International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I3P101

Volume 73 Issue 3, 1-15, March 2025

Review Article

An Ensemble Model Based on Multinomial Naïve Bayes and Lexicon for Sentiment Classification of Product Reviews

Gabriel V. Oliko[1], Calvins Otieno[2], Titus M. Muhambe[3]

[1]*Department of Information Technology, Maseno University, Kisumu, Kenya.*
[2]*Department of Computer Science, Maseno University, Kisumu, Kenya.*
[3]*Department of Mathematics, Physics & Computer Science, Alupe University, Busia, Kenya.*

[1]*Corresponding Author : goliko@maseno.ac.ke*

Received: 08 January 2025 Revised: 18 February 2025 Accepted: 02 March 2025 Published: 15 March 2025

Abstract *- In the emerging trend, product developers and their customers use internet reviews as the primary tool for evaluating products. Online communities, blogs, and public review websites provide a multitude of data about customers' overall viewpoints, experiences, and opinions about goods. Product developers can harvest data on users' perceptions about their preferred features and use that information to boost revenue and profit by planning and monitoring business strategies and improving the overall quality of products. The reviews also assist prospective purchasers in making informed decisions on the suitability of a product and pricing while reducing time and effort. Machine learning algorithms are used to identify and categorize product evaluations. This paper presents an ensemble machine learning approach that integrates results drawn from two base learners to improve accuracy in classification, which is the percentage of correctly classified product evaluation. Multinomial Naïve Bayes and Unsupervised Lexicon were the base learners utilized to model the ensemble that was used to classify consumer reviews as positive, neutral or negative. Feature extraction methods N-gram, Part of Speech, and features from the lexical library TextBlob were used. The proposed model was evaluated on the real dataset for two items: the "Samsung Galaxy A12" smartphone and the "Nissan Sentra" automobile brand and series. The experimental results indicate that the MNB Lexicon Pooled Ensemble outperformed the individual MNB and Lexicon classifiers in rating prediction, with respective accuracy, precision, recall and F1 measurements of 0.8250, 0.8932, 0.7970 and 0.8325.*

Keywords *- Product, Reviews, Sentiment analysis, Multinomial Naïve Bayes, Lexicon.*

1. Introduction

The importance of customer reviews in determining In the internet age, the significance of consumer feedback in assessing satisfaction has grown dramatically. The possibilities for applying sentiment evaluation to consumer evaluations are enormous.

Sentiment Analysis of product reviews is highly helpful to both product developers looking to gather consumer or public opinions from online sources about their offerings and prospective buyers looking to learn from what previous customers have to say before committing to a purchase.

1.1. Product Failure

Product failure can be attributed to several reasons. Some of the most frequent causes of product failures, in addition to a flawed concept or design, often fit into one or several of the seven classifications [1,2,3]: incorrect positioning of the product, ineffective packaging, deceptive or unclear advertising message about the good or service, its characteristics and features, or its use, a lack of understanding of the target segment of the market and the branding strategy that would best serve that segment, incorrect pricing—both too high and too low—excessive research and/or design and development costs, and an incorrect or underestimated understanding of the market. These researchers contend that many novel new items fail in the markets because businesses don't put enough effort into comprehending how consumers assess products and make purchasing choices.

A clear pattern can be drawn from the above reasons: a majority of the failure is a result of either little or wrong information reaching the product developers, or this information is received and adopted rather late after the market has evolved. In other words, from the beginning of the product's lifecycle to the end, customers ought to be deliberately and continuously involved in the creation of new products if they are to succeed.

Concept, design and production are the three main phases of developing new products [4]. Preliminary stages of product development are often information-intensive and consume a lot of time. Any successful product development and reengineering process depends on collecting pertinent and current information: information made up of well-organized facts and statistics that make sense in the context in which product engineers are supposed to understand it. Information is, therefore, a significant resource and a key component of product development success.

1.2. Market Research and Feedback

Market research is a process in product development that helps product developers gather information vital in the preliminary stages of product development. Customer monitoring, focus groups, interviews, surveys, and other traditional market research techniques have presented problems such as slower responses, high cost of implementation, poor customer insight, poor reach, poor targeting, poor respondent selection, higher dropouts, less relevancy and diminishing returns on research investments [5,6].

Customer reviews are assessments of a good or service written by someone who has used it before. Consumer review sites, which are websites that are specifically created for customers to upload their reviews on products or services, include a multitude of data regarding the overall viewpoint, encounters, and comments that customers have about goods. They provide a different, albeit excellent, approach for customers to get honest feedback on what a company can or cannot offer in relation to their demands.

With the vast number of information resources available today, a critical challenge is locating, retrieving and processing information so that consumers' aspirations, needs or wants are captured in real-time.

Artificial intelligence can be employed to harness information that is specific and timely from user review sites and Social Media to reduce product development time and help in coming up with products that satisfy the needs of consumers and address the concerns that consumers have with the current offerings.

1.3. Artificial Intelligence

Artificial intelligence makes it easier to acquire current information through a superior communications system [7] [8] that is helpful to notice and react to shifts in technology, marketing tactics, competition policies and customer needs.

The incorporation of cutting-edge technology, novel approaches, and improved and significantly greater quality production and marketing tactics by manufacturing businesses is made possible by artificial intelligence in order to satisfy the changing requirements. AI can extract insights regarding human-product interactions and the user experience.

1.4. Sentiment Analysis

Sentiment analysis, a branch of NLP, assesses data's neutrality, positivity, and negativity. It is frequently applied to text to assist companies in tracking how consumers view goods and companies in consumer reviews and understand what consumers want. In sentiment analysis, natural language processing, text analysis, and statistics are used to locate, extract, and assess subjective data. A successful customer survey will understand the what, how, and why the respondents express themselves. The sentiment dataset may primarily be composed of X tweets, comments, and reviews. Sentiment analysis utilizes software to comprehend emotions, becoming more prevalent in contemporary sectors.

Any content or object that contains a customer's voice, such as reviews or answers, can be subjected to sentiment analysis. For instance, before making an online purchase, a consumer typically checks reviews of the product or service in question, enabling them to make the best decision possible [9,10]. To obtain the sentiment, Sentiment Analysis narrows down on an object, a feature, an opinion bearer, an opinion, and an opinion orientation. Sentiment analysis deals with several topics, such as object recognition, feature extraction, and opinion orientation.

Traditionally, different techniques are employed to evaluate sentiments. Lexicon or machine-learning-based categorization may be used to group Sentiment Analysis approaches. According to Zhang et al. [11], supervised learning approaches have a high accuracy but need significant data and a long-running training period. Lexicon-based approaches, on the other hand, offer a fast classification speed but present a low recall rate. These benefits and drawbacks have led to ensemble or hybrid approaches being utilized to maximize the benefits and reduce the drawbacks. It has been demonstrated that combining the two approaches using ensemble or hybrid methods can mask the drawbacks of each strategy and increase the accuracy of the outcome [11].

Ensemble models combine single classification algorithms and techniques, albeit slightly differently. Ensemble classifiers merge many yet homogenous models. Usually, merging is performed at the output of individual base learners through various combined approaches. These techniques can be divided into "trained" and "fixed" and trained combiners [12]. Majority voting is an example of a fixed method. In contrast, hybrid methods incorporate many heterogeneous machine learning techniques, with the output of one classifier becoming the input of the next classifier [13, 14]. Ensemble modeling is one of the methods of increasing the accuracy of a prediction or classification.

The study's primary goal was to create and assess the performance of an ensemble model for sentiment categorization of reviews, using Multinomial Naïve Bayes and Lexicon methods. This study contributes as follows: (1) a

product review classification ensemble learning model; (2) an evaluation of the ensemble model utilizing product review features concerning classification accuracy, recall, precision, and F1 score. The study is arranged as follows: The literature of related investigations is reviewed in Section 2, the proposed ensemble model is discussed in Section 3, the experimentation data and a detailed explanation are provided in Section 4, and the findings and recommendations for further research are discussed in Section 5.

