Customer Opinion Mining by Comments Classification using Machine Learning

Moazzam Ali[1], Farwa yasmine[2], Husnain Mushtaq[3], Abdullah Sarwar[4]
Adil Idrees[5], Sehrish Tabassum[6] , Dr.BaburHayyat[7], Khalil Ur Rehman[8]
Department of CS and IT University of Lahore
(UOL) Gujrat Campus, Gujrat, Pakistan

***Abstract*—In this era of digital and competitive market, every business entity is trying to adopt a digital marketing strategy to get global business benefits. To get such competitive advantages, it is necessary for E-commerce business organizations to understand the feelings, thinking and seasons of their customers regarding their products and services. The major objective of this study is to investigate customers' buying behavior and consumer behavior to enable the customer to evaluate an online available product in various perspectives like variety, convenience, trust and time. It performs data analysis on the E-commerce customer data which is collected through intelligent agents (automated scripts) or web scrapping techniques to enable the customers to quickly understand the product in given perspectives through other customers' opinion at a glance. This is qualitative and quantitative e-commerce content analysis in using various methods like data crawling, manual annotation, text processing, feature engineering and text classification. We have employed got manually annotated data from e-commerce experts and employed BOW and N-Gram techniques for Feature Engineering and KNN, Naïve Bays and VSM classifiers with different features extraction combinations are applied to get better results. This study also incorporates data mining and data analytics results evaluation and validation techniques like precision, recall and F1-score.**

***Keywords*—Customer comments; behavior mining; data mining; machine learning**

I. Introduction

A. Motivation

This study emphasis on need to develop some mechanism which ensures to get advantages form E-commerce users' generated data. A general approach reveals that to get opinion of other people before buying any product is common in online and offline shopping. But in this digital era, each customer has hundred or sometimes thousands of people readily available to provide valuable opinion and largely effect the decision-making process of new customers. Each customer looks for best product in lowest possible price. Actually, each customer tries to find the best commodity within his/her financial range along with surety of the justifiable quality attributes. Therefore, it is a normal practice to get neutral and genuine opinion of general public that is not generated by selling organization also not tempered by anyone else [1][2][3].

B. Consumer

A customer is an entity or person having an ability or will to buy the products and offered ventures available for purchase by advertising organization so that it may fulfil needs or want of an individual, family or a particular group of similar interests. A famous definition of consumer done by Mahatma Gandhi says, "A consumer is the most important visitor on our premises. He is not dependent on us. We are dependent on him. He is not an outsider to our business. He is part of it. We are not doing him a fulfil by serving him. He is doing us a fulfil by giving us an opportunity to do so" [4][5].

C. Consumer Behavior

Consumer behavior is characterized as "psychological, physical and social acts of potential clients as they have made their minds to access, assess, purchase and inform others about any item and its attributes". Consumer behavior is the study of single buyer, a group or organization about their course of selection, buying, utilization and disposing of commodities, services, ideas or experiences to satisfy his/her needs and also impacts of such process on the consumers and whole society. Consumer behavior varies from individual to the groups (like class students in school or college wear same uniform/dressing) and from group to firms (various groups at same place working together horizontally and vertically and decides collectively whether a product must be user by the firm or not) [6][7]. The customer opinion is often very important for advertising agency/marketer because it influences the market position as well as consumption of the product. Consumer behavior involves services and ideas as well as tangible products.

D. Internet Marketing

Internet marketing is utilization of internet as a medium to assess the showcasing and potential sale of merchandise. It has been proved highly beneficial by apply standard fundamental promotion systems on e-commerce applications [8]. Contrary to physical business strategies, online promotion and advertisement strategies have been proved far better with little hazards comparatively. Web showcasing process not only convenient for business community it also supports green solutions across the globe [9].

E. Purchase Decision

Buying decisions are defined as: "Several stages carried out by consumers before making a purchase decision on a product" [10]. According to the [11] buying behavior means activities of an individual who is involved in exchange of money for goods or services and also it involves some decision-making process to determine those activities. Consumer's decisions in buying a product involves physical and mental activities. The former

refers to the direct activities for decisions making process while later involves assessment of product using some particular criteria.

F. Data Mining and E-commerce

It is a substantial undertaking to build a system which take advantages from mined knowledge. Studies revealed that some applications of data mining techniques on e-commerce data is comparatively less challenging as compared to other sort of data. For example, we can develop data mining system in E-commerce with much convenience rather than translating and correcting the data to make it suitable for data mining purpose. As data set is not collected manually or by survey but accessed electronically so it comparatively less noisy or sometimes contains no noise. Moreover, data set contains variety of vast and varied information as shown in "Fig. 1" [12] 1[3].

It has been analyzed that public information over E-commerce platforms play a vital role in success of regression models due to justifiable quantity of varied information. Therefore, E-commerce platforms provide very useful data and its inferences to produce a platform that is trustworthy for E-commerce customers [14]. They have discussed various applications of clustering and fuzzy set theory to determine issues in E-commerce platform through data mining application.

