Computer Systems Science & Engineering
DOI:10.32604/csse.2022.022579
Article

Tech Science Press

An Improved Method for Extractive Based Opinion Summarization Using Opinion Mining

Surbhi Bhatia* **and Mohammed AlOjail**

Department of Information Systems, College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia
*Corresponding Author: Surbhi Bhatia. Email: sbhatia@kfu.edu.sa
Received: 12 August 2021; Accepted: 13 September 2021

Abstract: Opinion summarization recapitulates the opinions about a common topic automatically. The primary motive of summarization is to preserve the properties of the text and is shortened in a way with no loss in the semantics of the text. The need of automatic summarization efficiently resulted in increased interest among communities of Natural Language Processing and Text Mining. This paper emphasis on building an extractive summarization system combining the features of principal component analysis for dimensionality reduction and bidirectional Recurrent Neural Networks and Long Short-Term Memory (RNN-LSTM) deep learning model for short and exact synopsis using seq2seq model. It presents a paradigm shift with regard to the way extractive summaries are generated. Novel algorithms for word extraction using assertions are proposed. The semantic framework is well-grounded in this research facilitating the correct decision making process after reviewing huge amount of online reviews, considering all its important features into account. The advantages of the proposed solution provides greater computational efficiency, better inferences from social media, data understanding, robustness and handling sparse data. Experiments on the different datasets also outperforms the previous researches and the accuracy is claimed to achieve more than the baselines, showing the efficiency and the novelty in the research paper. The comparisons are done by calculating accuracy with different baselines using Rouge tool.

Keywords: Sentiment analysis; data mining; text summarization; opinion mining; principal component analysis

1 Introduction

The opinions are mushrooming on the Internet and it has now become a trend to post reviews online due to its easy availability. It has further lead to abundance of data that is readily available for researchers to come out with novel solutions in opinion and text mining. Information Extraction (IE) isolates the relative information from the text in the form of entities and relationship among them for building structured databases. The availability of cheaper and advance Internet services made the consumers to use it to express their feeling in the form of opinions, feedbacks, comments or blogs. This lead to interest in

developing the automatic system which can extract, classify and summarize the data [1] generated on "Consumer generated media" automatically. A large number of recommender systems want to dig into the summaries to predict the recommendations and the followings of the user in which he is interested depending on his posts and opinions interests [2]. Opinion mining has subsided the traditional feedback system and impacted the web users to use the online review or opinions for making their decision for a particular product or service. Society can be benefited by these opinions or reviews available online in number of ways. It provides fusing the sentiments for a complete involvement of scope regarding opinion classification. People who are interested can easily get the summary of the reviews as per the interests in features of the product. There can be a cumulative count that can be taken up collection of both positive and negative reviews completely on the basis of features. According to the time spent on internet reviews that are reliable with less efforts are taken up by people. Summarization has become an essential service to remove the unwanted content from the web. Producing summary from different set of reviews can be as a process called as opinion summarization. The summarization has two broad categories: extractive and abstractive. Extractive summarization agonizes with a problem called dangling problem as it is mainly concerned with the summary content. So, sometimes not so important parts of sentences also get included and also sentences which are extracted out of the sentence, like pronouns lose important information those otherwise needs to be conserved [3]. To produce grammatical summary for advanced language generation techniques, a strong emphasis on the form is required. Abstractive summaries of opinions, review etc. gives better result for summarization than extractive summaries for the above listed reasons and the research demonstrated earlier [4]. Under the umbrella of the term Natural Language Processing (NLP), summarization of text is an important part and it's an intersection of computer science and linguistic towards the evolution of NLP. It can be resulted in swapping of notion and thoughts between computer and human by applying NLP on the data which is processed. NLP comprising the analytics of data which is processed at lexical, semantic, and syntactic level along with discourse processing. The task of summarizing the opinions includes selecting feature [5], rating those feature, [6] and sentence identification that contain features [7] on multiple range of opinions present in web document. Primarily summarization is grouped into two classes. There may be other approaches such as general graph based, and hybrid techniques [8]. The proposed work covers extractive approach where summary of sentences is achieved by considering features. The extractive summary generated on the feature 'food and service' with dataset containing 100 reviews as Food from opinions dataset. Extractive approach uses the principle of applied Artificial Intelligence (AI) models. Two models are used, one is using Principal Component Analysis (PCA) and the other one is Bidirectional Recurrent Neural Networks and Long Short-Term Memory (RNN-LSTM) deep learning model. PCA is a statistical technique that helps in transforming an array of data values which are associated or correlated in some form into values that are linearly uncorrelated data sets known as principal components using orthogonal transformation. PCA is considered as an important tool for analysing data as it finds interesting patterns in the data by highlighting the differences and similarities in the patterns. The work includes the application of PCA in summarization of text by reducing the number of dimensions in data the (words or entities as features) and reasonably abstracting of the reviews on ranking the most pertinent ones, considering the prime features devoid of any deprivation of information respective of a distinct domain. This method will help in removing those features into consideration which are not in the top priority with respect to the particular domain. The second uses the encoder-decoder architecture a method of building Recurrent Neural Networks (RNNs) for grouping forecasts. Here encoder takes the total information grouping by encoding in a fixed length vector and decoder takes the encoded input from the encoder and defines the yield grouping. It is built on a character-level seq2seq model for text rundown.

Instead of traversing and searching the reviews on different sites, the decision can be taken by the user quickly by finding the sentences which are taken to be the representative of the corpus. These are selected as

the most relevant sentences since their vector representations are approximated in a best way by expecting them into the limits of the sparse principal components [9]. The result of the above explained Text Summarization using Opinion Mining (ETSUOM) is evaluated and compared using Rouge tool. ROUGE is a widely adopted, automatic evaluation measure for text summarization. It has been shown to correlate well with human judgments and works well with extractive summaries where surface lexical similarities are not considered. This tool is used evaluation tool for multi-document summarization and has great advantages in the areas of extractive summarization evaluation. The main contributions of the paper are listed as follows:

- To study the background on the different summarization techniques.
- To develop a novel extractive method for summarizing opinions that will greatly reduce computational costs for fast and iterative exploration.
- To propose the algorithms required for extracting words using assertions, given then compressing and merging information based on AI models.
- To evaluate the work on different datasets and measure using on the basis f Information retrieval metrics using Rouge tool.

