Amazon Reviews Classifications

Vellore Institutes of Technology Amaravati Campus

Abstract

Sentiment Analysis (SA), which is also known as Opinion Mining, is a hot-fastest growing research area, making it challenging to follow all the activities in such areas. It intends to study people's thoughts, feelings, and attitudes about topics, events, issues, entities, individuals, and their attributes in social media (e.g., social networking sites, forums, blogs, etc.) expressed by either text reviews or comments. Amazon is an example of the world's largest online retailer that allows its customers to rate its products and freely write reviews. Analyzing these reviews into positive or negative; will assist customers' decision making, which varies from purchasing a product like a camera, mobile phone, etc., to writing a review about movies and making investments - all of these decisions will have a significant impact on the daily life. Sentiment analysis draws the attention of both scientific and market research in Natural Language Processing and Machine Learning fields. In general, the machine learning approach consists of supervised and unsupervised algorithms. In this research study, a detailed typical workflow process often adopted by the researchers is presented. Moreover, traditional supervised machine learning classification techniques have been investigated on various categories of Amazon product reviews to find the best method that provides a reliable result of sentiment analysis and assists future research in this newly emerging area.

Keywords— *Sentiment analysis, Sentiment classification, opinion mining, machine learning, polarity classification, supervised algorithms, Amazon classification.*

I. INTRODUCTION

In modern times, social media and online shopping play a vital role in connecting people around the world and help them express their feeling and opinion about any topic on the globe. Amazon.com, as an example, is one of the reputed online retailers nowadays. It allows its users to post and share their opinions about its products freely. As a result, a massive amount of structured and unstructured data generated. Sentiment analysis is utilized to study and analyse these data and finds valuable insights beyond people's thinking and feeling about anything that can help in decision making. Sentiment analysis or SA as a shortcut is considered as a text classification task, which is a subfield of natural language processing (NLP). NLP is used by machine learning techniques to understand, analyse, and gain in-depth meaning from a human language with intelligence [1]. Recent supervised classification approaches that have been used to find sentiment analysis in Amazon product reviews are surveyed in this paper to discover the best one that can provide reliable and accurate results. This approach can then be used as a baseline for Amazon reviews, classification tasks, recommendation systems [2], etc.

This research study has been structured as follows: Section 2 presents an overview of Amazon. Section 3 introduces sentiment analysis, its levels, and approaches .Section 4, 'literature review' summarizes some related work on sentiment analysis, followed by section 5, the typical sentiment analysis methodology. The last parts 6 and 7 cover the discussion and the conclusion of our study as well as address some future directions for research, respectively.

II. *AMAZON BACKGROUND*

Amazon is one of the world's largest online retailers. It had overgrown since 1994 when it launched as an online platform. Currently, It provides over 12 million different products and had 150.6 million active mobile users (Statista, 2019), so it is considered a microcosm for excellent usersupplied reviews. Amazon sells various products like books, phone apps, movies, clothes, electronics, toys, etc. , and makes use of the universal rating system -from 1 to 5 stars; lowest to highest score, respectively, and writing product review text. This scoring system has no user guide on how customers should use it; besides, the product reviews are subjective and personal. Therefore a user can give a good product a score of "1", but a bad buying experience, such as late delivery, or high price, and vice versa. The absence of guidelines makes it difficult to identify the user's sentiments about

different product aspects and parts of a shopping experience [3]. Another challenge is the J-shaped distribution, with an overwhelming majority being positive reviews of the sampled reviews crawled by Amazon's standard identification number (the ASIN).

III. *SENTIMENT ANALYSIS*

Sentiment analysis or subjectivity analysis is the top research field under NLP. It utilizes textual data available almost on social media like Amazon to analyse people's opinions and concentrates on the subject part of the text (phrase or sentence) that leads to a good or bad direction [4][5]. SA plays an essential role in the business domain by providing the businesses with in-depth insight into the attitude of buyers' feedback about their product; therefore, they can improve their strategy to meet the customer's expectations & needs and avoid loss. On the other hand, it is helpful for potential customers to decide about the products they are going to buy.

Star Level	Example
★	I hate it.
★★	I don't like it.
★★★	It's ok.
★★★★	I like it.
★★★★★	I love it.

Fig. 1 Product rating & Review example

The typical sentiment analysis task involves taking a piece of text, whether it's a word, sentence, or an entire document; classify it into either binary classification or multi-classification using a Machine Learning classification model that results in a score that measures the text polarity. That is a big challenge zone, due to the structure of the text that may include negation, comparative, slangs, domain, multi-language [6], etc.

A. Sentiment analysis levels

In general, sentiment analysis is mainly explored into three different levels based on the text scope, namely, A) document, B) sentence, and C) feature level [5][7]. (A) The task at the document level is to determine whether the overall opinion of the document expresses a positive or negative sentiment about a single entity. In contrast, the Sentence level of analysing (B) is concerned with determining whether each sentence in the document(s) holds a positive, negative, or neutral opinion. Whereas the previous analysis levels cannot perform, (C) Entity and Aspect level analysis since they are focusing directly on identifying the people's opinion about specific aspects either preferred or not. It is also called phrase-level sentiment analysis and feature level analysis [5] [8]. It is employed when conducting sentiment analysis on reviews like mobile phones and restaurants.

