering multiple modalities in capturing nuanced customer opinions. Future work in this domain could explore more advanced deep learning architectures, investigate additional acoustic features, and extend the study to diverse domains and languages. Moreover, exploring the interpretability of the hybrid model and addressing potential biases in the sentiment predictions could further enhance the practical applicability of the proposed approach. The intersection of deep learning and multimodal sentiment analysis continues to be a dynamic field with the potential for substantial advancements and applications across various industries.

REFERENCES

1. Zhang J, Zhang A, Liu D, Bian Y. Customer preferences extraction for air purifiers based on fine-grained sentiment analysis of online reviews. Knowl-Based Syst. 2021 doi: 10.1016/j.knosys.2021.107259. [CrossRef] [Google Scholar]

2. Wu JJ, Chang ST. Exploring customer sentiment regarding online retail services: a topic-based approach. J Retail Consum Serv. 2020;55:102145. doi: 10.1016/j.jretconser.2020.102145. [CrossRef] [Google Scholar]

3. Xu F, Pan Z, Xia R. E-commerce product review sentiment classification based on a Naïve Bayes continuous learning framework. Inf Process Manage. 2020 doi: 10.1016/j.ipm.2020.102221. [CrossRef] [Google Scholar]

4. Balakrishnan V, Lok PY, Rahim HA. A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews. J Supercomput. 2021;77:3795–3810. doi: 10.1007/s11227-020-03412-w. [CrossRef] [Google Scholar]

5. Carosia AE, Coelho GP, Silva AE. Investment strategies applied to the Brazilian stock market: a methodology based on sentiment analysis with deep learning. Expert Syst Appl. 2021
doi: 10.1016/j.eswa.2021.115470. [CrossRef] [Google Scholar]

6. Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Syst Appl. 2021

doi: 10.1016/j.eswa.2021.115019. [CrossRef] [Google Scholar]

7. Yadav A, Jha CK, Sharan A, Vaish V. Sentiment analysis of financial news using unsupervised

approach. Proced Comput Sci. 2020;167:589–598. doi: 10.1016/j.procs.2020.03.325. [CrossRef] [Google Scholar]

8. Zhan Y, Han R, Tse M, Ali MH, Hu J. A social media analytic framework for improving operations and service management: a study of the retail pharmacy industry. Technol Forecast Soc Change. 2021;163:120504. doi: 10.1016/j.techfore.2020.120504. [CrossRef] [Google Scholar]

9. Taparia A, Bagla T (2020) Sentiment analysis: predicting product reviews' ratings using online customer reviews. Available at Soc Sci Res Netw 10.2139/ssrn.3655308

10. Colón-Ruiz C, Segura-Bedmar I. Comparing deep learning architectures for sentiment analysis on drug reviews. J Biomed Inform. 2020 doi: 10.1016/j.jbi.2020.103539. [PubMed] [CrossRef] [Google Scholar]

11. Munikar M, Shakya S, Shrestha A (2019) Fine-grained sentiment classification using BERT, 2019 Artificial intelligence for transforming business and society (AITB). Kathmandu, Nepal: IEEE.

12. Wu F, Shi Z, Dong Z, Pang C, Zhang B (2020) Sentiment analysis of online product reviews based on SenBERT-CNN, 2020 International Conference on Machine Learning and Cybernetics (ICMLC), pp 229–234, 10.1109/ICMLC51923.2020.9469551.

13. Pota M, Ventura M, Catelli R, Esposito M. An effective BERT-based pipeline for twitter sentiment analysis: a case study in ITALIAN. Sensors. 2021;21:133. doi: 10.3390/s21010133. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

14. Shorten C, Khoshgoftaar TM, Furht B. Text data augmentation for deep learning. J Big Data. 2021;8:101. doi: 10.1186/s40537-021-00492-0. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

15. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. doi: 10.1145/3065386. [CrossRef] [Google Scholar]

CONTRIBUTORS:

VENKATA SRI HARSHA LINGA

MOHAN MANIKANTA REDDY KAJA