The paper has been decomposed into three parts. Related work includes the previous research already done on recent abstractive summarization techniques. All the previous work followed on summarization is explained in brief with its advantages and drawbacks. In the Methodology section, the technique we are implementing along with the algorithm and its example with all the manual formulas are explained. Results and analysis shows all the values evaluated and the comparison is done using manual consideration. Last, Conclusion part, all the possible and relevant areas that can be worked out further are detailed out.

2 Literature Review

A range of literature is available on successful abstractive based technique summarization of text from web documents. It includes rule based approach [10], sentence compression [11–13] merging sentence based on their semantics [14,15]. The authors [16] proposed construction of graph using hash algorithms for effective ranking system. Directed graph has been used by the researchers [17] for generating abstractive summary. They have reduced and replaced the redundant sentences as opinions by using the connectors graphically in the form input. This has been concluded as readable, fairly well-formed and concise summaries. This method has a drawback which mark that pre-existing of connectors are not available which is unable to consolidate and integrate the sentences that can be tied up together. The emphases on surface order of words increases the complexity of proposed methodology. A complete structure of mining opinions has been presented by Bhatia et al. [18] where opinions are classified using Naïve Bayes classifier and summarization of opinions is done. The unified framework of mining opinions is the strength and the novelty of the work. The authors [19] proposed a novel method of summarization at document level adding up the benefits of conventional lexical chain method to retrieve summaries. The method relies on unique features or various keyword in a text by presenting the probability distribution model upon each feature.

The authors [20] built the word graph and scoring the path by imposing POS constraints to generate the graphical based summary. They have used Rouge tool for fusion of sentences to incorporate the rules of NLP. Their work is limited to fusion of sentences. A soft computing based automatic text summarization using semantic similarity has been proposed by Tayal et al. [8]. To summarize the text, they have used the human summarization rules of Subject, Verb and Object as well as it is based on this approach designed an NLP parser. But the rules stated in parser were not enough for summarizing the complex sentences

and leads to tag based ambiguity. It was not enough to support the errors. The authors [10] used PCA to summarize the large corpora and contrived the approximation related to low rank of a Salton matrix to form a sparse PCA. This set up they have applied on collection of news articles. They did not explain manually how the encoding of sparse PCA is achieved. Also left a doubt on how low variance features are not affecting the accuracy. This research will fill this gap. Agrawal et al. [21] used dependency relations based on part of speech constraints to find features and calculated polarity of all these features are using both super-vised and unsupervised methods. The author evaluated the proposed method on the dataset of Movie reviews. The authors [22] extracted the common sense knowledge from ConceptNet based ontology for sentiment analysis.

Whereas authors of [23] proposed supervised learning based technique to classify sentences using dictionary based approach. They have analyzed the sentiments to deduce the semantic priority to generate the summary of text. They have provided the detailed computing polarity using SentiWordNet[1]. They have used different dataset to provide empirical result comparison. The paper [24] discussed detailed overview of the concepts and applications of deep learning. The author discussed the deep and shallow deep NN and presented the concepts of machine learning techniques behind deep learning. The researcher proposed the Word2Vec model obtain word embedding of unprecedented quality and perform scaling of naturally to very large data sets. The paper [25] describes the background implementation of Dynamic CNN named with CNN and experimented in four things on Twitter sentiment prediction. The paper [26] trained a classifier using corpus to detect emotions by using syntactic structure of text. The empirical experiments of proposed method showed the good accuracy and efficiency as compared with the existing approaches. The future direction shows towards implementing a Multi lingual sentiment. The authors [27] presented a model related to graph showing opinion which focuses the content which is domain oriented domain. The existing user graph based approach and the opinion graphs comparison has been described. Combination of both the approaches, taking a hybrid approach is shown as a future scope adding the temporal and time analysis feature in the opinion. The authors [28] proposed the extractive summarization of text using two main aspects; reviewing the content and the creditability of the author. The necessity of a sentence is evaluation by considering the semantics in the couple of sentences. The paper [29] discusses the email classification of email using novel method of identifying features though the interconnections made between the words of classes and words having higher probability have been taken for representing informative features. The authors [30] discussed summarization method by presenting the comparison between abstractive and extractive summarization techniques, the work is carried on Rouge tool.

3 Proposed Methodology

Proposed technique covers feature extraction and summarization. In the first part, Word Extraction Using Assertions Given by ConceptNet is discussed. Then a novel algorithm to extract the words from text reviews has been explained. One method is considered as a baseline that aims to determine all relevant concepts with the help of assertions given by ConceptNet related to the domain. A simple ontology only for a specific domain (not for its synsets) given by ConceptNet is considered as input to word extraction algorithm [31]. To improve the performance of baseline method, novel approach for word extraction is proposed in which the importance of the extended domain ontology is accessed in improving precision, recall and F-scores. The second baseline is taken is a combination of unsupervised learning with NLP [32]. The words are assigned a high weight for keyword extraction in a sentence. The word's weight is adjusted by the salience scores of the sentences to leverage the sentence information. The sentence score is calculated based on its cosine similarity for getting the results. The overall methodology is given in Fig. 1.

[1]https://github.com/aesuli/SentiWordNet.

