International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST52310672

OPEN-AMZPRE : Optimized Preprocessing with Ensemble Classification for Amazon Product Reviews Sentiment Prediction

Prof. Aparna Hote [*1], Dr. Dev Ras Pandey[2]
[*1]Department of Computer Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India
[2]Department of Computer Science & Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India

ARTICLEINFO

Article History:
Accepted: 01 Dec 2023
Published: 20 Dec 2023

Publication Issue
Volume 10, Issue 6
November-December-2023
Page Number
385-401

ABSTRACT

Customer feedback plays a vital role in helping consumers make informed purchasing decisions. Understanding customer opinions and preferences through sentiment analysis is crucial. However, existing sentiment analysis methods face challenges when dealing with noisy, unstructured text data, leading to limitations in accuracy, precision, recall, and F1-score. To address these limitations, this paper introduces OPEN-AMZPRE, an innovative solution for sentiment prediction in Amazon Product Reviews. Unlike traditional approaches that rely on standard techniques like tokenization and stopword removal, OPEN-AMZPRE utilizes a comprehensive preprocessing pipeline. This pipeline includes various steps such as text normalization, lowercasing, handling rare words, expanding contractions, removing HTML tags, tokenization, removing stopwords, replacing slang words, removing digits, stemming, lemmatization, punctuation and special character removal, white space removal, spell checking and correction, and removal of duplicate text. Additionally, the proposed algorithm employs an ensemble classification approach by combining optimized versions of K-Nearest Neighbors (KNN), Naive Bayes, J48 (C4.5 decision tree), and Random Forest classifiers. The hyperparameters of each classifier are tuned to achieve optimal accuracy and performance. By combining the outputs of these classifiers, the algorithm produces robust sentiment predictions. The methods of accuracy, precision, recall, and F1-score are utilized to improve sentiment prediction and provide valuable insights for both consumers and businesses in relation to Amazon Product Reviews.

Keywords: Sentiment Analysis, Amazon Product Reviews, Ensemble Classification, Optimized Preprocessing, Hyperparameter Optimization

I. INTRODUCTION

Amazon, a leading global e-commerce platform, offers a wide range of products. The Amazon Product Reviews are a valuable resource for understanding consumer opinions and sentiments. These reviews play a crucial role in assessing product quality, usefulness, and customer satisfaction, benefiting both consumers and businesses.

Although existing studies have made progress in extracting valuable insights from Amazon Product Reviews, they face inherent limitations that hinder their effectiveness. The challenges stem from the unstructured nature of the text data, which includes informal language, misspellings, contractions, and other forms of textual noise. Consequently, current methods struggle to handle these issues, leading to less-than-optimal results in sentiment prediction accuracy, precision, recall, and F1-score.

To address these limitations, this paper introduces the "Optimized Preprocessing with Ensemble Classification for Amazon Product Reviews Sentiment Prediction" algorithm, also known as OPEN-AMZPRE. This algorithm offers a comprehensive solution by effectively preprocessing the Amazon Product Reviews data and harnessing the power of ensemble classification to enhance sentiment prediction accuracy.

The OPEN-AMZPRE algorithm comprises several essential components to provide a comprehensive sentiment analysis tool. It initiates with meticulous text preprocessing, tackling informal language, spelling errors, and textual inconsistencies.

In the subsequent steps, sentiment predictions are generated by utilizing a combination of optimized classifiers such as K-Nearest Neighbors, Naive Bayes, J48 (C4.5 decision tree), and Random Forest. To achieve the best possible results, these classifiers are further improved through hyperparameter optimization.

This research paper introduces several noteworthy contributions. Firstly, it presents a unique preprocessing pipeline that specifically addresses the challenges faced in analyzing Amazon Product Reviews. This pipeline effectively enhances the quality and consistency of the data. Secondly, by incorporating ensemble classification with optimized classifiers, the accuracy of sentiment predictions is significantly improved. Moreover, the paper thoroughly evaluates these methods and provides compelling evidence of OPEN-AMZPRE's exceptional performance across various metrics, including accuracy, precision, recall, and F1-score. Lastly, it highlights the extensive application potential of OPEN-AMZPRE, showcasing its value for both consumers and businesses in product evaluation and decision-making processes.

OPEN-AMZPRE algorithm's novelty lies in its combination of intricate preprocessing steps specifically tailored to address the unique challenges posed by Amazon Product Reviews dataset, including the standardization of abbreviations and acronyms, expansion of contractions, removal of rare words and slang, spell checking and correction, and the utilization of an ensemble classification approach featuring optimized KNN, Naive Bayes, J48 (C4.5 decision tree), and Random Forest classifiers with hyperparameter optimization. This holistic approach ensures data quality and uniformity. It enhances sentiment prediction accuracy, making OPEN-AMZPRE a valuable tool for consumers and businesses seeking to make informed product evaluations and decisions based on Amazon Product Reviews.

The purpose of this paper is to showcase the effectiveness of a novel algorithm in improving the accuracy of sentiment analysis. By implementing a comprehensive preprocessing pipeline and an ensemble classification approach with optimized classifiers, the algorithm enhances data quality and sentiment prediction accuracy. This, in turn, contributes to a better understanding of consumer sentiment and supports more informed product evaluations and decision-making.

The benefits of the OPEN-AMZPRE algorithm extend beyond sentiment analysis of Amazon Product Reviews. It can be applied to various applications such as e-commerce, market research, and consumer insights. Businesses can use it to make informed decisions regarding product development, marketing, and customer satisfaction improvement. On the other hand, consumers benefit from more accurate and insightful product evaluations when making purchasing decisions.

The paper is organized as follows: Section 2 provides an overview of previous research on sentiment analysis of Amazon Product Reviews. Section 3 dives into the methodology of the OPEN-AMZPRE algorithm, explaining its preprocessing steps and ensemble classification approach in detail. Section 4 presents the experimental results, highlighting the algorithm's superior performance compared to existing methods. Lastly, Section 5 concludes the paper and suggests future research directions in sentiment analysis for Amazon Product Reviews.

II. RELATED WORKS

Amazon Product Reviews sentiment analysis has garnered substantial attention in recent years, with various approaches and techniques explored by researchers. This section provides an overview of related works in the field, highlighting their contributions and limitations.

Wassan et al. [6] employed machine learning techniques to conduct Amazon Product Sentiment Analysis. Their work significantly contributed to the field, laying the foundation for using machine learning for sentiment analysis of Amazon product reviews. However, their research faced challenges related to data preprocessing. The accuracy of sentiment prediction is highly dependent on the quality of the input data, and their work highlighted the importance of addressing data preprocessing issues to improve the overall performance of sentiment analysis.