B. Approaches of sentiment analysis

In general, two main conventional approaches are employed in solving sentiment analysis challenges; (A) Machine Learning (ML) methods based, (B) Lexicon based [9], Fig. 2. All these approaches applied to a piece of text- a product review or comment, to distinguish between positive and negative opinions appear in that text.

1. Machine learning Approaches

These approaches tackle the problem of how the computer program itself can learn to identify complicated patterns and create intelligent-decisions based on data – text review. It mainly comprises supervised learning and unsupervised learning approaches Fig. 2. The supervised learning employs ML classification techniques, whereas the unsupervised methods employ clusters that imply Lexicon approaches.

A. Supervised machine learning

In the supervised learning approach, they mainly focus on the classification of data. It often requires an extensive labelled training dataset to teach an algorithm- how each word (sequence) in a text corresponds to the overall sentence's outcome is negative or positive in a supervised manner. Examples of benchmark supervised techniques are Decision Tree DT, Naïve Bayes NB, Maximum Entropy ME, Support Vector Machine SVM, etc. This approach requires manually labelled data, which is not always possible, and often time-consuming [10].

B. Unsupervised machine learning

In contrast, in the unsupervised approach, they concentrate on the grouping of unsorted data according to the similarities or differences without any prior knowledge about the data that is not labelled- no training of data given to the machine.

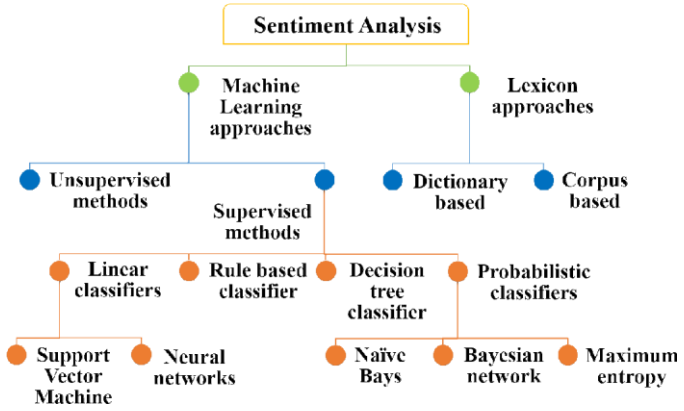


Fig. 2 Sentiment Analysis Approaches

It allows the application of traditional unsupervised clustering types like Hierarchical, K-means, K Nearest Neighbours (KNN), Principal Component Analysis (PCA), etc. , to that information without human interference. This approach is beneficial once a lack of labelled data [11].

ML algorithms are used in sentiment analysis to recognize the sentiment that appears in the given text according to the word patterns, their order, and a sentimentlabelled training set in supervised learning, and similarities and differences in unsupervised learning case. These approaches aim to automate expensive manual tasks or timeconsuming ones.

2. *Lexicon-based*

This approach aims to find the lexicon containing the opinion and then analyse it by using, for example, a dictionary of antonyms and synonyms for the opinionated words and phrases with their corresponding sentiment scores. Furthermore, it is classified into the corpus and dictionarybased approaches [12].

A. Dictionary-based

Generally, opinionated dictionaries such as WordNet, SentiWordNet, or online dictionaries contain a list of negative and positive sentiments. This approach intends to search the terms that hold subjective meaning if found in the text, match them with the words listed in that dictionary, and return their corresponding scores. This approach is unable to find the domain or context-specific opinion [13].

B. Corpus-based

It identifies the opinionated words in the corpus and assigns the polarity to these words to find the domain or context-specific opinion that dictionary-based cannot. It needs a dictionary of all English words .

IV. LITERATURE REVIEW

Since this research is intended to study supervised approaches to find the sentiment of Amazon reviews, the works related to analysing the sentiment using these approaches are only considered in this section. These researches are reviewed in terms of pre-processing techniques, feature extraction methods, proposed methodologies (classification methods), and evaluation metrics.

Several works have focused on identifying users' opinions of different Amazon product reviews. In the research study by [14], a general sentiment polarity classification process is presented that implies a negation phrase identification (a phrase that conveys opposite sentiment), mathematical sentiment score computation, and a feature-vector generation method. The experiments were done on the data collected from Amazon (www.amazon.com) in 2014, about online product reviews on both review and sentence levels categorization. They used a max-entropy POS tagger and three ML classification models (i.e., Random Forest, SVM, and NB). The results were promising, and the Random Forest model outperforms Naïve Bayes and SVM, table1.