G. Consumer Behavior Mining

Consumer behavior mining is deals with web content mining which is concerned with valuable information extraction regarding users/customers opinion. Consumer behavior mining is almost a new subject in field of data mining as part of web mining and growth of E-commerce business has accelerated its growth significantly across the E-commerce applications, blogs and forums [15][16]. Consumer behavior towards online shopping instead of physical visiting markets and shops has been changed due to growth and profitability of E-commerce business. So along with online shopping habits, many people also prefer to get knowledge about public opinion regarding a particular merchandise before placing online order.

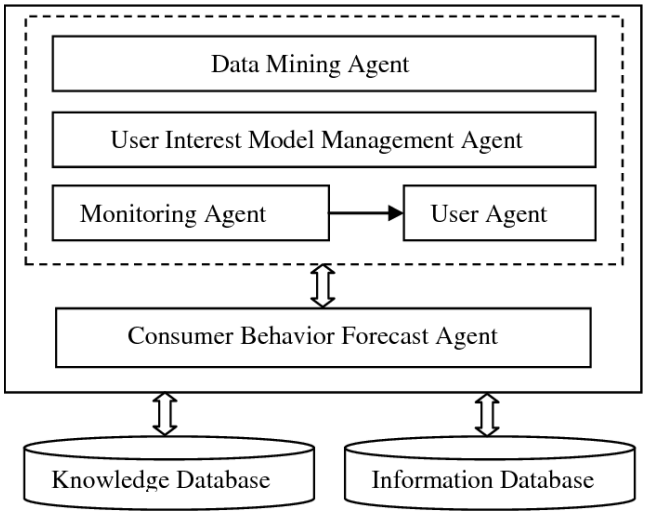

Fig. 1. Social Commerce as Sub Set of E-commerce and Social Media.

Consumer behavior is highly affected by public opinion but at same time availability of desired, accurate and quick information is a problem. Finding thousands of reviews regarding a product is not a big problem but finding summary of user opinion which is true reflection of customers' thought is a still a challenge across the E-commerce platforms [15]. Consumer opinion mining is not only useful at consumer end but also useful for manufacturing and advertisement companies the former get support about decision of transaction while the later are able to use opinions as customer feedback to improve their product accordingly while advertising agencies can easily find about individual customer opinion and post advertisements on his/her pages according to his/her choice [16].

H. Research Problem

Research objective is to identify the Consumer Behavior relevant elements related data in online E-commerce platforms. Validate Consumer Behavior relevant elements' related data with respect to customer opinions. To identity suitable text preprocessing techniques and selection of most relevant features for each Consumer Behavior relevant element. Develop a supervised machine learning based system which extract, classify into major predefined categories and preserve Consumer Behavior related knowledge in order to make this knowledge extensive, versatile, verified, easy to use and up to date.

II. Literature Review

A. Consumer Behavior

Customer experience is very important for every business organization where numerous products and vendors/brands are available. Customer experience optimizes and improves the online shopping experience of people which has been extensively increase after outbreak of pandemic Coid-19 across the world [18]. E-commerce companies and organization has experienced business prosperity due to increase in online demand of merchandise. Each business organization is string hard to improve their marketing by better understanding of customers' needs and priorities through analyzing customer buying behaviors[19][20]. Each organization trying its level best to retain their customers by offering the best shopping choices so that customers do not switch brands and shopping platforms. Organizations are also focusing to identify their potential customers segments by tracking their priorities, selections and expectations regarding products over the time [12]. It is necessary for business organizations to track, collect and organize consumer behavior possible data to develop business and analysis insight to take appropriate actions [13]. Recent studies revealed that enormous research has been conducted and a lot much more is underway to understand consumer behavior or to understand changes in customer activities over the time. Analysis of customer behavior is now an integral part of customer relations management strategies [17]. Data mining methods, tools and techniques are being incorporated to discover useful patterns using large amount of data collected by organizations. There is a variety of data mining models like clustering, association, classification, forecasting, regression, sequence discovery, visualization and machine learning models for data mining models like association rule mining, Logistic Regression, K-nearest

neighbor, Neural Networks, Decision Trees, etc. E-commerce and other business organizations are collecting massive amount of data on daily basis in form of sale purchase transactions, customers profiles, cart management and product search data [21]. Such large data owner organizations are keen to unleash the potential behind this data and are also interested in mining the association among different data segments as shown in "Fig. 2". These organization have firm believe that proper analysis of this large customer data can yield useful knowledge to get insight about consumer behavior [15]. They have proposed a customer segmentation system to discover and analyze frequent items searched by customers and their change over time. Authors employed association rule mining to discover useful and meaningful data patterns using database containing customers transactions records [22]. They also devised a strategy to automatically detect changes using customer profile and sales data of a particular period. The authors also proposed three types of changes in customer behavior: Unexpected Changes, Emerging Changes and Add/Perished Rule [19].