Using machine learning, Nandal et al. [7] focused on aspect-level sentiment analysis for Amazon products. They delved into a specific aspect of sentiment analysis, which is crucial for providing detailed insights into the different facets of a product's performance. However, their approach also encountered comprehensive data preprocessing and ensemble classification issues. While they addressed specific aspects of the analysis, the holistic improvement of the preprocessing pipeline and the combination of classifiers remained unexplored.

Alharbi et al. [8] evaluated sentiment analysis using word embeddings and recurrent neural network (RNN) variants. RNNs are known for their ability to capture contextual information in text data, and their work showcased the potential of neural network-based approaches. Their performance could be further enhanced by optimizing preprocessing techniques and integrating other classifiers. The need for a more robust ensemble classification method was evident.

Geetha and Renuka [9] improved aspect-based sentiment analysis using a fine-tuned BERT model. Their approach leveraged state-of-the-art deep learning techniques to achieve better results in aspect-based sentiment analysis. However, the algorithm's reliance on a single model can limit its adaptability to varying review structures and nuances, and it may not fully address the challenges posed by noisy or unstructured data.

Budhi et al. [10] introduced a framework for comparative analysis using machine learning. Their work provided valuable insights into comparing different aspects of products in reviews. However, it lacked a holistic preprocessing solution for unstructured Amazon Product Reviews data. The challenges related to handling unstructured data and improving the data quality for sentiment analysis remained a common theme in the field.

Rintyarna et al. [11] enhanced sentiment analysis by considering both local and global contexts, thereby acknowledging the importance of contextual information in understanding sentiment. Their

approach was a step in the right direction, but there was room for further improvement in sentiment prediction accuracy. Addressing data preprocessing and combining it with a more diverse set of classifiers could help achieve higher accuracy.

Zhou et al. [12] utilized machine learning for customer needs analysis in product ecosystems. Their work contributed to understanding customer preferences and needs. However, it did not adequately address the comprehensive preprocessing needs of Amazon Product Reviews data. Robust preprocessing techniques are essential to make the most of customer data for meaningful analysis.

Dang et al. [13] proposed integrating sentiment analysis into recommender systems. Integrating sentiment analysis into recommendation systems is valuable. Still, the accuracy and reliability of sentiment analysis can be enhanced by optimizing preprocessing techniques to improve the quality of sentiment analysis data.

AlQahtani [14] focused on product sentiment analysis for Amazon reviews. The paper presented valuable insights into this specific domain. However, there was a need for further improvements in sentiment prediction accuracy. Addressing data preprocessing challenges and exploring ensemble classification methods could yield more accurate results.

Dadhich and Thankachan [15] employed a hybrid rule-based approach for sentiment analysis of Amazon Product Reviews. Their approach showed promise in combining rule-based and machine-learning techniques. However, combining these two approaches can be further optimized to improve accuracy and reduce the impact of noisy text in reviews.

Rashid and Huang [16] conducted sentiment analysis on consumer reviews of Amazon products, focusing on consumer sentiment. While their work was valuable in understanding consumer perspectives, there was an opportunity to enhance sentiment prediction accuracy by implementing a robust preprocessing pipeline and ensemble classification. These elements could lead to more accurate sentiment analysis results for consumer reviews.

Dey et al. [17] proposed and compared the performance of two popular machine learning classifiers, Linear Support Vector Machine (SVM) and Naive Bayes, for sentiment analysis on Amazon product reviews. They aimed to determine which classifier better predicts the sentiment (positive, negative, or neutral) expressed in these reviews. This research helps understand which machine learning technique is more effective for analyzing sentiments in Amazon product reviews, providing valuable insights for sentiment analysis tasks. The authors concluded Linear SVM outperformed Naive Bayes regarding accuracy, precision, recall, and F1-score.

The related works have contributed significantly to Amazon Product Reviews sentiment analysis, and each has brought unique insights and techniques to the field. However, a common theme across these works is the challenge of data preprocessing, handling noisy text, and the need for more effective ensemble classification methods. OPEN-AMZPRE aims to address these challenges by introducing a comprehensive preprocessing pipeline and leveraging ensemble classification, resulting in superior accuracy, precision, recall, and F1-score in sentiment prediction. This paper presents a novel approach to Amazon Product Reviews sentiment analysis, ensuring consumers and businesses benefit from more accurate and insightful product evaluations when purchasing decisions.

III. OPTIMIZED PREPROCESSING WITH ENSEMBLE CLASSIFICATION FOR AMAZON PRODUCT REVIEWS SENTIMENT PREDICTION (OPEN-AMZPRE)

Amazon Product Reviews play a pivotal role in influencing consumer purchasing decisions. These reviews offer a wealth of information but come with inherent challenges, including unstructured text data, informal language, and spelling errors. The

"Optimized Preprocessing with Ensemble Classification for Amazon Product Reviews Sentiment Prediction" algorithm, OPEN-AMZPRE, addresses these challenges and aims to enhance sentiment prediction accuracy for Amazon Product Reviews.

Amazon Product Reviews are a valuable source of consumer sentiment, but existing sentiment analysis methods often fail to deliver accurate and reliable results. These methods commonly lack a robust preprocessing pipeline to handle unstructured and noisy text data, leading to suboptimal sentiment prediction accuracy. OPEN-AMZPRE is needed to provide a comprehensive solution that preprocesses Amazon Product Reviews effectively and leverages ensemble classification to significantly improve sentiment prediction accuracy, precision, recall, and F1-score.

The novelty of OPEN-AMZPRE lies in its holistic approach to Amazon Product Reviews sentiment analysis. It offers a unique combination of tailored text preprocessing and ensemble classification, addressing the intricacies of unstructured text data. By optimizing each step of the preprocessing pipeline and incorporating an ensemble of classifiers, OPEN-AMZPRE aims to provide a more accurate, robust, and comprehensive solution to sentiment analysis, setting it apart from conventional methods.

The OPEN-AMZPRE algorithm offers numerous significant advantages. Firstly, it enhances sentiment prediction accuracy by tackling data preprocessing challenges and optimizing ensemble classification techniques. It achieves this through a comprehensive preprocessing approach, encompassing text normalization, lowercasing, rare word handling, contractions expansion, and more, resulting in high-quality, standardized data. OPEN-AMZPRE employs a combination of optimized K-Nearest Neighbors (KNN), Naive Bayes, J48 (C4.5 decision tree), and RandomForest classifiers, ensuring optimized sentiment classification through hyperparameter tuning. It provides robust sentiment analysis, capable of categorizing sentiments as positive, negative, or neutral, offering a comprehensive understanding of customer opinions.