Also, traditional Logistic Regression, SVM, and deep learning models like Long Short Term Memory networks LSTM and convolutional neural network CNN have been applied on large-scale Amazon reviews datasets in [15]. They compared and analysed different pre-processing techniques that raise the accuracy and found that the best combination method operates stemming over lemmatization and does not include spell checking. They used various feature approaches like bag-of-words, and n-grams, and their TF-IDF. Also, paragraph vector, pre-trained word embedding like Word2Vec, and Glove to discover which combination works better. Among conventional techniques, Linear SVM classifier and bag-of-n-grams with TF-IDF features perform the best, whereas, among deep learning models, LSTM works better.

Another research [3], presents a comparative study of two text sentiment classification groups to analyse Amazon reviews datasets. Supervised machine learning techniques group i.e., LR, Gradient Boosting, and SVM algorithms and the lexicon based approaches group i.e., SentiWordNet, Pattern, and VADER lexicons. They used various NLP techniques, including word lemmatization, stop words removal, and TF-IDF vectorization. They conclude that supervised classification methods with minimum hyperparameter tuning outperform others, especially LR classifiers, and VADER works the best

among all lexiconbased approaches on all accuracy, recall, precision, and F1 score measurement metrics. They added that both groups performed better in identifying positive labels than negative ones. This performance could have been due to certain stop words associated with positive emotion and the inherent class imbalance problem of having a large proportion of one class review than the other in the dataset.

In this paper [16], the authors claim that the accuracy of the existing algorithm (NB and SVM) is not worthwhile. They proposed an ensemble model approach, which is combining two or more algorithms. The proposal contains three modules. They collected the dataset from the official product site using Amazon API in the first module, and unwanted data such as stop words, punctuations, and conjunctions were pre-processed in the pre-processing module. Whereas in the classification module, they combine NB and SVM and calculate the mode value based on the vote for every algorithm used. This approach provides better accuracy than could be obtained by the existing algorithms individually.

Furthermore, the performance of Multinomial Naïve Bayes (MNB), Linear (SVM) and LSTM algorithms were investigated in [17], to see if it is feasible to tag the polarity of (N = 60,000) Amazon product records randomly selected from four million Kaggle benchmark dataset. Besides, they scrapped some pages/ ~ 230000 real-time reviews related to Electronic devices, Toys, and furniture categories to use it along with the Kaggle rest records for the models' evaluation test. They applied data pre-processing, and the training set is fit on a 'maximum features' parameter. TF-IDF vectorizer is used for the traditional models, while the tokenization method is used for LSTM. The test shows that LSVM and MNB achieved satisfying results, but LSTM networks work the best on Amazon binary sentiment classification [17].

In another study by [18], they collected reviews from Amazon about Tablets, Mobile phones, Cameras, Laptops, and TVs. They proposed a hyper approach, dictionary-based, and supervised classification models, i.e., NB and SVM, to find the sentiment of each product reviews individually. They employed opinion lexicons that contain 4783 negative and positive words to calculate the sentiment scores of the sentences. They conclude that both NB and SVM got excellent accuracy on camera review, i.e., 98.17% and 93.54%, respectively.

In the research [19], the efficiency of applying KeywordBased, SVM, NB, and ME classifiers were tested for classifying online Amazon reviews- reviews extracted using API, using a web model. In feature extraction, unigrams and weighted unigrams features have been used to train these classifiers. They designed a framework to consider feature extractors and classifiers as different mechanisms. Query terms consequences have been normalized, emoticons were used as rough labels in the training phase, and feature reduction is applied. The results were acceptable on weighted unigrams, specifically with SVM compared to Unigrams. It also shows that using weights can increase accuracy.

In [20], NB and decision list classifiers, along with bagof-words and bigrams features, are compared in their effectiveness in correctly tagging reviews as positive and negative attitudes towards book products. NB gave better results than the decision list classifier in the all-books dataset; and found that the system performed well on small datasets even when trained and tested on entirely different product reviews.

Other researchers concern on Mobile phone reviews, in this paper [21], they proposed a framework with three modules (i.e., (1) data collection and pre-processing (2) feature selection and sentiment analysis, (3) classification, and cross-validation). They collected over 400,000 reviews of ~4500 mobile phones from the e-commerce giant Amazon.com. They cleaned and handled it to be balanced. They added the score to each record using the NRC dictionary. They then employed ML classification algorithms like NB, SVM, and DT to classify these reviews as positive or negative. Out of the three classifiers, they found SVM's predictive accuracy is the best, table1.

The work by [22], concentrates on mining and applies their experiments on Apple iPhone 5S, Samsung J7, and Redmi Note 3 reviews from Amazon. They proposed a system to retrieve the selected product reviews of the given URL automatically. They used ML algorithms such as NB classifier, and Logistic Regression LR, and also SentiWordNet algorithm to classify the retrieved text as negative, positive, or neutral). In the end, they have used metric parameters to measure the performance of each. The results found that NB proves to be the most efficient among all.

The research by [23], investigated three features of text representation approaches, i.e., Countvectorizer, TF-IDF, Ngrams, with logistic regression classifiers to find the sentiment of mobile phone review. They

found that the combination of TF-IDF with n-gram is the winner with 97 AUC.