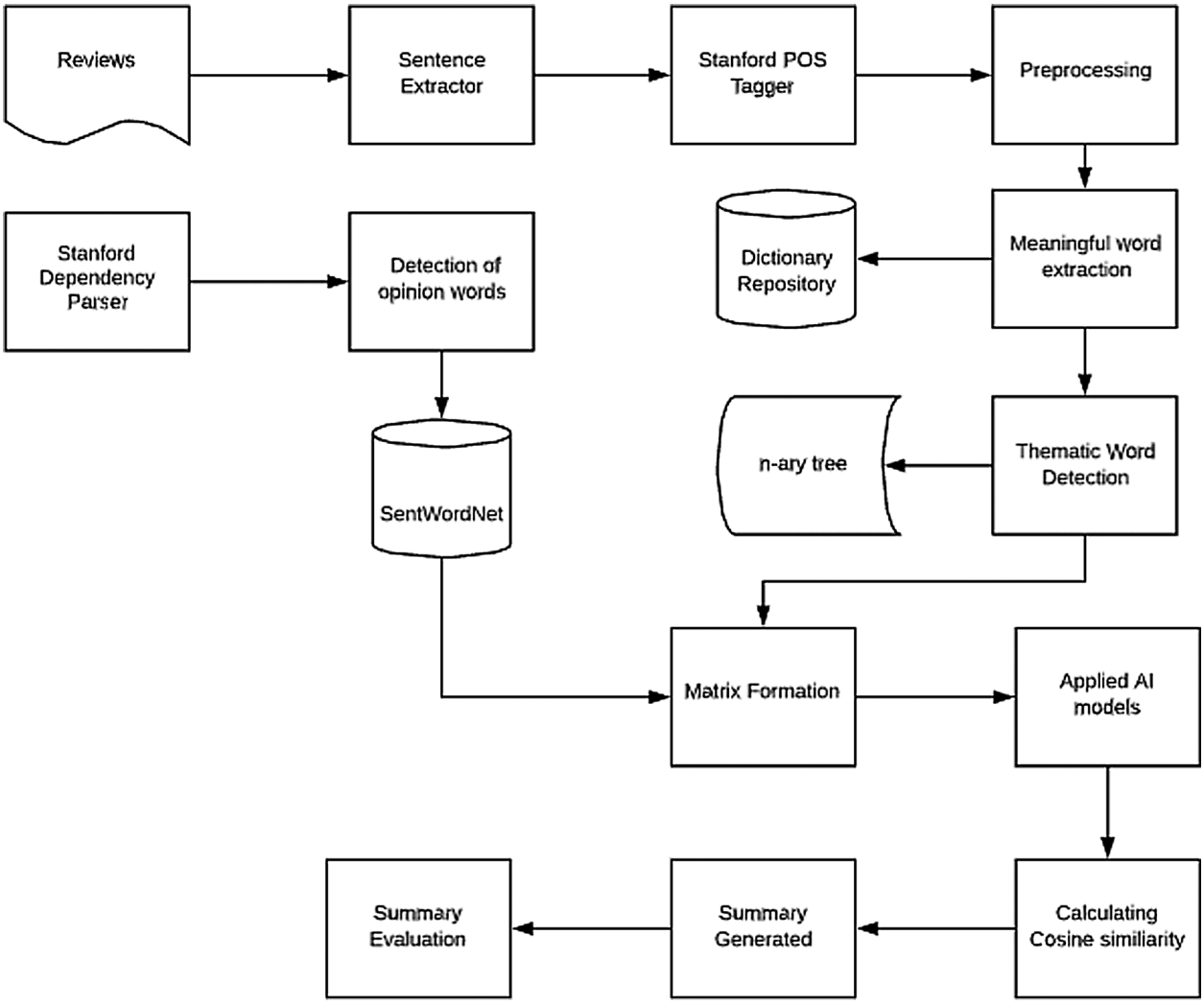

Figure 1: Framework design of the proposed model

In the second part, applied AI models are used for generating extractive summaries which includes two important models, one is PCA and the other is deep learning models. PCA will reduce the number of dimensions that will drop having irrelevant words from the set of reviews and uses the ranks of reviews to isolate the relative summary. Deep learning model will help to extract the concealed setting of the sentences. The encoder-decoder engineering used in this model comprises of two essential models: one peruses the information succession and encodes it to a fixed-length vector, and the second disentangles the fixed-length vector and yields the anticipated grouping. This engineering is intended for seq2seq issues. The prime features have considered for ranking of the reviews without any loss of information. The summary is compiled by checking the semantics of the sentences and a decision can be taken by filtering out the irrelevant opinions from the relevant ones. There are three components. The step by step flow is given in Fig. 2.

• User: The product for required summary is given as a query. The query vector is chosen such that the most frequently used words are extracted first by applying the counter in the reviews extracted online. The words with the highest score are taken as a query vector as they can influence one's decision in recommendation of a product or rejection of the product.

• Summarizer: The extractive summary on the basis of query vector. The components of summarizer will be explained further in proposed methodology.

• Corpus: Last is corpus from which summary is being generated.

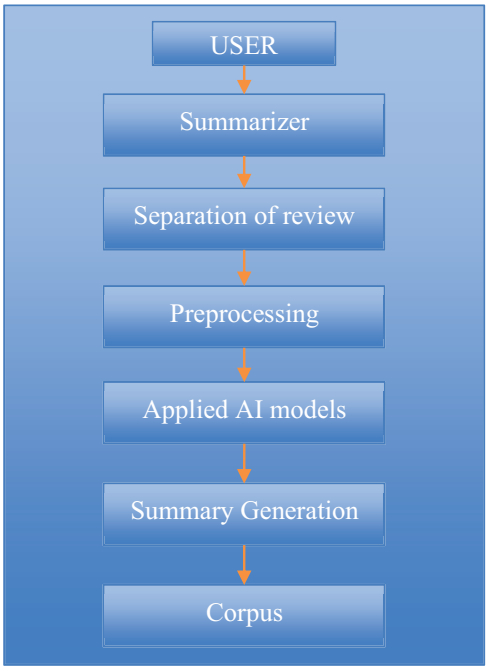

Figure 2: Flow diagram showing the interconnection between components

3.1 Methodology for Word Extraction

Equations Word in text describes the main idea of the review, for example, "Avengers is a fantastic movie". Here the sentiment has expressed for the word 'Avengers'. A novel approach for identifying the 'word' in review, an entire domain about word extraction has been proposed here. ConceptNet has been used to fetch the common sense open mind knowledge and then applied for word extraction from the reviews. Further, identification of relevant concepts and synonyms of the concepts from these ConceptNet assertions are taken as nodes, and then n-ary tree constructed using these nodes of the specific domain. All the synonyms have used to extend the ontology by constructing n-ary tree for complete understanding of the assertions explained above.