2. Related Works

In a world where information is abundant, it is challenging to cut through the clutter and find the most pertinent information on a certain product or market. Product developers and scholars get data from target markets and consumers through market research. Through market research, manufacturers identify gaps, assess product requests, enhance value propositions, and create marketing plans that appeal to their target audience. Both the traditional approach and the more contemporary AI-powered approach can be used to perform market research.

Traditional market research techniques, however, have drawbacks. First, traditional market research exhibits shortcomings [6] in measuring attitudes and emotions. Secondly, Traditional approaches may find it challenging to scale up or quickly adjust to new formats or data sources like social media platforms and other forms of user-generated content platforms. Lastly, traditional Market Research does not possess AI's forte of speed, as highlighted by Simona & Ramona [15]. Unlike traditional methods, which may take weeks or months, AI delivers real-time insights. This agility enables businesses to adapt promptly, staying steps ahead of the competition. Lastly, AI's capacity to analyze unstructured data, including social media and customer reviews, offers nuanced insights into consumer sentiment [16].

According to [8], AI-powered solutions enable organizations to track their competitors and product offerings in real time, assess their strategies, and extract insightful information from various digital sources in the field of competitive intelligence in product development.

Machine learning is a game-changer in the field of market research for product creation, influencing how companies glean insights from enormous datasets. Sentiment Analysis, a subset of artificial intelligence, is the premise for automating the examination of various data sources. This allows businesses to identify patterns, customer habits and market shifts with unmatched efficacy, essential for preserving competitiveness.

Sentiment analysis has been investigated by numerous scholars who have utilized different datasets, base models and analysis techniques in their works. Sinha and Narayanan [17] suggested an HLESV (Hybrid Lexicon Ensemble-based Soft Voting) model using a hybrid technique that combines lexicon and ensemble machine learning. Supervised and unsupervised learning were mixed during the ensemble learning process to boost classification and prediction performance. Consumer Electronic Product Reviews (CEPR) datasets, which were obtained from the Kaggle website, were used for the task. The suggested HLESV model was divided into two stages: 1) Using a Lexicon-based approach and 2) Using an Ensemble Learning approach. Soft voting was used for aggregation. Accuracy and Receiving Operating Characteristic Curve were the two primary criteria used in evaluating the proposed HLESV model to gauge its overall efficacy and performance over various datasets. Accuracy results for electronic devices, Kindles, and gift cards were 0.7, 0.72, and 0.87 for the suggested HLESV ensemble, respectively. The accuracy of boosting and bagging ensembles was 0.64 and 0.65 for electrical devices, 0.65 and 0.66 for Kindles, and 0.78 and 0.86 for gift cards; HLESV outperformed these ensembles. This model was unable to handle complicated sarcasm, negation, and spam reviews, and it did not have a way of continuously applying the weight patterns for the top-performing classifiers.

Geriska et al. [18] used Multinomial Naïve Bayes (MNB) to extract features from English lexicons, including unigrams, POS-tagging, and score-based features. The study determined that the lexicon pooled with lemmatization and the Adverb+Adjective and Adverb+Verb (AAAVC) algorithm effectively raised the accuracy of MNB by 0.016191, while when lexicon pooled together with the AAAVC algorithm and lemmatization was applied, the accuracy was raised by 0.010391. The performance achieved the highest possible accuracy of 0.707792, precision of 0.71833, recall of 0.859083, and F-measure 0.776291.

Barik et al. [19] suggested a lexicon-based classification algorithm based on an Improved VADER (IVADER) to assess consumer opinion across various domains. The approach entailed creating a domain-specific vocabulary derived from the VADER lexicon and categorizing reviews using that dictionary. For evaluating four multi-domain customer review datasets and compared to similar previous studies, the classification training duration of 44 seconds, the accuracy of 0.9864, precision of 0.97, recall of 0.94, and F1-measure of 0.92 were all attained using the IVADER model. Their model, nevertheless, was constrained by the vagueness of phrases and words in context. This was because, while VADER contains a vocabulary associated with certain sensations, the precise meaning of a word can change depending on the circumstances.

Trinh et al. [20] developed a strategy for sentiment evaluation that integrated learning-based and lexicon-based approaches for assessing the sentiment of Vietnamese-language reviews of products. A Vietnamese emotional

lexicon (VED) was created, and analytics of texts, linguistic examination and language-specific elements were showcased. The dictionary, which included five manual sub-dictionaries, was partially derived from the English Semantic Orientation CALculator dictionary. The study demonstrated higher accuracy than other Vietnamese systems for the same domain. This is due to the integration of the benefits of combining learning-based and lexicon-based methodologies when certain phrases were incorporated into their dictionary to make it compatible with Vietnamese grammar and harmonize the conciseness of spellings that people use on the internet. The findings of the randomization test showed that the subjective classification accuracy was 0.9430, while the sentiment classification accuracy was 0.8350. In the cross-validating test, the average accuracy of sentiment classification was 0.8193, while the average accuracy of subjective classification was 0.9149. However, because their system was not trained on domain-specific data, it could not assess domain-dependent implicit sentiments.

Researchers Ramadhony et al. [21] demonstrated sentiment analysis using an Indonesian Food and Drink Review (FDReview) dataset, which included more than 700,000 reviews. Two tasks were carried out: classification of consumer feedback into three categories (positive, negative and neutral) and prediction of ratings. The study approached opinion mining as a classification problem and used different classifiers: MNB, SVM, LSTM, and BiLSTM. The results showed that compared to conventional approaches, SVM outperformed by MNB in rating prediction, although SVM fared well in the test involving the classification of reviews. Furthermore, the BiLSTM technique surpassed all other approaches on both tasks. The findings from these experiments demonstrated that deep learning-based strategy worked well in big dataset settings. The results of a tiny balanced dataset showed that conventional machine learning techniques perform comparably to deep learning methods.

Fayaz et al. [22] employed an ensemble machine learning strategy to increase the classification accuracy of spam products by combining predictions from these three classifiers, namely: Random Forest (RF), multilayer perception (MLP), and K-Nearest Neighbor (KNN), which were chosen based on empirical study. With an accuracy of 0.8813, the results revealed that the proposed ensemble model performed superior to other classifiers with regard to classification.

Several scholars in the software product domain have presented several software prediction algorithms. Nevertheless, traditional software fault forecasts have continually shown poor classification accuracy. Dada et al. [23] recommended an innovative ensemble machine learning approach to software flaw detection using KNN, Generalized Linear Model with Elastic Net Regularization (GLMNet), and Linear Discriminant Analysis (LDA) with Random Forest as the base learner. Dada et al.'s ensemble technique achieved an accuracy of 0.8769 for the CM1 dataset, 0.8111 for the JM1 dataset, 0.9070 for the PC3 dataset, and 0.9474 for the KC3 dataset. The suggested model attained a mean accuracy of 0.8856 in prediction across all datasets tested in experimentation. The results showed that the ensemble method worked well for identifying errors in the well-known noisy feature-filled and vast dimensions of PROMISE datasets. This showed that ensemble machine learning has the potential to predict software defects in the future. Table 1 summarizes the methodologies and limitations of relevant studies.

Limitations of previous works can be summarized as follows: some of the works did not explore the entirety of the dataset. Secondly, while most works were innovative, they lacked a mechanism to handle negation. Thirdly, one of the works presented limited extraction methods for the MNB input due to a multiplier/word polarity with the POS "adverb," which incorrectly assigned a word's meaning to its polarity. Further, while innovative, most studies have focused on single or closely related product domains. Therefore, it was not possible to establish the robustness of their models across the different domains. This study fronted an MNB Lexicon Pooled Ensemble model to address the aforementioned gaps and to improve the accuracy of the figures posted.

3. Materials and Methods

Ensembles are a collection of a number of different base models whose separate and individual outputs are integrated in some way to get a final forecast [24]: a collection of independently trained classification algorithms whose predictions are merged to classify fresh cases. Figure 1 depicts the typical ensemble layout by Petrakova et al. [24]. Every member of the ensemble ought to collaborate and reinforce one another. When the ensemble methods utilized reinforce each other, the possibility of detecting a mistake in the forecast improves, and the inaccuracy can be corrected using other methods.