Sometimes researchers prefer to train and test their classifiers on Amazon polarity and Kaggle benchmark datasets. Also, LR, Stochastic Gradient Descent SGD, NB, and CNN performance is studied in [24], using balanced and unbalanced versions of Amazon unlocked mobile phones dataset and variety of feature extraction techniques as in [14]. They applied the Lime technique to analyse the classification results for the reviews being either negative, neutral, or positive (based on significant or most frequent words). The study found that negative reviews lengths are longer in general and reach a conclusion that the CNN model with word2vec performed the best among the other models on both versions of the data, 79.60%, and 92.72% accuracy, respectively.

This paper [25] aimed to tackle traditional representation methods such as BOW and TFIDF for sentiment analysis since these non-distributed vectorization methods extract sentiment from lexical or syntactic features. These techniques do not consider word order, semantic word relations, and contextual information appear in the review. Instead, they used word2vec approaches (Continuous Bag of Words (CBOW) and skip-gram models) representation with several parameter settings to capture the deep semantic relationship between words, followed by supervised classification algorithms like Logistic Regression, SVM, Random Forest, and Naïve Bayes. The experiments show that word2vec was efficiently incorporate contextual information and semantic word relations for the sentiment classification task, and exhibit the superior accuracy obtained by using Random Forest with CBOW.

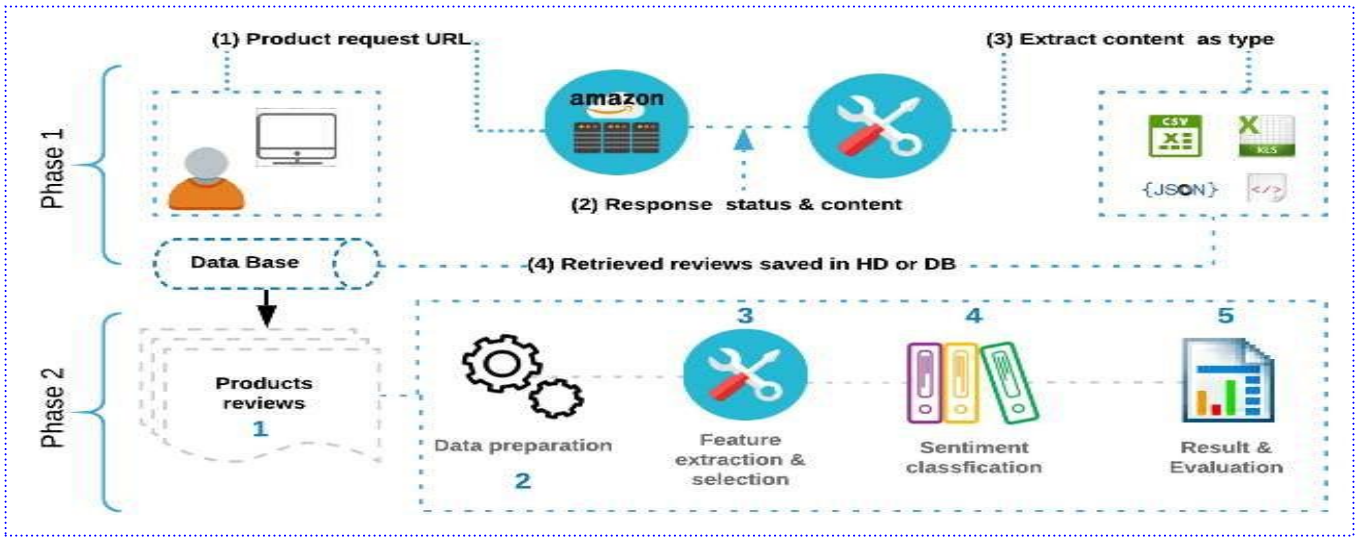
Another study conducted by A. Ejaz, et.al. [26], developed a lexicon dictionary-based algorithm along with n-grams to perform sentiment evaluation of only 500 MB of Amazon product reviews database. They compared it with three ML techniques with different text representations: Random Forest RF with word vector, DT with a document vector, and Random Forest with n-gram. They used KNIME software to clean ambiguous and missing rows and other unwanted words and characters. Their experiment concludes that their approach performs better than machine learning techniques used in terms of AUC-ROC accuracy measurement.

Another work to solve sentiment polarity classification is done by [27]. They developed a classification technique for a dataset of Office products and musical DVDs that crawled using python crawler. They considered five classes (Strongly Negative, Negative, Neutral, Positive, and Strongly Positive). The paper has applied RF, DT, NB, SVM, GB, LSTM classifiers along with three types of adverbs as features, i.e.,
Adverbs RB, Comparative adverbs RBR, Superlative adverbs RRS, and a combination of them to perform review level classification. The experiments show that a single RBR feature is suitable for most of the classifiers except LSTM and NB, and a combination of RBR-RBS features proved to be more efficient for all the classifiers.