Furthermore, its application versatility extends to various domains, including e-commerce, market research, and customer insights, benefiting both consumers and businesses. In terms of utilization, OPEN-AMZPRE serves multiple purposes. Consumers can make more informed product evaluations, aiding their decision-making processes. Businesses can leverage it to gain valuable customer sentiment insights, guide product development and marketing strategies, and enhance customer satisfaction. Market researchers can analyze Amazon Product Reviews to uncover market trends and consumer preferences. Finally, OPEN-AMZPRE enables detailed sentiment analysis of product reviews, assisting consumers and businesses in assessing product quality, functionality, and overall satisfaction. Figure 1 shows the system architecture of the proposed OPEN-AMZPRE algorithm.

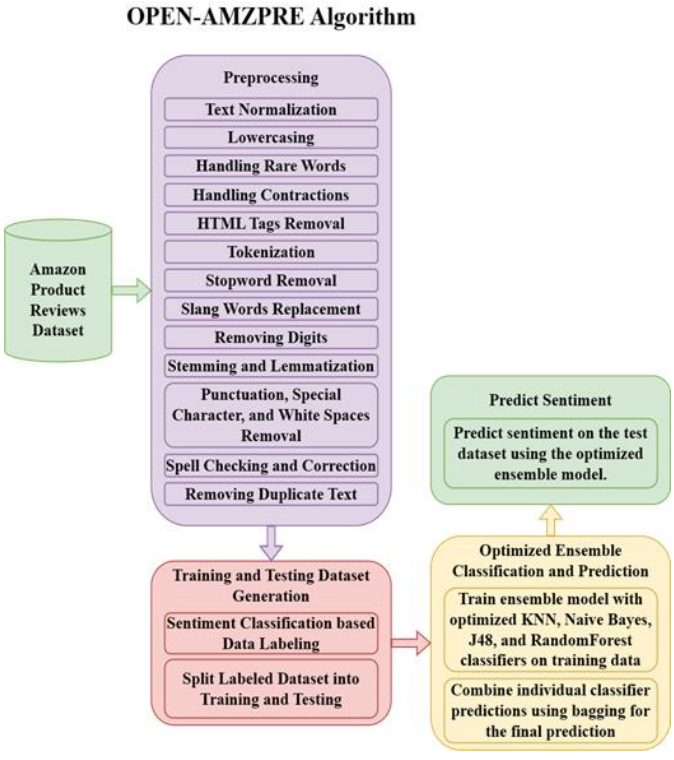

Figure 1: System architecture of OPEN-AMZPRE algorithm

The OPEN-AMZPRE algorithm can be divided into several key components, each serving a specific purpose, which is discussed in Algorithm 1.

Algorithm 1: Optimized Preprocessing with Ensemble Classification for Amazon Product Reviews Sentiment Prediction (OPEN-AMZPRE)

Input : Amazon Product Reviews Dataset (raw text data)

Output : Preprocessed and normalized dataset for sentiment analysis
Labeled training dataset (75%)
Labeled testing dataset (25%)
Trained Optimized Ensemble Classification model for sentiment prediction

Step 1 : Load Amazon Product Reviews Dataset.

Step 2 : Extract Reviews from the Amazon Product Reviews Dataset.

Step 3 : Text Normalization:
a. Standardize abbreviations, acronyms, and numbers to their full forms for consistency and better analysis.

Step 4 : Lowercasing: Convert all text to lowercase.

Step 5 : Handling Rare Words:
a. Identify and remove rare words that appear infrequently in the dataset.

Step 6 : Handling Contractions:
a. Expand all contractions to their full forms (e.g., "don't" to "do not").

Step 7 : HTML Tags Removal:
a. Remove any HTML tags present in the text.

Step 8 : Tokenization:
a. Tokenize the text into words and punctuation marks.

Step 9 : Stopword Removal:
a. Remove common stopwords (e.g., "the," "and," "is") from the text.

Step 10 : Slang Words Replacement:
a. Replace slang words with their standard equivalents.

Step 11 : Removing Digits:
a. Remove digits and numerical characters from the text.

Step 12 : Stemming:
a. Apply stemming to reduce words to their base or root form (e.g., "running" to "run"). Stemming is used to capture the general sense of words by stripping prefixes and suffixes. It can help reduce the data's dimensionality and improve processing speed.

Step 13 : Lemmatization:
a. Lemmatize words to convert them to their base form (e.g., "running" to "run"). Lemmatization is employed to ensure more accurate text normalization by considering the linguistic context and providing valid dictionary words. It helps retain the original meaning of words and improves the quality of sentiment analysis.

Step 14 : Punctuation, Special Character, and White Spaces Removal:
a. Remove punctuation, special characters, and excess white spaces.

Step 15 : Spell Checking and Correction:
a. Perform spell checking and correction to fix any spelling errors in the text.

Step 16 : Removing Duplicate Text:
a. Identify and remove duplicate or near-duplicate reviews to ensure data quality.

Step 17 : Sentiment Classification for data labelling:
a. Use positive and negative word lists to analyze a product review's sentiment by counting positive and negative word occurrences and classifying them as positive, negative, or neutral based on the word counts.

Step 18 : Split Labeled Dataset:
a. Split the labelled dataset into a

training dataset (75%) and a testing dataset (25%).

Step 19 : Optimized Ensemble Classification:

a. Train an ensemble classification model using optimized KNN, Naive Bayes, J48 (C4.5 decision tree), and Random Forest classifiers on the training dataset.

b. Combine the predictions from the individual classifiers to make a final prediction.

Step 20 : Predict Sentiment:

a. Use the trained, optimized ensemble classification model to predict the sentiment of the testing dataset.

3.1 Data Preprocessing

The data preprocessing phase in OPEN-AMZPRE is a crucial component that prepares the raw Amazon Product Reviews data for sentiment analysis. This phase ensures the data is standardized, noise-free, and well-structured, laying the foundation for accurate sentiment predictions. It encompasses a series of substeps, each serving a specific purpose in enhancing data quality and consistency.

3.1.1 Text Normalization

Text normalization is a pivotal preprocessing step within the OPEN-AMZPRE algorithm, dedicated to standardizing the textual data extracted from Amazon Product Reviews. This process serves the fundamental purpose of converting abbreviations, acronyms, and numerical representations into their full and explicit forms. Doing so aims to achieve consistency and uniformity in the text, thereby laying a solid foundation for more accurate sentiment analysis.