B. Consumer Behavior Mining

Big data technologies and their implementations are becoming cause of an immensely increasing information nowadays. Banking and insurance sector is seeking benefits of big data analytics and data mining to detect defaults and potential risks. In [23], the authors collected 22745 data samples and 14 attributes from Turkish Statistical Institution. This targeted to find the selection of best algorithm for classification to identify risk because of personal characteristics [24]. They incorporated and evaluated the several classification algorithms like Naïve Bayes, J48, Logistic Regression, Random Forest and Multilayer Perceptron were selected and their accuracies were evaluated using several evaluation techniques like Precision, Recall Roc Curve, etc. and most of them were found suitable to deal such type of data. Weka, a renown data mining tools was used to perform experimental methodologies as shown in "Fig. 3". Data mining is extensively being used in field of medicines to predict various diseases using patient medical records especially detection and prevention of Diabetes Mellitus as it has deteriorated human, social and economic fabrics due its boosting penetration in all societies across the globe. traditional data mining techniques have been integrated with clinical research to advertise adverse effects of various diseases like Diabetes Mellitus, cancer and other diseases. These researches generally incorporate plain combinations or single classifiers. There are many comprehensive efforts are conducted to enhance the accuracy of systems with combination of multiple classifiers. [25][26] classified the diabetes mellitus in individual patients by using a set of risk factors and applied Bagging techniques of natural language processing with Adaboost, Decision Tree as a baseline experiment. The experiment was carried with Canadian Primary Care Sentinel Surveillance and three different adult teams. The Adaboost outperformed the bagging techniques an individual Decision Tree J48. Anarkali [27] have enlisted number of data mining application fields which are popular in recent times. Knowledge Driven Databases (KDD) and data mining are being employed in numerous fields to collect and analysis large chunks of information.

Fig. 2. Social Commerce as Sub Set of E-commerce and Social Media.

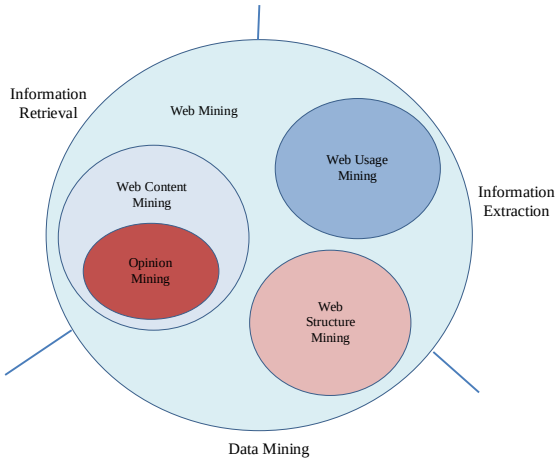

Fig. 3. Information Retrieval using Web Mining.

Integration of big data techniques in E-commerce web applications has also make it easier to collect structured, wide and large volumes of data regarding internal business process, marketing, supplier, venders and shipment. Therefore, big data analytics has open new avenues of opportunities for business companies related to E-commerce [28]. [29] presents many clustering, association and prediction techniques which are highly useful for E-commerce business. Sales forecasting, customer relationship management, customer retention management and basket analysis are common data mining models in E-commerce business. Major objective of this study is to review, implement and evaluate classification data mining techniques on user comments to classify them into different categories to assist online customer in decision making process.

C. Perceived Benefits

There are numerous benefits of internet shopping or online shopping but it there are also some risks involved which affect the consumer behavior towards online shopping [10]. The perceived benefits related to consumer satisfaction and belief that online shopping offers following benefits as compared to combinational shopping.

1) Convenience: Internet has made our life easier due to easy and quick access to the large number of desired products within seconds so we can buy almost everything with convenience. Seiders, Berry and Gresham (2000) says that convenience offers four benefits in buying process: search, access, possession and transaction as convenience is frontline attribute of online shopping which incites the consumer to go for online shopping. Furthermore, it is major predictor which prompts the consumer to go for online shopping. It also creates online buying willingness among customers. In online shopping, consumer does not need to leave his home/job or business and visit market personally for buying anything and consumer is not bound to keep cash in pocket for shopping for 24 hours [30].

2) Trust: Online shopping is characterized by time saving in terms of travelling to shops and then selection by visiting shops one by one. Therefore, online shopping saves time of customer but this statement does not stand true all the times due to late shipment of product. Generally, delivery time is often mentioned with the order and online customer confirms the order after verification of product delivery time [30].

3) Time: The time as perception of consumer that the vendor or seller will provide the best commodity in terms of price, quality, utilization and satisfactions. Most of people lack of trust in online shopping due to cheating and misconduct of e-commerce sellers. As there is lack of E-commerce cybercrime and needful legislation therefore it is sole responsibility of the seller to establish an environment of trust among people. Uncertainty in customers regarding reality of online shopping is one of the major obstacles in success of online business [30].