TABLE 1. SUMMARY OF SOME STUDIES

Paper	Dataset	Technique	Features	Accuracy	Notes
Nguyen, et.al. [3]	Product reviews	SVM, GB, LR	TF-IDF	[89, 87, 90]	-
X. Fang, et.al. [14]	Product reviews	NB, SVM, RF	POS	[98, 90, 97]	ROC
K.Tamara ,et.al. [15]	Amazon Polarity Dataset	LR, SVM, MNB & GB	BOW, B-ngrams, TFIDF,Word Embeddings	[92.88 , 92.90, 90.59, 82.03]	Bag-ofn-grams + TF-IDF Result
J.Sadhasivam, et.al. [16]	Product reviews	SVM, NB, Ensemble	POS with a weighted score	[30.19 , 34.74, 75.95]	-
R.S. Jagdale, et,al. [18]	Product reviews	SVM, NB	BOW	[93.54 , 98.17]	Camera product
A.Singh, et.al. [19]	Product reviews	SVM, NB, ME	unigrams and weighted unigrams	[81.20 , 77.42, 70.35]	weighted unigrams
Z.Singla, et.al. [21]	M.Phone reviews	SVM, NB, DT	Not mentioned	[81.77 , 66.95, 81.25]	-
Kumar, et.al. [22]	M.Phone reviews	NB, LR,	Not mentioned	[85.7, 66.1]	F1measure, Samsung j7
Aljuhani, et.al. [24]	Unlocked Mobile Phones reviews	LR, SGD, NB,	BOW,TF-IDF,N-grams ,Word Embeddings & Combination	[77.47 , 76.05, 74.90]	Balance d data, different word representation
B.Bansal, et.al. [25]	Unlocked Mobile Phones reviews	SVM, LR, NB, RF	Word2Vec (CBOW, Skip Gram)	[90.3, 90.9, 54.8, 90.6]	CBOW
S.Kausar, et.al.[27]	Product reviews	RF, DT, NB, SVM, GB,	Adverbs, Comparative & Superlative adverbs & Combination	[95.0, 95.0, 91.0, 94.0, 95.0]	F1measure, RBR_RBS

V. SENTIMENT ANALYSIS ON PRODUCT REVIEW



The following Fig. 3, along with its steps, presents the typical methodology workflow process that the researchers often adopted to conduct sentiment analysis. This methodology comprises of two phases as in the following.

1. Phase 1: Scraping A- Web scraping

It is a technique used to access and extract a large amount of data from any accessible websites, e.g., Amazon, as in our case and stores it for later used analysis.

B- Scraping process

Fortunately, most of the famous sites like Facebook, Twitter, and Amazon offer API that facilitates access to any

Fig. 3: Typical Sentiment Analysis Workflow Process

kind of information, such as posts, comments, or products available on the page for the researchers with some restrictions. These APIs require having an account or becoming an associate on these websites. As an alternative, Request and *Beautiful Soup* are python libraries that help in crawling and extracting all the structured/unstructured detailed data displayed on the product page.

The request library is responsible for getting the content of the desired URLs of the inspection element. This content will contain a status code and the web page content of that particular element. At the same time, *Beautiful Soup* accesses this response content, uses ASIN — Amazon Standard Identification Number to get all the details of it (in case of
Amazon reviews). It then converts it into a proper format (CSV, JSON, XLS, etc.), and saves it on the computer for later use. Also, the Amazon dataset contains fields like ASIN, Reviewer id, Reviewer name, Review text , Helpful, Rating, and Time.

2. Phase 2: Sentiment classification A. Data Collection

The previous phase serves as a data collection. However, some researchers prefer to use benchmark labelled datasets downloaded from the UCI machine learning repository or Kaggle websites to train their algorithms before they test it on real data. Loading this data, dropping columns, dealing with missing values, etc., to make data ready for further process, are the aim of this step - Pandas python library can serve a lot in this step. A proper dataset is an important step and needs to be defined for analysing and classifying the text.

B. Text preparation

After the text is obtained, the preparation step comes to make data ready to be used for further machine learning steps. Pre-processing employed for reducing the noise and removing the data that are irrelevant for sentiment classification, such as eliminating punctuations, numbers, accent marks, stop words, sparse terms, white spaces, and particular words. Other parts of this, convert words to lower case, tokenization, stemming, lemmatization, part of speech tagging, etc. This noisy data can affect the accuracy of the classifier [28]. Using Natural Language Processing Toolkit (NLTK) is preferable in this step.

C. Feature selection & extraction

Features must describe the data in a format required by a particular machine learning algorithm to be used to solve the task. Feature extraction is the process that combines and reformats these original features using a combination of (BOW, TF-IDF, N-grams, POS, Word Embedding, etc.) until it yields a new set of features to be utilized by the Machine Learning models. Then, selects the most relevant, useful, informative features and ignore the rest. It

prevents redundancy or gets a limited number of features to avoid overfitting and curse of dimensionality, i.e., too many features to describe insufficient samples. Proper feature extraction and selection play a crucial role in determining classifier accuracy [29]. Hence, the appropriate technique must choose for extracting the features. Scikit-learn library offers many built-in methods that can help a lot here.