Unlike many classic learning classifiers, which generate only one model, ensemble learning approaches generate several models.

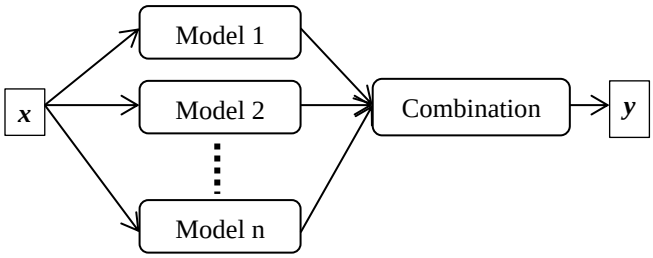

Fig. 1 The common ensemble architecture [25]

Table 1. Summary of related works

	Author	Method	Dataset	Study Limitations
1	Romadhony et al (2024).	MNB, SVM, LSTM and BiLSTM	e Large-Scale Arabic Review (LABR) dataset	Did not explore the entirety of the dataset.
2	Barik et al (2024)	IVADER Lexicon classification algorithm	Electronics, DVDs, books, and kitchens	VADER problems with complex negation patterns, double annulments, irony, sarcasm detection
3	Sinha & Narayanan (2023)	HLESV (Hybrid Lexicon Ensemble-based Soft Voting)	. Consumer Electronic Product Reviews (CEPR)	This model was unable to handle negation and spam reviews, and it did not have a way of continuously applying the weight patterns for the top-performing classifiers.
4	Dada, et al. (2021).	kNN, GLMNet, and LDA	NASA PROMISE (CM1, JM1, KC3 and PC3)	Lack of benchmark with other ensembles and deep learning approaches
5	Fayaz et al (2020)	(MLP, KNN, and RF	Yelp Dataset	Did not explore deep learning approach and LSTM with weighted TF-IDF
6	Geriska et al (2019).	MNB+Lexicon Pooled	Movie Review DB	Limited extraction methods for the MNB input due to a multiplier/word polarity with the POS "adverb," incorrectly assigned a word's meaning to its polarity.
7	Trinh et al. (2017).	SVM+Vietnamese emotional dictionary (VED)	Technology site's comments and reviews	We have not yet factored in linguistic analysis in Vietnamese, but the weight of affection affects results. The model was unable to evaluate domain-reliant emotions since it was not trained on data unique to a domain.

The ensemble design procedure consists of two major steps: (1) model training and (2) model combining. In general, an ensemble construction procedure comprises some additional steps: (1) selecting a technique for incorporating diversity into baseline models, (2) selecting a method for combining models, and (3) selecting which kind of baseline model to employ.

3.1. Proposed Method

In this study, the base learners chosen were two; MNB and Lexicon. MNB was trained on the provided training set, and then classification scores from both MNB and Lexicon were integrated using a combination technique named Lexicon Pooling. This strategy is intended to help learners improve their accuracy. Figure 2 displays the architecture of the proposed ensemble machine learning model.

There are two primary approaches to combining models: voting and averaging. While the voting approach is utilized for combining the nominal output, averaging is mostly employed for the numerical output combinations [24].

This study selected averaging because the expected outputs from each method were numeric. The flowchart displayed in Figure 3 explains the model implementation. The suggested method consists of two main steps: product feature extraction and polarity prediction.

3.1.1. Preprocessing and Feature Extraction

The studies used review datasets. Data normalization and transformation are examples of preprocessing processes performed before training the ensemble machine-learning model. Missing data and outliers were corrected during the preprocessing step. The relevant feature variables for the input model were extracted, and feature selection was performed. The proposed model was next trained to predict the polarity of reviews.

3.1.2. Ensemble Phase

The ultimate result is produced by integrating the results of the two base learner algorithms taught concurrently. The base and ensemble levels are the two levels that make up the architecture. The ensemble level has a result pooling layer, whereas the base level has MNB and Lexicon learners. The final forecast is produced by combining the classification results from the two machine-learning models (Lexicon and MNB) into one output. The results are also evaluated at this point.

3.2. The Dataset

Using two datasets from two real-world products, tests were run to objectively assess this system. The first dataset denoted as D1 — Samsung Galaxy A12s user reviews — consists of 37,724 unprocessed reviews, comprised of reviews from various phone review websites and YouTube, collected in 2023. The second dataset, denoted as D2 — Nissan Sentra Reviews — consists of 56913 unprocessed reviews collected in 2023 from various car review websites and YouTube. Both D1 and D2 contained English language reviews.

3.2.1. Choice of Products

The choice of these two products was informed by the following points: (1) Relevance - Selecting the cases was guided by the purpose and scope of the research. Relevance meant selecting the cases that would match the research questions and variables (2) Information richness and availability - which addressed whether the cases offer enough data and information to address the research question. Nissan Sentra has been one of the top-selling models according to Nissan Corporation for the fiscal year 2019-2023, while Samsung Galaxy A12 was the 6th bestselling smartphone for the year 2021 according to websites like thenationalnews.com, counterpointresearch.com and gadgets360.com. Best-selling global products are bound to be reviewed the most; therefore, they provide a good data source. Furthermore, the cases being complex products that contain many sub-components they provided rich information while maintaining broader applicability (3) Variation - The use of two diverse cases is important in enriching the analysis and providing a broader understanding of the phenomenon being studied. The use of two non-related product domains further validates the robustness of the model in handling diverse product domains (4) Significance: The cases are important and influential in the context of the study by virtue of commanding a large market share, and (5) Uniqueness – this study is being conducted in a niche area where there is the paucity of studies carried out before.

3.2.2. Handling Biases in Data

With any dataset, concerns arise from the biases associated with the dataset. Potential dataset biases identified with this study were Association bias, sample bias and Exclusion bias. During the data collection, inputs from a number of diverse sources were obtained to guarantee the diversity of data prior to preprocessing. Then, a thorough review of the collected and annotated data was conducted. This was done with the help of an independent research assistant from outside the team of authors who could identify any biases the authors had overlooked. The research assistant was able to verify the annotations for accuracy.

3.2.3. Ethics in Handling Data

This study was conducted with due caution, taking into account proper data protection measures to prevent data loss or leakage. This research took into account the following specific ethical considerations throughout the study: 1) data de-identification through the removal of names, account information, contacts, addresses, geolocations and all elements of dates (except year) for dates directly related to an individual and their accounts 2) confidentiality and privacy was ensured by having all the authors sign a Confidentiality and Non-Disclosure Agreement.