Its primary objective is establishing a consistent and standardized text dataset, which is crucial for sentiment analysis. This process performs three essential functions: firstly, it tackles the issue of consistency within the dataset by addressing the challenge of inconsistent representations of identical concepts, which can emerge due to the use of diverse abbreviations and acronyms in platforms like Amazon Product Reviews, thereby ensuring uniform representation of sentiment-related terms. Secondly, it enhances interpretability by converting these abbreviations and acronyms into full forms, making the text more understandable, benefiting both automated sentiment analysis algorithms and human readers relying on these insights for decision-making. Lastly, text normalization improves accuracy by reducing ambiguity and minimizing potential misinterpretations stemming from expression variations, ultimately fostering more reliable sentiment predictions.

Text normalization offers many advantages that significantly enhance the effectiveness of the OPEN-AMZPRE algorithm and elevate the quality of sentiment analysis for Amazon Product Reviews. First and foremost, it cultivates improved consistency, ensuring that synonymous terms and concepts are uniformly represented, simplifying the identification and analysis of sentiment-carrying terms within the reviews. Furthermore, it bolsters accuracy by converting abbreviations and acronyms into their full forms, reducing ambiguity, and empowering sentiment analysis algorithms to confidently identify and classify relevant keywords and phrases within standardized text. In addition, the standardized text streamlines subsequent preprocessing and analysis steps, facilitating a more efficient and streamlined sentiment analysis process for both researchers and algorithms. Moreover, it enhances interoperability with other natural language processing (NLP) tools and sentiment analysis libraries, making the OPEN-AMZPRE algorithm adaptable and versatile across various datasets and domains. Overall, text normalization plays a pivotal role in the preparation of text data for sentiment analysis in the context of Amazon Product Reviews, promoting internal consistency, boosting the precision of sentiment interpretation, simplifying analysis, and fostering interoperability with NLP tools, all of which collectively fortify the OPEN-AMZPRE algorithm in

delivering accurate and dependable sentiment predictions.

3.1.2 Lowercasing

The Lowercasing step within the OPEN-AMZPRE algorithm is a foundational text preprocessing technique designed to convert all text in the Amazon Product Reviews dataset to lowercase. Its primary objective is to establish uniformity in the letter case of the text. By accomplishing this, Lowercasing aims to mitigate discrepancies and inconsistencies that may arise from variations in letter casing, ultimately enhancing the accuracy and reliability of sentiment analysis.

Lowercasing offers a range of significant benefits, contributing to the consistency and quality of the text data in the context of sentiment analysis. Firstly, it ensures uniformity in the dataset, as all words are represented in the same lowercase format. This consistent letter case simplifies the sentiment analysis, allowing the algorithm to identify and categorize sentiment-related terms more effectively.

Additionally, Lowercasing prevents discrepancies that can result from case sensitivity. For instance, without lowercasing, “Great” and “great” might be treated as distinct terms in sentiment analysis, potentially leading to fragmented sentiment vocabulary. Lowercasing eliminates such disparities, allowing the algorithm to capture sentiment signals comprehensively. In this way, converting all text to lowercase promotes uniformity and enhances sentiment analysis accuracy, contributing to more reliable sentiment predictions for Amazon Product Reviews.

3.1.3 Handling Rare Words

In the context of the OPEN-AMZPRE algorithm, “Handling Rare Words” refers to identifying and removing words that occur infrequently within the Amazon Product Reviews dataset. This step recognizes that not all words are equally crucial in determining sentiment and that the presence of infrequent terms can introduce noise, potentially hindering the accuracy of sentiment analysis. Therefore, handling rare words involves systematically curating the dataset to retain the most relevant and commonly occurring terms while discarding less significant ones due to their rarity.

The primary purpose of the “Handling Rare Words” step is to improve the quality of the text data used for sentiment analysis. By identifying and eliminating words with limited occurrence, this process reduces the impact of noise. Rare words, being infrequent and often irrelevant to sentiment, can introduce variability and ambiguity in the dataset. Removing such terms enhances the overall clarity and focus of the text data, making it more suitable for sentiment prediction. This curation of the dataset streamlines the analysis. It contributes to the accuracy of sentiment interpretations, as the algorithm can devote more attention to the most meaningful and commonly used words.

The benefits of handling rare words are manifold and pivotal to the effectiveness of the OPEN-AMZPRE algorithm. Firstly, this process reduces noise within the dataset by removing infrequent and less relevant terms. Noise reduction is essential for maintaining the integrity of sentiment analysis results, as uncommon words may not carry substantial sentiment information. By eliminating these outliers, the text data becomes more precise and more focused, ensuring that the analysis concentrates on the most meaningful content.

Secondly, removing rare words contributes to improved accuracy in sentiment analysis. Less frequent terms may not be reliable indicators of sentiment and can potentially lead to false interpretations. The algorithm can make more accurate predictions about the sentiment expressed in the reviews by focusing on the more common and informative words. In essence, handling rare words streamlines the data, enhances the signal-to-noise ratio, and ultimately leads to more precise and reliable sentiment predictions for Amazon Product Reviews.

3.1.4 Handling Contractions

In the OPEN-AMZPRE algorithm, "Handling Contractions" is a specific data preprocessing step designed to address contractions found within the text of Amazon Product Reviews. Contractions are combinations of words where one or more letters or characters are replaced with an apostrophe, such as "don't" for "do not" or "it's" for "it is." This step identifies and expands these contractions to their full, uncontracted forms. For instance, "don't" is expanded to "do not," "it's", to "it is," and so on. This process is typically achieved through pattern matching and replacement, ensuring that all contractions within the text are fully expanded.

The primary purpose of the "Handling Contractions" step is to enhance the clarity and readability of the text within Amazon Product Reviews. Contractions, while commonly used in informal writing, can introduce ambiguity, especially in the context of sentiment analysis. Expanding contractions into their full forms makes the text more explicit and unambiguous. This expansion helps automated sentiment analysis algorithms and human readers better understand the reviews' content, ultimately promoting more accurate sentiment predictions.

Handling contractions in the OPEN-AMZPRE algorithm offers several notable benefits that significantly enhance the quality and interpretability of the text data. First, it improves readability by expanding contractions, making the full forms of words easier to comprehend for sentiment analysis algorithms and human readers relying on these insights for informed decision-making. Second, it enhances clarity and unambiguity by eliminating the potential for multiple interpretations associated with contractions, ensuring more precise and accurate sentiment analysis results.

Overall, "Handling Contractions" is a crucial preprocessing step in the OPEN-AMZPRE algorithm, executed by expanding contractions to their full forms. This process serves the fundamental purpose of enhancing text clarity and readability while eliminating ambiguity that contractions may introduce. The benefits of improved readability and unambiguous text ultimately contribute to the quality and reliability of sentiment predictions for Amazon Product Reviews.