D. Sentiment classification

In this step, various sentiment classification mechanisms are applied to determine the polarity of the review documents- supervised learning methods commonly used for SA to assign the sentiment label for a given text. Generally, the problem of SA is of two types; one is a binary with positive and negative labels. Another is multi-class, i.e., more than two labels (very positive, positive, neutral, negative, and very negative) [30]. Scikit-learn library provides various classes that assist in this process.

E. Results evaluation

This final step intends to evaluate the performance of ML techniques used to determine the overall accuracy of the sentiment analysis. The result of the classification consists of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True positives and true negatives accurately predict actual labels, while false positives and false negatives are misclassifications [3]. Accuracy (1), precision (2), recall (3), and F-measure (4) are commonly used statistical metric parameters to measure the performance of each algorithm, formed from a confusion matrix in the Scikit-learn library. After that, the final output is analysed to decide whether it should be considered or not, and then it can be displayed in a pie chart, bar/line graph using the *Matplotlib* python library.

$$\text{Accuracy} = ((TP) + (TN)) / \text{Total of observations} \quad (1)$$

$$\text{Precision} = (TP) / ((TP) + (FP)) \quad (2)$$

$$\text{Recall} = (TP) / ((TP) + (FN)) \quad (3)$$

$$\text{F1 Score} = 2 * (TP) / 2 * (TP) + (FP) + (FN) \quad (4)$$

VI. Discussion and Notes

In this section, we discussed the results of some classifiers with different feature representation approaches that were achieved by the experiments in the literature review section to determine which models perform better than others. We used Excel software to draw various charts for comparison.

Firstly, we compared the widely used supervised learning algorithms in our literature review: Naïve Bayes and Support Vector Machine in five selected papers [16] [18] [19] [21] [25] and found that NB got better results in [16] [18], whereas SVM achieved the best in the others. As shown in Fig. 4, both classifiers perform very poorly in terms of accuracy when using POS with weighted scores as features [16]. These experiments also revealed that NB and SVM with BOW as a feature representation provide above 93% accuracy [18], and SVM also works well with CBOW features while the NB not [25] with a dataset consideration.

Then, we compared the results of SVM, NB, and RF classifiers, which achieved in [14] [25] [27], Fig. 5. These papers showed that RF works better than the others in terms of POS features [14] [27]. Moreover, RF also works well with the CBOW word vector [25], and n_grams [26].

Lastly, the accuracy of the SVM and LR in [3] [15] [25] papers are compared, Fig. 6. We can observe that SVM performed equally to the LR, with TF-IDF, BO_N-grams + TF-IDF, and CBOW approaches. On the other hand, NB is compared with LR in [22] [24] [25], Fig. 7. It is found that using these classifiers with a combination of Tri-grams with BOW or TF_IDF features can achieve only less than 80% of accuracy in balanced data [24]. Moreover, the CBOW word vector again works with the LR classifier [25].

The main findings of this analysis are mentioned below:

- Generally, the traditional supervised machine learning models perform well on Amazon reviews, especially SVM, NB, RF, and LR. These approaches would inquire collecting labelled data first, which is not always possible and often time-consuming. These methods remain strong candidates with acceptable accuracy for the size of dataset up to several hundreds of thousands of training records [15].
- Researchers often scrape Amazon reviews and use stars scoring corresponding to each review as labels to teach the supervised algorithms.
- The collected data can be imbalanced - observations per class is not equally distributed in this data, which should be processed to be balanced or use a suitable metric measure to consider the model bias; otherwise, researchers use balance benchmark datasets.

- The inherent class imbalance problem causes ML algorithms to achieve better in classifying positive labels and perform poorly in the negative ones.
- The traditional feature extraction methods like BOW,

TF-IDF, N-grams, and their combination are still standing as a baseline for classification tasks and produce excellent results. However, it does not capture the meaning as in [3] [15] [18].