4) Trust: Online shopping is characterized by time saving in terms of travelling to shops and then selection by visiting shops one by one. Therefore, online shopping saves time of customer but this statement does not stand true all the times due to late shipment of product. Generally, delivery time is often mentioned with the order and online customer confirms the order after verification of product delivery time. But sometimes, due to disaster, natural calamity, power break down, strikes or transportation issues order does not arrive on time. Late delivery of product harms to trust of customer so timely arrival of product boost up customer confidence. Therefore there is strong relationship between time and online shopping but there must be some more attributes to better understand this relationship [30].

III. Methodology

This research aims to develop an E-commerce customer comments knowledge classification system based on consumer behavior attributes (listed as Convenience, Variety, Time and Trust), which depends upon tasks to be customers' comments given under any product on any E-commerce platform. We have targeted the social E-commerce platforms of online shopping to acquire data set where versatile people post variety of comments and answer each other's comments. One cannot only easily find type of his comments and discussion activity by using scrolling on web applications. By using proposed classification system, one can find the consumer behavior (Convenience, Variety, Time and Trust) relevant data and its overall status. Moreover, proposed system is not confine to classify all consumer behavior related comments, rather it is also ready to perform classification of other e-commerce customer comments under the umbrella of other classes or categories.

This study strives to build up an E-commerce customer comment data classification/categorization framework which heavily rely on the customer behavior attributes (listed as Convenience, Variety, Time and Trust). The propose system is developed by incorporating data mining and supervised machine learning text classification methods. This study employs various kinds of machine learning models like K Nearest Neighbor, Support Vector Machine, Naïve Bays, Character Based BOW and N-Gram on the state-of-the-art available data set.

A. Dataset Collection

First of all, data set is collected from E-commerce platforms using web agents or automated scripts by apply X query scripts using XML X paths. It requires up to mark professional expertise to extract exactly required and contextual data from large data dumps. E-commerce platforms contains large quantity of consumer behavior knowledge/comments and it can be utilized for consumer behavior mining (reference). Data set regarding consumer reviews about products which ranges across period from 2018 to 2020 contains almost thousands of comments about single product. But as per research requirement only 2000 total and 500 posts of each activity were included in the experiment.

B. Manual Annotation

E-commerce platforms or knowledge sharing communities and forums provide platform to fresh, experienced and professional online buyers/people to share their knowledge with people of their domain across the world and round the clock without any barrier of distance, language and level of expertise. All the members of an E-commerce platform cannot tag their post/comment with a specific category according to the chart mentioned above. If they do so then it will help the people to easily identify each comment under the umbrella of consumer behavior. By doing so, classification of each new comment will be started and data search and segregation will become easy for people who want to learn about opinions of other customers who have already utilized the product from the same platform. To visualize the comments data and better understanding of data set, it is categorized into 4 major categories and each category contain associated attributes related to consumer behavior. Every comment in selected data set is manually annotated and verified by E-commerce experts.

C. Attributes Associated with Each Class

1) Convenience: Convenience, Best, Appropriate, Easy, Great, Perfect, Useful, Functionality, Effective, Desire.

2) Variety: Variety, Forever, Compatible, Collection, Intended, Different, Diversity, Pack 3-1, Specific, Warranty.

3) Time: Recommend, Comfortable, Reliable, Excellent, Complains, Described, Quality, Satisfaction, Honest.

4) Trust: Time, Period, Delivered, Received, Pair, Come up with, Duration, Deal, Successfully.

D. Text Preprocessing

Customer comments classification process is actually started with data collection from E-commerce platform and then the most important function of data mining is performed that is called preprocessing of data. Data preprocessing is important because machine learning models are very sensitive to the features in each document. So pure and more relevant feature can be extracted by feeding the data set to some programming application. Before providing data to machine learning models, it is preprocessed and in case if data set is obtained from social media or E-commerce platform then it contains useless and noisy data which is mandatory to remove to get good results from machine learning classification models. This noisy data includes hashtags, symbols, domain terminologies, trends, smilies, web link, social media promotions, etc. so such noisy material is mandatory to remove to make the training data set clean. Preprocessing of E-commerce data set that has been manually annotated is made free form punctuations, semicolons, quotes, notations and the above said useless elements of language [31][32].

1) Tokenization: Tokenization process splits the data set text into the single elements or tokens by using a specify delimiter like space as shown in Table I. In this study we have used space in words as delimiter. As a text document is composed of linguistics elements and structures called as sentences and each sentence is further made up of grammatical units which are separated from each using space, full stop, hyphen, comma, slash. etc. Tokenization is first step in text preprocessing in which grammatical unit is converted into the tokens using some delimiter and output of this process is tokenized document [33].