3.1.5 HTML Tags Removal

HTML Tags Removal, a significant phase in the OPEN-AMZPRE algorithm, entails the elimination of HTML tags that may be present within the text of Amazon Product Reviews. HTML tags are used to format and structure web content. This process uses techniques such as regular expressions to identify and remove all HTML tags from the text. The result is text-free from any HTML-related formatting or structural elements, leaving only the textual content for analysis.

The core purpose of HTML tag removal is to ensure that the text data used for sentiment analysis lacks any HTML-related artefacts. When Amazon Product Reviews are scraped from web sources, they often contain HTML tags used for styling and layout, but these tags can interfere with sentiment analysis. Removing HTML tags results in cleaner and more focused text, preparing it for precise sentiment predictions.

By eliminating HTML tags, this step safeguards the accuracy and reliability of sentiment analysis. The benefits are two-fold: it prevents the interference of HTML formatting with sentiment interpretation and ensures that the text is pure and ready for subsequent analysis. It contributes to improved data quality and enhances the effectiveness of sentiment predictions for Amazon Product Reviews.

3.1.6 Tokenization

Tokenization is a crucial process in the OPEN-AMZPRE algorithm, executed to split the text within Amazon Product Reviews into individual words and punctuation marks. This task is accomplished by applying specialized tokenization algorithms that identify and separate words, phrases, and punctuation within the text, effectively segmenting the text into discrete units for analysis.

The fundamental purpose of tokenization is to break down the text into analyzable units, preparing it for further sentiment analysis. By segmenting the text into tokens, it becomes more manageable for the algorithm to identify and analyze individual words and punctuation marks, which is essential for extracting sentiment-related information.

Tokenization offers the critical benefit of segmenting the text into manageable units, facilitating the subsequent steps of sentiment analysis. The algorithm can focus on individual terms and punctuation marks, enabling more accurate sentiment predictions. This division into tokens contributes to a more structured and interpretable dataset, ultimately enhancing the quality of sentiment analysis for Amazon Product Reviews.

3.1.7 Stopword Removal

Stopword Removal is a pivotal phase in the OPEN-AMZPRE algorithm executed to eliminate common stopwords, such as “the,” “and,” and “is,” from the text within Amazon Product Reviews. Stopwords are identified using predefined lists of common words and systematically removed to extract meaningful content.

The primary purpose of Stopword Removal is to filter out words common in the language but often lack significant sentiment-bearing information. By removing stopwords, the text data becomes more focused on meaningful terms, preparing it for accurate sentiment analysis.

The benefits of Stopword Removal are substantial. This process streamlines the text data by eliminating less informative terms, making it more concise and relevant. By focusing on words with sentiment-related information, the algorithm can make more accurate predictions about the sentiment expressed in the reviews. This results in enhanced precision and reliability in sentiment analysis for Amazon Product Reviews.

3.1.8 Slang Words Replacement

Slang word replacement in the OPEN-AMZPRE algorithm identifies and substitutes slang words within the text of Amazon Product Reviews with their corresponding standard equivalents. This execution involves the creation of a comprehensive list of slang words and their joint replacements, followed by systematic pattern matching and substitution within the text.

The primary objective of Slang Words Replacement is to ensure that colloquial language and informal expressions do not hinder the accuracy of sentiment analysis. Slang words often possess context-specific meanings that might not align with standard vocabulary, leading to potential misinterpretations. The text becomes more consistent and interpretable by replacing slang with its standardized counterparts.

Slang word replacement offers significant benefits, enhancing the reliability and clarity of sentiment analysis. This step ensures the text is free from colloquialisms that could introduce ambiguity or misinterpretation. It promotes uniformity and consistency in the language used in reviews, contributing to more accurate and dependable sentiment predictions for Amazon Product Reviews.

3.1.9 Removing Digits

Removing Digits is a vital preprocessing step executed within the OPEN-AMZPRE algorithm to eliminate digits and numerical characters present within the text of Amazon Product Reviews. This process uses regular expressions or pattern-matching techniques to identify and remove all numeric elements from the text.

The fundamental purpose of Removing Digits is to enhance the quality of text data used for sentiment analysis. Being numerical, digits are often unrelated to sentiment expression and can introduce noise. Removing these numerical characters makes the text cleaner and more focused, enabling more accurate sentiment predictions.

The benefits of Removing Digits are twofold. Firstly, it reduces the impact of noise within the dataset by eliminating unrelated numerical elements. Secondly, it enhances the clarity and interpretability of the text, as the focus is placed on linguistic content relevant to sentiment analysis. This results in improved data

quality and the ability to make more precise and reliable sentiment predictions for Amazon Product Reviews.

3.1.10 Stemming

Stemming, a crucial process in the OPEN-AMZPRE algorithm, involves reducing words to their base or root form. This reduction is achieved by applying stemming algorithms that identify and extract the common root of words, for example, converting "running" to "run."

The primary purpose of Stemming is to achieve a more compact vocabulary and account for variations in word forms. Stemming aims to simplify the dataset by reducing words to their essential forms, making them more consistent for analysis.

Stemming provides several notable benefits. It results in a more concise and focused dataset by reducing word variations to their common root. This simplification enables the algorithm to identify and categorize sentiment-related terms more effectively, ultimately enhancing the quality and precision of sentiment analysis for Amazon Product Reviews.

3.1.11 Lemmatization

Lemmatization is a critical process in the OPEN-AMZPRE algorithm executed to convert words to their base form. Unlike stemming, Lemmatization involves determining a word's canonical or dictionary form, such as converting "running" to "run." It is achieved through the application of lemmatization algorithms and linguistic knowledge resources.

The core purpose of Lemmatization is to ensure that words are represented in their most fundamental and consistent form for analysis. By converting words to their base form, the text data becomes more unified and more accessible to interpret for sentiment analysis.

Lemmatization offers substantial benefits. It ensures that the text data consists of words in their most fundamental forms, promoting uniformity and clarity. It enhances the accuracy and interpretability of sentiment analysis, as words are represented consistently, enabling more precise sentiment predictions for Amazon Product Reviews.

3.1.12 Punctuation, Special Character, and White Spaces Removal

Punctuation, Special Character, and White Spaces Removal, an essential preprocessing step within the OPEN-AMZPRE algorithm, involves systematically eliminating unnecessary characters, including punctuation marks, special symbols, and excess white spaces from the text of Amazon Product Reviews. This process is executed using pattern matching and replacement techniques.

The primary purpose of Punctuation, Special Character, and White space removal is to enhance the clarity and uniformity of the text. By eliminating these extraneous characters, the text data becomes more focused, interpretable, and well-prepared for sentiment analysis.

This preprocessing step offers multiple benefits. It enhances the clarity and readability of the text by removing distractions introduced by punctuation marks and special symbols. Additionally, it ensures that the text is consistent and focused, enabling more precise and reliable sentiment analysis for Amazon Product Reviews.