The significant words from dataset, obtained after applying preprocessing on each token and checking its existence in thesaurus dictionary are then matched against the nodes of each n-ary tree constructed. This approach helps to prune irrelevant meaningful term present in the dictionary, but not a word.

3.1.1 Extraction and Cleaning of Reviews

The reviews are extracted by crawling from E-forums, Social Media, and other review sites. The method to parse the HTML source for inspecting the keywords such as summary of reviews, comment, and opinion has been adopted from our earlier work [33]. The timestamp between the seed URLs is calculated using Update formula. The extracted opinions are stored in Review Repository. Fig. 3 describes the structure of cleaning opinions.

The removal of stop words is an important step of NLP, as it reduces the data size. In this preprocessing step, the dataset is reduced by approximately 30% (stop words) and indexing size by approximately

40%–50% using stemming. This mainly includes three phases *i.e*., review extractor, tokenizer and SSLN (acronym for stop word removal, stemming, lemmatization and normalization). The Reviews has been extracted by Review Extractor from review repository and split into sentences with sentence ID. Tokenizer generates tokens from sentences and send signal to SSLN.

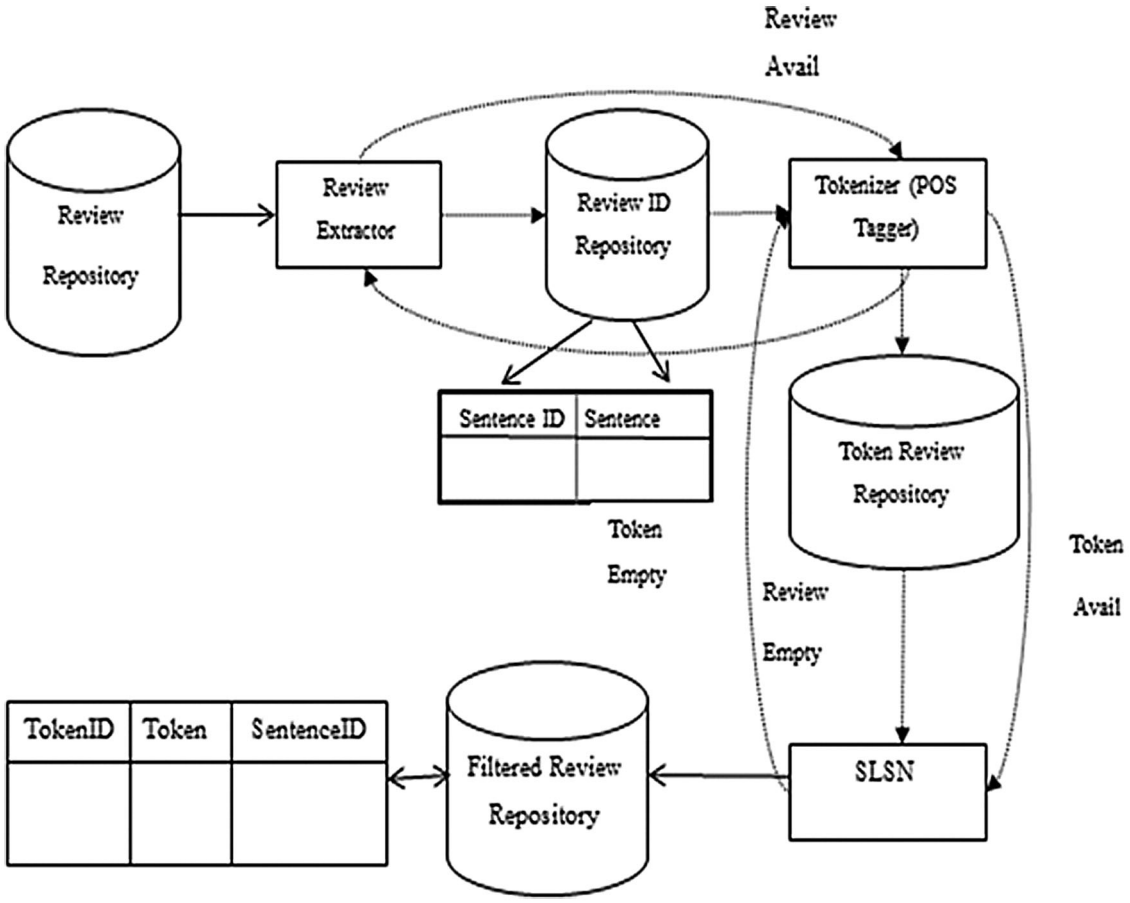

Figure 3: Structural model of preprocessing diagram

The reviews get extracted from the repository and got split up into the individual sentences. The ReviewIDRepository store all the sentences with their particular sentence ID. ReviewAvail signal to the Tokenizer is sent by ReviewExtractor to sentences tokenization. If all the sentences have been tokenized in the ReviewIDRepository, then Tokenizer sends ReviewEmpty signal to the ReviewExtractor. The Algorithm is given below in Algorithm 1.

Algorithm 1: Review extraction Algorithm

Extractor
{
Wait(*ReviewAvail*)
fetch review from the *ReviewRepository*
break review into individual sentences and allocate a
unique sentence ID to them
Store in *ReviewIDRepository*
signal (*ReviewEmpty*)
}

Tokenization process makes each extracted review to split up into tokens for POS tagging. POS tagging of all the documents is done by Stanford NLP Parser. The tokens generated got stored in the TokenReviewRepository. Tokenizer sends the TokenAvail Signal to SSLN and if no token is present, and then SSLN will revert back TokenEmpty signal to Tokenizer. The algorithm for the same is given in Algorithm 2.

Algorithm 2: Tokenization

```
Tokenizer
{
Wait (TokenAvail)
Fetch Sentences from ReviewIDRepository
Tokenize the extracted sentences into tokens
Store in TokenReviewRepository
Signal (TokenEmpty)
}
```

SSLN is a phase of pre-processing that reduces the data size and increases the efficiency. To achieve the same following four tasks has been carried out:

Removal of Stop Word: Stop words appears frequently in text and usually words are the, a, an, for, which, what, among and many more. Stop words are removed for review just to increase search performance.