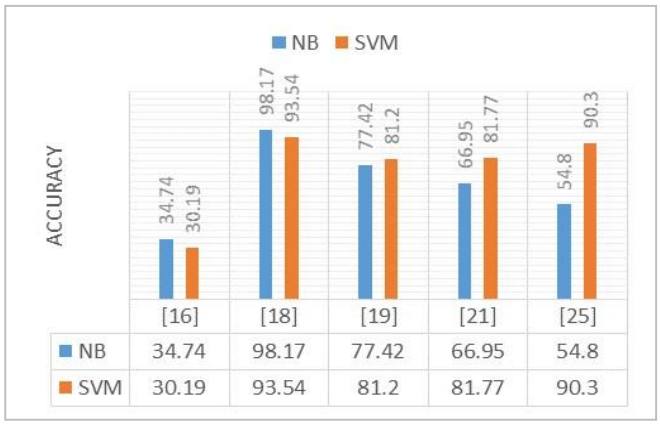


Fig. 4 Accuracy of NB & SVM with different features

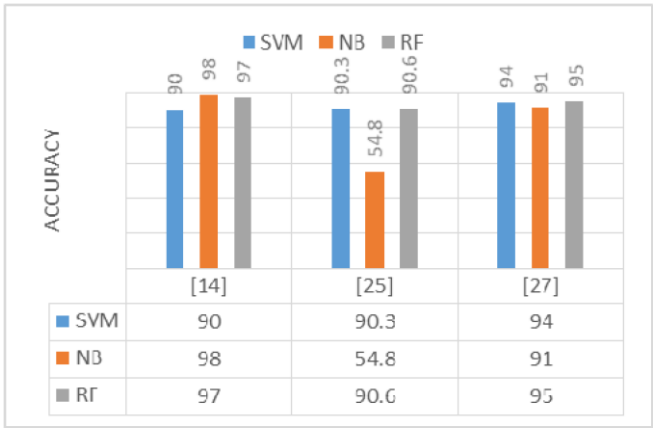


Fig. 5: Accuracy of SVM, NB & RF with different features

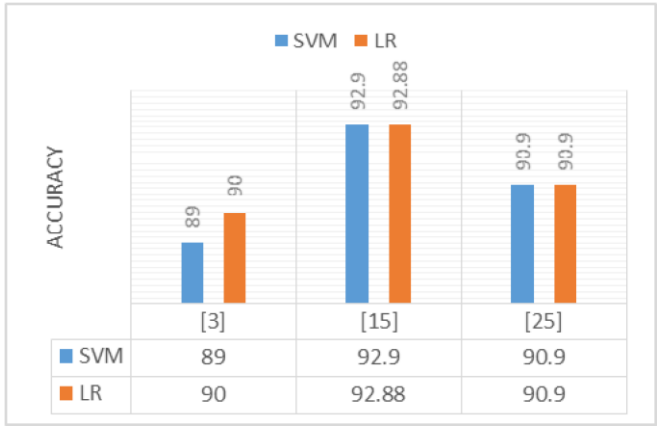


Fig. 6 Accuracy of SVM & LR with different features

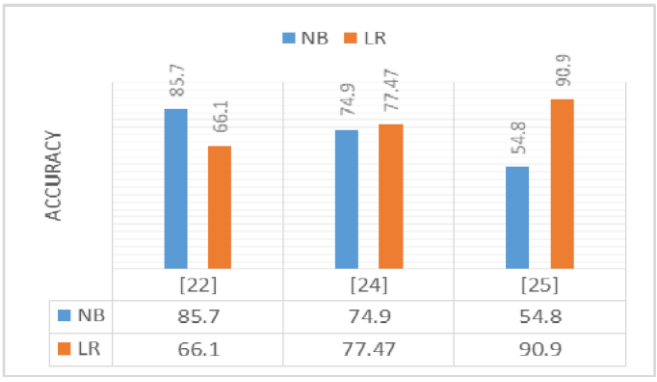


Fig. 7 Accuracy of NB & LR with different features

- Word embedding like Word2Vec and Glove can capture the meaning and semantic relationships and other words contexts usage [15]. Using these representations with traditional ML classifiers is not always efficient, as in [15] [24] [25].
- The performance of the models can vary according to the Amazon category [18] [20].
- Applying pre-processing techniques is necessary since it affects the classifier's performance and depends according to the problem on hand [28].
- Researchers often prefer to use SVM and NB for classification, though RF and LR classifiers prove their efficiency over SVM and NB in some tasks.

VII. Conclusion and Future Works

Sentiment analysis is the computational study of the subjective words in the text, which composes the user's opinion about entities on micro-blogging social media websites. Research work is continuing to find solutions with high accuracy for the challenges. This survey studied the use of conventional supervised learning methods used by many researchers to find the sentiment opinion hidden in amazon reviews data at the document level and presented a comparative study of various parameters like datasets, features, techniques, and accuracy. The results show that these approaches with excellent pre-processing and suitable feature representation methods can achieve accuracy, specifically RF, LR, SVM, and NB. This paper makes the research open to investigate the use of unsupervised and advanced deep learning methods in classifying Amazon reviews and other online shopping like Flipkart of different SA levels for future works.

REFERENCES

[1] D. Khurana, A. Koli, K. Khatter, S. Singh, and M. Rachna, "Natural Language Processing : State of The Art , Current Trends and Challenges Department of Computer Science and Engineering Accendere Knowledge Management Services Pvt . Ltd ., India Abstract," no. Figure 1.

[2] K. Anwar, J. Siddiqui, and S. S. Sohail, "Machine learning-based book recommender system: A survey and new perspectives," Int. J. Intell. Inf. Database Syst., vol. 13, no. 2–4, pp. 231–248, 2020, doi: 10.1504/IJIIDS.2020.109457.

[3] H. Nguyen, R. Al, and K. Academy, "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," SMU Data Sci. Rev., vol. 1, no. 4, 2018.

[4] B. Liu, "Sentiment Analysis and Subjectivity," pp. 1–38, 2010.

[5] B. Liu, "Sentiment Analysis and Opinion Mining," no. May, 2012.

[6] A. Joshi, P. Bhattacharyya, and S. Ahire, Sentiment Resources: Lexicons and Datasets. 2017.