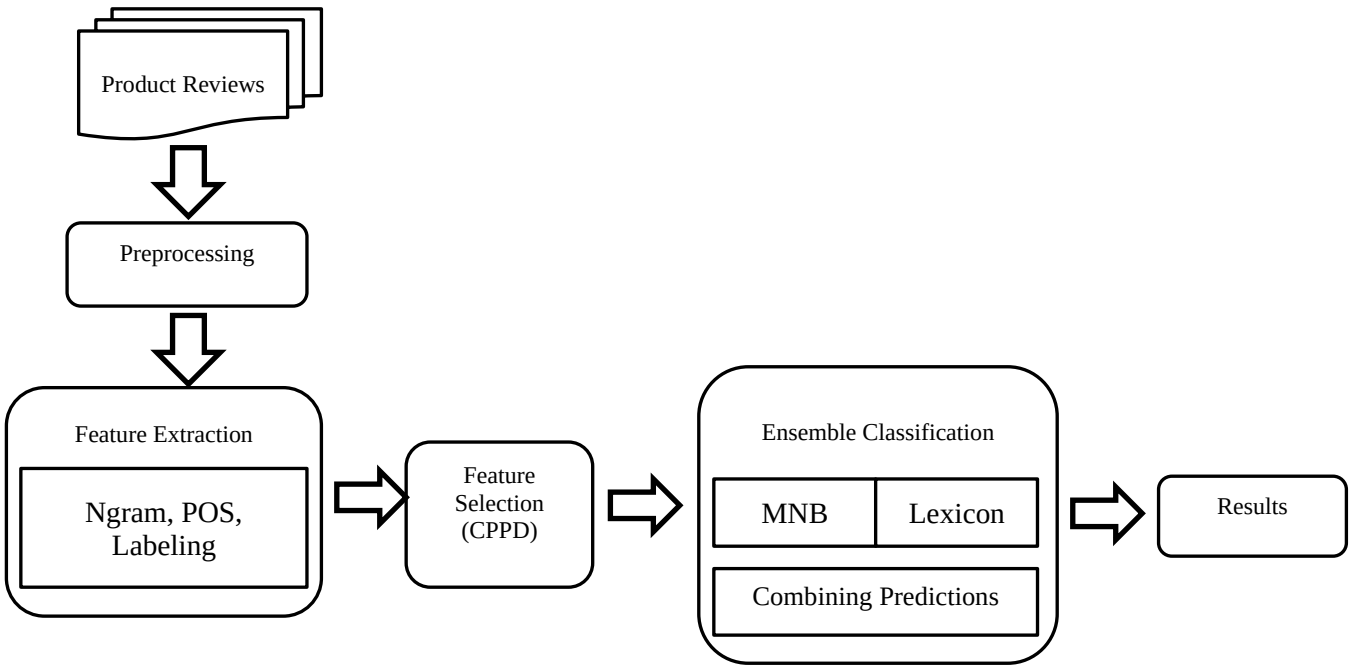

Fig. 2 Proposed ensemble architecture

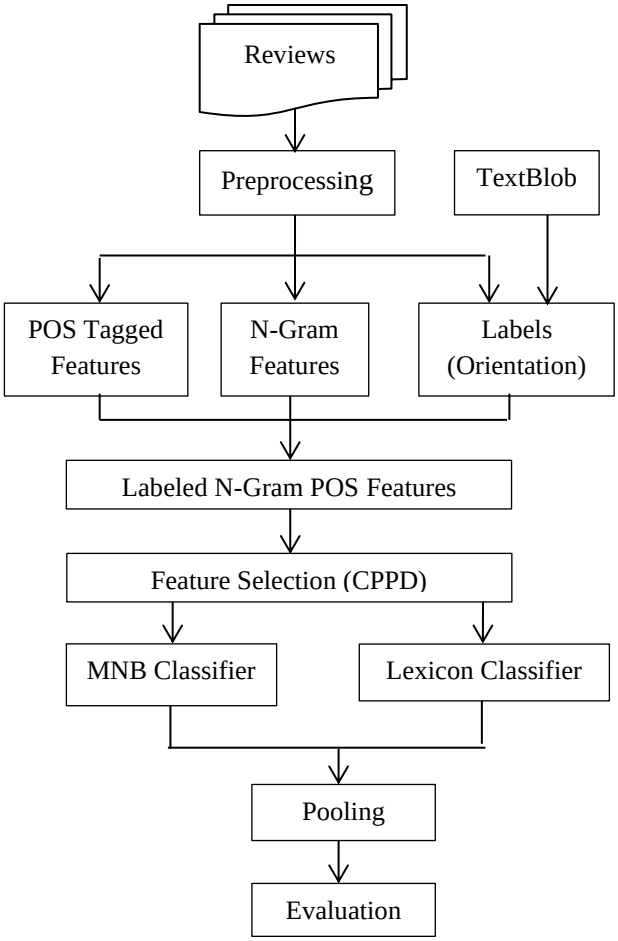

Fig. 3 Flowchart of proposed ensemble machine learning model

Table 2. Statistics of outcome of the cleaning process

Description	Samsung Galaxy (D1)	Nissan Sentra (D2)
Total Collected Reviews	37724	56913
Total Preprocessed Reviews	9431	25851

3.3. Preprocessing

First, a preliminary examination and processing of the data was carried out on the dataset before undertaking sentiment analysis tests. Given that consumer evaluation datasets are usually composed of colloquial English, a number of preparation processes were undertaken to normalize the text and make it ready for rating predicting and sentiment analysis. An outline of the pre-processing is shown below.

- Data Cleaning
- Contractions
- Lemmatization
- Removing non-English Reviews.
- Removing stopwords
- Correcting spellings
- Convert abbreviations
- Breaking attached words
- Negation handling

Pre-processing was a step that cleaned the dataset by eliminating extraneous letters, punctuation, other languages, difficult words, and stop words. It also prepared the datasets into formats that the base learners could consume.

The extracted data were saved in Excel format. Table 2 below shows the statistics of the outcome of the cleaning process.

3.4. Feature Extraction

After preprocessing, features were extracted to prepare the input for classification. Features such as Ngram (unigram, bigram, trigram), Part of Speech (POS) and features based on lexicon linguistic resources such as SentiWordNet, TextBlob, and Stanford coreNLP [25] are popular feature extraction techniques used in sentiment analysis tasks. Feature extraction techniques relied on POS tagging and Ngram as features, and NLTK natively supports input in the form of words. After pre-processing, the data was transformed; each row was enclosed in a tuple, the first index holding a dictionary containing individual words and their part of speech tag in the form of {unigram:pos_tag, unigram:pos_tag}. The second index of the tuple represents the class to which the review belongs, i.e. whether positive, negative or neutral. For instance:

$$original\ data = good\ phone$$
$$transformed\ data = (\{good: adj, phone: n\}, positive)$$

3.4.1. N-gram Features

N-grams, a continuous series of n consecutive symbols [26, 27], were extracted to a maximum weight of 3 grams. Further research [28, 29, 30] has demonstrated that N-grams are useful features for identifying the meaning of words in their context. As a result, this work utilized a word-based n-gram model to extract 1, 2 and 3-gram features in order to highlight relationships between words and the significance of specific phrases [30, 31].

3.4.2. POS Features

The Part of speech tag (Pos_tag) provides tags to every word in a sentence [32] which was used for classification. The output is in the form (unigram:pos_tag). The tags align with the standard POS groups found in the English language, including conjunction, interjection, noun, verb, adjective, adverb, POS prepositions, and pronouns. POS is useful in identifying candidate attributes that indicate sentiment orientation since it can identify sentiment expressions and their semantic connections. The following two factors make the POS tagger crucial: 1) most words, such as pronouns and nouns, are sentiment-less [33]. As a result, a POS tagger can filter out words like these; 2) A POS tagger can also help differentiate words that are used in distinct parts of speech. For

instance, as an adjective, the word "improved" could convey a different degree of sentiment than it does as a verb. Rule-driven POS tagging assigns tags for POS to phrases in a sentence using a collection of linguistic principles and patterns, the POS tagger employed in this work. A dictionary of words, associated POS tags and a predetermined set of grammatical rules are the foundations of this approach. In this study, pos_tag(), a function of the Python NLTK library that makes use of the Penn Treebank POS function, was used [34].

3.4.3. TextBlob Features

TextBlob, a popular library in Python, has an intuitive API that other applications can use to assess text sentiment and carry out other typical NLP operations like tokenization and POS tagging. There are two versions of its sentiment analyzer: one is drawn from a set of semantic patterns, while the other is drawn from a Naïve Bayes learning module. TextBlob provides sentiment analysis outcomes as a numerical polarity, with values ranging from -1 denoting highly negative to 1 denoting highly positive. A companion subjectivity score, ranging from 0, denoting extremely objective, to 1, denoting highly subjective, is also generated by it.