Stemming: Stemming refers to reducing the word to its fundamental form; mostly it chops off derivational affixes. For example, the word cars get reduced to *i.e.*, car. Lemmatization aims to form lemma which contains the base or dictionary form of a word using morphological analysis of words. For example, the word saw get reduced to the root form, *i.e.*, saw or see. Normalization is defined as the process of canonicalizing each token. All these tasks will be performed and tokens with its token id and sentence is stored in FilteredReviewRepository. It is given in Algorithm 3.

Algorithm 3: SSLN Algorithm

```
SSLN
{
Wait (TokenAvail)
Fetch Tokens from the TokenReviewRepository
Apply stop word removal
Apply Lemmatization
Apply stemming
Apply normalization
Store in FilteredReviewRepository
Signal (TokenEmpty)
}
```

3.1.2 Creating Knowledge Base

Tokens available in FilteredReviewRepository are extracted and then their existence is checked in the dictionary to find whether the term extracted is a meaningful term or just a dummy. If it is a meaningful

term, then will get added to the KnowledgeBaseRepository else pruning is done. The algorithm is given in Algorithm 4.

Algorithm 4: Creating knowledge base algorithm

```
Creating KnowledgeBase
{
Wait (TokenAvail)
Fetch the token from FilteredReviewRepository
Check the given token is a dictionary word or not
if (it's a dictionary word && not a polarity word)
Add into knowledgeBaseRepository
else Prune
signal (TokenEmpty)
}
```

The structural model of extracting features has been displayed in Fig. 4.

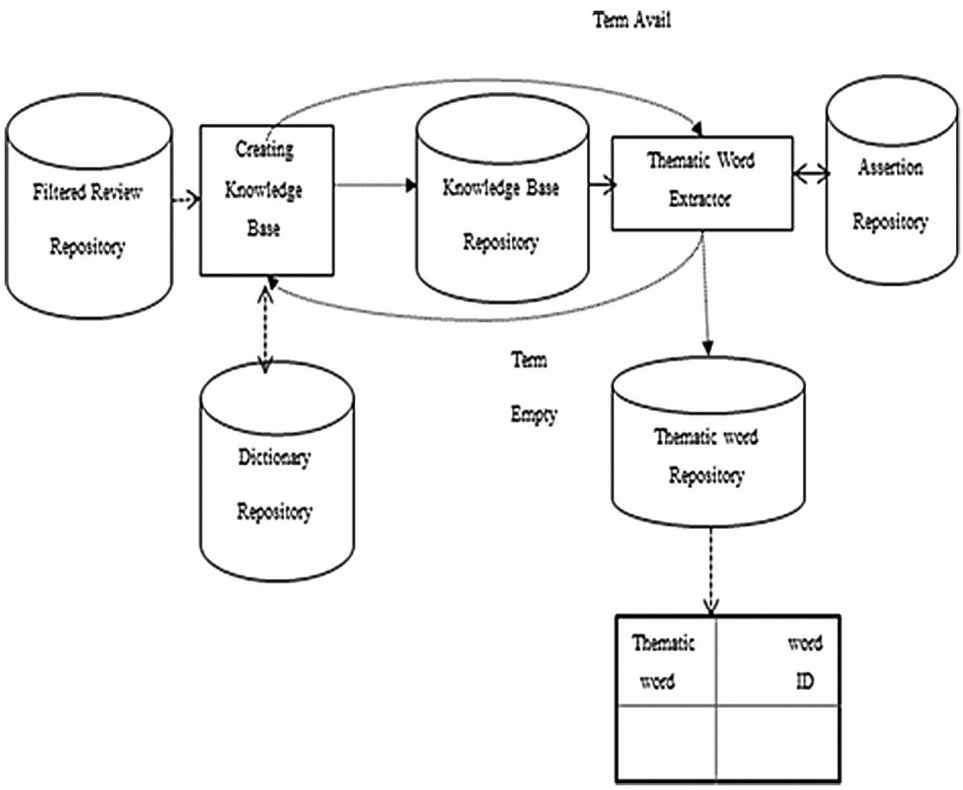

Figure 4: Structural model of feature extraction

Equations Raw Assertions fetched from ConceptNet for the specific domain are stored as concepts (nodes in n-ary concept tree for construction of n-ary tree of specific domain). Further, for greater coverage of concepts we expand the ontology by constructing n-ary tree of all synonyms related to the

specific domain. Connecting root with any vertices with an edge E gives the relation R. N-Ary tree for all domains will be constructed by fetching concepts (*i.e.*, vertices present in N-Ary tree) from ConceptNet Edges in N-Ary tree represent the relationship between domain and that concept.

3.1.3 Words Extraction

Meaningful Terms in KnowledgeBaseRepository are extracted and matched with the nodes of n-ary tree, if match becomes a hit then term is added in WordRepository else prune. In Fig. 4 FilteredReviewRepository contains the pre-processed tokens that are mapped from a dictionary and stored in KnowledgeBaseRepository. Terms from KnowledgeBaseRepository are searched in n-ary tree. The term is a word, if it is available, else prune. Algorithm 5 explains the details of the algorithm.

In this proposed approach all the words related to the domain and even to its synsets are detected. The standard evaluation measures are analyzed in the next section and the performance is compared further.