3.1.13 Spell Checking and Correction

Spell Checking and Correction is a crucial phase within the OPEN-AMZPRE algorithm, executed to identify and rectify spelling errors within the text of Amazon Product Reviews. It is achieved by applying an Edit Distance-based spell checker to detect and replace misspelt words. It calculates the edit distance between the input word and words in a dictionary to suggest corrections for misspelt words. The spell checker generates possible corrections by performing various edit operations such as deletion, insertion, substitution, and transposition on the input word and checks if the resulting words are present in a predefined dictionary. It then suggests the most likely correct word based on the edit distance and the frequency of the corrected words in the dictionary.

The fundamental purpose of spell-checking and correction is to ensure the text is accurate and free from spelling-related distractions. Misspelled words

can introduce confusion and ambiguity, which could hinder sentiment analysis.

The benefits of Spell Checking and Correction are significant. This step guarantees that the text is free from spelling errors, promoting accuracy and clarity in sentiment analysis. It prevents the introduction of unnecessary distractions that could affect the interpretability of the text, ultimately enhancing the quality of sentiment predictions for Amazon Product Reviews.

3.1.14 Removing Duplicate Text

Removing Duplicate Text, a crucial preprocessing step within the OPEN-AMZPRE algorithm involves systematically identifying and eliminating duplicate or near-duplicate reviews within the dataset of Amazon Product Reviews. This process is executed using deduplication techniques, such as text similarity measures.

The primary purpose of Removing Duplicate Text is to enhance the overall quality of the dataset. Duplicate reviews introduce redundancy and do not contribute unique information to sentiment analysis. By removing such duplicates, the dataset becomes more concise and focused.

Removing Duplicate Text offers significant benefits. It prevents data redundancy, leading to a more compact dataset that contains unique and meaningful content. It enhances the efficiency of sentiment analysis and ensures that the results are based on a representative and non-repetitive set of Amazon Product Reviews.

The data preprocessing phase of OPEN-AMZPRE is instrumental in preparing the Amazon Product Reviews data for accurate sentiment analysis. These substeps collectively contribute to data standardization, cleanliness, and quality, ensuring that subsequent sentiment analysis is reliable and insightful.

3.2 Sentiment Classification for Data Labeling and Split Labeled Dataset

In the OPEN-AMZPRE algorithm, the phase of sentiment classification serves a dual purpose: first, it labels the Amazon Product Reviews by assigning sentiment categories (positive, negative, or neutral) to each review, and second, it facilitates the subsequent division of the labelled dataset into training and testing datasets, a critical step in machine learning model development.

Sentiment Classification for Data Labeling:

The sentiment classification process is initiated using two specialized text files that contain carefully curated lists of positive and negative words. These lists are fundamental to the sentiment analysis task and enable the algorithm to gauge the sentiment of each product review efficiently. The procedure unfolds as follows:

- **Utilization of Positive and Negative Word Lists:** OPEN-AMZPRE deploys lists of words inherently associated with positive or negative sentiment. These word lists are essential for the algorithm to analyze the reviews effectively. The presence and frequency of words from these lists within the review text are meticulously counted, forming the basis for the sentiment classification.
- **Counting Word Occurrences:** The algorithm thoroughly examines each product review, counting the occurrences of the words from the positive and negative lists. This process provides quantitative insights into the sentiment-bearing content within each review. The occurrence count is crucial in categorizing reviews into their respective sentiment classes.
- **Classification as Positive, Negative, or Neutral:** Once the word occurrence counting is complete, OPEN-AMZPRE categorizes each review into one of three sentiment classes: positive, negative, or neutral. This classification is determined by assessing the preponderance of positive or negative words within the review text. Reviews exhibiting a dominant presence of positive words are categorized as positive, while those predominantly containing negative terms are classified as negative. Reviews with a balanced distribution of sentiment-bearing terms or those

lacking significant sentiment words are assigned to the neutral category.

The sentiment classification process ensures that each Amazon Product Review is systematically labelled with a sentiment category, enabling the subsequent analysis of the overall sentiment distribution within the dataset.

Split Labeled Dataset:

With the reviews successfully labelled based on sentiment, OPEN-AMZPRE proceeds to the next critical step: splitting the labelled dataset into two subsets, a training and a testing dataset. This division is integral to the training and evaluation of the machine learning model for sentiment prediction:

- **Splitting the Labeled Dataset:** The labelled dataset, now categorized into positive, negative, or neutral reviews, is divided into two subsets. The training dataset, representing 75% of the labelled data, is allocated for the training of the sentiment prediction model. The testing dataset, constituting 25% of the labelled reviews, is reserved for evaluating the model's performance.

The division of the labelled dataset into training and testing datasets is pivotal for assessing the predictive capability of the sentiment analysis model. It allows for training the model on a substantial portion of the data and, subsequently, the validation of its accuracy and generalizability on unseen data. The testing dataset is a benchmark for evaluating the model's effectiveness in predicting sentiment and provides valuable insights into the algorithm's performance.

The combination of sentiment classification and dataset division ensures that OPEN-AMZPRE can accurately predict and evaluate sentiment in Amazon Product Reviews while maintaining the integrity of the model's training and testing processes. This multi-step approach forms the foundation for the algorithm's sentiment analysis and prediction success.

3.3 Optimized Ensemble Classification

The Optimized Ensemble Classification component in OPEN-AMZPRE focuses on enhancing the accuracy and performance of sentiment prediction. It involves a multi-step process combining various classifiers' powers to make sentiment predictions collectively. This section outlines this phase's critical steps and details how each base classifier is optimized.

OPEN-AMZPRE initiates the Ensemble Classification process by training an ensemble classification model on the labelled training dataset. This model leverages the strengths of multiple classifiers to make accurate sentiment predictions. The following classifiers are employed within the ensemble:

- **Optimized K-Nearest Neighbors (KNN):** KNN is used as one of the base classifiers within the ensemble. In the OPEN-AMZPRE algorithm, the KNN classifier is initialized with specific parameters, including setting the value of 'k' (the number of nearest neighbors to consider) to 3. Additionally, an AdaBoostM1 classifier is created to optimize the KNN classifier further. AdaBoostM1 is configured with a total of 10 iterations, allowing it to boost the performance of the KNN classifier.
- **Optimized Naive Bayes (NB):** The ensemble incorporates the Complement Naïve Bayes classifier, a specialized version of the Naive Bayes classifier tailored for sentiment analysis. Unlike traditional Naive Bayes, Complement Naive Bayes effectively handles class imbalance issues commonly encountered in sentiment datasets. It achieves this by considering the absence of words in a given class, making it a valuable tool for sentiment prediction in scenarios where one class dominates the other. In OPEN-AMZPRE, it dramatically enhances the overall accuracy of sentiment analysis.
- **Optimized J48 (C4.5 Decision Tree):** J48, based on the C4.5 decision tree algorithm, is used for its ability to build a decision tree for classification. In the OPEN-AMZPRE algorithm, J48 is configured with specific options for pruning, handling unclassified instances, binary splits, and setting the

random seed to 1. These options ensure that J48 operates optimally for sentiment analysis.