2) Stop words removal: Stop words removal is next step to the tokenization. When all the text data set has been converted into individual tokens then unimportant words are removed because in a text data set, all the words are not important with respect to the document or class context. There are numerous words in the input text document which are less or least informative for the machine learning model [34]. Many words occurs multiple times in the document and such repeating words are also included in the noisy data. These less informative and repeating word are call stop words. Stop words removal is an important step in text preprocessing because stop words affect the features set extraction and ultimately efficiency of machine learning model.

As mentioned above, stop words gives little meaning of a word or context and are not helpful to the machine learning models. Stop words removal reduces the data set volume by removing useless or less important token or elements. Ultimately it reduces the computations coat and time. This technique is simple in use and helpful in increasing classification accuracy. Example of these words are ‘The’, ‘is’, ‘also’ etc which are frequently used in textual data. We remove all stop words in English language by using standard list of stop-words.

3) Identification of hash tags: Hash tags are also part and parcel of social media and E-commerce comments data set. Hash tags are commonly known a trends and it has specialty that it is stated with special symbol and it does not contains any kind of separator like space, comma or full stop. Hash tags are not dictionary words and also do not have any particular meaning so it is a good approach to remove Hash Tags to get useful and meaningful words which are easily recognized by dictionary or Word Net in Natural Language Processing.

4) Spell checking and correction: E-commerce text data set in form of comments or discussion in natural language contains multiple words which are not part of language dictionary but are understandable by the people belong to some specific domain. These words are not recognized by search engine optimization and also natural language processing techniques are confused of such words. Majority of such words is set of words which are not properly written and contain spelling mistakes. This problem of such words can be resolved up to some extent using regular expressions or mutually annotated data.

5) Stemming: Words are used in various forms in the English language text. It base might have different forms or behavior when it is used in different sentence structures like present, past, future tenses, singular, plural, adjectives, pronouns, etc [35][36]. English language has different laws for all these situations of a root word. No doubt each words comes from its root word of dictionary but it is used in different styles in the sentences. If we take an example of a word “Go” then we will come to know that it is used as “go”, “went”, “gone”, “going”, “goes”, etc. But dictionary has a single root word for all these words that is “Go”. These words have different posters or shapes but have same meaning as base word. To check the real frequency of each word in the document, stemming is applied. Stemming converts each word to its root word and so actual frequency of each word is calculated as shown in Table II. Stemming is a natural language processing based approach that is used in search engine optimization and information retrieval system. There are many stemming algorithms which can be used but in our study as E-commerce process activities are reported in English language by filling summary and description fields, we use Porter stemming technique to converts all tokens to their base stems.

TABLE I. EXAMPLES OF E-COMMERCE COMMERNTS

E-commerce Customers Comments
I got product in given time period
They promised me to send this in a duration of 2 months and they sent me

TABLE II. COMMENTS DATA AFTER STEMMING

Comments After Stemming
Get Product give time period
Promise send during month send

E. Feature Engineering

Feature selection and extraction that is collectively known as feature engineering employs multiple techniques to extract useful features from a given text document [37]. This study incorporates Bag of Words (BOW) and N-Gram (1 - 4) for the purpose of feature selection.

1) N-Gram: A text document is a connected sequence of n number of words/token/items/elements [40]. These items refer to some symbols, letters, and pairs of words so n-grams are combination of words patterns. Sequence of words to make a sense in a document is called N-gram where N is number of words in a pattern. Uni-gram describes single word, Bi-gram represents two, Tri-gram shows a sequence of three words and so on. Let's consider example of E-commerce comment: "this is a beautiful camera". Uni-grams of this text are 'this', 'is', 'a', 'beautiful', 'camera'. Bi-grams produced from this comment are 'this is', 'is a, 'a beautiful', 'beautiful camera' post. Tri-grams of this text piece are 'this is a, 'a beautiful camera as shown in Table III.

TABLE III. EXAMPLE OF CONVERTING TOKENS TO BI-GRAM, TRI-GRAM AND QUAD-GRAM

Tokens	**I got product in given time period**
Bi-grams	I got, got product, product in, in given, given time, time period
Tri-grams	I got product, got product in, product in given, in given time, given time period
Quad-grams	I got product in, got product in given, product in given time, in given time period

2) Bog of words: The Bag of Words (BOW) is also a features engineering model that counts all the useful feature without giving them weight with respect to document as well class corpora [38][39]. It counts the word number times it occur in the document and also does not accounts the sequence and order of words in the document. Each word in the Bag of Word model is independent of the next and previous words. Let's consider an example to understand the Bog of Words feature selection and extraction technique of Natural Language Processing. The cat is better than the dog and: The weather is better than yesterday.

F. Experimental Setup

All the above mentioned experiment are conducted on the same platform and IDE. Same text preprocessing techniques are applied on the whole data set prior experiment execution. Each experiment took 2400 customer comments as training input data set and 1600 customer comments as testing data et. Four classes (convenience, Variety, Time and Trust) were used to label the data set. Each class comprises of 1000 E-commerce customer comments which is further divided into 600 training and 400 testing instances/document/comment.