Algorithm 5: Word Exraction Algorithm

```
ThematicwordExtractor
{
Wait (TermAvail)
Fetch the term from knowledgeBaseRepository
For (i=1;i<=n; i++) { // where n is no of n-ary tree
if (Term is node in N-Ary tree )
Add in ThematicWordRepository
else
prune
signal (TermEmpty)
}
```

3.2 Methodology

Applied AI is the part of man-made consciousness that brings it out of the lab and into this present reality, empowering real world use cases into business problem and day-to-day activities. Applied AI improves programming applications and puts propelled AI to utilize, giving elevated levels of exactness and variation after some time. Applied AI is contextualizing plans of action and industry forms, just as improving the manner in which we collaborate with everything around us. In this paper we discussed use of Applied AI models for generating extractive text summaries. In this paper we discussed two Applied AI models which include:

PCA Based Summarization

Here a PCA based summarization is proposed to gather the main token of words from the set of reviews and generates extractive summary. PCA algorithm in combination with SVD is used by dropping extraneous words and unnecessary sentences to determine the relevant text. The dimensionality redactions are considered by explained the detailed algorithms in the next subsection. To generate the summary following steps are involved:

Sentiment Extraction

For each word, the adjectives and its phrases are extracted from sentences by feeding them into the Stanford dependency parser. For each extracted adjective and its phrases, (SentiWordNet), lexical resource, used for retrieving and scoring the words [34] using Stanford Parts-of-speech tagging. The

scores generated *via* SentiWordNet are used in scoring the overall sentence for Opinion analysis. The scoring for the words is calculated using the connectors 'and' and 'or' for the same features. If both the words are adjectives and connected by "and", then the scoring is proceeded by adding the scores of adjective 1 and adjective 2.

If either the words are adjectives and connected by "or", then the scoring is calculated by taking score of any one of the adjective with the higher score. The adjectives are denoted as Adv as shown below:

Adv1 and Adv2: Scoring (Adv1) + Scoring (Adv2) for connector 'and' (both of two adjectives)

Adv1 or Adv2: Scoring (Adv1) or Scoring (Adv2) for connector 'or'. (Either of the two adjectives)

Creation of Matrix for Summarization

The matric is created using reviews and aspect words, where m denotes reviews and n denotes the aspect words related to m. [T] denoting matrix = ab…n [ ], where a, b, c… ,n are the aspect words. [Tij] will be SentiWordNet score of sentiment extracted, denoting the values in matrix.

Each n aspect word will be associated with m number of reviews; these reviews are generally adjectives associated with each aspect.

Now, we create a matrix [T] with rows:

$$T = ab\ldots\ldots n \begin{bmatrix} ma & mb\ \cdots & m\ times \\ m'a & m'b\ldots & m\ times \\ m''a & m''b\ldots & m\ times \end{bmatrix}$$

The principal components of any matrix are its Eigen values, which are calculated using Singular Value Decomposition (SVD), explained in below section.

PCA implementation: Matrix created for PCA Algorithm is used as an input for Applying PCA.

Apply SVD on matrix created in previous step. Here, S, U and VT are the three matrices which is the output of SVD. U and V matrix will be used up to rank k (*i.e.*, ignoring lower sparse parts of those matrices) known as a rank k approximation. The value of k is a parameter and have critical importance. Apply the counter to find the frequency count of each word, known as term frequency. Choose the word with the highest count value as query vector as qv. The term qv is taken to benchmark the occurrence of a particular word, which is then used for similarity analysis of feed-in data. Next is to find the measure of similarity between qv and rv (review vector), *i.e.*, Cosine Similarities denoted as m

$$(q, r) = q.r/|q|.|r| \tag{1}$$

The values of m are sorted in descending order. Thereafter, applying the threshold on sorted values.

4 Results and Discussion

In this experimental setting, the meaningful terms that exist in the dictionary from the review documents were extracted and similar terms were selected which are only domain related words. The work has been implemented using the python library popularly known as sumy. It consists of Lexranksummarizer, LuhnSummarizer and LsaSummarizer.

4.1 Datasets

Two datasets have been taken for producing the summary: Opinosis dataset and Amazon food reviews dataset. Opinosis dataset has been used that contain reviews, gold standard summaries, script for Rouge and documentation. The topic "food in restaurant" containing 100 reviews are taken. Some reviews related to "food in restaurant" from Opinosis dataset are given below with their summarization results.

Amazon food review dataset [35] has been used that contain user id, username named as profile name, helpfulness score, summary and text review. Dataset consist of 10,000 rows out of which around 5000 rows are used for building the model. Some reviews and their summarized are mention below from the Amazon food review dataset in Tab. 1.

Table 1: Snapshot from Amazon food review dataset

Summary	Text
Good Quality Dog Food	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most.
Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts…the peanuts were actually small sized unsalted. Not sure if this was an error or if the vendor intended to represent the product as "Jumbo".
"Delight" says it all	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts–in this case Filberts. And it is cut into tiny squares and then liberally coated with powdered sugar. And it is a tiny mouthful of heaven. Not too chewy, and very flavorful. I highly recommend this yummy treat. If you are familiar with the story of C.S. Lewis' "The Lion, The Witch, and The Wardrobe"–this is the treat that seduces Edmund into selling out his Brother and Sisters to the Witch.
Cough Medicine	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer Extract I ordered (which was good) and made some cherry soda. The flavor is very medicinal.
Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.

4.2 Baseline Method

The two baselines are compared for demonstrating the work with better accuracy in terms of precision and recall. The first baseline (Baseline 1) taken for comparison is a simple ontology based method used for specific domain (not for its synsets) given by ConceptNet [28]. In this approach, ontology related to particular domain is considered and extraction algorithm is applied. Second approach (Baseline 2) is using unsupervised learning with the help of NLP toolkit [36]. The domain specific data from the dataset is extracted and then training is performed on the data using the deep learning model. For example: Out of different food categories, if there is a need to choose only one food category then filtration is carried out on the dataset in the initial stage and afterwards, deep learning model is built on the filtered dataset.

4.3 Evaluation Measures

Evaluating system generated extractive summary with gold summary by unit overlap measure. Unit Overlap measure, calculate overlap between set X and set Y, where set X is the words present in reference summary and Y is the set of words present in candidate summary.

$$(X, Y) = |X \cap Y|/|X| + |Y| - |X \cap Y| \tag{2}$$

Proposed opinion mining work is summarized with the comparison is obtained between two types of summaries. One is the system generated summaries as candidate summaries that is referred to as 'peer summaries' and the human generated summaries as reference summaries is referred to as 'gold standard

summaries' [37]. Abstractive and extractive techniques are being compared. Summary evaluation method used is N-gram Co-occurrence Statistics–ROUGE.