3.4.4. Feature Extraction Process

Feature extraction relied on Ngram, POS and features based on the lexical resource TextBlob. While the NTLK library supports Ngram, POS, and sentiwordnet, the TextBlob library was imported to support TextBlob and Stanford feature extraction. Figure 4 shows the importation of libraries.

```
import json

import nltk
from nltk import ngrams, pos_tag
from nltk.corpus import wordnet
from nltk.corpus import sentiwordnet as swn
from textblob import TextBlob
```

Fig. 4 Code snippet showing importation of libraries to support feature extraction

A POS tagger was crucial for sentiment classification for two reasons: 1) most words, such as pronouns and nouns, are sentiment-free [33]. Thus, the use of a POS tagger was able to filter out such terms; 2) A POS tagger can also help differentiate words that can be used in various parts of speech. For example, "improved" as an adjective may convey a different level of sentiment than it does as a verb. Rule-based POS tagging, a POS tagger that assigns POS tags to words in a phrase based on a set of linguistic rules and patterns, was the one employed in this study. This approach was based on a dictionary of words, associated POS tags, and a predetermined set of grammatical rules. The pos_tag() function from the Python NLTK library was used; it makes use of the Penn Treebank POS function [34]

For N-gram feature extraction, this work employed an n-gram model based on words to extract 2-gram and 3-gram combinations, revealing relationships between words and the significance of specific phrases [30][31]. Finally, the last step of feature extraction involves annotation. Extracted reviews were annotated and leveraged for the purposes of understanding the respective opinions on the two products and obtaining the best feature set combination for the study. Using Python libraries and functions, the study built a model to extract Ngram, POS features and labels using a TextBlob as a Senti-Analyzer. The labeled feature set served as input to the classifiers. Figure 5 shows a code snippet for implementation feature extraction. Figure 6 shows the output of the process and the respective labels from Stanford, VADER and TextBlob.

3.5. Feature Selection

A learning task's accuracy and efficiency are improved through feature selection [35]. The study used feature selection techniques based on Categorical Probability Proportion Difference (CPPD), which incorporates the best aspects of both PPD and CPD approaches while removing their drawbacks. The CPD approach has the advantage of measuring a term's degree of class-distinguishing quality, a critical component of a notable feature. Because CPPD is computationally very efficient [26], it improves classification performance above baseline results by filtering out features that are not relevant.

```
#feature extraction
try:
    fulldf_without_emoji = pd.read_csv("training/features.csv") #reads extracted files from memory
except: #if not in memory it extracts again
    fulldf_without_emoji['bigram'] = fulldf_without_emoji['review'].apply(lambda x: fe.get_gram(x, 2))
    fulldf_without_emoji['trigram'] = fulldf_without_emoji['review'].apply(lambda x: fe.get_gram(x, 3))
    fulldf_without_emoji['pos_tag'] = fulldf_without_emoji['review'].apply(lambda x: fe.get_postag(x))
    fulldf['textblob_polarity'] = fulldf['review'].apply(fe.textblob_sentiment)
    fulldf['textblob_sent']=fulldf['textblob_polarity'].apply(fe.polarity)
    fulldf_without_emoji.to_csv("training/features.csv", index=False)
```

Fig. 5 Code snippet for feature extraction implementation

	review	bigram	trigram	pos_tag	textblob_polarity	textblob_sent
2	phone lag much ofte	[('phone', 'lag'), ('lag', 'much	[('phone', 'lag', 'much'), ('la	[('phone', 'NN'), ('lag', 'NN'),	-0.066666667	negative
3	slowest phone ever	[('slowest', 'phone'), ('phon	[('slowest', 'phone', 'ever'),	[('slowest', 'NN'), ('phone', '	0	neutral
4	samsung galaxy a go	[('samsung', 'galaxy'), ('gala	[('samsung', 'galaxy', 'a'), ('g	[('samsung', 'NN'), ('galaxy',	0.5	positive
5	samsung make lot ga	[('samsung', 'make'), ('make	[('samsung', 'make', 'lot'), ('	[('samsung', 'NNS'), ('make',	-0.25	negative
6	can not wait see new	[('can', 'not'), ('not', 'wait'), (	[('can', 'not', 'wait'), ('not', '	[('can', 'MD'), ('not', 'RB'), ('w	0.136363636	positive
7	a not anymore get a	[('a', 'not'), ('not', 'anymore'	[('a', 'not', 'anymore'), ('not	[('a', 'DT'), ('not', 'RB'), ('anyn	0	neutral
8	do be_born series gc	[('do', 'be_born'), ('be_born'	[('do', 'be_born', 'series'), ('	[('do', 'VB'), ('be_born', 'JJ'),	0.1	positive
9	still lg stylo mediate	[('still', 'lg'), ('lg', 'stylo'), ('st	[('still', 'lg', 'stylo'), ('lg', 'sty	[('still', 'RB'), ('lg', 'JJ'), ('stylo	-0.15	negative

Fig. 6 Screen capture of Ngam, POS and TextBlob Sentiment Captured from D1

3.6. Classification

The study selected the MNB and Lexicon classification techniques since they are both popular and have demonstrated strong performance in sentiment analysis applications.

3.6.1. MNB

MNB, which is founded on Bayes' theorem and is typically employed for tasks such as text classification that require dealing with discrete data, has been shown to yield better results for sentiment analysis since it assumes feature independence, which implies the existence of one feature will not impact the existence of another. MNB is used to verify the multi-class classifier's categorization. This probabilistic strategy has two phases: training and testing. Formula (1) below computes the chance of each word in a class during training.

$$P(t|c) = \frac{T_{ct}}{\Sigma_{t\prime \epsilon V} T_{ct\prime}} \quad (1)$$

Where T_{ct} is the sum of the times a word t appears in class c of the training document, and $\Sigma_{t\prime \epsilon V} T_{ct\prime}$ is the sum of attributes in class c. The attributes are the sum of words in class C and the aggregate total of words in the lexicon. However, when T_{ct} is equal to zero or when words exist that are absent from the training set. By adding 1, Laplace smoothing eliminates zero values in formula (2).

$$P(t|c) = \frac{T_{ct}+1}{\Sigma t\prime \epsilon V (T_{ct\prime}+1)} = \frac{T_{ct}+1}{(\Sigma t\prime \epsilon V T_{ct\prime})+B\prime} \quad (2)$$

Data Splitting

Using the dataset, experimentation was run on the following tasks: sentiment categorization and evaluation. The classification of emotion seeks to forecast the polarity of reviews. The dataset was divided using the 80:10:10 ratio into training, test and validation sets for the MNB sentiment classification task. From the cleaned D1 dataset, 7,545 reviews were used for training, 943 reviews were used for validation, and 943 were used for testing. For the D2 dataset, 20,689 reviews were used for training, 2581 reviews were used for validation, and 2581 were used for testing.

3.6.2. Lexicon

Sentiment Analysis using the Lexicon-based method is a natural language analysis approach for determining the emotional polarity of a document. This technique derives sentiment orientations for the entire document or group of sentences from lexical-semantic orientations. The semantic orientation can be positive, negative, or neutral. It employs a dictionary and includes the polarity of the term. The score for sentiment is calculated whenever a phrase is found in a text and is contrasted to an analogous lexicon word. A lexicon-based technique is utilized to assess emotion and determine the sum of polarities encountered in an entire text work.

The three stages of the Lexicon-based assessment are word, sentence, and document-level calculation. A word in a review is compared to a term in a dictionary using word-level calculation, also known as lexical comparison, depending on the section of speech that each word contains. Formulae (3) and (4) are used to determine the sentence-level polarity:

$$\frac{1}{nk}\sum_{i=1}^{nk} PositiveSentence(i) \quad (3)$$

$$\frac{1}{nk}\sum_{i=1}^{nk} NegativeSentence(i) \quad (4)$$

Where:

PositiveSentence(i) and *NegativeSentence(i)* are word-ith positive and negative rankings according to a vocabulary lexicon, and the sum of words within a sentence is denoted by *nk*. The document polarization is determined by computing the document score. The overall average of phrases from each polarity is considered in document polarity. Formulae 4 and 5 are used to calculate the document polarity:

$$\frac{1}{ns}\sum_{i=1}^{nk} PositiveDocument(i) \quad (5)$$

$$\frac{1}{ns}\sum_{i=1}^{nk} NegativeDocument(i) \quad (6)$$

Where:

PositiveDocument(i) and *NegativeDocument(i)* are positive and negative sentence-ith scores. The sentence scoring is added up through the nth index. The overall number of sentences in a document is denoted by *ns*. Using a comparison of positive and negative document text ratings, the sum polarity of the text may be established. A piece of literature is given a positive polarity if its positive ratings exceed its negative ones and the other way around.