- **Optimized Random Forest:** Random Forest is another base classifier ensemble of decision trees. In the OPEN-AMZPRE algorithm, the Random Forest classifier is customized with settings such as the number of trees (500) and the maximum depth of trees (10). Additionally, an AdaBoostM1 classifier is created to optimize the Random Forest classifier further. AdaBoostM1 is configured with a total of 10 iterations, allowing it to boost the performance of the Random Forest classifier. These adjustments aim to enhance the performance of Random Forest for sentiment prediction.

Once the individual classifiers within the ensemble are trained and optimized, OPEN-AMZPRE combines their predictions to make a final sentiment prediction. Combining predictions from multiple classifiers helps mitigate potential biases or inaccuracies in any single classifier, resulting in a more robust and reliable sentiment prediction. To achieve this, OPEN-AMZPRE employs a Bagging ensemble, which includes the base classifiers - K-Nearest Neighbors (KNN), Naive Bayes (NB), J48 (C4.5 decision tree), and Random Forest (RF), each optimized for sentiment analysis. The Bagging ensemble leverages the diversity of these classifiers by aggregating their predictions. This ensemble approach further enhances the accuracy and robustness of sentiment predictions in OPEN-AMZPRE.

3.4 Sentiment Prediction

The Sentiment Prediction phase in OPEN-AMZPRE represents the culmination of the algorithm's capabilities. This critical step harnesses the trained and optimized ensemble classification model to predict the sentiment of Amazon Product Reviews, not only on the testing dataset but also on new and unseen product review data. The testing dataset, serving as a benchmark, is preprocessed similarly to the training dataset, ensuring consistency.

Within this phase, the model thoroughly analyzes the preprocessed text of the reviews and assigns sentiment labels, categorizing each review as positive, negative, or neutral. These sentiment predictions serve as the conclusive output of the OPEN-AMZPRE algorithm. They offer valuable insights into the sentiment distribution within Amazon Product Reviews, enabling businesses and users to make informed decisions based on the sentiment analysis results.

Various evaluation metrics, including accuracy, precision, recall, and F1-score, are employed to assess the effectiveness of the ensemble classification and sentiment prediction phases. These metrics provide a comprehensive understanding of the algorithm's performance in sentiment analysis, making it a powerful tool for testing datasets and predicting the sentiment of new, previously unseen product review datasets.

Overall, OPEN-AMZPRE's integrated preprocessing and ensemble classification approach offers a comprehensive and innovative solution to Amazon Product Reviews sentiment analysis, providing enhanced accuracy and valuable insights for consumers and businesses.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS:

In the rigorous evaluation of the OPEN-AMZPRE algorithm, a systematic analysis was conducted to assess its effectiveness and reliability in the context of sentiment analysis. The algorithm was rigorously compared with established methodologies, notably the Linear SVM and Naïve Bayes algorithms, as introduced by Dey et al. [17]. This comparison was executed within the framework of sentiment analysis, a field of paramount importance in deciphering the polarity of textual data. The specific aim was to evaluate OPEN-AMZPRE's performance across a range of vital performance metrics, including accuracy (%), precision, recall, and F1-score. This comprehensive assessment enabled us to gain insights

into OPEN-AMZPRE's capabilities and efficiency in relation to these well-regarded benchmarks. Furthermore, it is worth noting that the algorithm was tested on a substantial Amazon Product Reviews Dataset [18], representing a diverse and dynamic source of data that reflects real-world consumer sentiments, adding a layer of practical applicability to the assessment.

The results of this comparison are presented in Table 1, which clearly illustrates the performance metrics of Precision, Recall, and F1-score for the OPEN-AMZPRE algorithm alongside those of the Linear SVM and Naïve Bayes methods. These metrics offer a quantitative assessment of the strengths and weaknesses of each algorithm, shedding light on their respective capabilities and limitations in the context of sentiment analysis. The comparison in Table 1 is a valuable resource for evaluating the potential of OPEN-AMZPRE and understanding how it stacks up against existing state-of-the-art techniques.

Table 1: Precision, Recall, and F1-score Comparison

Classifiers	Precision	Recall	F1-score
Linear SVM	0.83990	0.83997	0.83993
Naïve Bayes	0.82853	0.82884	0.82662
OPEN-AMZPRE	0.87659	0.85992	0.86818

These metrics provide valuable insights into the algorithm's performance. The precision measures the accuracy of positive predictions, recall assesses the model's ability to identify actual positive cases, and the F1-score combines precision and recall into a single metric to gauge overall model effectiveness. Figure 2 shows the pictorial diagram of Precision, Recall, and F1-score Comparison.

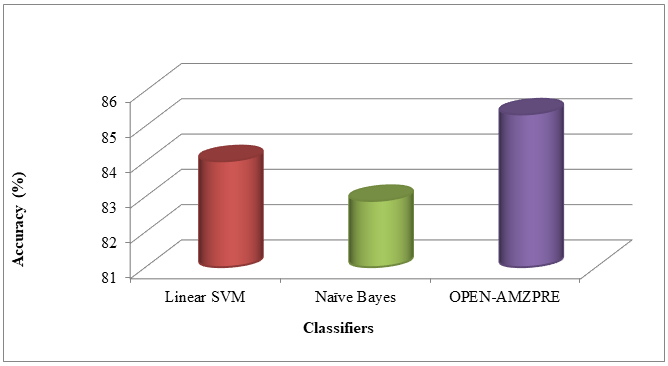

Figure 2 : Precision, Recall, and F1-score Comparison

Table 2 provides a detailed and informative accuracy comparison between the Linear SVM and Naïve Bayes classifiers, as introduced by Dey et al. [17], and the proposed OPEN-AMZPRE algorithm. This table is a valuable resource for discerning the relative performance of these sentiment analysis methods in terms of their ability to classify and analyze data accurately.

Table 2: Accuracy Comparison

Classifiers	Accuracy (%)
Linear SVM	84
Naïve Bayes	82.875
OPEN-AMZPRE	85.328

Accuracy (%) is a fundamental performance metric that quantifies the precision of a classification algorithm. It is computed by taking the number of correctly classified instances and dividing it by the total number of instances, expressed as a percentage. This metric is pivotal in assessing the reliability and effectiveness of a given model's ability to categorize data correctly. Figure 3 shows the pictorial diagram of accuracy comparison.