The quality of summary can be measured with ROUGE. Lin [38] introduced the concept of ROUGE metric which has been taken up by the Document Understanding Conferences (DUC) and conferences help on NLP.

$$\text{ROUGE} - \text{N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{3}$$

where, Countmatch(gramn) counts the overlapping units of n-grams that occur in candidate summary and gold summary and Count(gramn) is the number of n-grams in the reference summaries. ROUGE-1 refers to overlap of unigrams and ROUGE-2 refers to overlap of bigrams.

The IR measures, p as precision, r as recall and F-score as f evaluated and stated as below. Purity and entropy can also be considered as performance metrics [39]. The number of overlapping sentences/no of sentences in system generated summary is given by p. The number of overlapping sentences/no of sentences in gold summary denoted r. The harmonic mean of the above two is denoted as f mathematically [40].

4.4 Performance Analysis

The proposed model is fabricated which includes successive data. This incorporates Sentiment order, Neural Machine Translation, and Named Entity Recognition – some extremely basic utilizations of successive data. On account of Neural Machine Translation, the information is a book in one language and the yield is additionally a content in another dialect. In the Named Entity Recognition, the info is a succession of words and the yield is a grouping of labels for each word in the information arrangement. The major goal is to fabricate a book summarizer where the information is a long succession of words and the yield is a short outline. Proposed approach returns set T of words, and the set of terms labeled by human annotators be H. Tab. 2 shows the comparison of the results.

Table 2: Results for word extraction obtained after comparing for the sentences that contain nouns

S. No.	Approaches	Dataset 1			Dataset 2		
		Precision	Recall	F-Score	Precision	Recall	F-Score
1.	Baseline 1	0.83	0.71	0.76	0.65	0.62	0.63
2	Baseline 2	0.85	0.72	0.78	0.71	0.72	0.71
3.	Proposed method	0.90	0.81	0.84	0.72	0.70	0.70

The results have also been compared on Dataset 1 using different approaches of ROUGE given in Fig. 5. The results of the above three measures of IR using Rouge tool suggest the improvement in the efficiency up to a great extent over Unit overlap.

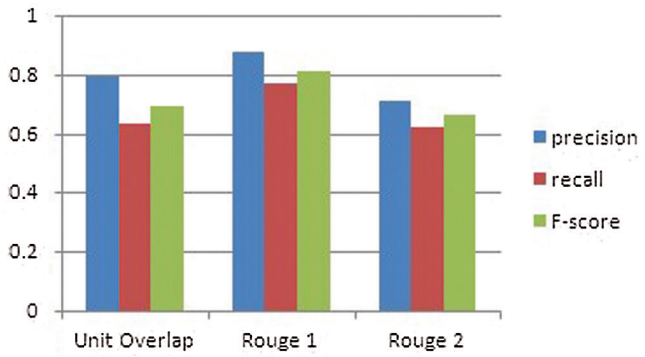

Figure 5: Graphical representation of measures of summary evaluated measures

5 Conclusions

The three main characteristics in summary generation are investigated in this research. Specific domain ontology, importance of the features, relevant sentences all together in summarization of reviews. The proposed methodology is implemented on the topic of Opinosis dataset and Amazon food review dataset. Applied AI models are discussed in this paper for extractive text summarization which include Deep learning algorithm and Principal Component Analysis (PCA). Deep learning is well known for imitating the human brain and when something imitates human brain then it can do all the task which can be accepted by a human to do. Deep learning model shows great result in extractive text summarization. While PCA reduces the space complexity by doping the irrelevant features and degree of freedom.

Extractive summary generated after implementation includes 81% accuracy in selecting relevant sentences. Finding the k value for Rank k approximation greatly speed up our research work by reducing time complexity. The results of experiments evaluate and shown analysis about the effectiveness approach proposed. This technique is a commendable addition to many text processing algorithms. In future, the research can be extended to different directions by using variations of PCA like Sparse PCA, Kernel PCA and Multi Linear PCA other Applied AI models and Unsupervised NLP. Also, feature extraction can be achieved by using different algorithm using NLP instead of ontology.

Funding Statement: The authors extend their appreciation to the Deanship of Scientific Research at King Faisal University for its financial support, with reference to the research grant number as 216082.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. Mansouri, M. M. Javidi and B. Mohammad Hasani Zade, "Using data mining techniques to improve replica management in cloud environment," *Soft Computing*, vol. 24, no. 10, pp. 7335–7360, 2020.

[2] M. K. Najafabadi, A. H. Mohamed and M. N. Mahrin, "A survey on data mining techniques in recommender systems," *Soft Computing*, vol. 23, no. 2, pp. 627–654, 2019.

[3] J. Steinberger, M. Poesio, M. A. Kabadjov and K. Ježek, "Two uses of anaphora resolution in summarization," *Information Processing & Management*, vol. 43, no. 6, pp. 1663–1680, 2007.

[4] S. Bhatia, P. Chaudhary and N. Dey, "Opinion mining in information retrieval," Springer Singapore, 2020. [Online]. Available: https://www.springer.com/gp/book/9789811550423.

[5] C. Kaushal and D. Koundal, "Recent trends in big data using hadoop," *Int. Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 8, no. 1, pp. 39–49, 2019.

[6] J. Liu, "Harvesting and summarizing user-generated content for advanced speech-based human-computer interaction," Doctoral dissertation. Massachusetts Institute of Technology, 2012.

[7] V. B. Raut and D. D. Londhe, "Opinion mining and summarization of hotel reviews," in *Int. Conf. on Computational Intelligence and Communication Networks*, IEEE, pp. 100–115, 2014.

[8] M. A. Tayal, M. M. Raghuwanshi and L. G. Malik, "ATSSC: Development of an approach based on soft computing for text summarization," *Computer Speech & Language*, vol. 41, no. 4, pp. 214–235, 2017.

[9] B. Gawalt, Y. Zhang and L. E. Ghaoui, "Sparse PCA for text corpus summarization and exploration," in *NIPS, 2010 Workshop on Low-Rank Matrix Approximation*, Whistler, Canada, pp. 215–210, 2010.