Preprocessing techniques are applied on the 70% of the data set and also have employed multiple preprocessing methods like tokenization, stop-words removal, word completion and spell checking, stemming are applied to each document of training data.

Initially, our data set was in form of raw comments which were obtained from E-commerce online shopping platforms. We applied tokenization as first step of preprocessing using natural language processing. To implement tokenization, we used 'space' between two words a delimiter. Following the tokenization, removal of stop words and noisy data is removed to sanitize our data set. At end of stop words removal, stemming is applied to get root words of each lingual element. For stemming, standard stemmers are used because each comment contains valid words after former preprocessing steps.

Machine learning model development and implementation steps are followed by preprocessing steps. Supervised machine learning algorithms are trained on the training data set which has undergone from text preprocessing steps. Feature selection and extraction is an important step which is characterized by the selection of most important and meaningful elements from each document with respect to each class. As we have mentioned in literature review and other sections that in supervised machine learning, algorithms are mostly learned on the probabilities. Unique words are selected and extracted from preprocessed training data etc. Frequency of each word is calculated using BOW model. This probability matrix measures the probability of each unique word in any class. Probability for each word is then calculated from training data. Square root of each probability is computed by calculate square root of each value.

1) Experiment 1: The first experiment is performed using Bag of Word Model feature extraction technique. The major goal behind performing the experiment using BOW model approach to illustrate the classification efficiency and accuracy of supervised machine learning model. The experiment is carried out using preprocessed training data set from which features are extracted Bow approach and algorithms are trained on BOW features. The hash map produced from this experiment is given in the following Table IV.

TABLE IV. RESULT OF EXPERIMENT 1 WITH BOW MODEL

Features	**No. of Features**	**Naïve Bayes**	**SVN**	**KNN**
Bag of Words	2317	71.52	**72.23**	68.95

2) Experiment 2: Experiment 2 is extension of experiment 1 and it yields better results by combining the token by N where N >= 1 & N = <4. The major objective of this experiment is to clearly demonstrate the difference of accuracy of supervise machine learning models from the experiment 1 by incorporating N-Gram technique with BOW approach. This experiment brings into use same data set as used in the previous experiment. In this experiment, the whole programming environment remain same as in previous experiment. It also brings into use preprocessed data set as discussed in previous section. Words with highest frequency

carry less information for a class as compared to least frequency. Supervised machine learning model is trained using training data set of 2400 comments and on the basis of this training, unlabeled and unseen comment (testing data set) is classified. Same features along with threshold value are applied as in last experiment. In feature extraction phase, different N-gram patterns are applied. We used N-Gram where value of N ranges from 1 to 4.

- Unigram pattern: consist of one word for extracting semantic information.
- Bigram pattern: consist of two word for extracting semantic information.
- Trigram pattern: consist of three word for extracting semantic information.
- Quad gram pattern: consist of four word for extracting semantic information.

Following the preparation of N-gram pattern matrix is used to compute the score of each individual class which is computed by total number of words in corpus divided by their individual frequencies as shown in Table V. Training model is generated and each classifier used in the experiment.

TABLE V. RESULT OF EXPERIMENT 2 WITH BOW AND N-GRAM MODEL

Features	No. of Features	Naïve Bayes	SVN	KNN
Unigram	2317	64.33	71.11	62.02
Bigram	13957	73.68	**82.32**	71.25
Trigram	11695	57.12	53.78	53.39
Quad Gram	17658	43.98	40.35	48.61

G. Comparison of Accuracies of Data Mining Techniques Followed in Experiment

Performance of the proposed classification system in terms of accuracy measures is depicted in Table V. It is showing accuracy measures of all the classifiers along with their N-Gram values. There are three classifiers employed with different N-Gram values from N=1 to N=4. Results reveal that Support Vector Machine (SVM) has outperformed with the best accuracy. SVM gives 71.11% and 82.32% accuracy with uni-gram and bi-gram respectively while KNN is better with tri-gram whereas the over performance of KNN is less than SVM. Naïve Bays stands between SVM and KNN in performance as shown in Tables VI, VII, and VIII.

H. Comparison of Precision and Recall Scores

Precision, recall and F1-score measures for all given algorithms using Uni-gram, Bi-gram, Tri-gram and Quad-gram. All three algorithms are applied one by one on the same data set to get accuracy, precision, recall and F1-score.

We have also performed K-fold (where k = 10) cross validation mechanism to validate the confusion matrix and data set authentications. We have divided our dataset into 10 folds (f1, f2, f3 . . . . f10) of equal size. We trained all the classifiers one by one to f1 to f9 folds and then from f1 to f8 and tested for f9 fold and so on.

The overall performance of all three algorithms are compared and their comparison report is given in the graph below. The Support Vector Machine which has been proved the best algorithm in our text classification system and it has better values for confusion matrix as compared to the KNN and Naïve Bays as shown in Table IX.