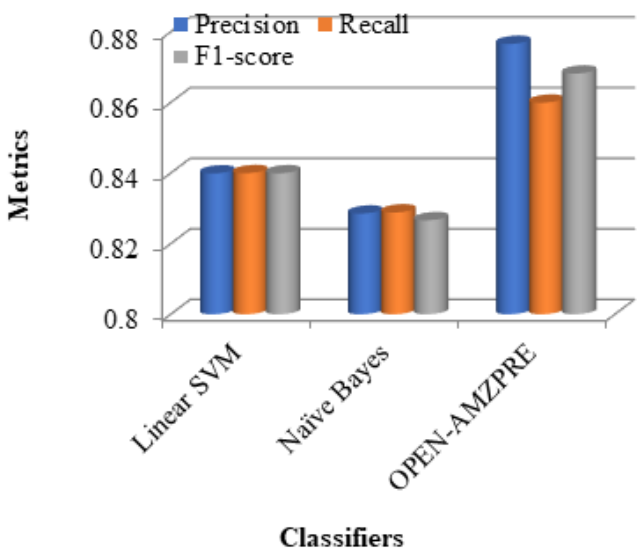

Figure 3 : Accuracy Comparison

The results demonstrate that OPEN-AMZPRE outperforms the other classifiers, achieving higher accuracy, precision, recall, and F1-score. It indicates that the proposed algorithm excels in sentiment analysis on Amazon product reviews and offers a more reliable and accurate solution than the existing methods.

These results highlight the effectiveness of OPEN-AMZPRE in accurately predicting sentiment and its potential for practical utilization in sentiment analysis tasks. The algorithm provides a valuable tool for businesses and users seeking insights from product reviews on the Amazon platform, aiding in data-driven decision-making. The experiments affirm that OPEN-AMZPRE is a robust and promising sentiment analysis solution, offering accuracy and overall performance improvements.

V. CONCLUSION AND FUTURE WORK

In conclusion, the OPEN-AMZPRE algorithm presented in this study demonstrates its effectiveness in sentiment analysis for Amazon product reviews. The comprehensive preprocessing, ensemble classification and sentiment prediction phases collectively contribute to its superior performance, surpassing existing methods such as Linear SVM and Naïve Bayes regarding accuracy, precision, recall, and F1-score. The algorithm's enhanced capacity to forecast sentiment with greater precision and dependability paves the way for numerous possibilities in the e-commerce sector, facilitating various applications and decision-making procedures. The OPEN-AMZPRE algorithm can be extended to other domains and platforms for future work. It provides valuable insights into sentiment analysis for diverse textual data, such as social media posts, news articles, and customer feedback in different industries, expanding its potential for broader utilization and impact.

VI. REFERENCES

[1]. Keung, P., Lu, Y., Szarvas, G., & Smith, N. A. (2020). The multilingual Amazon reviews corpus—arXiv preprint arXiv:2010.02573.

[2]. Ali, M. M., Doumbouya, M. B., Louge, T., Rai, R., & Karray, M. H. (2020). An ontology-based approach to extract product design features from online customers' reviews. Computers in Industry, 116, 103175.

[3]. Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2020). Sentiment analysis on online product reviews. In Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018 (pp. 559-569). Springer Singapore.

[4]. Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S. M., Sangwan, R. S., & Sharma, R. (2021). Effect of negation in sentences on sentiment analysis and polarity detection. Procedia Computer Science, 185, 370-379.

[5]. Alantari, H. J., Currim, I. S., Deng, Y., & Singh, S. (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. International Journal of Research in Marketing, 39(1), 1-19.

[6]. Wassan, S., Chen, X., Shen, T., Waqar, M., & Jhanjhi, N. Z. (2021). Amazon product sentiment analysis using machine learning techniques. Revista Argentina de Clínica Psicológica, 30(1), 695.

[7]. Nandal, N., Tanwar, R., & Pruthi, J. (2020). Machine learning-based aspect level sentiment analysis for Amazon products. Spatial Information Research, 28, 601-607.

[8]. Alharbi, N. M., Alghamdi, N. S., Alkhammash, E. H., & Al Amri, J. F. (2021). Evaluation of sentiment analysis via word embedding and RNN variants for Amazon online reviews. Mathematical Problems in Engineering, 2021, 1-10.

[9]. Geetha, M. P., & Renuka, D. K. (2021). Improving aspect-based sentiment analysis performance using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks, 2, 64-69.

[10]. Budhi, G. S., Chiong, R., Pranata, I., & Hu, Z. (2021). Using machine learning to predict the sentiment of online reviews: a new framework for comparative analysis. Archives of Computational Methods in Engineering, 28, 2543-2566.

[11]. Rintyarna, B. S., Sarno, R., & Fatichah, C. (2020). Enhancing the performance of sentiment analysis tasks on product reviews by handling local and global contexts. International Journal of Information and Decision Sciences, 12(1), 75-101.

[12]. Zhou, F., Ayoub, J., Xu, Q., & Jessie Yang, X. (2020). A machine learning approach to customer needs analysis for product ecosystems—journal of mechanical design, 142(1), 011101.

[13]. Dang, C. N., Moreno-García, M. N., & Prieta, F. D. L. (2021). An approach to integrating sentiment analysis into recommender systems. Sensors, 21(16), 5666.

[14]. AlQahtani, A. S. (2021). Product sentiment analysis for Amazon reviews. International Journal of Computer Science & Information Technology (IJCSIT) Vol, 13.

[15]. Dadhich, A., & Thankachan, B. (2022). Sentiment analysis of Amazon product reviews using a hybrid rule-based approach. In Smart Systems: Innovations in Computing: Proceedings of SSIC 2021 (pp. 173-193). Springer Singapore.

[16]. Rashid, A., & Huang, C. Y. (2021). Sentiment Analysis on Consumer Reviews of Amazon Products. International Journal of Computer Theory and Engineering, 13(2), 7.

[17]. Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020, February). A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In 2020 International Conference on Contemporary Computing and Applications (IC3A) (pp. 217-220). IEEE.

[18]. Yasser H. (2020). Amazon Product Reviews Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset

Cite this article as :

Prof. Aparna Hote, Dr. Dev Ras Pandey, "OPEN-AMZPRE : Optimized Preprocessing with Ensemble Classification for Amazon Product Reviews Sentiment Prediction", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 6, pp. 385-401, November-December 2023. Available at doi : https://doi.org/10.32628/IJSRST52310672
Journal URL : https://ijsrst.com/IJSRST52310672