[10] P. E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, Association for Computational Linguistics, vol. 2, pp. 354–358, 2012.

[11] J. Clarke and M. Lapata, "Models for sentence compression: A comparison across domains, training requirements and evaluation measures," in *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, Association for Computational Linguistics, pp. 377–384, 2006.

[12] K. Knight and D. Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artificial Intelligence*, vol. 139, no. 1, pp. 91–107, 2002.

[13] C. Kaushal and D. Koundal, "Big data application in medical domain," in *2017 Int. Conf. on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, IEEE, pp. 1936–1939, 2017.

[14] F. Liu, J. Flanigan, S. Thomson, N. Sadeh, N. A. Smith *et al.,* "Toward abstractive summarization using semantic representations," in *Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, NAACL, Association for Computational Linguistics, pp. 180–187, 2015.

[15] D. Wang, T. Li, S. Zhu and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proc. of the 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval–SIGIR '08*, Singapore, Association for Computing Machinery, New York, United States, pp. 307–314, 2008.

[16] K. Sankar and L. Sobha, "An approach to text summarization," in *Proc. of the Third Int. Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies–CLIAWS3 '09*, Boulder, Colorado, Association for Computational Linguistics, pp. 53–60, 2009.

[17] K. A. Ganesan, C. X. Zhai and J. Han, "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions," in *Proc. of the 23rd Int. Conf. on Computational Linguistics (COLING '10)*, Beijing, China, Coling 2010 Organizing Committee, 2010.

[18] S. Bhatia, M. Sharma, K. K. Bhatia and P. Das, "Opinion target extraction with sentiment analysis," *Int. Journal of Computing*, vol. 17, no. 3, pp. 136–142, 2018.

[19] H. M. Lynn, C. Choi and P. Kim, "An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms," *Soft Computing*, vol. 22, no. 12, pp. 4013–4023, 2018.

[20] R. Bhargava, Y. Sharma and G. Sharma, "ATSSI: Abstractive text summarization using sentiment infusion," *Procedia Computer Science*, vol. 89, no. 6, pp. 404–411, 2016.

[21] R. Agrawal, S. Rajagopalan, R. Srikant and Y. Xu, "Mining newsgroups using networks arising from social behavior," in *Proc. of the Twelfth Int. Conf. on World Wide Web - WWW '03*, ACM, Budapest Hungary, Association for Computing Machinery, New York, United State, pp. 529–535, 2003.

[22] B. Agarwal, N. Mittal, P. Bansal and S. Garg, "Sentiment analysis using common-sense and context information," *Computational Intelligence and Neuroscience*, vol. 2015, no. 6, pp. 1–9, 2015.

[23] F. H. Khan, U. Qamar and S. Bashir, "A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet," *Knowledge and Information Systems*, vol. 51, no. 3, pp. 851–872, 2017.

[24] A. Sharaff, H. Shrawgi, P. Arora and A. Verma, "Document Summarization by Agglomerative nested clustering approach," in *2016 IEEE Int. Conf. on Advances in Electronics, Communication and Computer Technology (ICAECCT)*, Pune, India, IEEE, pp. 187–191, 2016.

[25] N. Kalchbrenner, E. Grefenstette and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, Association for Computational Linguistics, vol. 1, pp. 655–665, 2014.

[26] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. of the Seventh Int. Conf. on Language Resources and Evaluation (LREC'10)*, vol. 10, pp. 1320–1326, 2010.

[27] A. Stavrianou, J. Velcin and J. H. Chauchat, "A combination of opinion mining and social network techniques for discussion analysis," *Revue des Nouvelles Technologies de l'Information, In FDO*, vol. RNTI-E-17, pp. 25–44, 2009.

[28] M. K. Najafabadi, A. H. Mohamed and M. N. R. Mahrin, "A survey on data mining techniques in recommender systems," *Soft Computing–A Fusion of Foundations, Methodologies and Applications*, vol. 23, no. 2, pp. 627–654, 2019.

[29] A. Sharaff and U. Srinivasarao, "Towards classification of email through selection of informative features," in *2020 First Int. Conf. on Power, Control and Computing Technologies*, Raipur, India, IEEE, pp. 316–320, 2020.

[30] S. Bhatia, "A comparative study of opinion summarization techniques," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 110–117, 2021.

[31] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. of the Tenth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '04)*, ACM, Seattle WA USA, Association for Computing Machinery, New York, United States, pp. 168–177, 2004.

[32] F. Liu, D. Pennell, F. Liu and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in *Proc. of Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*, Association for Computational Linguistics, Boulder, Colorado, pp. 620–628, 2009.

[33] S. S.Bhatia, M. Sharma and K. K. Bhatia, "A novel approach for crawling the opinions from World Wide Web," *Int. Journal of Information Retrieval Research*, vol. 6, no. 2, pp. 1–23, 2016.

[34] E. Cambria, B. Schuller, Y. Xia and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[35] V. Bhati and J. Kher, "Survey for Amazon fine food reviews," *Int. Research Journal of Engineering and Technology*, vol. 6, no. 4, pp. 601–603, 2019.

[36] S. Mukherjee and S. Joshi, "Sentiment aggregation using ConceptNet ontology," in *Proc. of the Sixth Int. Joint Conf. on Natural Language Processing*, Nagoya, Japan, IEEE, pp. 570–578, 2013.

[37] A. Sharaff and A. Soni, "Analyzing sentiments of product reviews based on features," in *2018 2nd Int. Conf. on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, IEEE, pp. 710–713, 2018.

[38] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Text Summarization Branches Out: Proc. of the ACL-04 Workshop*, vol. 8, pp. 1–8, 2004.

[39] M. Alojail and S. Bhatia, "A novel technique for behavioral analytics using ensemble learning algorithms in E-commerce," *IEEE Access*, vol. 8, pp. 150072–150080, 2020.

[40] S. Bhatia, M. Sharma and K. K. Bhatia, "Opinion score mining: An algorithmic approach," *Int. Journal of Intelligent Systems and Applications*, vol. 9, no. 11, pp. 34–41, 2017.