TABLE VI. RESULTS EVALUATION FOR NAÏVE BAYS ALGORITHM

Class – Naïve Bayes	Precision	Recall	f1-score
Analysis	0.76	0.75	0.76
Synthesis	0.81	0.66	0.75
Evaluation	0.72	0.57	0.82
Implementation	0.70	0.80	0.40

TABLE VII. RESULTS EVALUATION FOR SVM ALGORITHM

Class – SVM	Precision	Recall	f1-score
Analysis	0.61	0.76	0.69
Synthesis	0.74	0.71	0.10
Evaluation	0.75	0.69	0.51
Implementation	0.62	0.80	0.82

TABLE VIII. RESULTS EVALUATION FOR KNN ALGORITHM

Class – KNN	Precision	Recall	f1-score
Analysis	0.78	0.73	0.75
Synthesis	0.81	0.76	0.74
Evaluation	0.80	0.65	0.83
Implementation	0.72	0.68	0.39

TABLE IX. RESULTS EVALUATION FOR N-GRAM MODEL

Features	Class – Naïve Bayes	Precision	Recall	f1-score
Unigram	***Naïve Bayes***	0.71	0.72	0.71
	SVM	0.82	0.66	0.72
	KNN	0.75	0.68	0.64
Bigram	***Naïve Bayes***	0.88	0.82	0.84
	SVM	0.67	0.69	0.68
	KNN	0.68	0.68	0.62
Trigram	***Naïve Bayes***	0.54	0.66	0.60
	SVM	0.48	0.54	0.51
	KNN	0.46	0.48	0.53

IV. Discussion

This study incorporate data mining and machine learining technqiues along with NLP to develop an automated system to categrorize the products w.r.t given classes. We have used N-grams, BOW and TF-IDF technqies for features extratcion. There could be feature engineering tecniques which might improve the system relaibality. Our proposed system is good with proper grammar and well spelled words but in case of slangs in data set might confuse the system. The most important thig to discuss is that this is an initial approach to classify consumer comments under the given classes to assist both, seller and buyer. Therefore, we don't have any benchmark to compare our results.

V. Conclusion

This study has evaluated the proposed machine learning model with various data analytics techniques as mentioned in literature like accuracy, precision, recall and F1-score. Briefly, SVM algorithm along BOW and bi-gram features engineering techniques is proved winning classifier in the proposed E-commerce customer comment classifier. This work opens new avenues to E-commerce customers and sellers to get a quick status of customer opinion in four important contexts which helps many customers to decide the about purchase of product. At the same it assist sellers to improve their product or services in the given four context (convenience, variety, time and trust) using data mining and machine learning and Natural Language Processing. This wok not only demonstrates the usefulness of machine learning and data mining in E-commerce business development and customer assistance but also identify preprocessing techniques and the important features engineering methods.

VI. Future Work

This work is concerned to classification of E-commerce comments data using supervised machine learning model by incorporating BOW and N-gram feature engineering methods. Currently we have selected/preferred those words/features with highest value to its respective class. In future work, we shall employ semantic and syntactic features engineering techniques to select features with contextual relevance. In this way, we will get the improved percentage accuracy i.e. to consider different dimensions of vectors in E-commerce customer's comments classification and other aspects of E-commerce related text classification.

References

[1] Chen, X., Duan, S., & Wang, L. (2020, June). Comments Prediction Model on Emotional Analysis Based on Bayes Classification. In Journal of Physics: Conference Series (Vol. 1575, No. 1, p. 012020). IOP Publishing.

[2] Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: a study of consumer perceptions. Electronic commerce research and applications, 2(3), 203-215.

[3] Cirqueira, D., Hofer, M., Nedbal, D., Helfert, M., & Bezbradica, M. (2019, September). Customer purchase behavior prediction in e-commerce: a conceptual framework and research agenda. In International Workshop on New Frontiers in Mining Complex Patterns (pp. 119-136). Springer, Cham.

[4] Raorane, A., & Kulkarni, R. V. (2011). Data mining techniques: A source for consumer behavior analysis. arXiv preprint arXiv:1109.1202.

[5] Voinea, L., & Filip, A. (2011). Analyzing the main changes in new consumer buying behavior during economic crisis. International Journal of Economic Practices and Theories, 1(1), 14-19.

[6] Nayyar, T. (2019). Analyzing Customer Buying Behavior.

[7] Saeed, R., Lodhi, R. N., Rauf, A., Rana, M. I., Mahmood, Z., & Ahmed, N. (2013). Impact of Labelling on Customer Buying Behavior in Sahiwal, Pakistan. World Applied Sciences Journal, 24(9), 1250-1254.

[8] Pahwa, B., Taruna, S., & Kasliwal, N. (2017). Role of Data mining in analyzing consumer's online buying behavior. International Journal of Business and Management Invention, 6(11), 45-51.