Lexicon-based methods have low accuracy, poor recall and limited coverage of sentiment words for multiple domains [36]; however, they are quick to compute since they do not need data training. The foundation of the Lexicon strategy is the premise that the inclination of the emotion of each word, phrase, or part of speech feature (for instance, an adjective, an adverbial combination, or an adverbial verb combination) that appears in a given text adds up to the sentiment orientation context of that phrase, aspect, or document [37]. Lexical resources, POS Tagger, and an effective technique for calculating the contextual sentiment value of a feature are necessary to implement the lexicon-based method. This approach's accuracy depends on the dictionary, feature extraction method, and sentiment score calculation strategy. Sentiment detection and word labeling were accomplished by utilizing the NLTK-embedded native library, SentiWordNet.

3 6.3. MNB + Lexicon Ensemble

Ensemble models and hybrid models are used to combine single classification algorithms and techniques, albeit in slightly different ways, to mask the drawbacks of each strategy and increase the accuracy of the outcome [26].

3.6.4. Pooling

Lexicon pooled is a formula that combines the MNB and Lexicon-based probabilistic scores [18]. This study used the Lexicon pooled equation to aggregate the probability scores between the MNB and the Lexicon-based approach. By applying linear pooling [38], which is determined by using equation (7) below, the lexicon pooled equation.

$$P(t_i|c_p) = \alpha_{MNB} P_{MNB}(t_i|c_p) + \alpha_{LB} P_{LB}(t_i|c_p) \quad (7)$$

The approaches are MNB and LB (Lexicon Based), and $P(t_i|c_p)$ is the term i's likelihood in class c_p . $\propto_{MNB}$ and $\propto_{LB}$ are given weights for the MNB approach and the lexicon-based methods. The weight of each approach was determined using the formulae (8) and (9) below:

$$\alpha_{MNB} = \log\left(\frac{acc_{MNB}}{1-acc_{MNB}}\right) \quad (8)$$

$$\alpha_{LB} = \log\left(\frac{acc_{LB}}{1-acc_{LB}}\right) \quad (9)$$

Where acc_{LB} and acc_{MNB} are the accuracies of LB and NB, respectively, on the training set.

3.7. Evaluation and Analysis

Accuracy, precision, recall, and F-1 score were employed to evaluate performance on sentiment analysis and rating prediction tasks. Equation (10), which shows the percentage of the sum of things rightly classified as the aggregate sum of all objects, is used to calculate accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (10)$$

Recall is the proportion of accurately predicted positives relative to all positive class items, whereas precision measures the sum of properly classified positive class objects relative to the predicted positives. These measurements' formulas are provided in (11) and (12).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (11)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

Where:

- True positives (TP): The instances where the reviews are expected to be good, and they are.
- True negatives (TN): Instances where reviews are expected to be negative and turn out to be negative.
- False positives (FP): The instances where the reviews are expected to be negative but were very positive.
- False negatives (FN): Instances in which positive comments are expected notwithstanding the initial negative assessment.

The F1 score is a harmonized average of recall and precision shown in equation (13).

$$F1\ Score = 2\ x\ \frac{(Recall\ x\ Precision)}{Recall + Precision} \quad (13)$$

4. Results and Discussion

Results for the datasets were computed. The results are discussed in form or polarity classification, accuracy and precision, recall, and F1 as a cluster.

4.1. Polarity Classification

The sentiment polarity classification task of reviews was performed, focusing on subjectivity classification. Using the proposed model, the study proceeded to obtain the polarity of the reviews into positive, negative, or neutral polarities. Figures 7 and 8 show the overall polarity classification of the reviews.

4.2. Accuracy

Table 3 below displays the accuracy results of the individual classifiers MNB and Lexicon and the proposed Lexicon pooled MNB Ensemble when feature selection is not applied, while Table 4 presents the accuracy levels with the application of feature selection.

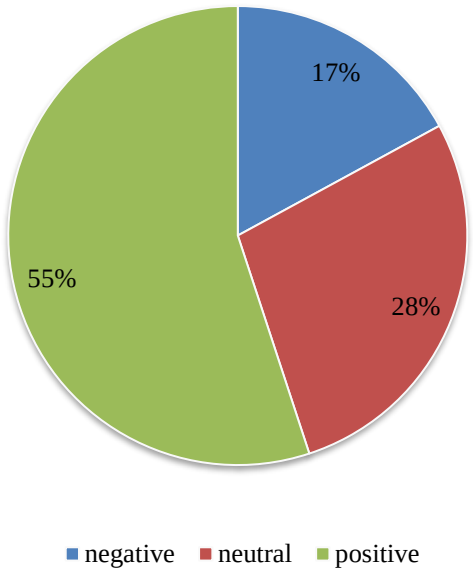

Fig. 7 Polarity classification for dataset D1

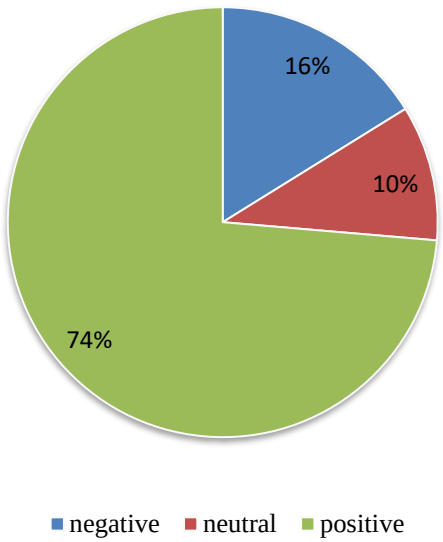

Fig. 8 Polarity classification for dataset D2

Table 3. Accuracy without Feature Selection

Dataset	Algorithm	Train Accuracy	Test Accuracy	Val Accuracy
D1	MNB	0.8598	0.7016	0.7172
	Lexicon	0.5828	0.5760	0.6028
	Proposed	**0.8604**	**0.7059**	**0.7225**
D2	MNB	0.8563	0.7823	0.7857
	Lexicon	0.6255	0.6141	0.6150
	Proposed	**0.8574**	**0.7842**	**0.7870**

Table 4. Accuracy with feature Selection

Dataset	Algorithm	Train Accuracy	Test Accuracy	Val Accuracy
D1	MNB	0.8381	0.8240	0.8395
	lexicon	0.7450	0.7515	0.7563
	Proposed	**0.8389**	**0.8250**	**0.8395**
D2	MNB	0.8314	0.8121	0.8178
	lexicon	0.7114	0.6886	0.6978
	Proposed	**0.8364**	**0.8174**	**0.8215**

Tables 3 and 4 demonstrate that in this experiment, Multinomial Naive Bayes outperforms Lexicon-based accuracy in terms of accuracy. Due to the fact that the n-grams utilized for this system were unigram, bigram, and trigram, which allowed for the calculation of several word combinations, multinomial Naive Bayes demonstrated greater accuracy than the lexicon-based approach. MNB's accuracy was shown to have increased as a result of the lexical pooled results. This is because the lexicon-based methods can handle words not necessarily part of the training data. While the lexicon-based model by itself does not yield particularly high accuracy values, it is observed that Lexicon pooling with MNB yields superior accuracy. For datasets D1 and D2, respectively, lexicon pooling without feature selection improved MNB test accuracy from 0.7016 to 0.7059 and from 0.7823 to 0.7842. MNB test accuracy increased for datasets D1 and D2 when feature selection was used, going from 0.8121 to 0.8174 and from 0.8240 to 0.8250, respectively. The best test accuracy obtained is 0.8250. The overall test accuracy improvement ranged from 0.0010 to 0.0053. The SentiWordNet lexical resource does not contain all words, which results in decreased accuracy when using the Lexicon technique. SentiWordNet has 155,327 terms, while the Oxford English Dictionary has 171,476 terms. This represents a 9.42% difference. Lexicon provides a zero value when a word cannot be located, which impacts the document's polarity computation.

During this phase, various methodologies were used to help the model improve its ability to classify reviews from text. These methods are explained as follows: 1) Preprocessing: The first method was through preprocessing. Under this, methods applied included negation handling, removing irrelevant reviews, and changing words to suit the context. For instance, on Sentra reviews, the text "I love this beast" was changed to "I love this car". 2) Hyperparameter Tuning: The second approach involved hyperparameter tweaking. These were external configurations used to handle the training process of the supervised classification model, such as the settings set before training began and remained constant throughout. During this phase, the alpha learning rate was adjusted to between 0 and 1. Default 1 was found to be performing better. 3) Resolving the Issue of Zero Observations: In situations when the test set and training data have different frequency distributions, the Naive Bayes classifier typically performs poorly. Values not represented in the training set have a notably negative impact on the classifier. A new category is given a probability of 0 if the model finds a categorical feature absent from the training set. This is undesirable since 0 will be the outcome of multiplying 0 by the probabilities of other attributes. The zero observation problems persist even when using the log probability. Due to the fact that log(0) = infinity and summation will eliminate all of the useful data from other characteristics. In cases where the test data set had zero frequency issues, Laplace Smoothing was applied to eliminate the Zero Observations Problem. In this technique, a parameter is added to both the numerator and denominator when calculating the class probabilities. It is ensured that the probability value is never 0 by the smoothing parameter. Using a smoothing technique, the Naive Bayes classifier is made more regular by giving such zero-frequency occurrences a very small probability value. This is explained by Equation 2. 4) Ensemble learning: Performance improved

with ensemble learning. Boosting, pooling, stacking, and bagging are among common ensemble techniques that combine the output of several models to get a new result. The primary goal of combining the data is to minimize variance. Comparing this proposed technique to other related researchers' methods, this proposed system performed comparatively well, as displayed in Table 5 and Figure 9.

Table 5. Accuracy results comparison

	Author	Method	Accuracy
1	Trinh et al [21]	SVM+VED	0.8193
2	Romadhony et al [22]	MNB, SVM), LSTM, BiLSTM)	0.7300
3	Dada, et al [24]	(kNN), (GLMNet), s (LDA)	0.8856
4	Barik et al. [20]	IVADER Lexicon	0.9864
5	Sinha & Narayanan [18]	HLESV on Electronics	.0.7000
	Sinha & Narayanan [18]	HLESV on Gift Cards	0.8700
6	Geriska et al. [19]	Lexicon-based+ without AAAVC	0.7078
	Geriska et al. [19]	Lexicon pooled+ + AAAVC	0.6371
7	Fayaz et al [23]	MLP, KNN, RF	0.8813
8	Proposed (D1)	MNB + Lexicon Pooled	0.8250
9	Proposed (D2)	MNB + Lexicon Pooled	0.8174

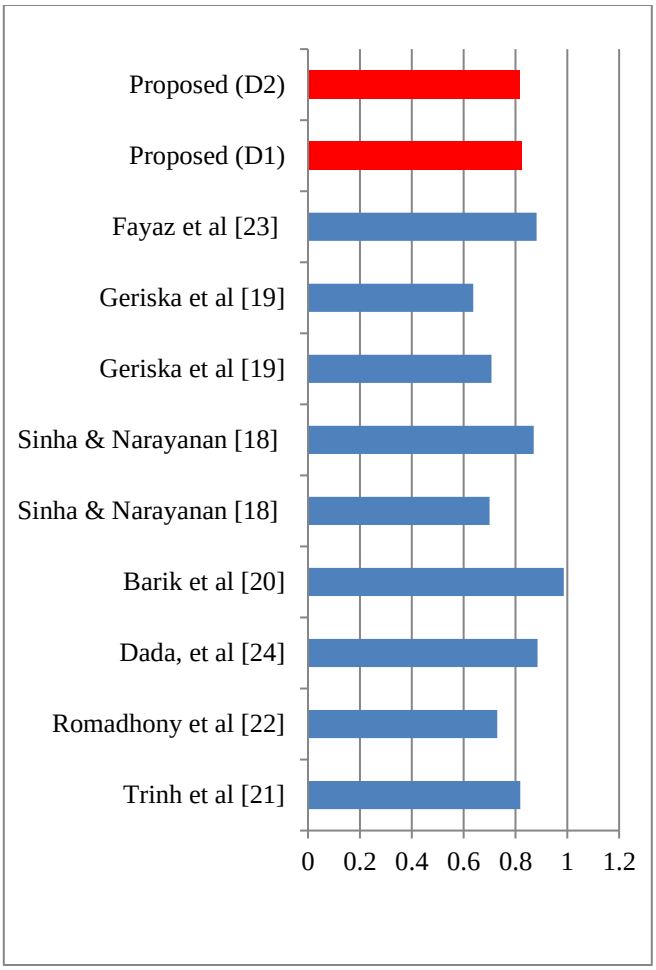

Fig. 9 Graph showing accuracy results comparison

Table 6. Precision, Recall and F1

	MNB Only			Lexicon Only			Proposed (MNB+Lexicon Pooled)		
Dataset	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
D1	0.8994	0.7994	0.8361	0.4180	0.4732	0.3913	0.8932	0.7970	0.8325
D2	0.8713	0.6322	0.689	0.3999	0.4556	0.3855	0.8703	0.6331	0.6882

It is believed that this ensemble produced very competitive results, even though a direct comparison between these systems and the proposed system is not possible due to the usage of different datasets. The best accuracy achieved was 82.5%.

4.3. Precision, Recall and F1

The study evaluated precision, recall and F1 score for the model, as presented in Table 6. Lexicon performance metrics are low because the lexical library used, SentiWordNet, does not apply the magnitude of the emotions. For example, "good" and "amazing" both have a positive polarity, but the latter has a higher intensity. Also, Polarity shifts, which changes in the emotional direction of a word or a phrase due to the presence of negators, intensifiers, diminishers, or contrastive conjunctions, are not considered.

An F1 score nearer 1 denotes a more successful model, where recall and precision are both high; on the other hand, an F1 value nearer 0 denotes a less successful model or the inability of the model to predict on at least one of the classes.

The proposed models had a satisfactory F1 score of 0.8325 for dataset D1 but a slightly lower F1 score of 0.6882 for D2. This is because the dataset D2 is too unbalanced, as evidenced by Figure 6, and the model was unable to learn perfectly how to predict one or more of the classes.

5. Conclusion

This research proposes an innovative Ensemble machine-learning model. Two datasets were used to assess the practicality of the suggested framework. In the first mode, feature selection was not used; however, in the second mode, it was. For the two datasets employed, the experimental findings of the Ensemble model demonstrate encouraging results in the sentiment prediction of product reviews. Despite the fact that the performance difference between Lexicon alone is several orders of magnitude lower than MNB alone, the most significant finding from the empirical data presented in Tables 3, 4, 5, and 6 was as follows:

1. The study's most pertinent conclusion was that the ensemble machine-learning approach improved sentiment analysis prediction accuracy for product reviews. Compared with similar run parameters of MNB and Lexicon alone, the Lexicon pooled MNB ensemble achieves greater accuracy.
2. Feature selection significantly enhanced the prediction accuracy of the proposed model by reducing overfitting and eliminating features that do not contribute to the predictive power of the model.

This study's findings are significant because they advance the argument that the use of data science and ensemble learning, to be specific, can provide a vital novel approach for enhancing the product engineering process and thereby reduce the product failure rate. Thus, this work contributes concisely to ongoing efforts to comprehend how consumers assess products and make purchasing choices and to ensure adequate information necessary for the product development life cycle reaches the product developers early enough before the market evolves. In other words, this work provides an avenue for deliberate and continuous engagement of consumers in creating new products.

Enhancements for this model could include applying polarity shift to improve Lexicon accuracy by detecting and adjusting the polarity of words and phrases based on these modifiers and combining a Word Sense Disambiguation (WSD) algorithm with SentiWordNet to get the most promising meaning. Further research work may concentrate on comparing this suggested model to ensemble deep learning models and offering more comprehensive comparisons of ensemble learning algorithms' performance. Future research could also explore the study's applicability to other relevant topics, such as applying sentiment analysis to online news sources and readers' comments to vet the suitability of candidates seeking public office. In addition, as a complimentary, future work can explore the impact of fake reviews and fake news detection, both of which are spreading at an alarming rate in public online places.

References

[1] Caroline Blais, and Raymond K. Agbodoh-Falschau, "An Exploratory Investigation of Performance Criteria in Managing and Controlling New Product Development Projects: Canadian SMES' Perspectives," *International Journal of Managing Projects in Business*, vol. 16, no. 6/7, pp. 788-807, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Edim Eka James, Altuğ Ocak, and Samuel Eventus Bernard, "Exploring the Dynamics of Product Quality and Failures in Export Trade: A Systematic Literature Review," *International Journal of Science and Research Archive*, vol. 12, pp. 272-306, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Garmt Dijksterhuis, "New Product Failure: Five Potential Sources Discussed," *Trends in Food Science & Technology*, vol. 50, pp. 243-248, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[4] Marianna Kazimierska, and Magdalena Gręboszc-Krawczyk, "New Product Development (NPD) Process – An Example of Industrial Sector," *Management Systems in Production Engineering*, vol. 25, no. 4, pp. 246-250, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[5] Dagmara Skurpel, "Advantages and Disadvantages of Internet Marketing Research," *World Scientific News*, vol. 57, pp. 712-721, 2016. [Google Scholar] [Publisher Link]

[6] Sertac Eroglu, and Nihan Tomris Kucun, "Traditional Market Research and Neuromarketing Research: A Comparative Overview," *Analyzing the Strategic Role of Neuromarketing and Consumer Neuroscience*, pp. 1-22, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Huimin Lu et al., "Brain Intelligence: Go Beyond Artificial Intelligence," *Mobile Networks and Applications*, vol. 23, no. 2, pp. 368-375, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Tarun Kumar Vashishth et al., "AI and Data Analytics for Market Research and Competitive Intelligence Final," *AI and Data Analytics Applications in Organizational Management*, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Federico Neri et al., "Sentiment Analysis on Social Media," *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, Turkey**,** pp. 919-926, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[10] Michael Etter et al., "Measuring Organizational Legitimacy in Social Media: Assessing Citizens' Judgments with Sentiment Analysis," *Business* & Society, vol. 57, no. 1, pp. 60-97, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Lei Zhang et al., "Combining Lexicon-Based and Learning Based Methods for Twitter Sentiment Analysis," *HP Laboratories, Technical Report*, vol. 89, pp. 1-8, 2011. [Google Scholar]

[12] Ludmila I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, pp. 1-300, 2004. [Google Scholar] [Publisher Link]

[13] Oscar Castillo, Patricia Melin, and Witold Pedrycz, *Hybrid Intelligent Systems: Analysis and Design*, Springer, pp. 1-433, 2007. [Google Scholar] [Publisher Link]

[14] Emilio Corchado, Ajith Abraham, and Andre de Carvalho, "Hybrid Intelligent Algorithms and Applications," *Information* Sciences, vol. 180, pp. 2633-2634, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[15] Simona Valentina Pascalau, and Ramona Mihaela Urziceanu, "Traditional Marketing versus Digital Marketing," *Agora International Journal of Economical Sciences*, vol. 14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Jyoti Thakur, and Bijay Prasad Kushwaha, "Artificial Intelligence in Marketing Research and Future Research Directions: Science Mapping and Research Clustering Using Bibliometric Analysis," *Global Business and Organizational Excellence*, vol. 43, no. 3, pp. 139-155, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sourav Sinha, and Revathi Sathiya Narayanan, "A Novel Hybrid Lexicon Ensemble Learning Model for Sentiment Classification of Consumer Reviews," *Journal of Internet Services and Information Security*, vol. 13, no. 3, pp. 16-30, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Geriska Isabelle, Warih Maharani, and Ibnu Asror, "Analysis on Opinion Mining Using Combining Lexicon-Based Method and Multinomial Naïve Bayes," *Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering*, pp. 214-219, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Kousik Barik, and Sanjay Misra, "Analysis of Customer Reviews with an Improved Vader Lexicon Classifier," *Journal of Big Data*, vol. 11, pp. 1-29, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Son Trinh, Luu Nguyen, and Minh Vo, *Combining Lexicon-Based and Learning-Based Methods for Sentiment Analysis for Product Reviews in Vietnamese Language*, Computer and Information Science, Springer, Cham, pp. 57-75, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Ade Romadhony et al., "Sentiment Analysis on a Large Indonesian Product Review Dataset," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, pp. 167-178, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Muhammad Fayaz et al., "Ensemble Machine Learning Model for Classification of Spam Product Reviews," *Complexity*, vol. 2020, no. 1, pp. 1-10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Emmanuel Gbenga Dada et al., "Ensemble Machine Learning Model for Software Defect Prediction," *Advances in Machine Learning & Artificial Intelligence*, vol. 2, no. 1, pp. 11-21, 2021. [Google Scholar] [Publisher Link]

[24] Aleksandra Petrakova, Michael Affenzeller, and Galina Merkurjeva, "Heterogeneous versus Homogeneous Machine Learning Ensembles," *Information Technology and Management Science*, vol. 18, no. 1, pp. 135-140, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[25] Triyanna Widiyaningtyas, Ilham Ari Elbaith Zaeni, and Riswanda Al Farisi, "Sentiment Analysis of Hotel Review Using N-Gram and Naive Bayes Methods," *2019 Fourth International Conference on Informatics and Computing*, Semarang, Indonesia**,** pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Basant Agarwal, and Namita Mittal, "Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification," *Proceedings of the 2nd Workshop on Sentiment Analysis where AI Meets Psychology*, Mumbai, India, pp. 17-26, 2012. [Google Scholar] [Publisher Link]

[27] Lai Po Hung, Rayner Alfred, and Mohd Hanafi Ahmad Hijazi, "A Performance Comparison of Feature Selection Methods for Sentiment Classification," *Computational Science and Technology*, vol. 488, pp. 21-30, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] Michel Généreux, Thierry Poibeau, and Moshe Koppel, "Sentiment Analysis Using Automatically Labelled Financial News Items," *Affective Computing and Sentiment Analysis*, vol. 45, pp. 101-114, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[29] Zhongwu Zhai et al., "Exploiting Effective Features for Chinese Sentiment Classification," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9139-9146, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[30] Sepideh Foroozan Yazdani et al., "NgramPOS: A Bigram-Based Linguistic and Statistical Feature Process Model for Unstructured Text Classification," *Wireless Networks*, vol. 28, pp. 1251-1261, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Yelena Mejova, and Padmini Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 546-549, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ayman S. Ghabayen, and Basem H. Ahmed, "Polarity Analysis of Customer Reviews Based on Part-of-Speech Subcategory," *Journal of Intelligent Systems*, vol. 29, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[33] Pankaj et al., "Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 320-322, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[34] Neri Van Otten, Part-of-speech (POS) Tagging In NLP: 4 Python How To Tutorials, 2023. [Online]. Available: https://spotintelligence.com/2023/01/24/part-of-speech-pos-tagging-in-nlp-python/

[35] Mohammad Salim Hamdard, and Hedayatullah Lodinx, "Effect of Feature Selection on the Accuracy of Machine Learning Model," *International Journal of Multidisciplinary Research and Analysis*, vol. 6, no. 9, 4460-4466, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] G. Vaitheeswaran, and L. Arockiam, "Combining Lexicon and Machine Learning Method to Enhance the Accuracy of Sentiment Analysis on Big Data," *International Journal of Computer Science and Information Technologies*, vol. 7, no. 1, pp. 306-311, 2016. [Google Scholar] [Publisher Link]

[37] Madhavi Devaraj, Rajesh Piryani, and Vivek Kumar Singh, "Lexicon Ensemble and Lexicon Pooling for Sentiment Polarity Detection," *IETE Technical Review*, vol. 33, no. 3, pp. 332-340, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[38] Rupika Dalal et al., "A Lexicon Pooled Machine Learning Classifier for Opinion Mining from Course Feedbacks," *Advances in Intelligent Systems and Computing*, vol. 320, pp. 419-428, 2015. [CrossRef] [Google Scholar] [Publisher Link]