[7] D. Mohey and E. M. Hussein, "ORIGINAL ARTICLE A survey on sentiment analysis challenges," J. King Saud Univ. - Eng. Sci., 2016, doi: 10.1016/j.jksues.2016.04.002.

[8] N. Nandal, R. Tanwar, and J. Pruthi, "Machine learning based aspect level sentiment analysis for Amazon products," Spat. Inf. Res., 2020, doi: 10.1007/s41324-020-00320-2.

[9] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," CEUR Workshop Proc., vol. 718, pp. 81–92, 2011.

[10] B. Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up ? Sentiment Classification using Machine Learning Techniques," no. July, pp. 79– 86, 2002.

[11] L. M. Chiappe, "P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 417–424, Morristown, NJ, USA, 2001. Ass," Society, vol. 12, no. 3, pp. 344–350, 2010, doi: 10.3115/1073083.1073153.

[12] R. Feldman, "Techniques and applications for sentiment analysis," Commun. ACM, vol. 56, no. 4, pp. 82–89, 2013, doi: 10.1145/2436256.2436274.

[13] S. S. Zia, S. Fatima, M. S. Ali, M. Naseem, and B. Das, "A Survey on Sentiment Analysis , Classification and Applications," vol. 119, no. 10, pp. 1203–1211, 2018.

[14] X. Fang and J. Zhan, "Sentiment analysis using product review data," J. Big Data, 2015, doi: 10.1186/s40537-015-0015-2.

[15] K. Tamara and N. Milićević, "Comparing Sentiment Analysis and Document Representation Methods of Amazon Reviews," SISY 2018 - IEEE 16th Int. Symp. Intell. Syst. Informatics, Proc., pp. 283–288, 2018, doi: 10.1109/SISY.2018.8524814.

[16] J. Sadhasivam and R. B. Kalivaradhan, "Sentiment analysis of Amazon products using ensemble machine learning algorithm," Int. J. Math. Eng. Manag. Sci., vol. 4, no. 2, pp. 508–520, 2019, doi: 10.33889/ijmems.2019.4.2-041.

[17] W. Tan, X. Wang, and X. Xu, "Sentiment Analysis for Amazon Reviews," Reseaechgate, no. March, pp. 0–9, 2019, doi: 10.13140/RG.2.2.13939.37920.

[18] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, Sentiment Analysis on Product Reviews Using Machine Learning Techniques : Proceeding of CISC 2017 Sentiment Analysis on Product Reviews Using Machine Learning Techniques, no. January. Springer Singapore, 2019.

[19] A. Singh, A. Agarwal, and P. Dimri, "ScienceDirect ScienceDirect Comparative Study of Machine Learning Approaches for Amazon Reviews," Procedia Comput. Sci., vol. 132, pp. 1552–1561, 2018, doi: 10.1016/j.procs.2018.05.119.

[20] C. Rain, "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning," Swart. Coll., 2013, doi: 10.1097/IOP.0b013e318213f5d9.

[21] Z. Singla, S. Randhawa, and S. Jain, "Sentiment Analysis of Customer Product Reviews Using Machine Learning," 2017.

[22] K. L. Santhosh Kumar, J. Desai, and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," 2016 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2016, 2017, doi: 10.1109/ICCIC.2016.7919584.

[23] V. Thada and U. Shrivastava, "Sentiment Mining of Product Opinion Data," no. 3, pp. 1218–1222, 2020, doi:

10.35940/ijitee.C8641.019320.

[24] S. A. Aljuhani and N. S. Alghamdi, “A comparison of sentiment analysis methods on Amazon reviews of Mobile Phones,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 608–617, 2019, doi: 10.14569/ijacsa.2019.0100678.

[25] B. Bansal and S. Srivastava, “ScienceDirect Sentiment classification of online consumer reviews using word vector representations,” Procedia Comput. Sci., vol. 132, pp. 1147–1153, 2018, doi: 10.1016/j.procs.2018.05.029.

[26] A. Ejaz, Z. Turabee, M. Rahim, and S. Khoja, “Opinion mining approaches on Amazon product reviews: A comparative study,” 2017 Int. Conf. Inf. Commun. Technol. ICICT 2017, vol. 2017-Decem, pp. 173–179, 2018, doi: 10.1109/ICICT.2017.8320185.

[27] S. Kausar, X. U. Huahu, W. Ahmad, and M. Y. Shabir, “A Sentiment Polarity Categorization Technique for Online Product Reviews,” IEEE Access, vol. PP, p. 1, 2019, doi: 10.1109/ACCESS.2019.2963020.

[28] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” Procedia Comput. Sci., vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[29] M. Trupthi, S. Pabboju, and G. Narasimha, “Improved feature extraction and classification - Sentiment analysis,” 2016 Int. Conf. Adv. Hum. Mach. Interact. HMI 2016, pp. 117–122, 2016, doi: 10.1109/HMI.2016.7449189.

[30] F. Sebastiani, “Text Classification Sentiment Analysis and Opinion Mining,” 2015.