[9] Familmaleki, M., Aghighi, A., & Hamidi, K. (2015). Analyzing the influence of sales promotion on customer purchasing behavior. International Journal of Economics & management sciences, 4(4), 1-6.

[10] Singh, M., Jyani, L., Verma, R., Rajpurohit, L., & Goswami, P. Analysis of Consumer Behavior on SCM Related Factors Using Data Mining: A Case Study of the Indian E-Commerce Industry.

[11] Altunan, B., Arslan, E. D., Seyis, M., Birer, M., & Üney-Yüksektepe, F. (2018, August). A data mining approach to predict E-Commerce customer behaviour. In The International Symposium for Production Research (pp. 29-43). Springer, Cham.

[12] Prabhakumari, K., & Silviya, M. T. ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[13] Anggoro, M. A., & Purba, M. I. (2020). The Impact of Attractiveness of Ads and Customer Comments Against to Purchase Decision of Customer Products on the User of Online Shop Applications in the City of Medan. Jurnal Ilmiah Bina Manajemen, 3(1), 1-9.

[14] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[15] Mattosinho, F. J. A. P. (2010). Mining Product Opinions and Reviews on the Web. Technische Universitat Dresden.

[16] Gowtamreddy, P. (2014). Opinion mining of online customer reviews (Doctoral dissertation).

[17] Martin, M. (2017). Predicting ratings of amazon reviews-techniques for imbalanced datasets.

[18] Prabhakumari, K., & Silviya, M. T. ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[19] Chen, X., Duan, S., & Wang, L. (2020, June). Comments Prediction Model on Emotional Analysis Based on Bayes Classification. In Journal of Physics: Conference Series (Vol. 1575, No. 1, p. 012020). IOP Publishing.

[20] Mahmud, B. U., Bose, S. S., Majumder, M. M. R., Arefin, M. S., & Sharmin, A. Ecommerce Product Rating System Based on Senti-Lexicon Analysis.

[21] Belém, F. M., Silva, R. M., de Andrade, C. M., Person, G., Mingote, F., Ballet, R., ... & Gonçalves, M. A. (2020). "Fixing the curse of the bad product descriptions"–Search-boosted tag recommendation for E-commerce products. Information Processing & Management, 57(5), 102289.

[22] Sahib, S. M. Customers Buying Behaviour towards Online Shopping-A Study of Rural People in Southern Western Region of Punjab Dr. Monica Bansal.

[23] Kunjithapatham, K. A., & Santhanakannan, A. A Study on Consumer Behaviour towards Online Shopping in Kanchipuram Town.

[24] Sahu, M. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh. Journal of Xi'an University of Architecture & Technology Issn No, 1006, 7930.

[25] Vishwakarma, A., Ojha, T., & Mohanty, D. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh.

[26] Prabhakumari, K., & Silviya, M. T. ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[27] Anggoro, M. A., & Purba, M. I. (2020). The Impact of Attractiveness of Ads and Customer Comments Against to Purchase Decision of Customer Products on the User of Online Shop Applications in the City of Medan. Jurnal Ilmiah Bina Manajemen, 3(1), 1-9.

[28] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[29] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[30] Altunan, B., Arslan, E. D., Seyis, M., Birer, M., & Üney-Yüksektepe, F. (2018, August). A data mining approach to predict E-Commerce customer behaviour. In The International Symposium for Production Research (pp. 29-43). Springer, Cham.

[31] CLARK, M., RUTHVEN, I., HOLT, P., SONG, D., & WATT, S. (2012). OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University.

[32] RANDERSON, K., BETTINELLIB, C., FAYOLLE, A., & ANDERSON, A. OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University.

[33] JENNINGS, B., TSATTALIOS, K., & CHAKRAVARTHI, R. OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University. Scientific Reports, 6, 20504.

[34] Johnson, I. M., & Copeland, S. M. (2008). OpenAIR: The Development of the Institutional Repository at the Robert Gordon University. Library Hi Tech News.

[35] SACHSE, S., SILVA, F., IRFAN, A., ZHU, H., PIELICHOWSKI, K., LESZCZYNSKA, A., ... & KUZMENKO, O. OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University.

[36] Basani, Y., Sibuea, H. V., Sianipar, S. I. P., & Samosir, J. P. (2019, March). Application of Sentiment Analysis on Product Review E-Commerce. In Journal of Physics: Conference Series (Vol. 1175, No. 1, p. 012103). IOP Publishing.

[37] Li, N., & Zhang, P. (2002). Consumer online shopping attitudes and behavior: An assessment of research. AMCIS 2002 proceedings, 74.

[38] Nagra, G. K., & Gopal, R. (2014). Consumer Online Shopping Attitudes and Behavior: An Assessment towards Product Category. International Journal of Marketing and Technology, 4(5), 54.

[39] Sahu, M. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh. Journal of Xi'an University of Architecture & Technology Issn No, 1006, 7930.

[40] Vishwakarma, A., Ojha, T., & Mohanty, D. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh.